CIIDS 2022

14th Asian Conference on Intelligent Information and Database Systems

November 28-30, 2022, Ho Chi Minh City, Vietnam

Book of Abstracts

Krystian Wojtkiewicz, Rafał Palak, Marcin Jodłowiec

14TH ASIAN CONFERENCE ON INTELLIGENT INFORMATION AND DATABASE SYSTEMS

Book of Abstracts

Collected works edited by
Krystian Wojtkiewicz, Rafał Palak,
and Marcin Jodłowiec

Wrocław University od Science and Technology Publishing House
Wrocław 2023

Funded by the Ministry of Science and Higher Education
as part of the “Doskonała Nauka” (Exellent Science) project

Editors
Krystian WOJTKIEWICZ
Rafał PALAK
Marcin JODŁOWIEC

Editorial layout and proofreading
Krystian WOJTKIEWICZ
Rafał PALAK
Tapas KAR

Technical editing
Marcin JODŁOWIEC

Cover design
Wiktoria WOJTKIEWICZ

Printed in the camera ready form

Wrocław University of Science and Technology Publishing House
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław
http://www.oficyna.pwr.edu.pl, https://ksiegarnia.pwr.edu.pl
e-mail: oficwyd@pwr.edu.pl, zamawianie.ksiazek@pwr.edu.pl

ISBN 978-83-7493-289-9
https://doi.org/10.37190/ACIIDS-2022_BoA

Preface

ACIIDS 2022 was the 14th event in a series of international scientific conferences on research and applications in the field of intelligent information and database systems. The aim of ACIIDS 2022 was to provide an international forum for research workers with scientific backgrounds in the technology of intelligent information and database systems and their various applications. The ACIIDS 2022 conference was co-organized by the International University - Vietnam National University HCMC (Vietnam) and the Wrocław University of Science and Technology (Poland) in cooperation with the IEEE SMC Technical Committee on Computational Collective Intelligence, the European Research Center for Information Systems (ERCIS), Al-Farabi Kazakh National University (Kazakhstan), the University of Newcastle (Australia), Yeungnam University (South Korea), Quang Binh University (Vietnam), Leiden University (The Netherlands), Universiti Teknologi Malaysia (Malaysia), Nguyen Tat Thanh University (Vietnam), BINUS University (Indonesia), the Committee on Informatics of the Polish Academy of Sciences (Poland) and Vietnam National University, Hanoi (Vietnam). ACIIDS 2022 was held in Ho Chi Minh City, Vietnam, and was conducted as a hybrid event during 28-30 November 2022.

This volume contains abstracts of peer-reviewed papers selected for presentation from 406 submissions, with each submission receiving at least three reviews in a single-blind process. It is the first time the Book of Abstracts was published, but we will continue this in the future.

We would like to express our sincere thanks to the honorary chairs for their support: Arkadiusz Wójs (Rector of Wrocław University of Science and Technology, Poland) and Zhanseit Tuymebayev (Rector of Al-Farabi Kazakh National University, Kazakhstan). We would like to thank the keynote speakers for their world-class plenary speeches: Tzung-Pei Hong from the National University of Kaohsiung (Taiwan), Michał Woźniak from the Wrocław University of Science and Technology (Poland), Minh-Triet Tran from the University of Science and the John von Neumann Institute, VNU-HCM (Vietnam), and Minh Le Nguyen from the Japan Advanced Institute of Science and Technology (Japan).

We cordially thank our main sponsors: International University - Vietnam National University HCMC, Hitachi Vantara Vietnam Co., Ltd, Polish Ministry of Education and Science, and Wrocław University of Science and Technology (Poland), as well as all of the aforementioned cooperating universities and organizations.

We thank the Special Session Chairs, Organizing Chairs, Publicity Chairs, Liaison Chairs, and Local Organizing Committee for their work towards the conference. We sincerely thank all the members of the international Program Committee for their valuable efforts in the review process, which helped us to select the highest quality papers for the conference. We cordially thank all the authors and the other conference participants for their valuable contributions. The conference would not have been possible without their support. Thanks are also due to the many experts who contributed to the event being a success.

December 2022

Krystian Wojtkiewicz
Rafał Palak
Marcin Jodłowiec

Table of Contents

DOI: https://doi.org/10.1007/978-981-19-8234-7_1

Keywords: fake news detection, graph embedding, influence graph

G-Fake: Tell me how it is shared and I shall tell you if it is fake

Nawfal Abbassi Saber, Rachid Guerraoui, Anne-Marie Kermarrec, Alexandre Maurer

nawfal.abbassi@um6p.ma, rachid.guerraoui@epfl.ch, anne-marie.kermarrec@epfl.ch, alexandre.maurer@um6p.ma

SIMPLIFIED TITLE

G-Fake: Tell me how it is shared and I shall tell you if it is fake.

ABSTRACT

The propagation of fake news is an increasingly serious concern in social platforms, and designing methods to automatically detect them and limit their spread is an important research challenge. Most existing methods rely on inspecting the content of news to decide on their veracity, but this information is not always available. In this paper, we present G-Fake (Graph-Fake), the first fake-news detection method that is entirely network-based. G-Fake only relies on the sharing history of news items. It does not assume any information on the content of these items (e.g. text or pictures), nor on the trustworthiness of users. In fact, G-Fake does not even require access to the underlying social graph, nor to the interactions between users. Our experimental evaluation conducted on real-world data shows that G-Fake can limit the spread of fake news in the earliest stages of propagation with an accuracy of 96.8%.

I INTRODUCTION

Social networks have become the first information exchange platforms, gathering billions of active users. On these platforms, anyone can share and propagate anything, including false information. For instance, Twitter was subject to this kind of manipulation to influence the 2016 US presidential election [1]. The use of a human workforce to fact-check the news and counter the propagation of fake ones is generally inefficient, and does not scale to the immense volume of content shared on social media platforms. Clearly, this calls for automatic methods to solve this problem - and they can be used to assist humans in charge of detecting false information.

Fake news can be seen as a data-science problem. Most approaches rely on analyzing the content of shared news items, but this approach has significant limitations. A complementary approach involves exploiting the underlying social graph, i.e., the graph of friendships or followers.

In this paper, we explore a novel approach that does not require access to either the underlying social graph or the content of news items. As an alternative, we propose G-Fake, a new protocol that builds an influence network from the history of shared news items, and uses this graph to effectively identify fake news early in the dissemination process.

In other words, the data (1) we consider is simply the history of sharing actions of news items and their labels (2). Sharing actions (3) are represented by a list of user identifiers, ordered by the time of sharing. Labels are binary variables “fake”/“not fake”.

$$data = \{item_1, item_2, \dots, item_N\} \tag{1}$$

$$item_i = (actions_i, label_i) \tag{2}$$

$$actions_i = (user_i^1, user_i^2, \dots, user_i^K) \tag{3}$$

The goal of G-Fake is to use a set of actions (tweeting, in the context of Twitter) related to some news items, to classify the latter as fake or not fake.

II STATE OF THE ART

There are primarily three approaches to perform automatic fake-news detection. Most prevalent are content-based approaches, which rely mainly on the content analysis of news items (e.g. text and pictures). Less common are strategies based on reputation and credibility. They exploit the history of user involvement w.r.t. to fake news, and attempt to estimate the reliability of news items based on the reputation of users who choose to share or not share them. Finally, network-based approaches attempt to classify news items using the social-graph features and the dissemination process. Numerous methods are hybrid, that is: they blend content-based methods with network- or credibility-based methods.

III Original Contribution

Our primary contribution is a novel, content-independent method for detecting fake news. We were able to obtain good performance in identifying fake news despite having limited access to data, and we demonstrated that unsupervised learning is helpful in extracting rich and meaningful user representations from social graphs.

IV Methodology

G-Fake assumes a realistic view of the environment. It only accesses the history of sharing actions. This can, of course, be achieved by the social network owner but also by third parties (external observers) that have access to very limited data. G-Fake is itself composed of three sub-protocols, each interesting in its own right. The first protocol is a fast and lightweight social influence graph construction method. The second is a deep-learning-based graph-embedding method that derives rich representations of users, reflecting their sharing patterns. The third protocol is a content-agnostic, network-based approach for fake-news detection.

V Results

Using only the first 30 sharing actions (for each news item), G-Fake is able to achieve excellent results: 96.8% in accuracy, 85.5% in F1-score and 92.0% in precision for the fake-news class, but also 98.2% in F1-score and 97.5% in precision for the real-news class. We also demonstrated that our method for recovering influence graphs is as effective as the Netrate [2] algorithm, but significantly faster for detecting fake news.

VI Evaluation

We test G-Fake using a real-world dataset, FakeNewsNet, which contains 21,595 annotated news articles, over 1.7 million tweets, and over 500,000 users. We divided the dataset into two comparable subsets (training set and test set). Since the dataset is unbalanced, we evaluate the performance of the classifier using not just accuracy, but also precision, recall, and F1-score by class to determine how well the model can detect fake news.

VII Conclusions

Our findings strongly indicate that content-agnostic and privacy-preserving machine learning is feasible, particularly for automatic fake-news detection. Our findings pave the way for third parties to effectively combat the propagation of fake news.

References

[1] Bovet, A., et al. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications 10*, 1 (2019), 1–14.

[2] Rodriguez, M. G., et al. Uncovering the temporal dynamics of diffusion networks. *arXiv preprint arXiv:1105.0697* (2011).

DOI: https://doi.org/10.1007/978-3-031-21967-2_27

Keywords: Super-resolution, Multi-image super-resolution, Deep learning, Convolutional neural networks, Data selection

Toward Understanding the Impact of Input Data for Multi-Image Super-Resolution

Jakub Adler, Jolanta Kawulok[0000-0002-8321-7760], Michal Kawulok[0000-0002-3669-5110]

michal.kawulok@ieee.org

Simplified Title

How to prepare the input images showing the same scene to generate a single image of high resolution?

Abstract

Super-resolution reconstruction is a common term for a variety of techniques aimed at enhancing spatial resolution either from a single image or from multiple images presenting the same scene. While single-image super-resolution has been intensively explored with many advancements proposed attributed to the use of deep learning, multi-image reconstruction remains a much less explored field. The first solutions based on convolutional neural networks were proposed recently for super-resolving multiple Proba-V satellite images, but they have not been validated for enhancing natural images so far. Also, their sensitiveness to the characteristics of the input data, including their mutual similarity and image acquisition conditions, has not been explored in depth. In this paper, we address this research gap to better understand how to select and prepare the input data for reconstruction. We expect that the reported conclusions will help in elaborating more efficient super-resolution frameworks that could be deployed in practical applications.

I Introduction

Super-resolution is a process of creating a high-resolution image with more details visible from a low-resolution image (or from many images) that show the same scene. If there are multiple low-resolution images available, then each of them carries a different part of the information necessary to create a high-resolution image that we need.

Such techniques may be beneficial, especially if we do not have a camera that can capture an image that is large enough. For example, such limitations exist for satellite images—we cannot get closer to the Earth to take a picture at higher resolution, but we can take many pictures of the same area to generate a higher-resolution image with more details visible afterward.

II State of the Art

There have been many methods proposed for super-resolution which are based on deep learning [3]. At first, they were developed for enlarging single images, as this is an easier task to do using a deep neural network, but later the methods that operate from many images were also proposed [1]. Such techniques use a neural network to do the job—a low-resolution image (or images) are presented to the network input, and the network is expected to return an enlarged image. Of course, the network must be trained to do so using a large number of image pairs. Each pair contains a low-resolution image together with a high-resolution one that shows the same scene. In this way, the network learns how to enlarge the images and it should be able to do it well for images which it has not seen during training.

III Original Contribution

In this paper, we wanted to answer the question of what to do if we have a large number of input images available. Is it better to use all of them or to exploit just some of them? How to select the good ones? How many images do we need to say we have enough of them? If it is possible, how should we capture such small images? Should they be captured in the same lighting conditions, or each one using a different light?

We have taken an attempt to answer the above questions, and here we present our conclusions. They may be helpful for those who want to apply super-resolution techniques in practice, and they may also help in developing better super-resolution methods in the future.

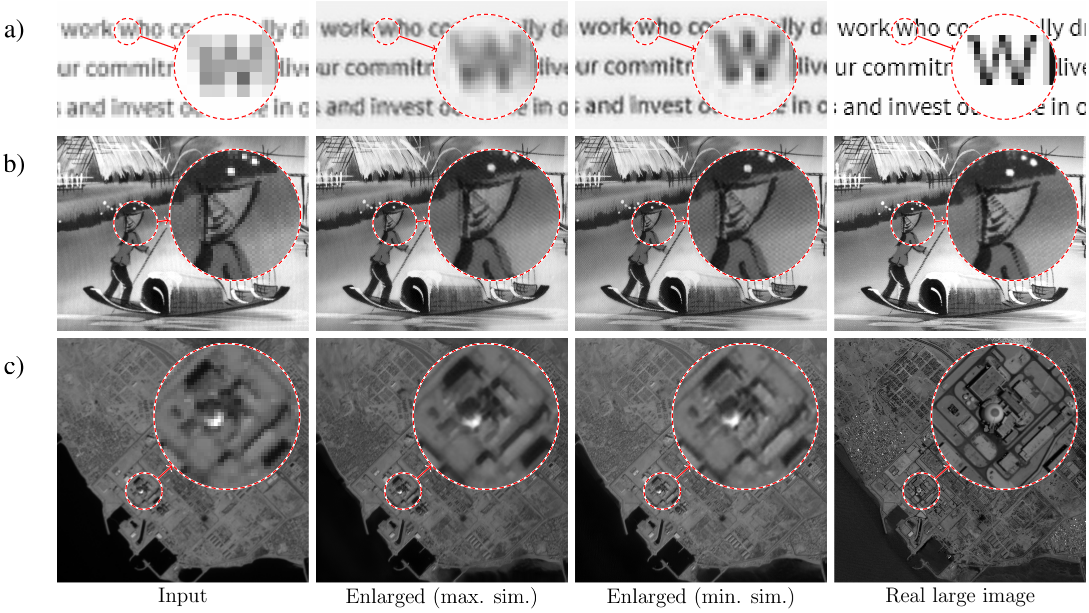


Figure 1: Outcome of RAMS for the scene with simulated low-resolution input (a), and for two scenes with real low-resolution input (b, c). The input images were selected based on maximized and minimized mutual similarity.

IV Methodology

In the reported research, we use one of the recent techniques for super-resolution that process multiple images—a residual attention model (RAMS) [2] which was developed for enlarging satellite images. We use a model trained using the images captured by the Proba-V satellite, and we use it for enlarging different kinds of images, including synthetic, natural, and satellite ones. We have considered several cases in which we checked the outcome depending on whether the input images are most similar to each other or most different from each other. Also, we considered different lighting conditions, including regular light, flash lamp, and a sidelight.

V Results

An example of the obtained results is shown in Figure 1. We present the original small input image, enlarged images from most similar (max. sim.) and most different (min. sim.) input images, and the real large image. It can be seen that the best outcome was obtained from most different images—for example, note the details in the hat in (b). Also, we found out that if the differences among the input images are high, then increasing the number of input images improves the final outcome. On the other hand, it occurred that combining input images acquired under different lighting conditions is not beneficial and results in obtaining a blurred outcome.

VI Conclusions

The reported study showed that the selection of input images has a significant impact on the obtained outcome of super-resolution. This definitely requires further investigation, however, our initial results suggest that the best outcome is obtained from images that differ between each other, provided that they were acquired in similar lighting conditions. This observation is important for practical applications of super-resolution, and it may be helpful in developing methods for preprocessing the input images before they are subject to the enlargement process.

Acknowledgements

This research was supported by the National Science Centre, Poland, under Research Grant 2019/35/B/ST6/03006 (MK) and co-financed by the SUT grant for maintaining and developing research potential (JK).

References

[1] Kawulok, M., Benecki, P., Piechaczek, S., Hrynczenko, K., Kostrzewa, D., and Nalepa, J. Deep Learning for Multiple-Image Super-Resolution. *IEEE GRSL 17*, 6 (2020), 1062–1066.

[2] Salvetti, F., Mazzia, V., Khaliq, A., and Chiaberge, M. Multi-image super resolution of remotely sensed images using residual attention deep neural networks. *Remote Sens. 12*, 14 (2020), 2207.

[3] Wang, Z., Chen, J., and Hoi, S. C. H. Deep Learning for Image Super-Resolution: A Survey. *IEEE TPAMI 43*, 10 (2021), 3365–3387.

DOI: https://doi.org/10.1007/978-3-031-21967-2_25

Keywords: neural networks, connections grouping, selective attention

Layer-wise Optimization of Contextual Neural Networks with Dynamic Field of Aggregation

Marcin Jodłowiec[0000-0001-7387-9210],Adriana Albu[0000-0003-1579 -6163], Krzysztof Wołk[0000-0001-7751-7591], Nguyen Thai-Nghe[0000-0002-9127-2778], Adrian Karasiński[0000-0001-9513-5394]

marcin.jodlowiec@pwr.edu.pl,adriana.albu@aut.upt.ro,
krzysztof.mateusz.wolk@gmail.com,ntnghe@cit.ctu.edu.vn,adrian.karasinski@cern.ch

Simplified Title

Layer-wise Optimization of Contextual Neural Networks with Dynamic Field of Aggregation

Abstract

This paper includes a presentation of experiments performed on Contextual Neural Networks with a dynamic field of view. It is checked how their properties can be affected by the usage of not-uniform numbers of groups in different layers of contextual neurons. Basic classification properties and activity of connections are reported based on simulations with H2O machine learning server and Generalized Backpropagation algorithm. Results are obtained for data sets with a high number of attributes (gene expression of bone marrow cancer and myeloid leukemia) as well as for standard problems from UCI Machine Learning Repository. Results indicate that layer-wise selection of numbers of connection groups can have a positive influence on the behavior of Contextual Neural Networks.

I Introduction

Common neural network model types analyze all inputs of their neurons for every data vector. Such a solution forces the model to store all internal connections active whenever the model is utilized. This feature limits their capabilities of analyzing contextual relations which can be found in data. Contextual Neural Networks (CxNNs) bypass this problem by introducing a novel architecture.

The schema of input signals processing by contextual neurons in CxNNs reflects the idea of the scan-path theory which was developed by Stark to mimic the human sensory system. This is achieved by the usage of neuronal aggregation functions which perform conditional, multi-step accumulation of input signals. Therefore every single contextual neuron can calculate its output based on different subsets of inputs in relation to different data vectors. As a result, data sets can be processed more accurately and with lower time and energy costs. The paper covers the analysis of Contextual Neural Networks (CxNN) with different numbers of connection groups in different layers of neurons. Especially, three patterns of numbers of groups in hidden layers have been considered (from inputs to outputs): increasing, decreasing and constant (uniform).

II State of the Art

The main elements of CxNNs are neurons with multi-step aggregation functions. Such neurons can aggregate signals from their all or only selected inputs depending on the data vector being processed. Such possibility makes CxNN generalization of Multilayer Perceptron (MLP) model of which neurons analyze signals from all their inputs and without any relation to data. It was shown that conditional, multi-step aggregation functions – in comparison to MLP – can notably increase classification accuracy and decrease its computations cost.

Current knowledge about Contextual Neural Networks covers mostly models in which every hidden neuron in a CxNN network uses the same count of groups K. The architecture presented in [1] considers layer-wise assignments of groups for neurons with Constant Field of Aggregation (CFA).

III Original Contribution

The paper introduces the results and their analysis from the experiment which verifies if the properties of Contextual Neural Networks with a dynamic field of aggregation (implemented by Sigma-if neurons) can be modified by layer-wise optimization based on mentioned patterns of numbers of groups. We extend our previous analysis using not-uniform patterns of numbers of groups in CxNNs with Sigma-if aggregation function.

IV Methodology

We have used experimental methodology to verify our hypotheses. The experiment has been conducted on a dedicated H2O machine learning server for standard problems from UCI Machine Learning Repository, as well as for real-life data sets with a high number of attributes (gene expression of bone marrow cancer and myeloid leukemia). Six datasets have been analyzed: *Breast cancer*, *Crx*, *Golub*, *Heart disease*, *Sonar* and *Soybean*. We have simulated four-layer CxNN with Sigma-if neurons to observe how their properties change for different layer-wise patterns of connection groups. All networks had two hidden layers. Each of the layers was built of ten Sigma-if units. One-hot and one-neuron-per-class unipolar encodings were used to construct input and output layers. The following patterns of numbers of groups were considered: increasing (G1471), decreasing (G1721, G1741) and constant (G1771).

V Results

Obtained results show that nonuniform number patterns of connection groups can significantly modify results achieved by CxNN models. Especially, the decreasing pattern of number of connection groups seems to increase the chances of creation of CxNNs with higher average classification accuracy. This has been observed both for G1721 and G1741 patterns, but the former leads to a stronger effect. Additionally, for G1721 pattern, four out of six data sets generated outcomes with decreased standard deviation of classification accuracy. It can be also noticed that usage of layer-wise decreasing patterns of numbers of groups (G1741 and G1721) decreases activity of hidden connections of CxNNs. The effect is stronger for G1721 – in this case lowered activity of connections can be observed for four out of six analyzed data sets.

Similar behavior was earlier observed for CxNNs with CFA aggregation function and in both cases it was unexpected. This is because single contextual neurons with lower number of groups typically present higher activity of input connections. The most evident case of such relation is contextual neuron with only one group, which by default has maximal, 100% activity of connections.

VI Evaluation

Measurements of classification accuracy obtained for test data are presented in Table 1.

Table 1: Average activity and its standard deviation of hidden connections of considered CxNNs with Sigma-if neurons and various patterns of numbers of groups in the hidden layers.

Dataset	**G1441 [%]**	**G1771 [%]**	**G1471 [%]**	**G1741 [%]**	**G1721 [%]**
Breast cancer	3.2 ± 1.7	4.8 ± 1.8	4.6 ± 2.9	3.8 ± 3.1	3.7 ± 2.8
Crx	12.7 ± 4.6	12.5 ± 4.7	12.4 ± 4.3	11.6 ± 4.4	10.2 ± 1.6
Golub	7.2 ± 0.4	7.6 ± 0.3	7.3 ± 0.4	6.7 ± 0.2	6.3 ± 0.3
Heart disease	11.5 ± 4.8	12.8 ± 4.7	11.2 ± 3.9	11.6 ± 4.3	11.2 ± 4.3
Sonar	7.4 ± 2.0	6.7 ± 1.4	6.8 ± 1.8	7.8 ± 1.6	5.4 ± 1.2
Soybean	4.6 ± 1.7	4.2 ± 1.5	4.5 ± 1.3	4.7 ± 1.6	3.1 ± 1.1

.

VII Conclusions

To our best knowledge, this study is the first to present in practice that GBP algorithm can successfully build CxNN models constructed of Sigma-if neurons when layer-wise setting of numbers of connection groups is used. This suggests that GBP method can be further extended with automatic optimization of numbers of connection groups – both in 'per-neuron' and layer-wise modes.

References

[1] Mikusova, M., Fuchs, A., Karasiński, A., Baruah, R. D., Palak, R., Burnell, E. D., and Wołk, K. Towards layer-wise optimization of contextual neural networks with constant field of aggregation. In *Intelligent Information and Database Systems* (Cham, 2021), Springer International Publishing, pp. 743–753.

DOI: `https://doi.org/10.1007/978-3-031-21967-2_50`

Keywords: AES, block cipher, ECB, CBC, OFB, CFB, mode of operation, ciphertext, periodicity, multiple encryption

Analysis of ciphertext behaviour using the example of the AES block cipher in ECB, CBC, OFB and CFB modes of operation, using multiple encryption

Zhanna Alimzhanova[0000-0001-6282-5356], Dauren Nazarbayev[0000-0001-7505-0422]
Aizada Ayashova[0000-0003-1893-5835], Aktoty Kaliyeva[0000-0003-2198-5630]

d.a.nazarbayev@gmail.com

Simplified Title

Ciphertext behaviour in modes of operation using multiple encryption

Abstract

This paper explores the Advance Encryption Standard (AES) block cipher in Electronic Code Book (ECB), Cipher Block Chaining (CBC), Output Feedback (OFB) and Cipher Feedback (CFB) modes of operation to compare the characteristic properties of ciphertext, and to compare the block complexity level of building ciphertext schemes using the methodology of periodic regularities.

This paper investigates the features of four block modes of operation, which includes two analytical principles: the first principle, which defines periodicity with respect to the ciphertext; and the second, which includes the principle of repeated cipher iterations, to react the characteristic manifestations of the ciphertext, under certain control input data. In accordance with the above principles, the results of analysis of the regularities of ciphertext with respect blocks and with respect encryption iterations were shown in tables and respectively in obtained formulae. Package Matplotlib of the Python programming language was used for graphical visualization ciphertexts of first iteration of encryption on all investigated modes of operation under different key sizes. The implementation of AES algorithm and obtaining encryption results were performed using package Crypto.

I Introduction

This paper investigates the AES block cipher in four modes of operation: ECB, CBC, OFB, and CFB [1, 2]. Block ciphers handle a text of a constant length, called the block size. Modes of operation are such special constructions designed to enhance the cryptographic strength of block ciphers and ensure the confidentiality of information [3]. In this paper, the following encryption modes were investigated: ECB, CBC, OFB and CFB, using the example of the AES block cipher. The study was conducted by the method of multiple encryptions to study the schemes for constructing modes of operation. The paper consists of an introduction, a central part: an analysis of behavioral characteristics in the modes of operation and a conclusion.

II State of the Art

In this paper, analysis of ciphertext behaviour was done for four modes of operation: ECB, CBC, OFB and CFB under certain control data. The analysis consists of identifying the periodicity of the ciphertexts. The authors applied method of multiple encryptions to analyze of ciphertexts in four modes of operation for a 128-bit key. The visualization technique in Python programming language using the package Matplotlib was applied to visualize the behaviour of changes in the dynamics of the ciphertexts under different key sizes. The implementation of the AES algorithm and obtaining encryption results were carried out using the package Crypto.

III Original Contribution

The scientific novelty of the research consists in identifying the main characteristics of the ciphertext behaviour and in the detect the periodicity of ciphertexts by method of multiple encryptions in the analysis of schemes for the construction of modes of operation on the example of the AES block cipher in investigated modes of operation.

IV METHODOLOGY

To determine the level of complexity and identify the distinctive characteristics of modes of operation, the method of multiple encryption iterations was applied as an exploratory technique under certain control input data to obtain reactions of characteristic manifestations of the encrypted text. The implementation of AES algorithm and obtaining encryption results were performed using the package Crypto and performed work related to experimental research.

V RESULTS

In the paper, the first object of research was selected the ECB mode, where all encryption results are periodic sequences relative to each block. Then for research, the CBC, OFB and CFB modes were selected, respectively. Under certain control data in the three modes of operation: CBC, OFB and CFB, the periodicity relative to the blocks was not detected, and a multiple encryption approach was applied for them. As a result of the exploration of the CBC, OFB and CFB modes, certain relations were identified, which were represented in tables and in formulae. In CBC mode, with multiple encryption iterations, a symmetric matrix was obtained, the elements of the main diagonal of which correspond to the result of encryption in ECB mode. We observe a connection between ECB and CBC modes: the ciphertext of the first block of n iterations in ECB mode corresponds to the ciphertext of the first iteration of n blocks in CBC mode. In the OFB mode, there is an obvious periodicity with respect to encryption iteration, where at odd iterations, the ciphertext corresponds to the ciphertext of the first iteration. At even iterations, the ciphertext corresponds to the plaintext, which is reflected in the formula. In this mode, the encryption algorithm matches the decryption algorithm, so we can encrypt and decrypt messages with the encryption algorithm. In CFB mode, a sequence of repetitions was determined, which changed in accordance with the geometric progression formula for detecting the period (absolute cycle) for n block ciphertext.

VI EVALUATION

In accordance with the above principles, the results of analysis of the regularities of ciphertext with respect to blocks and with respect to encryption iterations were shown in tables and respectively in obtained formulae, which will allow to take a new look at the problems level of complexity of schemes of modes of operation and strength of encryption algorithms.

VII CONCLUSIONS

All encryption results were obtained using the Python programming language using the packages: Matplotlib and Crypto. From the received results of the analysis of the ciphertext behaviour in investigated modes of operation using the example of the AES block cipher, the authors plan further research and development of an expert system model to analyze and estimate the level of complexity of schemes of modes of operation, to detect periodicity of ciphertexts and recognize different modern cryptographic ciphers with the assistance machine learning tools.

REFERENCES

[1] DWORKIN, M. *Recommendation for Block Cipher Modes of Operation. Methods and Techniques. National Institute of Standards and Technology* 2001.

[2] ROGAWAY, P. *Evaluation of some block cipher modes of operation, Cryptography Research and Evaluation Committees (CRYPTREC) for the Government of Japan* 2011.

[3] SMART, N.P. *Block Ciphers and Modes of Operation. In: Cryptography Made Simple. Information Security and Cryptography.* Springer, Cham, 2016.

DOI: https://doi.org/10.1007/978-981-19-8234-7_2

Keywords: LDA, Topic modeling

LDA+: An Extended LDA Model for Topic Hierarchy and Discovery

Amani Drissi [0000-0002-6863-9463], Ahmed Khemiri [0000-0002-4973-5396], Salma Sassi [0000-0002-9893-1158], Anis Tissaoui [0000-0003-0144-6365], Richard Chbeir [0000-0003-4112-1426] and Abderrazek Jemai [0000-0003-4033-2969]

drissiamani19@gmail.com, ahmedkhemiri24@outlook.fr, salma.sassi@fsjegj.rnu.tn, anis.tissaoui@fsjegj.rnu.tn, rchbeir@acm.org, Abderrazak.Jemai@insat.rnu.tn

Simplified Title

An Extended LDA Model for Topic Hierarchy and Discovery.

Abstract

The success of topic modeling algorithms depends on their ability to analyze, index and classify large text corpora. These algorithms could be classified into two groups where the first one is oriented to classify textual corpus according to their dominant topics such as LDA, LSA and PLSA which are the most known techniques. The second group is dedicated to extract the relationships among the generated topics like HLDA, PAM and CTM.
However, each algorithm among these groups is dedicated to a single task and there is no technique that makes it possible to carry out several analyses on textual corpus at the same time.
In order to cope with this problem, we propose here a new technique based on LDA topic modeling to automatically classify a large text corpora according to their relevant topics, discover new topics (sub-topics) based on the extracted ones and hierarchy the generated topics in order to analyse data more deeply.
Experiments have been conducted to measure the performance of our solution compared to the existing techniques. The results obtained are more than satisfactory.

I Introduction

n recent years, there has been an exponential growth in the number of complex documents and texts that require a deeper understanding of machine learning methods to be able to accurately classify texts in many applications[1]. So, it requires using new techniques or tools that deals with automatically organizing, searching and indexing a large collection of textual documents, in order to have a better way of managing the explosion of electronic document archives[2].
Topic modeling has demonstrated a good performance in this task. However, the existing topic modeling techniques are dedicated to a single task (Topic generation and text classification, Topic hierarchy and extracting correlated topics) and there is no technique that makes it possible to carry out several analyses at the same time. In order to overcome these challenges, we provide a new approach consisting of automatic large text corpora classification based on the LDA [3] topic model which provides a powerful tool for discovering and exploiting the hidden topics structure in large text archives and it can easily used as a module in more complicated models for more complex goals[2].

II State of the Art

The Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) topic models can only find topics in a flat structure, but fail to discover the hierarchical relationship among topics. These topic model categories cannot discover structural relations among topics, thus losing the chance to explore the data more deeply. An alternative model that not only represents the topic correlations, but also learns them, is the Correlated Topic Model (CTM).
Thus, in CTM topics are not independent, however, note that only pairwise correlations are modeled and the number of parameters in the covariance matrix grows as the square of the number of topics.
Hierarchical topic modeling has the capability of learning topics, as well as discovering the hierarchical topic structure of text data.

III Original Contribution

Our approach named LDA+ and it is able to (i) automatically classify documents from a heterogeneous corpus according to their dominant topics, (ii) automatically discover new topics (sub-topics) from the extracted ones in order to analyse the data more deeply, and (iii) automatically hierarchy the generated topics which facilitate the data analysis and interpretation tasks.

IV Methodology

In order to achieve our objectives we evaluated our approach based on experimental study.
Moreover, we used LDA Topic Model not only to classify documents into topics but also the originality of our approach resides in discovering and generating new topics from existing ones, which allows to discover the sub-topics of each extracted one and analyze the data in greater depth. Then we integrated the hierarchical clustering to find the hierarchical relationships between topics.
The originality of our approach reside also in the automation of LDA parameters initialization based on coherence measure. So we used the optimum parameters values for each used corpus.

V Results

With the application of our approach we can have several results: in a first place we generate topics from the used corpus in order to classify documents according to their relevant topics which give the user an overview of his documents topics and the terms which represents each generated one.
In a second place, we will have the possibility to discover new topics from the already extracted ones by and finally we will have a hierarchical representation of the topics.

VI Evaluation

Our experiments aimed at comparing well-known topic modeling techniques such as LSA, PLSA, LDA, CTM, HLDA, PAM and our approach (LDA+) in order to evaluate the quality of the extracted and discovered topics as well as the performance of each method with a different corpus size.
We implement and experiment these techniques on the same dataset and with the same parameters setting. We evaluate the results using the coherence measures which help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference and run-times.
The evaluation study shows that our approach LDA+ absolutely outperforms the existing techniques.

VII Conclusions

Our approach can be used in several fields in order to properly manage large amounts of textual data. This management can be summarized in the probabilistic classification of documents by topic and by the discovery of new ones as well as the extraction of the hierarchy between the extracted themes. All these results can be obtained with the use and the application of our LDA+ approach.

References

[1] Kamran Kowsari, Kiana Jafari Meimandi, M. H. S. M. L. B., and Brown., D. Text classification algorithms: A survey.

[2] Rubayyi, A., K. A. A survey of topic modeling in text mining.

[3] Zhiyuan Liu, Y. L., and Sun, M. *Representation Learning for Natural Language Processing*. Springer, ISBN 978-981-15-5573-2 (eBook) https://doi.org/10.1007/978-981-15-5573-2., 2020.

DOI: https://doi.org/10.1007/978-3-031-21743-2_1
Link to repository: https://github.com/Dennis1989/textClust-experiments

Keywords: stream clustering, text mining, concept drift

Textual One-Pass Stream Clustering with Automated Distance Threshold Adaption

Dennis Assenmacher[0000-0001-9219-1956], Heike Trautmann 2[0000-0002-9788-8282]

dennis.assenmacher@gesis.org,trautmann@uni-muenster.de

Simplified Title

Textual One-Pass Stream Clustering with Automated Distance Threshold Adaption

Abstract

Stream clustering is a technique capable of identifying homogeneous groups of observations that continuously arrive in a digital stream. In this work, we inherently refine a TF-IDF-based text stream clustering algorithm by the introduction of an automated distance threshold adaption technique for document insertion and cluster merging, improving the performance during distributional changes in the data stream. By conducting a thorough evaluation study, we show that our new fast approach outperforms state-of-the-art one-pass and batch-based stream clustering algorithms on various existing benchmarking datasets as well as a newly introduced dataset that poses additional challenges to the community. Moreover, we find that current evaluation approaches in the field of textual stream clustering are not adequate for a sound clustering performance assessment of evolving distributions. We thus demand the utilization of time-based evaluation.

I Introduction

Stream clustering is concerned with identifying homogeneous groups of observations in an unsupervised manner. While traditional offline algorithms allow iterating over incoming data objects multiple times, this is not feasible for large, potentially unbounded data streams. While different algorithms were presented in the past, they suffer from susceptible parameter settings that profoundly impact the algorithm's performance. Within this work, we enhance an existing text-based stream clustering algorithm that works on TF-IDF representations. We propose a technique to automatically set a crucial distance threshold and adjust it over time to account for distributional changes.

II State of the Art

II.1 Algorithms

Recently a plethora of different stream-clustering algorithms have been proposed. Typically they are model-based or belong to the vector-space category and allow either batch processing or are restricted to one-time passes over the data. Most existing algorithms require parameters that highly influence the final cluster solution quality. Often those parameters are not trivial to set and highly depend on the underlying data distribution.

II.2 Datasets

There currently are only a handful of well-established benchmarking datasets available. Existing datasets such as `News-T` or `Tweets-T` are small in size and provide clear class boundaries.

III Original Contribution

We extend `textClust`, a text-based stream clustering algorithm that works on TF-IDF representations of cluster centers. The algorithm assigns a new observation to an existing cluster if it is close to it (according to the cosine distance); otherwise, a new cluster is created. The distance threshold determining if a cluster is close enough is a sensitive method parameter that can profoundly impact the final cluster performance. We introduce a mechanism to automatically determine and adapt this parameter based on the current data stream and cluster distribution. Additionally, we enhance the existing cluster merging approach similarly by combining those clusters that are in close proximity, considering the current and previous inter-cluster distances. Besides algorithmic improvements, we also publish a new Twitter benchmarking dataset that poses additional challenges to existing alternatives in terms of size and complexity.

IV Methodology

Our study is of theoretical nature (algorithm development). To evaluate our new technique's performance, we conduct several experiments.

V Results

The new textClust method outperforms competitors on most of the tested datasets. The newly introduced option to adjust the distance threshold automatically proved robust in different settings. Therefore it can be considered a superior alternative to manual specification. Our new dataset Trends-T challenged existing stream-clustering methods because of its high number of observations and diffuse class boundaries.

VI Evaluation

We evaluate the proposed method in a series of experiments on different benchmarking datasets, including our newly introduced Twitter corpus. We test for various external clustering performance metrics such as *Homogeneity*, *Completeness*, and *Normalized Mutual Information*. The algorithms are evaluated globally and interval-based, providing additional insights into how they react and adapt to distributional changes over time.

VII Conclusions

Our new, improved version of textClust[1] can compete with existing text-based one-pass stream clustering algorithms and outperform them on various datasets. At the same time, we inherently simplify the practical utilization of the algorithm by removing the burden of selecting a sensitive distance threshold manually. Our automatic approach proved to perform appropriately under different domain and stream settings. Our results indicate that similar to previous observations, TF-IDF-based stream clustering algorithms, despite their simplicity, are tough, hard-to-beat baselines. With the introduction of our new dataset, we found that existing approaches still struggle in more complex scenarios and with larger streams. Therefore, future research should focus on the development of benchmarking suites.

[1] The algorithm is now part of the Python framework River - https://github.com/online-ml/river

DOI: `https://doi.org/10.1007/978-3-031-21743-2_42`

Keywords: Abstractive text summarization, Pretrained Language Model

A Survey of Abstractive Text Summarization Utilizing Pretrained Language Models

Ayesha Ayub Syed[0000-0002-3113-8980], Ford Lumban Gaol[0000-0002-5116-5708], Alfred Boediman, Tokuro Matsuo, Widodo Budiharto[0000-0003-2681-0901]

ayesha.syed@binus.ac.id,fgaol@binus.edu

Abstractive Text Summarization

Abstract

We live in a digital era - an era of technology, artificial intelligence, big data, and information. The data and information on which we depend to fulfil several daily tasks and decision-making can become overwhelming to deal with and requires effective processing. This can be achieved by designing improved and robust automatic text summarization systems. These systems reduce the size of text document while retaining the salient information. The resurgence of deep learning and its progress from the Recurrent Neural Networks to deep transformer based Pretrained Language Models (PLM) with huge parameters and ample world and common-sense knowledge have opened the doors for huge success and improvement of the Natural Language Processing tasks including Abstractive Text Summarization (ATS). This work surveys the scientific literature to explore and analyze recent research on pre-trained language models and abstractive text summarization utilizing these models. The pretrained language models on abstractive summarization tasks have been analyzed quantitatively based on ROUGE scores on four standard datasets while the analysis of state-of-the-art ATS models has been conducted qualitatively to identify some issues and challenges encountered on finetuning large PLMs on downstream datasets for abstractive summarization. The survey further highlights some techniques that can help boost the performance of these systems. The findings in terms of performance improvement reveal that the models with better performance use either one or a combination of these strategies: (1) Domain Adaptation, (2) Model Augmentation, (3) Stable finetuning, and (4) Data Augmentation.

I Introduction

With the current information and communication technologies in place, more and more people are getting connected through the world wide web. This has resulted in a tremendous explosion of unstructured data via emails, online workspaces, shopping websites, blogs, journals, articles, digital libraries, and social media websites. This data needs to be handled effectively to allow for the useful and efficient consumption of data. Automatic summarization systems can alleviate this issue by compressing the size of data while retaining only the significant information from it.

With neural network architectures like the Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and other variants, the performance of existing summarization models has improved to a reasonable extent, but the downside is that these models require large datasets to be trained on. If sufficient data is not available, the performance of the summarization model is degraded. However, large resource training datasets are not always available. It becomes very expensive and time-consuming to manually annotate a large dataset.

Based on the transfer learning paradigm of deep learning, the current Pretrained Language Models (PLM) are based on the Transformer architecture and possess massive parameters. These models have been pre-trained with specific objectives on huge resources in an unsupervised manner and possess a good deal of real-world and common-sense knowledge. PLMs-based ATS models work well on low-resource datasets. However, certain challenges are still associated with these models.

This research work analyses PLMs for abstractive summarization tasks and identifies challenges for PLMs-based summarization models through a literature review while indicating some improvement strategies from state-of-the-art models that can be opted to design future summarization models with better performance.

II State of the Art

Neural Abstractive text summarization is inspired by neural machine translation. It is a sequence-to-sequence modeling problem where an input sequence (source text) is processed to produce an output sequence (target summary). The initial models were built on top of RNN/LSTM/GRU architectures and the attention mechanism.

Later, these models evolved with the emergence of Transformer architecture. The Transformer architecture is at the core of current PLMs.

II.1 Transfer Learning

Abstractive text summarization using PLMs occurs via Transfer Learning. It is a process where large pre-trained language models are finetuned on the downstream summarization datasets. This results in the transfer of knowledge from the PLM to improve the target task.

III Original Contribution

As a literature review, this study contributes an analysis of recent pre-trained language models including UNILM, BART, MASS, PEGASUS, T5, LED, and BIGBIRD specifically for abstractive summarization tasks on four standard datasets (CNN/Daily mail, XSum, Gigaword, ArXiv) using ROUGE scores, identifies some issues with these systems, and highlights some strategies for performance improvement of PLM-based summarization systems.

IV Methodology

This study is a literature review and follows the methodology for finding and collecting the articles for review. Later, it uses quantitative and qualitative analysis methods to analyze and present its findings.

V Results

The results of the analysis of the PLMs on abstractive summarization datasets using ROUGE scores reveal that BART and PEGASUS perform competitively on CNN/Dailymail dataset, PEGASUS gives high performance on both XSum and Gigaword datasets while BIGBIRD and LED perform equally well in terms of R1 on the ArXiv dataset.

The identified challenges in the fine-tuning process include Catastrophic Forgetting, Representational Collapse, Overfitting/underfitting, and Few-shot domain transfer.

The indicated strategies for designing better abstractive summarization systems are, domain adaptation, model augmentation, better/regularized finetuning, and data augmentation.

VI Evaluation

Evaluation of the reviewed literature has been done using critical reading. The assumptions are reliable sources, useful, relevant, and recent research papers.

VII Conclusions

This article presents a comprehensive overview of neural abstractive text summarization. The emphasis is on the latest trends - the use of pre-trained language models for the task of abstractive text summarization. The survey presents multiple outcomes, including, an analysis of PLMs on abstractive summarization tasks, issues in the finetuning process, and performance improvement techniques. The problems in the finetuning process include catastrophic forgetting, representational collapse, model overfitting/underfitting, few-shot domain transfer, and low-quality generalization. Future research efforts may be directed toward tackling these challenges effectively. The performance improvement techniques include domain adaptation, model augmentation, stable finetuning, and data augmentation. We expect this survey will provide the abstractive summarization models to the research community with the latest trends, useful insights, and future research directions to design and develop better abstractive summarization models.

References

[1] Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J., Yuan, J., Zhao, W. X., Zhu, J., *Pre-Trained Models: Past, Present And Future*. AI Open, Volume 2, pp. 225-250, 2021.

[2] Zhang, J., Zhao, Y., Saleh, M., And Liu, P. J., *PEGASUS: Pre-Training with Extracted Gap-sentences for Abstractive Summarization*. ICML, pp. 11328–11339, 2020.

[3] Cao, Y., Wan, X., Yao, J., And Yu, D., *MultiSumm: Towards a Unified Model for Multi-Lingual Abstractive Summarization*. Proc. AAAI, Vol. 34, No. 01, pp. 11–18, 2020

DOI: https://doi.org/10.1007/978-3-031-21743-2_39

Keywords: Trust Reasoning, Trust Relationships

An Extension of Reciprocal Logic for Trust Reasoning: A Case Study in PKI

Sameera Basit [0000-0003-4931-6113], Yuichi Goto [0000-0003-2015-0340]

sameera@aise.ics.saitama-u.ac.jp, gotoh@aise.ics.saitama-u.ac.jp

Simplified Title

Application of extended reciprocal logic based on strong relevant logic in public key infrastructure for trust reasoning

Abstract

Trust relationship is one of the kinds of reciprocal relationship and basis of communications among agents, especially in open and decentralized systems, e.g., Public Key Infrastructure (PKI). In such systems, it is difficult to know whether an agent that is required to communicate with us can be trusted or not. Thus, it is indispensable to calculate the degree of trust of the target agent by using already known facts, hypotheses, and observed data. Trust reasoning is a process to calculate the degree of trust of the target agents. Although the current extension of reciprocal logic is an expectable candidate for a logic system underlying trust reasoning, it has a limitation when we deal with trust messages from other agents as a proposition. From the viewpoint of predicate logic, the current extension of reciprocal logic deals with messages from other agents as countable objects and are represented as individual constants. However, Demolombe represents messages from other agents as a proposition. From the viewpoint of expressive power, Demolombe's approach is better and the current extension of reciprocal logic is not enough. Following the Demolombe's approach, we introduced modal operators *Bel* and Inf and add several axioms to the current extension of reciprocal logic and a case study of trust reasoning based on the proposed extension in PKI is also presented.

I Introduction

The trust relationship is one of the important reciprocal relationships in our society and cyber space. Trust reasoning is one of the calculation methods used to calculate the degree of trust or to decide which target can be regarded as a trust one. It is an indispensable process to establish trustworthy and secure communication under open and decentralized systems that include multi-agent systems. The logic-based trust reasoning method can be applied to various target domains because formalization based on logic is qualitative and abstract.

Although we need a logic system underlying trust reasoning, there is no such logic system. The logic system underlying trust reasoning should satisfy the following two conditions: 1) the logic system is suitable for reasoning rather than proving, and 2) the logic system can deal with trust properties that are pointed out in previous works about trust reasoning [3]. A logic system that satisfies the two conditions is demanded to realize logic-based trust reasoning.

The purpose of our study is to propose the logic system underlying trust reasoning. We proposed extended reciprocal logic by introducing trust properties suggested by Demolombe [3] into reciprocal logic [2]. We also showed a case study in Public Key Infrastructure (PKI) to explain the usage of the proposed logic system.

II State of the Art

Several logic systems for trust reasoning were proposed. Their advantage is to deals with various trust properties. Especially, Demolombe [3] represents several trust properties, i.e., sincerity, validity, vigilance, credibility, cooperativity, and completeness, as logical formulas by using two modal operators *Bel* and Inf. Although the expressive power is enough, the logic systems are not suitable for reasoning because they are based on classical mathematical logic and its conservative extension. Through studies about relevant logics, as known, classical mathematical logic and its conservative extension are not suitable for reasoning.

Reciprocal logic [2] is an expectable candidate for a logic system underlying trust reasoning because it can deal with simple trust relationship and is suitable for reasoning. Reciprocal logic is a family of strong relevant logic that is suitable for logic systems underlying reasoning rather than classical mathematical logic and its conservative extensions. However, reciprocal logic cannot deal with the trust properties that Demolombe represented. The expressive power of reciprocal logic is not enough for a logic system underlying trust reasoning.

We proposed an extension of reciprocal logic by introducing trust properties [1]. However, the expressive power of the extension is not enough to represent trust reasoning in multi-agent systems.

III Original Contribution

We proposed a new extension of reciprocal logic, named "extended reciprocal logic", by introducing trust properties proposed by Demolombe [3] into reciprocal logic [2]. We also showed a case study in Public Key Infrastructure (PKI) to explain the usage of the proposed logic system.

IV Methodology

At first, we add a predicate "$TR(pe_1, pe_2, PROP)$" where pe_1 and pe_2 are agents, and *PROP* is an individual constant that represents trust properties: sincerity, validity, completeness, cooperativity, credibility, and vigilance into reciprocal logic. Secondly, we introduced two modal operators "$Bel_i(A)$" (an agent i believes that a proposition A is true) and "$Inf_{i,j}(A)$" (an agent i has informed an agent j about A) used in Demolombe's logic system into reciprocal logic to represent the trust relationship between agents and messages that comes from other agents.

Finally, we add new axioms into reciprocal logic. The axioms are as follows. ERcL1: $\forall i \forall j(TR(i,j,sincerity) \Rightarrow (Inf_{j,i}(A) \Rightarrow Bel_j(A)))$. ERcL2: $\forall i \forall j(TR(i,j,validity) \Rightarrow (Inf_{j,i}(A) \Rightarrow A))$. ERcL3: $\forall i \forall j(TR(i,j,vigilance) \Rightarrow (A \Rightarrow Bel_j(A)))$. ERcL4: $\forall i \forall j(TR(i,j,credibility) \Rightarrow (Bel_j(A) \Rightarrow A))$. ERcL5: $\forall i \forall j(TR(i,j,cooperativity) \Rightarrow (Bel_j(A) \Rightarrow Inf_{j,i}(A)))$. ERcL6: $\forall i \forall j(TR(i,j,completeness) \Rightarrow (A \Rightarrow Inf_{j,i}(A)))$.
BEL: $\forall i(Bel_i(A \Rightarrow B) \Rightarrow (Bel_i(A) \Rightarrow Bel_i(B)))$.

Let *RcL* be all axioms of reciprocal logic. Our new extension, named extended reciprocal logic, is $RcL \cup \{\text{ERcL1}, \ldots, \text{ERcL6}, \text{BEL}\}$.

V Results

Through an application of a case study in the domain of PKI, we showed that our approach is consistent when dealing with messages from other agents as a proposition because it is based on strong relevant logic.

Modal operators and new trust axioms aid in the reasoning out beliefs of agents in public key infrastructures (PKI). One of the advantages of our approach is generality. Trust reasoning based on a new extension of reciprocal logic is general in terms that messages from other agents in public key infrastructures (PKI) could be dealt with and represented as individual constants as well as propositions. Using extended reciprocal various complex scenarios can be described. Thus, we believe there is an improvement in our new extension of reciprocal logic.

VI Evaluation

We have evaluated our research by using a case study. We applied our extended reciprocal logic in the domain of public key infrastructure (PKI). The obtained results from the method show that the change in the trust relationships also causes the change in our reasoning results.

VII Conclusions

Results obtained by applying our extended approach for trust reasoning will be used to maintain the belief of the agent and also to deal with trust relationships with time-related constraints.

References

[1] Basit, S., and Goto, Y. An extension of reciprocal logics for trust reasoning. *In: Nguyen, N., Jearanaitanakij, K., Selamat, A., Trawiński, B., Chittayasothorn, S. (eds) Intelligent Information and Database Systems LNCS* (2020), pp. 65–75.

[2] Cheng, J. Reciprocal logic: logics for specifying, verifying, and reasoning about reciprocal relationships. *In: Khosla R., Howlett R.J., Jain L.C. (eds) Knowledge-Based Intelligent Information and Engineering Systems LNCS, vol. 3682* (2005), pp. 437–445.

[3] Demolombe, R. Reasoning about trust: a formal logical framework. *In Jensen, C., Poslad, S., Dimitrakos, T., eds.: Trust Management LNCS, vol. 2995* (2004), pp. 291–303.

DOI: https://doi.org/10.1007/978-3-031-21743-2_40

Keywords: graph,graph methods,XBRL,similarity graph

Common graph representation of different XBRL taxonomies

Artur Basiura[0000-0002-1034-6078] , Leszek Kotulski[0000-0002-0164-0048], Dominik Ziembiński

abasiura@agh.edu.pl,kotulski@agh.edu.pl,dominik.ziembinski@bft24.com

Simplified Title

Common graph representation for processing documents in different standards (taxonomy) XBRL

Abstract

Information nowadays plays a critical role in our lives, and its misinterpretation or lack of data, makes decisions wrong. It is important to systematize it, not only in the local context but globally. Finance is one of the key areas where standardization and normalization are attempted. One of the attempts is the XBRL format which is widely used in finance. However, the problem is the nature of local implementations. There are many different taxonomies that are implemented independently by countries and organizations. Currently there are no attempts to combine them and create a single standard. The paper presents a formal model for storing data in graph structures and the concept of using graph grammar to search financial indicators in big data storage. It provides a basis for the future construction of a common graph representation and thus the accumulation of cross-cutting knowledge.

I Introduction

The growing global market requires us to make quick and precise decisions. One of the format that is used for storing financial information is XBRL. It is used widely in many countries. The problem is that it usually store information that is hard to process and for different taxonomy in different form. The main concern raised in the publication is an attempt to map graph structures and convert them to a uniform format using the existing data processing algorithms. The goal is not to present a new standard, but to show a tool that will easily transform data from one taxonomy.

II State of the Art

Many publications in the literature deal with the problem of mapping real structures into graph structures and then processing them. An example may be publications introducing formal methods of introducing agent systems using graph grammars[3] [4] , or creating similarity graphs [1]. Those examples are used to form a formal graph representation of XBRL financial files.

III Original Contribution

This article introduces formal graph notation that is dedicated to the storage of data from the XBRL format. The notation is independent of taxonomy. Importing reports to a graph database allows us to perform operations related to identifying structures with the same similarity. It can be used in practice for speedy data retrieval. The last part shows a practical representation of the report parts in the US GAAP taxonomy. The compilation and storing of those data in graph structure allow to search quickly for the needed data. Not only based on one taxonomy.

IV Methodology

The article introduces XBRL graph. That has been used to present practical report and process them to obtain missing information. The definition of structures has been used to form practical example. To create the final model, the bft24.com [2] platform was used, which helped to obtain and model data.

V Results

The paper presents a formal model which can be used for storing XBRL reports and a case study of processing a group of 46,736 individual XBRL reports. With the proposed approach, it took about 130 seconds to create the ad-hoc report, and in comparison, the standard method reduced the preparation time by more than 10 times.

VI Evaluation

The primary assumption and purpose of the article were to create foundations and methods that can be extended and expanded in the future. The main goal was to design a method that, by converting reports to graph structures, allows their analysis to be independent of the type of taxonomy.

VII Conclusions

The concept presented can be used for developing an agent-based system that would offer even faster estimation. The use of parallel processing would allow initial estimates to be obtained in time comparable to real-time.

References

[1] Basiura, A., Sędziwy, A., and Komnata, K. Similarity and conformity graphs in lighting optimization and assessment. *International Conference on Computational Science* (2021).

[2] BFT24. BlockChain Financial Tools (XBRL processing). `https://prod.bft24.com`.

[3] Flasiński, M., and Kotulski, L. On the use of graph grammars for the control of a distributed software allocation. *The Computer Journal* (1992).

[4] Kotulski, L., and Sedziwy, A. Parallel graph transformations supported by replicated complementary graph. *ICANNGA* (2011).

DOI: `https://doi.org/10.1007/978-3-031-21967-2_45`

Keywords: C-V2X, Performance, Vehicles Speed, PDR

Performance of Packet Delivery Ratio for Varying Vehicle's Speeds on Highway in C-V2X Mode 4

Teguh Indra Bayu[1], Yung-Fa Huang[2], Jeang-Kuo Chen[3]

s10814906@gm.cyut.edu.tw[1], teguh.bayu@uksw.edu[1],
yfahuang@cyut.edu.tw[2], jkchen@cyut.edu.tw[3]

SIMPLIFIED TITLE

Performance evaluation using Packet Delivery Ratio metric for cellular vehicle to everything communication network technology under varying vehicle's speed and density conditions.

ABSTRACT

The 3GPP's Cellular Vehicle-to-Everything (C-V2X) specifications cover both short-range Vehicle-to-Vehicle (V2V) communications, which utilize an air interface called sidelink/PC5, and wide-area Vehicle-to-Network (V2N) communications, which enable vehicles to communicate with base stations (referred to as eNodeB in 3GPP). The primary contribution of this work is to investigate the performance of C-V2X Mode 4 in realistic highway scenarios by altering the vehicle's speed, vehicle density, resource keep probability (P_{rk}), and Modulation Coding Scheme (MCS) parameters. Simulation scenarios in this work will use three types of vehicle's speed. Each vehicle's speed type can be described as: Type 1 is 40 km/h, Type 2 is 80 km/h and Type 3 is 120 km/h. The simulation results show that the MCS = 9, and P_{rk} = 0.8 configuration achieved the best 95% Packet Delivery Ratio (PDR) performance distance breaking point with 380, 300, and 280 meters for the density of vehicles is 0.1, 0.2 and 0.3 vehicles/meter respectively.

I INTRODUCTION

Vehicle-to-Everything (V2X) communications are used broadly to describe methods for transmitting information flows in various Intelligent Transportation System (ITS) applications related to traffic safety and efficiency, automated driving, and infotainment. Vehicle-to-Vehicle (V2V), Vehicle-to-Network (V2N), Vehicle-to-Pedestrian (V2P), and Vehicle-to-Infrastructure (V2I) communications are all included in V2X. In short-range C-V2X, two techniques for resource allocation are defined: one that is network-controlled named Mode 3, and another that is entirely distributed among nodes, named Mode 4. To address this constraint and present an in-depth examination of Mode 4, we concentrate on the key factors mentioned, such as vehicle speed, vehicle density, resource keep probability (P_{rk}), and MCS. Previous work in [1] stated that the C-V2X performance for very slow vehicles speed depends on the number of vehicles and the value of P_{rk}. Where the higher P_{rk} showed better PDR performance in the higher vehicle's density. Considering similar vehicle's speed, vehicles tend to group and cause a significant loss in overall PDR performance.

II STATE OF THE ART

To investigate an in-depth exploration of C-V2X Mode 4, we enhance the OpenCV2X simulator with the corresponding vehicle's mobility scenarios. The SUMO framework is used to design the vehicle's mobility and road traffic control. Within the OmNet++ simulator, INET and SimuLTE modules are used to simulate the cellular network communication and packet data generator.

III ORIGINAL CONTRIBUTION

The major contribution of this work is the design and implementation of vehicle's speed variation within a single simulation with different vehicle classifications and particular traffic configurations. The Packet Delivery Ratio (PDR) and the distance breakpoint are used to emphasize the system's performance.

IV METHODOLOGY

We use OpenCV2X simulator as mentioned in [2]. Some known simulation parameters and configurations are used by authors in [3] investigated. OpenCV2X uses SUMO as the vehicle's mobility engine and OmNet++ for as the networking simulator. In this work, we design the SUMO’s mobility script to simulate the three types of vehicle’s speeds; Type 1 is 40 km/h, Type 2 is 80 km/h and Type 3 is 120 km/h. Each vehicle’s mobility type will be applied to particular highway road lanes as pictured in the report. The vehicle’s density (β) used in this work are: 0.1 (200 cars), 0.2 (400 cars) and 0.3 (600 cars) per meter. We evaluate the Packet Delivery Ratio (PDR) as the evaluation metric and the distance breakpoint to investigate the system stability.

V RESULTS

The simulation result shows six simulation conditions for MCS = 7, P_{rk} = 0; MCS = 7, P_{rk} = 0.8; MCS = 7, P_{rk} = 1; MCS = 9, P_{rk} = 0; MCS = 9, P_{rk} = 0.8; MCS = 9, P_{rk} = 1 as shown in Figure 1. Each of the six conditions was also tested with three different densities β = 0.1, 0.2 and 0.3. From Figure 1, it can be seen that the best overall performance is with the MCS = 9, P_{rk} = 0.8 conditions. Moreover, the 95% PDR performance chart is shown in Figure 2, which shows the superiority from MCS = 9 with P_{rk} = 0.8 configurations.

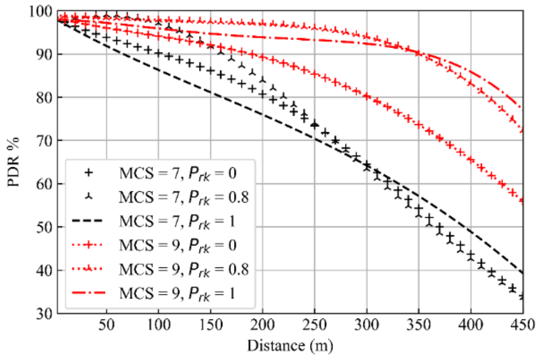


Figure 1. PDR Comparison for β = 0.3 vehicles/meter.

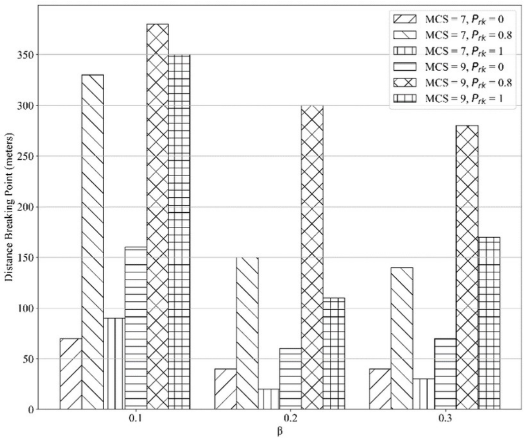


Figure 2. 95% PDR Performance Distance Breaking Point.

VI EVALUATION

From the corresponding configurations and conditions, the Packet Delivery Ratio (PDR) is observed. To investigate the stability of each configuration, a 95% PDR threshold is set. Then each condition is again observed until how far the distance that the PDR can stay above the 95% threshold. The distance number in the 95% PDR threshold means that the configuration can achieve above 95% PDR till corresponding distances.

VII CONCLUSIONS

This work emphasized mixed vehicle's speed inside the road environment. This mixed vehicle's speed comprises the varying vehicle's speed to create a realistic highway situation. Simulation results show that the MCS = 9 with P_{rk} = 0.8 achieves the overall best performance compared to the other configurations and achieve the highest result for 95% PDR threshold distance breaking point.

REFERENCES

[1] T. I. Bayu, Y. F. Huang, and J. K. Chen, "Performance of C-V2X Communications for High Density Traffic Highway Scenarios." pp. 228-233.

[2] B. McCarthy, A. Burbano-Abril, V. Rangel Licea, and A. O'Driscoll, "OpenCV2X: Modelling of the V2X Cellular Sidelink and Performance Evaluation for Aperiodic Traffic," https://ui.adsabs.harvard.edu/abs/2021arXiv210313212M, [March 01, 2021, 2021].

[3] A. Bazzi, G. Cecchini, A. Zanella, and B. M. Masini, “Study of the Impact of PHY and MAC Parameters in 3GPP C-V2V Mode 4,” *IEEE Access*, vol. 6, pp. 71685-71698, 2018.

DOI: https://doi.org/10.1007/978-981-19-8234-7_3

Keywords: Graph, Summarization,heterogeneous data, structure summarization, based content summarization

Schema formalism for semantic summary based on labeled graph from heterogeneous data

Amal Beldi[0000-0002-7768-6691], Salma Sassi[0000-0002-9893-1158] Richard Chbeir[0000-0003-4112-1426] Abderrazek Jemai

amal.beldi@fst.utm.tn, salma.sassi@fsjegj.rnu.tn, richard.chbeir@univ-pau.fr, abderrazekjemai@yahoo.co.uk

No simplified title given.

Abstract

The problem of graph summarization has been studied in the literature and many approaches for static contexts are proposed to summarize the graph in terms of its communities. These approaches typically produce groupings of nodes which satisfy or approximate some optimization function. Nevertheless, they fail to characterize the subgraphs and do not summarize both the structure and the content in the same approach. This means that there is no framework that provides summarization of mixed-source and information with the goal of creating a dynamic, syntactic, and semantic data summary. In this paper, we address the aforementioned problems by proposing an appropriate approach able to model heterogeneous sources based in a single graph based on a schema-driven approach, providing a personalized summary model capable of synthesizing graphically the content and finally summarizing the structure of the graph in order to reduce its size, minimize its complexity and keep the important nodes and relations. We illustrate this approach through a case study on the use of E-health domain.

I Introduction

Recently graph summarization has become a hot topic in the database research community in recent years. It facilitates the identification of structure and meaning in data. A summary is a concise representation of the original graph, whose objectives can significantly vary from reducing the number of bits needed for encoding the original graph, to the more complex database-style operations that summarize graphs where the resolution could be scaled-up or scaled-down interactively. Furthermore, existing approaches are only suitable for a static context, and do not offer direct dynamic counterparts. Some algorithms like [1], [2] work in a dynamic setting, but focus only on finding static patterns that appear over multiple time steps. This means that there is no framework that provides a summarization of mixed-source and information with the goal of creating a syntactic and semantic data summary. Given the above problems, the proposed paper focus on how we can best describe both structure and content in one summary and thus not just generate succinct summaries for the mixed-sources, but also understand its corresponding interactions and relationships with the past. Thus, towards building a semantic and dynamic summary, the following challenges emerge: (i) How to provide multi-sources-based summary, due to multi-modality of data (e.g.,text, video, and image) that can be encoded in different formats? (ii) How to provide user-oriented semantic based summary, due to the difficulty of retrieving information according to user'needs? and (iii) How to incorporate the dynamic nature of real data in computation and perform analysis efficiently? Our work aims to generate a concise semantic summary of heterogeneous sources to better understand their underlying characteristics.

II State of the Art

To compare existing approaches, we define here nine criteria such as input data, data type, represented standard, summarization technique, summarization approach, medical knowledge based summarization, output type, context-aware, and user-oriented. So, we observe that most of the existing studies do not consider real data in their analysis and do not consider the context of creating the summary and they rely only on the time property. Thus, existing systems can still not contextually interpret and reason on the transferred knowledge among real data, and, consequently, cannot synthesize data to provide accurate desired results. All existing systems focus on one objective, while none provide various functionalities in the same framework despite its importance in supporting users' preferences to find the data according to diverse needs. All objectives should be an integral part of a summarization-based system. Most of the above studies can only satisfy a certain aspect of user needs. Finally, another important part of this study is the output type of summarized data. They do not propose dedicated tools that make the summary accessible to the user nor provide them with appropriate perceptions of their needs. Users are more and more concerned about security, confidentiality, understanding of their data, and the accuracy and completeness of their data.

III Original Contribution

The proposed approach aims to summarize data into a single graph model. It's a schema-driven approach based on labeled graphs and links the graph model to the relevant domain knowledge to find relevant concepts to provide meaningful and concise summary. Last but not least, it provides a personalized visualization model capable to summarize graphically both the structure and the content of the data from databases, devices and sensors to reduce cognitive barriers related to the complexity of the information and its interpretation. To achieve this goal, our framework architecture is composed of four main modules: Data Pre-Processing module, Data Graph generation module, Data Summarization module and Data Post-Processing module.

Data Graph Formalism (DGM): Here, we define the DGM graph to efficiently represent important heterogeneous data as Data Node (DN) and the Relationship (DR) between them.

Summarization based structure model: This model aims to summarize the graph based on its topology to reduce the size and minimize the complexity of the graph. Here we used some summarizing techniques such as compression, grouping, simplification and visualization.

Summarization-based content model: We provide a user-centered summarization model depending on the user preferences. Our goal is to provide data model adjusted based on user preferences and needs. To this end, we define a new node to allow users to personalize the content according to the analysis needs and preferences. We provide here a calculation process aiming at calculating a mathematic function from any number of incoming numeric values from one of many Data Nodes from the GDM.

IV Methodology

In this study, we propose an appropriate approach able to model heterogeneous sources based in a single graph based on a schema-driven approach, providing a personalized summary model capable of synthesize graphically the content-based and finally summarizing the structure of the graph in order to reduce its size and minimize its complexity and keep the important nodes and relations. We propose a summary-driven formalism based on labeled graphs, which provides exciting data summary conserving better data integrity.

V Results

The result of summary graph visualization contains only linked nodes for structured summarization an attribute that expresses user needs such as a particular disease, or summarize medical prescriptions of X-ray interpretation. For the based content summarizes, the result of this type of synthesis shows us either the maximum value or the minimum value, the average of these measurements during the period mentioned (1 month) or a curve that interprets the variations of the measurements.

VI Evaluation

For our benchmarking, we used a medical database. We have proposed two metrics, the running time and the loss of information to evaluate our approach . We compared our algorithm in term based content summarization to other ones based on the structure in the literature. We choose the approach proposed using the execution time metric. We considered the execution time of our algorithm always remains low, and this guarantees its applicability to large graphs (nodes, relationships) and that shows good performance for our approach. We compare our algorithms from aggregation nodes, and aggregation relationships with algorithm of approach [2] using a loss of information metric. We noted that our approach is more efficient keep a large percentage of the content graph (nodes and relationships).

VII Conclusions

In this work, we study utility-driven graph summarization in-depth and make several novel contributions. We present a new, lossless graph summary, the first structured and the second content-based. In future works we will be interested for other types of node summarization like textual and image.

References

[1] Shah, N., Koutra, D., Zou, T., Gallagher, B., and Faloutsos, C. Timecrunch: Interpretable dynamic graph summarization. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (2015), pp. 1055–1064.

[2] Tian, Y., and Patel, J. M. Tale: A tool for approximate large graph matching. In *2008 IEEE 24th International Conference on Data Engineering* (2008), IEEE, pp. 963–972.

DOI: https://doi.org/10.1007/978-981-19-8234-7_4

Keywords: neural networks, image processing, imaging sensor identification

Individual Source Camera Identification with Convolutional Neural Networks[1]

Jarosław Bernacki[0000-0002-4488-3488], Kelton A.P. Costa[0000-0001-5458-3908], Rafał Scherer[0000-0001-9592-262X]

jaroslaw.bernacki@pcz.pl, kelton.costa@unesp.br, rafal.scherer@pcz.pl

Simplified Title

Digital Camera Identification with Neural Networks

Abstract

In this paper we consider the issue of digital camera identification which matches the area of digital forensics. This problem is well-known in the literature and many algorithms based on camera's fingerprint have been proposed. However, one may find that there is a little number of methods providing a fast and accurate digital camera identification. This problem is especially observed in terms of today's digital cameras, producing images of big sizes. In this paper we discuss several existing approaches based on convolutional neural networks (CNN). We try to find out whether it is possible to speed up the process of learning the networks by the images. One of the findings include replacing the ReLU with SELU activation function. We experimentally show that using SELU speeds up significantly the process of learning. We also compare the identification accuracy of all considered methods. The experiments are held on extensive image dataset, consisting of many images coming from modern cameras.

I Introduction

Digital forensics is a field that has attracted much attention in recent years. One of the most popular topics in digital forensics is the identification of imaging sensors in digital cameras. Nowadays, digital cameras are widely accessible and affordable, which makes them very popular. Smartphones and mobile devices are even more popular. Today's smartphones are equipped with built-in digital cameras what encourage people to take photos and share them on social media networks. However, the possibility of establishing whether an image was taken by a given camera may expose users' privacy to a serious threat. Hence, a number of papers in recent years are dedicated to the study of imaging device artifacts that may be used for digital camera identification. Digital camera identification can be realized in two approaches: individual source camera identification (ISCI) and source model camera identification (SCMI). The ISCI is capable of distinguishing a certain camera model among cameras of both the same and different camera models. On the other hand, the SCMI distinguishes a certain camera model among the different models but is not able to distinguish an individual copy of the camera from other cameras of the same model. Our main goal is to find such a convolutional neural network (CNN) structure that will make it possible a fast learning process.

II State of the Art

The state-of-the-art algorithm for the ISCI aspect was proposed by Lukás et al.'s [3] which serves as a unique ´ camera's fingerprint. Lukás et al.'s algorithm [3] is based on the calculation of the noise residual **N**

$$\mathbf{N} = \mathbf{I} - F(\mathbf{I}) \tag{1}$$

where F is a denoising filter, **N** stands for a noise residual of the image **I**. Thus, this procedure should be repeated for a certain number of images from a camera (authors propose using at least 45 images). Images are denoised using a wavelet-based denoising filter F. The camera's noise residual is finally calculated as an average of a particular number of noise residuals. Images are processed in their original resolution.

III Original Contribution

We propose a shallow convolutional neural network that may be used for an individual source camera identification. A first convolutional layer with 8 filters of size $3 \times 3 \times 3$ and stride 1 with a max-pooling layer with kernel size 2 and stride 2, ReLU as activation function. The second convolutional layer with 8 filters of size $3 \times 3 \times 8$ and stride 1 with a max-pooling layer with kernel size 2 and stride 2, ReLU as activation function. The third convolutional layer with 16 filters of size $3 \times 3 \times 8$ and stride 1 with a max-pooling layer with kernel size 2 and stride 2, ReLU as activation function. Fully connected layers for classification.

IV Methodology

We compare our network with the proposed by Bondi et al.'s [1], Tuama et al.'s [5], Mandelli et al.'s [4] and Kirchner and Johnson SPN-CNN [2].

V Results

We conduct two experiments. The first includes a comparison of the classification accuracy of Lukás et al.'s, proposed CNN, Bondi et al.'s and Tuama et al.'s CNNs. The second experiment compares CNNs with various hyper-parameters and their impact on the training accuracy. We use a set of 17 modern cameras `https://kisi.pcz.pl/imagine/`.

The results clearly indicate that all methods ensure very high identification accuracy. In all cases, the overall identification accuracy obtains 92%; the particular TPRs are not lower than 90% for each camera. Interestingly, Lukás et al.'s algorithm achieves almost the same results compared to CNN-based methods. This clearly confirms that all methods ensure a reliable individual source camera identification. In the case of the discussed CNNs, the results are very similar to each other both in terms of identification accuracy and speed of learning. All the CNNs require a similar number of training epochs to obtain the desired level of identification accuracy.

We also analyzed the usage of different activation functions for each convolutional layers. Analyzed functions include the ReLU, Leaky ReLU, and SELU activation functions. Since the ReLU, Leaky ReLU, and SELU activation functions are well-known, we will not recall them. Let us note that the discussed CNNs use ReLU as the activation function. Instead, we propose to use the Leaky ReLU and SELU functions. We have analyzed the training accuracy using all three activation functions.

The results clearly indicate that using the SELU activation function allows for much faster learning than ReLU or Leaky ReLU functions. Training already of 10 epochs for all considered CNNs gives 80% train accuracy, while training for 20 epochs obtains 90% of train accuracy. The Leaky ReLU function needs at least 10 epochs more for the same results, while the ReLU achieves the lowest results, demanding even two times more of training epochs than the SELU function. Therefore, achieving at least 90% accuracy for the ReLU activation function requires learning all the networks for at least 35 epochs.

VI Conclusions

In this paper, we discussed individual source camera identification based on images. We analyzed several approaches for digital camera identification in terms of identification accuracy and speed of learning. Experiments conducted on an extensive image dataset confirmed that modern algorithms obtain very high identification accuracy. We also showed that in the case of convolutional deep learning methods, different activation functions have a significant impact on the speed of network learning. Experiments confirmed that using the SELU activation function significantly speeds up the network learning process. The Leaky ReLU activation function obtains a visibly lower learning speed, while the very common ReLU is the slowest.

Acknowledgements

The project financed under the program of the Polish Minister of Science and Higher Education under the name ”Regional Initiative of Excellence” in the years 2019 - 2023 project number 020/RID/2018/19 the amount of financing PLN 12,000,000.

References

[1] Bondi, L., Baroffio, L., Guera, D., Bestagini, P., Delp, E. J., and Tubaro, S. First steps toward camera model identification with convolutional neural networks. *IEEE Signal Process. Lett. 24*, 3 (2017), 259–263.

[2] Kirchner, M., and Johnson, C. SPN-CNN: boosting sensor-based source camera attribution with deep learning. *CoRR abs/2002.02927* (2020).

[3] Lukás, J., Fridrich, J. J., and Goljan, M. Digital camera identification from sensor pattern noise. *IEEE Trans. Information Forensics and Security 1*, 2 (2006), 205–214.

[4] Mandelli, S., Cozzolino, D., Bestagini, P., Verdoliva, L., and Tubaro, S. Cnn-based fast source device identification. *IEEE Signal Process. Lett. 27* (2020), 1285–1289.

[5] Tuama, A., Comby, F., and Chaumont, M. Camera model identification with the use of deep convolutional neural networks. In *IEEE International Workshop on Information Forensics and Security, WIFS 2016, Abu Dhabi, United Arab Emirates, December 4-7, 2016* (2016), IEEE, pp. 1–6.

Link to arXiv: https://arXiv.org/abs/2104.12473
DOI: https://doi.org/10.1007/978-3-031-21743-2_15

Keywords: SQL Injection Attacks, XSS Attacks

An Ensemble based Deep Learning Framework to Detect and Deceive XSS and SQL Injection Attacks

Waleed Bin Shahid[0000-0001-9113-4490], Baber Aslam, Haider Abbas[0000-0002-2437-4870], Hammad Afzal[0000-0001-9583-5585], Imran Rashid[0000-0001-8958-672X]

{waleed.shahid,ababer,haider,hammad.afzal,irashid}@mcs.edu.pk

Simplified Title

A smart framework based on advanced deep learning techniques to detect and mitigate SQL injection and Cross Site Scripting Attacks. The framework also deceive the attackers launching these Web Application Attacks.

Abstract

Safeguarding websites is of utmost importance nowadays because of a wide variety of attacks being launched against them. Moreover, lack of security awareness and widespread use of traditional security solutions like simple Web Application Firewalls (WAFs) has further aggravated the problem. Researchers have moved towards employing sophisticated machine learning and deep learning based techniques to counter common web attacks like the SQL injection (SQLi) and Cross Site Scripting (XSS). Lately, keen interest has been taken in tackling these attacks through cyber deception. In this paper, we propose an ensemble based deep learning approach by combining Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) models. This detection framework also contains a Session Maintenance Module (SMM) which maintains user state in an otherwise stateless protocol by analyzing cookies thereby providing further optimization. The proposed framework detects SQLi and XSS attacks with an accuracy of 99.83% and 99.47% respectively. Moreover, in order to engage attackers, a deception module based on dockers has been proposed which contains deceptive lures to engage the attacker. The deceptive module has the capability to detect zero-days and is more efficient when compared to other similar solutions

I Introduction

The main purpose of this research is to propose an ensemble-based deep learning framework using CNN and LSTM classifiers for detecting and mitigating SQLi and XSS attacks and then to deceive and engage the attackers with the help of a light-weight high interaction docker based deception module which comprises of a docker daemon. The prime motivation was not just to stop the attacks but to study attacker behavior, motives, attack tactics and techniques [2]. Moreover, attackers are also profiled with the help of a state maintenance module.

II State of the Art

Most existing web attack detection solutions lack the key feature of analyzing the state of the incoming traffic, thereby lacking the ability to maintain attacker's profile over time [1]. Few research works have focused on profiling attackers in order to enhance the attack detection capability. Moreover, these techniques only counter attacks and do not have any deception module to engage the attacker. Most of the deception solutions do not offer customized deception and are static in nature. Moreover, they are not coupled with attack detection modules. Since the proposed technique is based on an ensemble approach using CNN and LSTM models along with a Session Maintenance Module which helps in maintaining the attacker's state, thereby augmenting the performance and efficacy of the deception module.

III Original Contribution

Our framework has improved the capability of web attack (SQLi and XSS) detection engines by adding the key feature of attacker categorization and profiling which helps us in providing customized deception for attackers launching these two attacks.

IV Methodology

Our framework is based on real-time experiments and has been deployed in a controlled environment. The incoming web requests are analyzed in order to extract state-related information that helps in profiling and categorizing the attacker(s). Later the request is passed to the ensemble based deep learning model that comprises of CNN and LSTM classifiers trained on a publicly available benchmark dataset. The requests if found to be SQLi and XSS attacks are diverted to the attack specific dockers designed to engage the attacker. Benign requests are forwarded to the website.

V Results

The proposed framework was deployed and tested in front of actual websites which were exposed to simulated red teaming attacks along with benign traffic. Both the CNN and LSTM models were evaluated on the test traffic comprising SQL and XSS traffic (attack and benign requests) for ten epochs in order to maximize accuracy and minimize validation loss. Later the ensemble classifier made the final classification decision by combining the CNN and LSTM classifications. Moreover, it was observed that by using the docker-based approach, no major delay was observed in the response time of the deception module as HTTP responses by both the attack dockers were very minutely greater than the response time of the actual website making the framework suitable for real world deployment [3].

VI Evaluation

The proposed framework primarily focused on SQLi and XSS attacks as these are the two most common web application attacks. The main assumption was users do not discard, evade and ignore the Web Cookies before visiting a website. The main outcome that can be drawn from this research work is that the key feature of attack detection and deception can be combined in a single solution so that attacks are detected, and attackers are profiled, and their tactics are analyzed in a single solution.

VII Conclusions

This research work proposes an ensemble-based deep learning approach comprising of CNN and LSTM classifiers along with a State Maintenance Module for detecting XSS and SQL injection attacks with very high accuracy. Deception is being carried out with the help of a full-fledged deception framework based on docker containers. In this module, the highly secure docker daemon controls the attack dockers that have lures to engage the attacker.

References

[1] Clincy, V., and Shahriar, H. Web application firewall: Network security models and configuration. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (2018), vol. 1, IEEE, pp. 835–836.

[2] Shahid, W. B., Aslam, B., Abbas, H., Afzal, H., and Khalid, S. B. A deep learning assisted personalized deception system for countering web application attacks. *Journal of Information Security and Applications 67* (2022), 103169.

[3] Valicek, M., Schramm, G., Pirker, M., and Schrittwieser, S. Creation and integration of remote high interaction honeypots. In *2017 International Conference on Software Security and Assurance (ICSSA)* (2017), IEEE, pp. 50–55.

DOI: 10.1007/978-3-031-21967-2_49
Keywords: Blockchain, Decentralized storage system, Digital content sharing model

Blockchain-based Decentralized Digital Content Management and Sharing System

Thong Bui[0000-0001-8709-7113], Tan Duy Le[0000-0001-6597-0209], Tri-Hai Nguyen[0000-0002-2132-2290], Bogdan Trawinski[0000-0002-2956-6388],
Huy Tien Nguyen[0000-0002-9948-1048], Tung Le[0000-0002-9900-7047]

bhthong@fit.hcmus.edu.vn, ldtan@hcmiu.edu.vn, haint93@seoultech.ac.kr, bogdan.trawinski@pwr.edu.pl,
ntienhuy@fit.hcmus.edu.vn, lttung@fit.hcmus.edu.vn

SIMPLIFIED TITLE

Blockchain-based Decentralized Digital Content System

ABSTRACT

With the explosion of the big data era, storing and sharing data has become a necessity. However, sharing and using misleading data can lead to serious consequences. Therefore, it is necessary to develop safe frameworks and protocols for storing and sharing data among parties on the Internet. Unfortunately, the current systems struggle against the typical challenges to maintain confidence, integrity, and privacy. This paper proposes a blockchain network in a decentralized storage system that supports the management and shares digital content preserving its integrity and privacy. The system allows users to access authenticated data and develops protocols to protect the privacy of data owners. The detailed analysis and proof show that the proposed system is considered a promising solution for sharing and storing data in the big data era.

I INTRODUCTION

Along with the increase in the need to share data, the demand for data storage and retrieval is also increasing. However, storing data in traditional data centers is expensive and lacks scalability [1]. Cloud storage service is much cheaper and easier to expand the solution. It offers unlimited storage space, convenient sharing and accessing services, and offsite backup [9]. We can categorize storage services into centralized and decentralized storage systems based on the architecture. A centralized storage system has been proven to lack availability and data privacy.

Decentralized storage systems that utilize blockchain technology are promising solutions to overcome these limitations. Data uploaded to Decentralize storage systems are encrypted, split, and stored on a different node; only the Data Owner has the authorization to share or access. Since the decentralized storage system is a peer to-peer network, there is no administration required, then the risk of losing control of data to a higher authority is removed. Decentralized storage systems also possess the availability property. If there are unavailable nodes due to being attacked, other nodes can still maintain the whole system operation.

II STATE OF THE ART

II.1 Blockchain

A blockchain network consists of multiple nodes in which each minor node holds a copy of the ledger. Nodes in the blockchain network communicate together directly through a peer-to-peer network. Minor node's ledgers are synchronized through a consensus algorithm [6]. Two most well-known consensus algorithms can be mentioned are *Proof-of-work (PoW)* [5] and *Proof-of-stake (PoS)* [4].

II.2 Decentralized storage system

Decentralized storage systems solve the remaining problems of centralized storage systems, such as data availability and data privacy [10]. Storj [8] supports end-to-end encryption as well as shards and distributes data to nodes around the world for storage. Sia [7] divides the document into pieces and encrypts them before delivering each piece to the storage nodes through smart contracts. IPFS [2] is a peer-to-peer distributed storage system that builds on content-based addressing.

II.3 Decentralized data sharing

Huynh et al. [3] proposed a model consisting of data producing scheme, a data-storing scheme, and a data-sharing scheme that guarantees the privacy of the Data Owner and prevents fraud from sharing data.

III Original Contribution

We propose a data-storing model integrated with a blockchain network and decentralized storage system to form a secure storage system. We design a data-sharing method to manage the sharing process that maintains the privacy and integrity of shared data. Finally, we analyze the proposed model and method to show their strong properties.

IV Methodology

The involvement of the mentioned entities can be observed through multiple real-life data-sharing circumstances. For example, a candidate (DO) takes part in an IELTS test from an assessment center (DP). An employer (DU) needs the IELTS result (digital certificate generated by DP) for qualification evaluation. Though, the result must meet accuracy, integrity, and privacy. Another realistic situation is that a scientist (DU) requires data from sample(s) taken by hospitals (DP) from the patient(s) (DO) with a rare disease for research purposes. These sample data must also be highly correct and private. We introduce two phases: (1) Data sharing between the Digital content agency and Data Owner, and (2) Data sharing between the Data Owner and Data User.

V Results

The proposed model meets the following characteristics: Confidentially; Integrity; Non-repudiation; Scalability; Availability.

VI Evaluation

This research utilizes the mathematical evaluation via the blockchain characteristics.

VII Conclusions

This paper proposes a model integrated with a blockchain network and decentralized storage system that supports the management and shares digital content preserving its integrity and privacy. We analyze the model to prove its properties, such as confidentially, integrity, anonymity, non-repudiation, scalability, and availability.

References

[1] BARI, M. F., BOUTABA, R., ESTEVES, R., GRANVILLE, L. Z., PODLESNY, M., RABBANI, M. G., ZHANG, Q., AND ZHANI, M. F. Data center network virtualization: A survey. *IEEE Communications Surveys & Tutorials 15*, 2 (2013), 909–928.

[2] BENET, J. Ipfs-content addressed, versioned, p2p file system. *arXiv preprint arXiv:1407.3561* (2014).

[3] HUYNH, T. T., NGUYEN, T. D., HOANG, T., TRAN, L., AND CHOI, D. A reliability guaranteed solution for data storing and sharing. *IEEE Access 9* (2021), 108318–108328.

[4] KING, S., AND NADAL, S. Ppcoin: Peer-to-peer crypto-currency with proof-of-stake. *Self-Published Paper* (2012).

[5] NAKAMOTO, S. Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review* (2008), 21260.

[6] NGUYEN, G.-T., AND KIM, K. A survey about consensus algorithms used in blockchain. *Journal of Information processing systems 14*, 1 (2018), 101–128.

[7] VORICK, D., AND CHAMPINE, L. Sia: Simple decentralized storage. *Blockchain Lab White Paper* (2014).

[8] WILKINSON, S., BOSHEVSKI, T., BRANDOFF, J., AND BUTERIN, V. Storj a peer-to-peer cloud storage network, 2014.

[9] YANG, P., XIONG, N., AND REN, J. Data security and privacy protection for cloud storage: A survey. *IEEE Access 8* (2020), 131723–131740.

[10] ZHANG, C., SUN, J., ZHU, X., AND FANG, Y. Privacy and security for online social networks: challenges and opportunities. *IEEE network 24*, 4 (2010), 13–18.

DOI: https://doi.org/10.1007/978-3-031-21967-2_23

Keywords: combining classifiers, ensemble of classifiers, classifier selection, feature generation

Relearning ensemble selection based on new generated features

Robert Burduk[0000-0002-3506-6611]

robert.burduk@pwr.edu.pl

Simplified Title

Relearning ensemble selection based on new generated features

Abstract

The ensemble methods are meta-algorithms that combine several base machine learning techniques to increase the effectiveness of the classification. Many existing committees of classifiers use the classifier selection process to determine the optimal set of base classifiers. In this article, we propose the classifiers selection framework with relearning base classifiers. Additionally, we use in the proposed framework the newly generated features, which can be obtained after the relearning process. The proposed technique was compared with state-of-the-art ensemble methods using three benchmark datasets and one synthetic dataset. Four classification performance measures are used to evaluate the proposed method.

I Introduction

The purpose of the supervised classification is to assign to a recognized object a predefined class label using known features of this object. Therefore, the goal of the classification system is to map the feature space of the object into the space of class labels. This goal can be fulfilled using one classification model (base classifier), or a set of base models called an ensemble, committee of classifiers or multiple classifier system. The multiple classifier system is essentially composed of three stages: 1) generation, 2) selection and 3) aggregation or integration. The generation phase aims to create basic classification models, which are assumed to be diverse. In the selection phase, one classifier (the classifier selection) or a certain subset of classifiers is selected (the ensemble selection or ensemble pruning) learned at an earlier stage. The final effect of the integration stage is the class label, which is the final decision of the ensemble of classifiers.

II State of the Art

For about twenty years in the literature related to classification systems, there has been considered the problem of using more than one base classifier at the same time to decide whether an object belongs to a class label. During this period, multiple classifier systems were used in many practical aspects, and ensemble pruning significantly impacted the performance of recognition systems using an ensemble of classifiers. The taxonomy of the selection methods distinguishes static and dynamic selection. The static pruning process selects one or a certain subset of base classifiers that is invariable throughout all feature space or defined feature subspaces. In the case of dynamic selection, knowledge about the neighborhood of the newly classified object is used (most often defined by a fixed number of nearest neighbors) to determine one or a certain subset of the base classifiers for the classification of a new object.

III Original Contribution

The main objectives of this work can be summarized as follows:

- A proposal of a new relearning ensemble selection framework.
- A proposal of the feature generation that is used to learn second-level base classifiers used in ensemble selection process.
- An experimental setup to compare the proposed method with other multiple classifier system approaches using different classification performance measures.

IV Methodology

This study is experimental research. The experiments were conducted on four real datasets from UCI repository and one synthetic dataset. The experiments were conducted to compare the classification performance metrics of the proposed relearning ensemble selection based on new generated features algorithm with referential ensemble techniques: majority voting and sum rule without ensemble selection.

V Results

We treat our research as a preliminary study. The directions of further research include:

- evaluation on more datasets with performing statistical analysis,
- evaluating the computational complexity of the proposed algorithm,
- evaluation of larger groups base classifiers,
- validating the method against other classifier selection methods,
- development of new features dedicated to the semi-supervised problem,
- development of a new feature dedicated to the problem of decomposition of a multi-class task that eliminates the problem of an incompetent binary classifier,
- development of new features dedicated to imbalanced dataset problem.

VI Evaluation

A performance classification metric such as the area under the curve, the G-mean, the F-1 score and the Matthews correlation coefficient have been used for evaluating the proposed method.

VII Conclusions

The proposed method may constitute a new scientific research topic on classifier ensemble problems.

DOI: https://doi.org/10.1007/978-981-19-8234-7_5

Keywords: Vehicle retrieval, Nature language, Appearance, Motion

Exploring the effect of vehicle appearance and motion for natural language-based vehicle retrieval

Quang-Huy Can, Hong-Quan Nguyen, Thi-Ngoc-Diep Do, Hoai Phan, Thuy-Binh Nguyen, Thi Thanh Thuy Pham[0000-0003-3985-3599], Thanh-Hai Tran, Thi-Lan Le[0000-0001-9541-3905]

thanh-thuy.pham@mica.edu.vn

No simplified title given.

Abstract

Vehicle searching from videos by textual descriptions is one of the most important tasks in traffic management towards smart cities. This paper proposes a method for retrieval of vehicles using a natural language-based query. Our method consists of two main components of textual extractor based on Bi-LSTM and visual extractor using ResNet-50 model. Both components extract hidden features from different modalities and then match them in a common space. This end-to-end process tries to build a textual-visual alignment model that will be utilized for the search phase. Our particularities in this framework are two-fold. In the video stream, we evaluate in detail the role of vehicle appearance compared to its motion. In the textual stream, we apply back-translation systems to enrich the textual dataset for the training phase. Experiments are conducted on AI City Challenging, showing the efficiency of each contribution in the overall framework. It confirms that not only appearance but additional motion cues are promising for vehicle retrieval, which provides the results of MRR, Rank@5 and Rank@10 are 0.2333, 0.3587 and 0.4837, respectively.

I Introduction

Vehicle searching from videos by textual descriptions has emerged recently, thanks to its wide applications in different factual situations. In this work, we explore vehicle motion information in addition to vehicle appearance information for the vehicle searching problem based on textual descriptions. An appearance alignment model is proposed to learn the common space of the visual and textual features extracted from these information cues. In order to improve the model performance, we perform vehicle motion analysis and re-ranking step.

II State of the Art

In natural language-based object search, person search has recently attracted more attention. The standard method of person search is embedding the images and descriptions into shared feature space and then ranking the object images based on the cross-modal similarities. This approach can be applied for vehicle searching by exploiting not only the vehicle appearance information (color or shapes) but also the vehicle movement information. In vehicle search, motion information could become the most relevant cue to distinguish one vehicle from the others.

III Original Contribution

An overall framework is proposed for vehicle search based on natural language descriptions. A joint training strategy is applied by combining both instance loss and triple loss. In addition, vocabulary enrichment for textual embedding is done by translating back-and-forth vehicle descriptions from English to French. The achieved experimental results show that the proposed method can effectively exploit the vehicle appearance and motion information for vehicle searching by textual descriptions.

IV Methodology

IV.1 Appearance feature model

For the visual branch, a Resnet-50 backbone model is used on cropped vehicle images from full video frames. This model maps an appearance image to a 256 dimension vector. In the textual branch, the descriptions are preprocessed and embedded to high dimension space. Each textual vector is then fed into BiLSTM network to extract a feature vector of $(1024 \times 1 \times 1)$. Then an additional linear layer maps the output of BiLSTM to the feature vector with the same dimension as the visual feature vector.

IV.2 Visual-Textual Alignment

In order to learn the mapping of visual and feature space to a common space, a contrast learning method with triplet loss is utilized. Instance loss is also used for discriminative learning in each appearance model of image and text.

Triplet loss: Inspired by [2], a logistic form of the triplet loss function used is expressed as 1. For contrast learning, the positive samples (a pair of descriptions and images belonging to the same class) and the negative samples (the remaining pairs of other classes) are exploited.

$$L_{align} = \frac{1}{N}\sum_{i=1}^{N}\left\{log\left[1+e^{-\tau_p(cos\theta_i^+ - \alpha)}\right] + log\left[1+e^{\tau_n(cos\theta_i^- - \beta)}\right]\right\} \quad (1)$$

where $cos\theta_i^+$ and $cos\theta_i^-$ are cosine similarity of a positive pair and a negative pair, respectively. τ_p and τ_n are hyper-parameters that control the slope of gradient. α defines as the lower bound for a positive score and β defines as the upper bound for a negative score.

Instance loss: In this work, this loss is implemented for both separate models as a cross entropy loss with the number of class being the number of instances.

The final loss for the training cross-appearance model is the sum of instance loss and triplet loss.

IV.3 Movement Analysis

In this step, we classify the description and the vehicle track into six classes: ”Straight”, ”Left”, ”Right”, ”Stop”, ”Slow” and ”Other”. For the description, the verb phrases are extracted from the description by a built-in NLP tool and classified based on a pre-defined set of keywords. For the vehicle trajectory inspired by [1], we convert track positions into GPS values. The velocity and the turn angle of the vehicle are computed based on the speed of movement through each frame and the angle between the first and last vectors of trajectory in GPS coordinates.

Re-ranking the results of the appearance model: The ranking lists from the appearance model go through the motion analysis phase for re-ranking. In this phase, the vehicle motion is classified based on both the query description and the trajectory of tracks in the input list. If the class of the query and the class of the track's trajectory is not matched, the track's ranking will be decreased.

V Results

Experimental results show that thanks to the proposed visual-textual alignment framework, the vehicle retrieval performance obtained using a combination of appearance and motion cues is very promising.

VI Evaluation

In order to evaluate the performance of a vehicle, Mean Reciprocal Rank (MRR) and Recall@5, Recall@10 are employed as the main evaluation metric.

Three different experimental scenarios are set in this work to evaluate the performance of the proposed system. In the first scenario, the impact of the number of vehicle images and the descriptions on the quality of the ranking accuracy is evaluated. The experimental results show that the more images and descriptions used for vehicle representation, the higher accuracy obtained with MRR gain from 0.1848 to 0.2272.

In the second scenario, in comparison with the case of using the whole description and augmented textual data (Appearance with augmented descriptions), the case of using appearance noun phrase with augmented descriptions gains lower results at MRR, Rank@5, Rank@10 (0.1668 compared to 0.2195 at MRR, 0.2663 against 0.3098 at Rank@5, and 0.3804 versus 0.4728 at Rank@10).

Finally, in the third scenario, when using motion analysis, we obtain a light improvement with the gains of 0.0138, 0.0489 and 0.0109 for MRR, Rank@5 and Rank@10 respectively (compared to 0.2195 at MRR, 0.3098 at Rank@5, 0.4728 at Rank@10).

VII Conclusions

The proposed method can effectively exploit the vehicle appearance and motion information for vehicle searching by textual descriptions. The obtained results are promising but it still exists a large margin between the performance of the proposed method with others teams that participated in AI City challenge.

References

[1] Park, E.-J., Kim, H., Jeong, S., Kang, B., and Kwon, Y. Keyword-based vehicle retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (June 2021), pp. 4220–4227.

[2] Wang, Z., Fang, Z., Wang, J., and Yang, Y. Vitaa: Visual-textual attributes alignment in person search by natural language. In *European Conference on Computer Vision* (2020), Springer, pp. 402–420.

Link to arXiv: https://arXiv.org/abs/2104.12473
DOI: 10.1007/978-3-030-73280-6
Keywords: data mining, genetic algorithm, genetic-fuzzy mining, large itemsets, graphics processing unit

Using GPUs to Speed Up Genetic-Fuzzy Data Mining with Evaluation on All Large Itemsets

Chun-Hao Chen[0000-0002-1515-4243], Yu-Qi Huangc and Tzung-Pei Hongc[0000-0001-7305-6492]

chchen6814@gmail.com, cream08111230@gmail.com, tphong@nuk.edu.tw

Simplified Title

Using GPUs to Speed Up Genetic-Fuzzy Data Mining.

Abstract

In fuzzy data mining, the membership function significantly influences exploration performance. Therefore, some scholars have proposed genetic-fuzzy mining to determine a set of good membership functions for effectively mining fuzzy association rules. Some scholars proposed evaluating the membership functions using both the number of large 1-itemsets and the suitability of chromosomes. They only considered large 1-itemsets instead of all large item-sets because of the time-consuming problem. We analyzed the time-consuming reason and found that there are many independent calculations in the mining process. Given this, we adopt the GPU devices and propose a GPU-based mining algorithm with evaluation on all large itemsets to improve obtained membership functions and reduce time cost. Experimental results also show the efficiency of using GPUs on genetic-fuzzy data mining.

I Introduction

Due to the Apriori algorithm is only suitable for binary data, many scholars have proposed various algorithms to mine fuzzy association rules by adopting the concept of fuzzy sets [2]. They used the membership functions to convert quantitative values into fuzzy expressions and combined the fuzzy expression to mine fuzzy association rules. In addition to the predefined membership functions used by previous scholars, some researchers have also proposed to evolve the membership function through the genetic algorithm [3]. The time-consuming problems plague existing methods, so some scholars have proposed some strategies to accelerate the process of evolution and exploration. Hong et al. proposed a divide-and-conquer algorithm to speed up the mining process [3].

With the rise of the Graphics Processing Unit (GPU), we propose an algorithm to fasten the evolution process using GPUs, named the genetic-fuzzy mining (GFM) using GPUs on large all itemsets (GFM-GPU-LAll). It first randomly generates the initial population, configures the transaction database and chromosomes into GPUs, converts all the transactions into fuzzy values, and calculates the fuzzy support. The fuzzy support of each item is compared with the minimum support to get the number of large 1-itemsets (L1). In the previous research [1], it used the GPU to accelerate the exploration process of L1 and got a good performance. Therefore, in this paper, we want to complete the mining process to consider all large itemsets. After the L1 is obtained, all itemset candidates will be generated, calculated, and filtered. Finally, all frequent itemsets (LAll) are collected. When those operations are processed, better chromosomes will be selected into the next generation until the termination condition is met for the next iteration.

The massive and independent calculations are the reason that causes the time-consuming problem. Thus, we take those calculations into the GPUs and process these calculations through the massive parallel characteristic of the GPUs, thereby significantly reducing the execution time. Due to time constraints, we completed GFM-GPU-LAll within an acceptable time and successfully broke through all previously impossible items to explore. Through the experiments, the number of rules obtained by GFM-GPU-LAll is better than that by GFM-GPU-L1.

II State of the Art

The state of the art approach in literature is that one utilizes GPU to speed up the evolutionary process to find number of large 1-termsets in GFM.

III Original Contribution

The two main contributions of this work are listed as follows: The GFM-GPU-LAll approach is proposed to speed up the evolutionary process. Using GFM-GPU-LAll, number of derived fuzzy association rules is better than existing approach.

IV Methodology

Table 1: Pseudo code of the proposed approach.

```
Input: A transaction data from IBM generator; population size ($P_{size}$); the number of generations; minimum support (MS); random
    number d.
Output: A set of fuzzy association rules; the best membership functions set from evaluation.
1     Generate the initial population randomly with the number of $P_{size}$.
2     Calculate the fitness value of each chromosome.
    2.1     Transform the purchased quantity of item 1 to $q$ into three fuzzy values of low, mid and high, where the q is the maximum
            quantity of the purchased item. Organize those fuzzy values into an index table, called fuzzy regions conversion table
            (FRCT).
    2.2     Convert the purchased quantity of each item $I_{ij}$ is in each transaction data, where j is the amount of item from 1 to $n$, and $i$
            is the transaction data of 1 to $m$, into the fuzzy set through the FRCT represented as $fuzzy\ support = FS(I_j) = \frac{\Sigma_{i=1}^{m} FRCT(I_{ij})}{m}$,
    2.3     Compare the fuzzy support and minimum support to get the 1-item frequent itemset, call L1. $L_1 = \{I_j \mid FS(I_j) \geq MS\}$,
    2.4     Determine whether there are unprocessed chromosomes. If there are, go to Step 2-6; otherwise, go to Step 2-5.
    2.5     Go to Step 2-12. Output the large itemsets and fuzzy support.
    2.6     Determine whether there are unprocessed candidates. If there are, go to Step 2-7; otherwise, go to Step 2-4.
    2.7     Generate candidate itemsets from the large itemset string, allocate and copy the data to the GPUs.
    2.8     Scan the database and calculate the fuzzy support.
    2.9     Select the large itemset by minimum support.
    2.10    Copy the data to the CPU and concatenate it to the corresponding string.
    2.11    Go to the Step 2-6.
    2.12    Calculate the suitability and the fitness value of each chromosome ($C_p$).
3     Determine the termination condition is met or not. If not, continue the execution; otherwise, output the result.
4     Select the next generation.
5     Execute the MMA crossover on the population.
6     Execute the one-point mutation on the population and go to Step 2.
```

It first randomly generates an initial population and then calculates the fitness function of each chromosome. The fitness calculation includes many details, such as many independent calculations like fuzzy region conversion table, fuzzy support, and chromosome suitability. We all perform those operations on the GPU in parallel to reduce time costs. Since there is no fixed amount of data after L2, we allocate threads to each chromosome dynamically through a nested loop. After the LAll calculation is completed, the fitness value of each chromosome could be calculated and compared. The algorithm then outputs the best chromosome and mines the association rules through the best chromosome.

V Results

The proposed GFM-GPU-LAll approach provides a more efficient way to find membership functions and fuzzy association rules for analyzing the quantitative transaction dataset. In other words, it can be used for basket analysis.

VI Evaluation

Comparison result of the proposed approach (GPU-GFM-LAll) and the existing approach (GPU-GFM-L1) is shown in Table 2. The result shows that GPU-GFM-LAll is better than GPU-GFM-L1 in terms of number of rules.

Table 2. Number of extracted association rules by the two methods.

	GPU-GFM-L1	*GPU-GFM-LAll*
Minimum Support	*Association rules #*	*Association rules #*
0.00015	34	48
0.00020	20	27
0.00025	3	15

VII Conclusions

When users that want to find fuzzy association rules without the predefined membership functions in a limited time, the proposed GFM-GPU-LAll approach can be utilized to reach the goal by preparing the transaction dataset and the equipment that has GPU in it.

References

[1] C. H. Chen, Y. Q. Huang and T. P. Hong, "An effective approach for genetic-fuzzy mining using the graphics processing unit," *International Conference on Advances in Information Mining and Management*, pp. 7-11, 2021.

[2] T. P. Hong, C. S. Kuo and S. C. Chi, "Trade-off between computation time and number of rules for fuzzy mining from quantitative data," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* Vol. 9, No. 5, pp. 587-604, 2001.

[3] T. P. Hong, C. H. Chen, Y. C. Lee and Y. L. Wu, "Genetic fuzzy data mining with divide-and-conquer strategy," *IEEE Transactions on Evolutionary Computation*, Vol. 12, No. 2, pp. 252-265, 2008.

Link to arXiv: `https://arXiv.org/abs/2104.12473`
DOI: `https://doi.org/10.1007/978-3-031-21967-2_26`

Keywords: content-based video analysis, surveillance video, people counting, security control, missing persons, dominant color hue, silhouette size

Automatic Counting of People Entering and Leaving Based on Dominant Colors and People Silhouettes

Kazimierz Choroś 1[0000-0001-6969-976X], Maciej Uran
kazimierz.choros@pwr.edu.pl

Simplified Title

Automatic Counting of People Entering and Leaving Based on Dominant Colors and People Silhouettes.

Abstract

Counting of people crowd is an important process for video surveillance, anomaly warning, public security control, as well as ongoing protection of people and public facilities. People counting methods are also applied for controlling the numbers of people entering and leaving such places as a tourist bus, an office or an university building, a public building, a supermarket or shopping mall, a culture or sport center, etc. The problem arises when the number of exiting people is not equal to the number of entering people. How many people are missing and who is missing? The paper presents an approach for people counting and detection of missing persons. This approach includes two procedures. First, the dominant color of people detected on video was analyzed. Next, the silhouette sizes were used. Both procedures finally allow us to define specific features distinctive for missing people. The results of tests performed with video recordings of people entering and exiting through the door are promising. In this approach individual's identities are not registered; therefore, privacy violation is avoided.

I Introduction

People counting is useful in monitoring of people entering or leaving some places. In such a case the direction of a person movement is important, not only the appearance in a camera. In many applications these entering and leaving people are observed at the door or some gate. Such a situation is noticed in the case of counting people entering or leaving a tourist bus, an office or an university building, a public building, a supermarket or shopping mall, a culture or sport center, etc. It is important to know whether the number of leaving people is the same as the number of people who have entered. Then it is important to know which person or persons are missing.

In such surveillance applications the people counting is performed in a relatively narrow passage like a tourist bus door, when the faces are frequently not clearly visible. Moreover, it happens that it is not allowed to register individual's identities. So, in this paper a solution is proposed for this problem taking into account, that we cannot violate people privacy by collecting the personal information.

II State of the Art

Counting of people crowd from a single still image or video sequence frames has become an essential part of computer vision systems [1]. It is an important process for video surveillance, anomaly warning, security control, as well as ongoing protection of people and public facilities. Automatic crowd counting is frequently used to estimate the number of people participating in some outdoor events such as processions, political meetings or manifestations, protest events, subway stations, tourist attractions, open-air concerts, sports competitions, and in many other mass activities. In past many methods of crowd counting have been proposed based on the detection and calculation of objects (faces, heads, human bodies) in images or videos. Then these methods tried to estimate crowd density and on the basis of the crowd density estimate the number of people. In recent years methods based on convolutional neural network (CNN) models have become very popular. Crowd counting is applied for images presenting almost static crowd during the mass events such as music concerts, sports events, or meetings, etc. People counting is also used for people walking (pedestrian detection), for example during political demonstrations, or simply on streets or passages in a supermarket or shopping mall, etc. Then counting of people is also useful for the people entering and exiting buses [2] or some buildings, crossing the gate observed by a digital video camera, etc.

III Original Contribution

The proposed and experimentally examined approach is based on three steps: detection of people, people counting, and then people annotation using the dominant color and the silhouette size.

IV Methodology

The people detection method applied in the proposed solution is based on the algorithm of point tracking that is a centroid of a rectangular shape of an object detected by a convolutional neural network. People detection is achieved by the already trained model MobileNet-SSD. The software used in the experiments is based on the computer program code available in the Internet [3], however, it was modified to ensure not only counting people but also recognizing the direction of movement (entering or exiting). Moreover, to enable the detection of the dominant color and the silhouette size of a passing person, then used in people annotation process, a camera could not be placed vertically down at the door as in majority of people counting algorithms. If a camera is placed vertically down the people heads are clearly visible and people counting is much easier, however, the detection of the dominant color of their dresses is rather impossible. Also the dimensions of the human body including human height cannot be observed and measured.

The dominant color was found by analyzing the frequencies of colors of a detected person. The HSL (Hue Saturation Lightness) color model was used and three most frequently observed H values were registered for each detected person in an entering people registration list. In this list also the size of a detected person was saved. The dimensions of the rectangle were taken as the size of a person silhouette. The similarity of two persons on the basis of the dominant colors was calculated using the Euclidean measure.

V Results

The standard efficiency measures of detecting missing people based on dominant colors such as recall, precision, and F-measure achieve the level of 0.66 to 1.00 for recall, 0.50 to 0.75 for precision, and 0.60 to 0,86 for F-measure. In the case of the analyses of human silhouettes, these values ranged from 0.33 to 0.75, so they were much lower. The measuring of silhouette sizes should be improved in further experiments.

VI Evaluation

The results of tests performed with video recordings of people entering and exiting through the door are promising. However some unexpected adversities have been encountered. To achieve good results a camera should be placed not above the entry but in front of entering people. When detecting dominant colors of people entering and leaving the dominant colors of background are at the beginning recognized and then should not be included in the set of colors characterizing a given person. Unfortunately, the most dominant colors are not always exactly the same or even not evidently similar, although they were registered for the same person entering and then leaving the room, because the lighting conditions are not the same for the both events. It happens that they are significantly different. The best situation is when clothes are of expressive colors. In practice many people wear clothes that are usually quite muted colors, so, only color may be not significantly discriminative. Because it happens that one leaving person is similar to more than one entering person the matching procedure of entering people and those leaving may have a great influence on final results.

VII Conclusions

The approach presented in the paper can be applied for controlling the numbers of people entering and leaving such places as a tourist bus, an office or an university building, a public building, a supermarket or shopping mall, a culture or sport center, etc. The problem arises when the number of exiting people is not equal to the number of entering people. How many people are missing and who is missing? The paper presents an approach for people counting and detection of missing persons.

Both procedures, i.e. the analyses of dominant colors and people silhouettes, finally allow us to define specific features distinctive for missing people. The results of the tests performed with video recordings of people entering and exiting through the door are promising. What is very important in this approach is the fact that individual's identities are not registered; therefore, privacy violation is avoided. This is critical factor because the protection of personal data and the protection of privacy is a priority in the present day.

References

[1] Jingying, W.: A survey on crowd counting methods and datasets. In: Advances in Computer, Communication and Computational Sciences, Advances in Intelligent Systems and Computing, AISC 1158, Springer, Singapore, pp. 851–863 (2021).

[2] Zhao, J., Li, C., Xu, Z., Jiao, L., Zhao, Z., Wang, Z.: Detection of passenger flow on and off buses based on video images and YOLO algorithm. Multimedia Tools and Applications, 81(4), 4669–4692 (2022).

[3] Rosebrock A.: OpenCV people counter with Python, Object Tracking Tutorials (2021) https://pyimagesearch.com/2018/08/13/opencv-people-counter/

DOI: https://doi.org/10.1007/978-3-031-21743-2_20
Link to repository: https://github.com/jderen/Covid-19-GAN-Results

Employing Generative Adversarial Network in COVID-19 diagnosis

Jakub Dereń, Michał Woźniak[0000-0002-4224-6709]

jakub.deren97@gmail.com,michal.wozniak@pwr.edu.pl

Simplified Title

Employing Generative Adversarial Network in COVID-19 diagnosis

Abstract

In recent years, many papers and models have been developed to study the classification of X-ray images of lung diseases. The use of transfer learning, which allows using already trained network models for new problems, could allow for better results in the COVID-19 disease classification problem. However, at the beginning of the pandemic, there were not very large databases of SARS-CoV-2 positive patient images on which a network could perform learning. A solution to this problem could be a Generative Adversarial Network (GAN) algorithm to create new synthetic data indistinguishable from the real data using the available data set. It would allow training a network capable of performing classification with greater accuracy on a larger and more diverse number of training data. Obtaining such a tool could allow for more efficient research on how to solve the global COVID-19 pandemic problem. The research presented in this paper aims to investigate the impact of using a Generative Adversarial Network for COVID-19-related imaging diagnostics in the classification problem using transfer learning.

I Introduction

Since the first infections were reported in November 2019 in Wuhan, China, SARS-CoV-2 has spread worldwide, including reaching Europe in early 2020. The World Health Organization (WHO) has declared ca the center of the pandemic. From the very beginning, a critical element in the fight against the pandemic has become its correct diagnosis and the choice of the correct therapeutic approach for the individual course of the disease . Chest X-ray has been selected as one of the diagnostic tools, although it should be noted here that it is relatively insensitive in detecting pulmonary abnormalities in the early stages of the disease. However, it can be very useful to monitor the progression of pulmonary involvement in COVID-19, especially in patients with advanced disease.

To obtain a good quality diagnostic system, it was necessary to collect a sufficiently large database containing images of COVID-19 sufferers and healthy patients. At the beginning of the pandemic, the main problem was obtaining images of infected individuals, so the so-called problem of building a diagnostic system based on imbalanced data was encountered., i.e., the disproportion between the number of observations from different classes. One of the techniques used to balance the distributions is data augmentation, which involves adding synthetic minority class objects to the data set. Among the various techniques that can be chosen to implement such a process, it was decided to use a generative adversarial network (GAN). It allows two competing generator and discriminator modules to generate synthetic chest x-ray images of individuals infected with Sars-Cov-2. With such a technique, a sufficient volume of images can be obtained to adequately train a deep network allowing diagnosis in the indicated range.

The main goal of this work is to evaluate whether the use of Generative Adversarial Network for generating synthetic images and transfer learning techniques is helpful in the problem of learning deep models dedicated to the COVID-19 classification problem.

II Original Contribution

We how the data was prepared, how synthetic images were generated, and how transfer learning techniques were used in the task of classifying lung images of COVID-19 patients.

III Methodology

The pipeline of the proposed framework is presented in Fig. 1.

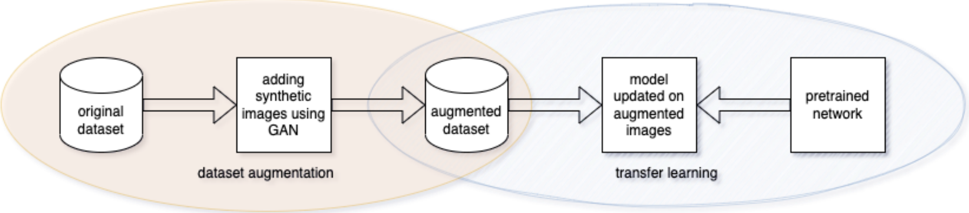


Figure 1: The pipeline of the proposed framework.

IV Results

Goal. The experimental study aimed to answer whether the augmentation of minority class images using GAN and transfer learning can improve the quality of x-ray image classification toward COVID-19 diagnosis.
Used datasets. The collection *COVID-19 Radiography Database* from the kaggle repository was used to investigate the posed thesis. It consists of several thousand X-ray images of 299x299 pixels, divided into four classes: (i) COVID (3616 images), (ii) Lung Opacity - (6012 images), (iii) Normal - (10192 images), (iv) Viral Pneumonia (1345 images).

As the aim of this study was to propose an effective tool for diagnosis of covid-19 patients, hence only images of healthy Normal) subjects and those diagnosed with covid-19 (COVID) were selected from the above database. This choice is also justified by the need for extensive computational resources when performing the GAN model learning process on high-resolution images. For this reason, the number of images was limited to 1000 COVID class images and 2819 Normal class images (the proportions of the number of images present in the selected data set were preserved). [1].

Implementation and reproducibility. Complete source code, sufficient to repeat the experiments, was made available at public avalialble repository[2]. The proposed algorithm, as well as the experiments described in this work, were implemented in the Python programming language. We also used *Keras* library.

IV.1 Using GAN to generate synthetic images

The resulting dataset was imbalanced, i.e., the imbalance ratio was ca. 1:3. Hence it was decided to choose a method how to balance it. One of the popular data augmentation methods is oversampling, i.e., adding the number of minority class objects (COVID) to the original set equals the number of majority class objects (Normal). We may distinguish two main approaches, i.e., random oversampling or generation of synthetic objects. We decided to generate synthetic minority class X-ray images from the used dataset to provide more diversity in the learning set One of the problems that had to be solved was the appropriate choice of generation parameters. Since it was impossible to determine in advance what number of epochs was needed to generate realistic images of covid-19 patients, the subjective expert evaluation was used every ten epochs to determine whether the generated images were already of satisfactory quality.

According to expert judgment, it was considered that satisfactory quality was obtained after 180 epochs, and these images were used in the next step to learn the classifier.

IV.2 Transfer Learning

The final step of the research was to perform classification with transfer learning, using real and synthetic data. For this purpose, the following research method was proposed: (i) performing a training of the transfer learning model with real data only to proceed with the classification; (ii) performing a training of the transfer learning model with real and generated data, to proceed with the classification; (iii) comparison of the results obtained from the two experiments. Two pre-trained models were selected to perform transfer learning: VGG16 and VGG19.

V Conclusions

The purpose of the study was to answer the question of whether the use of transfer learning and data augmentation based on the generation of synthetic images using GAN can improve the quality of classification. Based on the results obtained, the answer to both questions is positive. It was possible to obtain synthetic images that resembled real images in expert opinion, and adding them to the dataset resulted in an increase in classification quality for both trained networks. However, we have to emphasize that it required more epochs to obtain the final model.

[1] `https://www.kaggle.com/tawsifurrahman/covid19-radiography-database`, June 2021
[2] `https://github.com/jderen/Covid-19-GAN-Results`

DOI: `https://doi.org/10.1007/978-3-031-21743-2_40`

Complement Naive Bayes Classifier for Sentiment Analysis of Internet Movie Database

Christine Dewi$^{12*[0000-0002-1284-234X]}$, and Rung-Ching Chen$^{1[0000-0001-7621-1988]}$

christine.dewi@uksw.edu, crching@cyut.edu.tw

Simplified Title

Complement Naive Bayes Classifier for Sentiment Analysis of Internet Movie Database

Abstract

Sentiment analysis (SA), often known as opinion mining, is the subjective examination of a written text. Moreover, SA is a critical technique in today's artificial intelligence (AI) field for extracting emotional information from huge amounts of data. The study is based on the Internet Movie Database (IMDB) dataset, which comprises movie reviews and the positive or negative labels that relate to them. Our research experiment's objective is to identify the model with the best accuracy and the most generality. Text preprocessing is the first and most critical phase in a Natural Language Processing (NLP) system since it significantly impacts the overall accuracy of the classification algorithms. The experiment implements the Term Frequency-Inverse Document Frequency model (TFIDF) to feature selection and extractions. The following classifiers are used in this work: Linear Model and Naïve Bayes. Besides, we explore the possible options of loss functions such as *square_hinge*, *huber*, *modified_huber*, *log*, *epsilon_insensitive*, *perceptron*, and *modified_huber.* ComplementNB achieves the highest accuracy, 75.13%, for both classification reports based on our experiment result.

I Introduction

Sentiment analysis (SA) is a critical technique in today's field of artificial intelligence (AI) for extracting emotional information from vast amounts of data [1]. Sentiment analysis has improved over the years using several machine learning and dictionary-based algorithms to improve accuracy. In the industrial context, businesses primarily use SA to collect and assess client feedback. The fields of natural language processing (NLP) and SA are inextricably linked. The Internet Movie Database (IMDB) dataset comprises 50,000 reviews, equally split between 25,000 train and 25,000 test reviews. Positive or negative movie reviews are categorized, and the task is to guess the sentiment of an unseen review. The sentiment of a movie review is usually associated with a different rating, which can be used for classification dilemmas. It can be used as a reference instrument for movie preference.

II State of the Art

Researchers have been working on various recommendation algorithms based on text data supplied by internet users over the last couple of years. Zirn et al. [2] developed a completely automated system for fine-grained SA at the sub-sentence level, incorporating several sentiment lexicons, neighborhood links, and discourse linkages. Appel et al. [3] established a hybrid strategy based on ambiguity management, semantic rules, and a sentiment lexicon using Twitter sentiment and movie review datasets. The authors evaluated the performance of their suggested hybrid system to that of conventional supervised algorithms such as Naive Bayes (NB) and Maximum Entropy (ME). The recommended approach outperforms supervised methods in terms of precision and accuracy. Naive Bayes (NB) is a well-known classification technique in data mining.

III Original Contribution

The following are the significant contributions of this work: (1). The study is based on the IMDB dataset, which comprises movie reviews and the positive or negative labels that relate to them. (2). The goal of our study experiment is to find the model with the highest accuracy and the greatest generality. (3). The following classifiers are used in this work: Linear Model and Naïve Bayes. Different techniques, including Count Vectorizer, Term Frequency-Inverse Document Frequency model (TFIDF)Vectorizer, minimum-maximum number of words, and max features, are implemented. We can observe that the complement naive Bayes model performs well compared to other methods.

IV Methodology

The following methods are used to resolve the classification problem: Logistic Regression (LR), Bernoulli Naïve Bayes (BernoulliNB), Complement Naïve Bayes (ComplementNB), and Linear Support Vector Machine

(Linear SVM) with SGDClassifier and different loss type (*square_hinge, huber, modified_huber, log, epsilon_insensitive, perceptron, modified_huber*). Finally, the classification procedure is carried out, sentiment analysis will show the positive or negative result, and the methods used are analyzed.

V Results

ComplementNB achieves the highest accuracy, 75.13%, for both classification reports. ComplementNB was created to address the "severe assumptions" imposed by the ordinary Multinomial Naive Bayes classifier. It is especially well-suited for unbalanced data sets. Naive Bayes is often used as a baseline in text classification because it is fast and easy to implement. Therefore, we apply the sklearn class as follows: *sklearn.naive_bayes.ComplementNB (*, alpha=1.0, fit_prior=True, class_prior=None, norm=False).* On the other hand, BernoulliNB only achieves 54.71% accuracy for both classification reports.

VI Evaluation

Some advantages of the Naive Bayes Algorithm are: (1). The NB method is efficient and can save a significant amount of time. (2). NB is well-suited for tackling challenges involving multi-class prediction. (3). If the premise of feature independence remains true, NB can outperform other models and require less training data. (4). NB is more appropriate to categories than to numerical input variables.

VII Conclusions

Based on our experiment result, we can conclude: (1) ComplementNB achieves the highest accuracy, 75.13%, for both classification reports. (2) In the group of Linear classifiers SVM with SGD training, SGDClassifiers with loss functions "log" achieve the best performance 75.07% accuracy in the classification report for the bag of words. (3) In the classification report for TFIDF features, SGDClassifiers with loss function "perceptron" exhibit the optimum accuracy of 74.99%. We will explore the other loss function in future research to increase our performance results. We also want to combine sentiment analysis with Shapley Additive Explanations (SHAP) for explainable artificial intelligence (XAI).

References

1. Kumar, S., Gahalawat, M., Roy, P.P., Dogra, D.P., Kim, B.G.: Exploring impact of age and gender on sentiment analysis using machine learning. Electron. 9, (2020). https://doi.org/10.3390/electronics9020374.
2. Zirn, C., Niepert, M., Strube, Heiner Stuckenschmidt, M.: Fine-Grained Sentiment Analysis with Structural Features. Proc. 5th Int. Jt. Conf. Nat. Lang. Process. (2011).
3. Appel, O., Chiclana, F., Carter, J., Fujita, H.: Successes and challenges in developing a hybrid approach to sentiment analysis. Appl. Intell. 48, (2018). https://doi.org/10.1007/s10489-017-0966-4.

DOI: `https://doi.org/10.1007/978-981-19-8234-7_6`

Shapley Additive Explanations for Text Classification and Sentiment Analysis of Internet Movie Database

Christine Dewi[0000-0002-1284-234X], Bing-Jung Tsai, and Rung-Ching Chen[0000-0001-7621-1988]

christine.dewi@uksw.edu, s11014617@gm.cyut.edu.tw, crching@cyut.edu.tw

Simplified Title

Shapley Additive Explanations for Text Classification and Sentiment Analysis of Internet Movie Database

Abstract

The application of Artificial Intelligence (AI) is increasing in areas like sentiment analysis and natural language processing (NLP). Automatic sentiment analysis provides a guide to capture the user emotions and classify the reviews into positive or negative. One of the challenges of using general lexicon analysis is its insensitivity to all domains. There arises a need for the interpretability of the output predicted from the AI sentiment analysis models. This paper developed a Shapley Additive Explanations for Text Classification (SHAP) based model to classify the user opinion texts into negative or positive labels. Our sentiment analysis model is evaluated on the Internet Movie Database (IMDB) datasets which have rich vocabulary and coherence of the textual data. Results showed that the model predicted 89% of the user reviews correctly. This model is very flexible for extending it to unlabeled data.

I Introduction

Sentiment analysis (SA) is a rapidly expanding field of research due to the massive growth of digital information. In the field of artificial intelligence (AI), SA is a crucial tool for extracting emotional information from huge amounts of data [1]. In the Internet Movie Database (IMDB) dataset, there are equal numbers of 25,000 train and 25,000 test reviews, for a total of 50,000 reviews. Whether a film review is positive or negative, users must try to deduce the author's intent from the context in which it was written. The sentiment of a movie review is usually associated with a different rating, which can be used for classification dilemmas. It can be used as a reference instrument for movie preference. Shapley Additive Explanations *(*SHAP) is a highly valuable technique for dealing with one of the most difficulties associated with machine learning models: interpretability and explaining ability. It can be utilized both during the development and verification stages. SHAP can aid in the creation of an ML model by detecting outliers and missing values, segmenting data, selecting variables, and doing variable interaction analysis. In the validation governance stage, SHAP provides a clear means of explaining the interactions and consequences of the components.

II State of the Art

Moreover, there is various related research about SA, including Taboada et al. [2] proposed an approach to extract the sentiments from text using a semantic orientation calculator and assign positive or negative labels based on the polarity and strength. Bandhakavi *et al.* [3] proposed a generative unigram mixture model to extract the emotions of weekly labelled data using the word emotion association. SHAP is the state-of-the-art Machine Learning explanation ability, and it is available for free. Developed by Lundberg and Lee in 2017 [4], this method provides a great approach to reverse-engineer the output of any prediction algorithm. The purpose of SHAP is to provide an explanation for the prediction of an instance *x* by calculating the contribution of each characteristic to the prediction of the instance *x*.

III Original Contribution

The following are the significant contributions of this work: (1). The study is based on the IMDB dataset, which comprises movie reviews and the positive or negative labels that relate to them. (2). The goal of our study experiment is to do text classification and sentiment analysis by Shapley Additive Explanations (SHAP).

IV Methodology

SHAP Values dissect a prediction to reveal the relative importance of its various components. how much each player contributed to the success of a collaborative game can be determined using this method. To put it another way, each SHAP value measures how much of a role each feature in our model plays in our prediction. In our experiment, we use 100 data of IMDB and implement SHAP for text classification and sentiment analysis.

V RESULTS

The statistical performance of SHAP is described in Table 1. From the 100 data of IMDB dataset the sentiment analysis by SHAP got 89% accuracy. The total of positive data is 31 and Negative 58 for class "Yes". In other hand, class "No" have 4 positive sentiments and 7 negative sentiments.

Table 1. Statistic Performance of SHAP.

Data	Yes %	No %	Total
Positive	31	4	35
Negative	58	7	65
Total	89	11	100

VI EVALUATION

Some great benefits of the explanation by SHAP are as follows: (1). At the global level, the SHAP values together contribute to the interpretation and understanding of the model. Specifically, they demonstrate how much each predictor contributes to the target variable, either favorably or negatively. (2) At the local level, each observation receives its own set of SHAP values, which are then combined (one for each predictor). Transparency is substantially increased because of this, as contributions to predictions are shown on a case-by-case basis, something that standard variable significance algorithms are unable to achieve. There are some limitation of SHAP as follows: (1) An expected value is inferred by SHAP using data from the surrounding environment. (2) Predictions are explained by SHAP by comparing them to the training dataset's expected value.

VII CONCLUSIONS

In this paper, we introduced SHAP based movie sentiment analysis model for user comments. The model learns the word representation and gives insight into the captured sentiments, classifying the comments into positive and negative categories based on the text analysis. IMDB dataset is used for evaluating the model. The reviews are labelled with a positive/negative rating. A total of 100 samples with 35 positive and 65 negative ratings are evaluated using the model. The model correctly predicted 31 positive and 58 negative ratings. The user comments largely influenced the text perturbations and correlated well with our model analysis. In the future, this model can be extended to characterize a wide variety of text data, and thus can be applied in various areas of sentiment analysis.

REFERENCES

1. Kumar, S., Gahalawat, M., Roy, P.P., Dogra, D.P., Kim, B.G.: Exploring impact of age and gender on sentiment analysis using machine learning. Electron. 9, (2020). https://doi.org/10.3390/electronics9020374.
2. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-Based Methods for Sentiment Analysis DRAFT DRAFT DRAFT! Comput. Linguist. 37, (2011).
3. Bandhakavi, A., Wiratunga, N., Padmanabhan, D., Massie, S.: Lexicon based feature extraction for emotion text classification. Pattern Recognit. Lett. 93, (2017). https://doi.org/10.1016/j.patrec.2016.12.009.
4. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems. pp. 4766–4775 (2017).

Link to arXiv: `https://arXiv.org/abs/2104.12473`
DOI: `https://doi.org/10.1007/978-3-031-21743-2_53`

Keywords: Recommender system, user polarization, personalization, user preference, alternating least square

Polarization in Personalized Recommendations: Balancing Safety and Accuracy

Zakaria El-Moutaouakkil, Mohamed Lechiakh, Alexandre Maurer

{zakaria.elmoutaouakkil,mohamed.lechiakh,alexandre.maurer}@um6p.ma

Simplified Title

Polarization in Personalized Recommendations

Abstract

Recommender systems are beneficial to both service providers and users, but may also have unintended side effects. In this paper, we address the societal impact of recommender systemss on user polarization resulting from over-personalized recommendations. First, we model the user preference gap (uPG) as the distance between two timeseries, representing the user's consumption per content categories. Second, we map the uPG score onto the not-yet-rated items, and consider two approaches to minimize uPG per user, when they go beyond a certain threshold. In particular, we propose post and pre-constrained versions of the alternating least square algorithm that reduce the user's uPG score, hence avoiding to reinforce her polarization over time. Interestingly, these newly constrained algorithms still maintain a high level of recommendation accuracy. Our simulation results, derived from real datasets, show that our solutions enable personalization with a reasonable level of polarization.

I Introduction

Recommendation systems have recently attracted high interest from academic and industrial actors. This is mainly due to their impact in guiding user behavior through proposed services and products, which often result in significant business profit and user satisfaction. However, due to over-personalization and algorithmic bias, these algorithms are vulnerable to unfavorable societal side effects, raising serious safety concerns regarding these systems. In this paper, we focus on mitigating users' polarization by controlling their preference gap (uPG) while interacting with the RS, and propose a simple way to avoid its widening throughout the user-RS interaction process.

II State of the Art

A large number of recent research works have considered users' latent vectors derived from the user-item matrix factorization for encoding the user's preference toward items. As a result, they proposed different approaches for limiting their amplification over time as a potential solution to mitigate user polarization in RSs.

II.1 Item Polarization Score and User Bias Amplification

Some research works assigned polarization scores to items based on their ratings in order to detect the most polarizing ones among the top-N recommendations. They consider that high polarization scores are associated with items receiving opposite ratings and thus having a U-shaped rating histogram. However, these techniques fail when applied to items that do not have a U-shaped rating histogram and are still considered polarizing. Some other research works looked at users' bias as a potential cause of polarization behavior that the recommender system then amplifies through their interaction loop with the user. Several research works tried to ensure diversity, serendipity, and novelty of recommendations to decrease their polarizing effect. However, doing so can result in what is called "exposure bias", and reinforce user polarization even further.

II.2 RS Algorithmic Bias Amplification

The RS algorithm itself can be polarizing, as it encourages users, through recommendations, to adopt the views and opinions of like-minded users, and keeps them away from the views of opposite-minded users. This may result in the formation of echo-chambers, consisting of isolated and polarized sets of users sharing similar preferences, opinions and/or views. Therefore, item- and user-based polarization detection techniques should be combined with a non-polarizing RS algorithm too, so that the user preference is not significantly amplified over time.

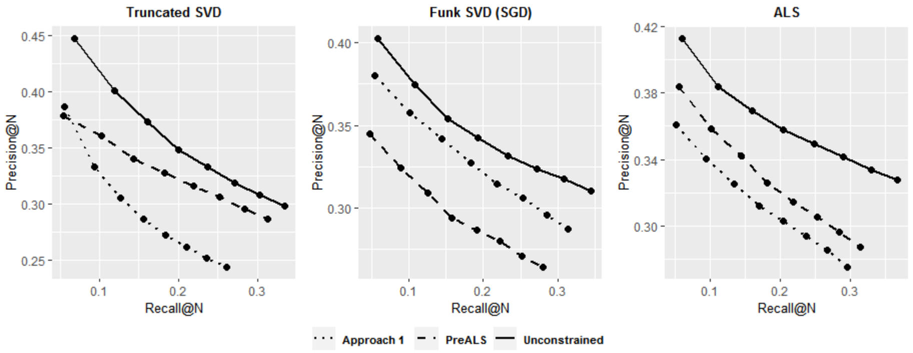


Figure 1: Precision vs Recall of our proposed polarization-constrained RS algorithms, compared to some unconstrained baselines: Truncated SVD, Funk SVD and ALS.

III Original Contribution

In this work, we introduce two methods for minimizing user polarization in RSs. To this end, we model the user preference gap (uPG) between a predefined set of item categories. Then, by controlling this gap, the RS balances between accuracy and user polarization reduction in its list of top-N recommendations.

IV Methodology

To minimize user polarization in RS, we propose a post-constrained (i.e., applicable to any unconstrained RS algorithm) and pre-constrained alternating least square (PreALS) algorithms, that minimize the uPG while maintaining an acceptable accuracy compared to some unconstrained RS algorithm baselines.

IV.1 Approach 1: Post-constrained uPG

Based on our modeling of uPG, our first low-complexity yet effective methodology to minimize user polarization is referred to as Approach 1. It relies on post-constraining the uPG score in the top-N list of recommendations. In particular, we directly operate on the output data of the RS model, and favor top-N items that reduce uPG scores if consumed by the user.

IV.2 PreALS Algorithm

Our second simple (but more involved) approach to minimize uPG scores is to consider that the not-yet-rated items (if consumed by the user) leading to high uPG scores have virtually been rated by the user, and fictitiously assign them a rating with a minimal value. Consequently, the RS model treats these ratings as training data, and will not recommend them to the active user. Obviously, assigning a low rating to all not-yet-rated items with high uPG scores may result in accuracy loss for the RS, because not all of these items will necessarily be among the top-N.

V Results

Our proposed methods reveal that ensuring the safety of recommender systems does not necessarily compromise their accuracy performance. As shown in Fig. 1, the Recall and Precision performance of our proposed polarization-constrained methods (Approach 1 and PreALS) are compared against the unconstrained RS algorithms Truncated SVD, Funk SVD, and ALS. Interestingly, Funk SVD and PreALS (middle of Fig. 1) are close in terms of performance, and yet, PreALS is safe in terms of user polarization.

VI Evaluation

Our results are evaluated on the MovieLens 100k dataset for N ranging from $N = 10$ to $N = 80$, with an increment of 10; k and the number of iterations of SGD and ALS are set to 10. For our PreALS algorithm, the value of the pre-constrained items' ratings is set to 3.5.

VII Conclusions

Our work addressed the problem of mitigating user preference amplification in RSs, through new modeling and measurement of its evolution. To minimize its effect in enforcing user polarization, we first introduced a simple uPG post-constraining approach and then our pre-constrained PreALS RS algorithm.

References

[1] Celis, L. E., Kapoor, S., Salehi, F., and Vishnoi, N. Controlling polarization in personalization: An algorithmic framework. In *Proceedings of ACM FAT* (2019).

DOI: https://doi.org/10.1007/978-981-19-8234-7_7

Keywords: Resource Prediction, Light Gradient Boosting Machine (LightGBM), SHapley Additive exPlanations (SHAP), k-nearest neighbor (KNN)

Explaining Predictive Scheduling in Cloud

Muhammad Fahimullah[0000-0003-1307-4570], Rohit Gupta, Shohreh Ahvar, Maria Trocan,

muhammad.fahimullah@ext.isep.fr, bindasrohit161@gmail.com, shohreh.ahvar@isep.fr, maria.trocan@isep.fr

SIMPLIFIED TITLE

Using Explainable Artificial Intelligence (AI) such as SHapley Additive exPlanations (SHAP) explanations for interpolation of missing values for predicting efficient scheduling in cloud computing.

ABSTRACT

The importance of cloud computing has been rapidly growing due to the increasing number of users' requests for diverse sets of resources. Although clouds have rich resources to handle these incoming requests, under or over-provisioning of resources can lead to failure. Therefore, it is important to provision cloud resources appropriately. Machine-learning-based techniques have been proven to be effective in the management of resources along with maintaining a Service Level Agreement (SLA). These techniques require complete data to produce better prediction results. In practice, it may happen that the data is incomplete and data with more missing attribute values can negatively affect the outcome of the predictions. Therefore, interpolation of missing attribute values is crucial for better predictions. However, the existing methods for interpolation of missing attribute values are heavy in terms of computation. This paper first predicts resource usage in terms of CPU by applying the lightGBM model to a real dataset. Furthermore, using the explanations of SHapley Additive exPlanations (SHAP) in combination with the K-Nearest Neighbor (KNN) to interpolate missing values in the dataset for CPU usage prediction. The experimental results show that SHAP explanations can be helpful for cloud providers in the selection of important features for the interpolation of missing values. This SHAP-based interpolation results in lower computational time along with acceptable accuracy in comparison with KNN-based interpolation.

I INTRODUCTION

For prediction, the datasets having missing attribute values of more than 15% can negatively affect the outcome of the predictions [1]. Therefore, it is essential to have complete data for the prediction of optimal resources in cloud computing. However, it is difficult to have complete data in practice [2]. Therefore, interpolation of missing values is required to achieve better prediction results. In this work, we considered the publicly available dataset from Delft University of technology GWA-T-12 Bitbrains (tudelft.nl) that provides the performance matrices of VMs. To tackle with missing attribute values problem for prediction, first, we created missing values in a complete dataset then we consulted SHAP explanations and K-Nearest Neighbor (KNN) for interpolation of missing values. Lastly, we used the predictions of interpolated data values with the predictions of the original completed dataset for comparisons.

II STATE OF THE ART

SHAP is one of the methods used in literature for complex model explanations from local and global perspectives. SHAP explanations have been used to explain black-box models such as Random Forest and Convolutional Neural Network (CNN) to get important features based on SHAP values and also the impact of features on workload prediction output. These explanations can be helpful to practitioners in selecting the most important features for resource management. The existing studies uses SHAP only for explaining complex models. However, SHAP explanation provides important features, which can be used for the interpolation of missing attribute values to reduce the computational time of interpolation while achieving acceptable prediction accuracy.

III ORIGINAL CONTRIBUTION

The existing studies considered SHAP for explanations of complex models. However, in this work, we considered SHAP explanations for the interpolation of missing values to reduce interpolation computational time along with acceptable prediction accuracy from the interpolation. Therefore, we used the SHAP explanations along with K-Nearest Neighbour (KNN) for reducing interpolation computational time.

Table 1: Experimental results

Actual		RMSE: 805.15 Test Data Accuracy: 0.9396								
Percentage of Null values		10 %			30 %			50 %		
K-nearest neighbors		K=3	K=5	K=7	K=3	K=5	K=7	K=3	K=5	K=7
KNN	RMSE:	851.33	856.07	869.78	966.59	952.73	975.49	1042.74	1022.83	1074.18
	Accuracy:	0.9324	0.9317	0.9295	0.9129	0.9154	0.91136	0.8987	0.9025	0.8925
	Computational time:	1h 9min 26s	1h 16min 41s	1h 14min 14s	3h 15min 58s	3h 19min 58s	3h 14min 20s	4h 7min 54s	4h 20min 11s	4h 38min 24s
KNN with Most Important Attribute	RMSE:	907.00	901.40	906.34	1044.82	1095.30	1129.89	1159.38	1169.75	1189.46
	Accuracy:	0.9233	0.9243	0.9234	0.8983	0.8882	0.8810	0.8748	0.8725	0.8682
	Computational time:	1h 5min 1s	1h 11min 59s	1h 17min 52s	2h 46min 8s	3h 1min 14s	3h 3min 6s	4h 39s	4h 13min 21s	4h 15min 57s
KNN with Multiple Attributes	RMSE:	874.38	893.33	898.37	1017.24	1037.77	1039.21	1116.30	1136.53	1156.82
	Accuracy:	0.9287	0.9256	0.9248	0.9036	0.8996	0.8994	0.8839	0.8796	0.8753
	Computational time:	1h 10min 12s	1h 14min 41s	1h 38min 32s	2h 52min 55s	3h 16min 33s	3h 26min 33s	4h 5min 9s	4h 19min 16s	4h 30min 12s
KNN with Least Important Attribute	RMSE:	958.88	922.17	953.40	1175.37	1211.14	1195.62	1330.29	1394.31	1404.14
	Accuracy:	0.9143	0.9207	0.9153	0.8713	0.8633	0.8668	0.8351	0.8189	0.8115
	Computational time:	1h 5min	1h 7min 21s	1h 9min 54s	3h 23s	3h 4min 58s	3h 10min 5s	3h 58min 33s	4h 10min 20s	4h 12min 50s

IV Methodology

We performed several experiments using the following steps as shown in Table 1.

– Using publicly available dataset from Delft University of technology GWA-T-12 Bitbrains (tudelft.nl).

– Using SHAP for model explanations and extracting important features.

– Using KNN and SHAP explanations for interpolation of missing values.

– Using Light Gradient Boosting Machine (LightGBM) model for predicting CPU resource with the completed dataset, KNN interpolated dataset and SHAP-based KNN interpolated dataset.

V Results

The results in Table 1 show that interpolation of missing values with one or fewer features compared to a complete set of features leads to less computational time. The proposed approach can be helpful to practitioners in the selection of important attributes for the interpolation of missing values in test datasets to reduce computational time along with achieving acceptable prediction accuracy.

VI Evaluation

In this work, we performed experiments in multiple parts. First, the dataset is trained on the LightGBM model. Secondly, we used KNN and SHAP explanations based KNN for the interpolation of null values. Lastly, the results of all the experiments were compared as shown in Table 1. Using the most important features for interpolation suggested by SHAP compared to using all the features in KNN not only provides acceptable accuracy but using the SHAP important features greatly reduces the computational time of interpolation.

VII Conclusions

The approach can be used by practitioners for the interpolation of datasets with missing values with SHAP-provided important features to reduce computational time along with acceptable prediction accuracy compared to other interpolation methods.

References

[1] Acuna, E., and Rodriguez, C. The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications*. Springer, 2004, pp. 639–647.

[2] Tsai, C.-F., and Hu, Y.-H. Empirical comparison of supervised learning techniques for missing value imputation. *Knowledge and Information Systems* (2022), 1–29.

DOI: https://doi.org/10.1007/978-3-031-21967-2_4

Keywords: Lung cancer, machine learning, metastasis, radiomic features

Machine learning approach to predict metastasis in lung cancer based on radiomic features

Krzysztof Fujarewicz[0000-0002-1837-6466], Agata Wilk[0000-0001-7554-1803], Damian Borys[0000-0003-0229-2601], Andrea d'Amico[0000-0003-4632-2139], Rafał Suwiński[0000-0002-3895-7938], Andrzej Świerniak[0000-0002-5698-5721]

krzysztof.fujarewicz@polsl.pl

Simplified Title

Radiomics-based prediction of metastasis

Abstract

Lung cancer is the most common cause of cancer-related death worldwide. One of the most significant negative prognostic factors is the occurrence of metastasis. Recently, one of the promising way to diagnose cancer samples is to use the image data (PET, CT etc.) and calculated on the basis of these images so called radiomic features. In this paper we present the attempt to use the radiomic features to predict the metastasis for lung cancer patients. We applied and compared three feature selection methods and two classification methods: logistic regression and support vector machines. The obtained accuracy of the best classifier confirms the potential of the radiomic data in prediction of metastasis in lung cancer.

I Introduction

Lung cancer is the most deadly malignancy, causing almost 20% of cancer-related deaths worldwide. The main cause of cancer death is associated with metastases, which are mainly incurable. We present the attempt to use radiomic features to predict metastasis for lung cancer patients.

II State of the Art

Medical imaging has been the subject of many cancer-related studies. It is considered a great potential source of data for diagnostic and prognostic classifiers and models thanks to the low invasiveness of their acquisition. Radiomics proved to be a reasonable compromise between extracting as much information as possible from the images and preventing model overfitting for usually limited cancer patient cohorts [1]. Radiomic features were demonstrated to be effective for detection and prediction of local metastasis, and even in some cases, distant metastasis.

III Original Contribution

Most studies focus on binary prediction of whether or not metastasis will occur. However, lung cancer is so invasive that the eventual appearance of metastases is almost certain. Therefore we also predicted the time of metastasis onset, in particular whether it will appear within a year or not. Additionally, we provide a detailed discussion of challenges related to use of radiomics in a machine learning context.

IV Methodology

We extracted 105 radiomic features from PET/CT images of 131 patients treated for non small cell lung cancer. For the majority of patients multiple regions of interest (ROI) were available. We performed clustering and principal component analysis to investigate any trends in the data. We used two classification models, logistic regression and support vector machine, to predict metastasis and metastasis within one year. We applied three feature selection methods, fold change, t test and Wilcoxon test.

V Results

Several issues had to be addressed to process the data. In addition to the multiple, very different, ROI per patient, the features were highly correlated, and the differences between classes were relatively small. For most features, the value distributions overlapped, there were also strong imbalances between classes.

The accuracy values achieved by both models were comparable, with a slight advantage of SVM classifier. As for selection methods, the feature rankings for t test and Wilcoxon test were highly similar and consisted primarily of first order features. In the fold change method, which performed slightly better, texture features were at the top of the ranking, including Contrast, ClusterShade and LargeAreaLowGrayLevelEmphasis.

VI Evaluation

The models were evaluated using the Monte Carlo cross-validation method with 500 iterations, preventing information leakage. The obtained results show, that although radiomic data present many challenges, they have a potential for predicting metastasis.

VII Conclusions

The ability to accurately predict that a cancer patient would soon develop metastasis would be a breakthrough in therapy planning. Using such knowledge the treatment could be intensified to prevent or slow down the process of cancer dissemination. In contrast, treatment for lower risk patients could be spread in time to minimize unnecessary adverse effects.

References

[1] Gillies, R. J., Kinahan, P. E., and Hricak, H. Radiomics: Images are more than pictures, they are data. *Radiology 278*, 2 (2 2016), 563–77.

DOI: https://doi.org/10.1007/978-981-19-8234-7_8

Keywords: Sentiment Analysis, Tunisian Romanized alphabet comments, Word Embedding, Deep Learning, Sentiment Classification

Exploring Word Embedding For Arabic Sentiment Analysis

Sana Gayed, Souheyl Mallat, Mounir Zrigui

sana.gaied@gmail.com,souheyl.mallat@gmail.com,mounir.zrigui@fsm.rnu.tn

Simplified Title

Try the representation of words, typically in the form of a real-valued vector that encodes the meaning of the word, for Arabic text analysis to determine whether data is positive or negative.

Abstract

In Natural Language Processing (NLP), the manual features (part-of-speech tagging, stemming…) might not be helpful sometimes to deciding the feeling expressed in a sentence. That more properties need to be considered. Word embedding, which is the key component for learning the text features, has just started to appear in Arabic sentiment analysis. On the other hand, Deep Neural Networks were widely used recently for this task, especially for the English language. In this paper, we focus on the Tunisian dialect sentiment analysis used on social media using a Convolutional Neural Network and Bidirectional Long Short-Term Memory. The results show that our models on the publicly available TUNIZI dataset achieved superior performances than the other models applied for the same dataset.

I Introduction

Sentiment analysis, the field of study, that analyzes people’s opinions, sentiments, evaluations…towards entities such as products, services, organizations, individuals...has become one of the essential research areas whose application is clearly visible in many domains (politics, health, tourism…) due to the proliferation of reviews, recommendations and other forms of expression.

Most of studies have been applied on English and some other Latin languages successfully. However, very few studies have focused on sentiment analysis in Arabic, due to its complexity and rich morphology.

The growing number of Arab Internet users and the exponential growth of Arabic content online pushed the attention of many researchers to this task. Moreover, Arabic can be written in both scripts, Arabic and Arabizi. The last one, which was defined as the newly emerged Arabic variant written using the Arabic numeral system and roman script characters, was a reason that encourage researchers to deal with this variant of Arabic.

We try to improve sentiment classification results. We perform a Tunisian dialect used on social media using the representation of words in the case of binary classification (positive, negative).

II State of the Art

A lexicon-based sentiment analysis system was used to classify the sentiment of Tunisian tweets. A Tunisian morphological analyzer developed to produce linguistic features. 800 Arabic script tweets are used (the TAC dataset) and achieved an accuracy of 72.1%.

The support vector machine was presented for Tunisian Arabic script tweets. It achieved the best results for binary classification with an accuracy of 71.09% and an F-measure of 63%.Different bag-of-word schemes used as features, binary and multiclass classifications were conducted on a Tunisian Election dataset (TEC) of 3,043 positive/negative tweets combining MSA and Tunisian dialect.

A study is conducted on the impact of the Tunisian sentiment classification performance when it is combined with other Arabic based preprocessing tasks (Named Entity Tagging, stopwords removal…). A lexicon-based approach and the support vector machine model were used to evaluate the performances on two datasets; TEC (Tunisian Election dataset) and TSAC (Tunisian Sentiment Analysis Corpus).

Three deep learning methods (convolutional neural network (CNN), Long short-term memory (LSTM), and Bidirectional LSTM (BiLSTM)) were evaluated on a corpus containing comments posted on the official Facebook pages of Tunisian supermarkets to conduct to an automatic sentiment analysis. In their evaluation, authors wanted to show that the gathered features could lead to very encouraging performances through the use of CNN and BiLSTM neural networks.

A robustly optimized BERT approach was used to establish sentiment classification for a Tunisian corpus. A Tunisian Robustly optimized BERT approach model called TunRoBERTa was proposed, which outperformed Multilingual-BERT, CNN, CNN combined with LSTM and RoBERTa. The proposed model was pretrained on seven unlabeled Tunisian datasets publicly available.

To produce document embeddings of Tunisian Arabic and Tunisian Romanized alphabet comments, the doc2vec algorithm was used. The generated embeddings were fed to train a Multi-layer Perceptron (MLP) classifier where both the achieved accuracy and F-measure values were 78% on the TSAC (Tunisian Sentiment Analysis Corpus) dataset. This last dataset combines 7,366 positive/negative Tunisian Arabic and Tunisian Romanized alphabet Facebook comments.

Syntax-ignorant n-gram embeddings representation composed and learned using an unordered composition function and a shallow neural model was proposed, it helps to relieve hard work due to the hand-crafted features. A proposed model, called Tw-StAR, was evaluated to predict the sentiment on five Arabic dialect datasets including the TSAC dataset.

We observe that, word embeddings are used for learning the text features. This type of work, the representation learning, has just started to appear in Arabic sentiment analysis. Therefore, fastext word embedding will be used.

III ORIGINAL CONTRIBUTION

These new representations of textual data have made it possible to improve the performance of automatic language processing methods (or Natural Language Processing), such as Sentiment Analysis.

The word vector representations proved to be efficient and successful technique in the applications of NLP due to its capability to take under consideration the morphology of words.

IV METHODOLOGY

CNN and BiLSTM deep neural network classifiers were used. Convolutional Neural Network (CNN) is traditionally used in the application of image processing, and is good at capturing the patterns. The use of CNN is efficient for Natural Language Processing (NLP) on various benchmark tasks. Bidirectional LSTM (BiLSTM) is a class of RNN models. BiLSTMs are used for sequential processing of the data and are efficient at capturing long-range dependencies. Our work consists of experimented these two classifiers; a CNN with a number of filters equal to 100 and a BiLSTM. Our study is experimental.

V RESULTS

We can notice that the representation of words; the fastText embedding combined with CNN leads to the best performance with an 84.25% of accuracy compared to 83.69% scored by BiLSTM. This is also the case for the F1.micro and F1.macro performances with values 84.25% and 84.16%, respectively.

VI EVALUATION

The dataset includes more than 9k Tunisian social media comments written only using Latin script [1]. Performance metrics are needed to evaluate how well our models do. In sentiment analysis area, accuracy and F-score are usually the most used. We can notice that the representation of words used combined with CNN leads to the best performance of accuracy compared to BiLSTM. This is also the case for the F1.micro and F1.macro performances. Compared to other models applied on the tackled dataset, our fastText models based CBOW outperformed other models. This is one of the main strength in our models, which demonstrates their capacity for handling this type of Dialectal Arabic emerged in social media.

VII CONCLUSIONS

We achieved good results on the TUNIZI dataset. This last helped us to better understand the nature of the Tunisian dialect.

REFERENCES

[1] FOURATI, C., MESSAOUDI, A., AND HADDAD, H. *TUNIZI: a Tunisian Arabizi sentiment analysis Dataset*. arXiv, 2020.

DOI: https://doi.org/10.1007/978-981-19-8234-7_9
Link to repository: https://github.com/iitis/ClusteringAE

Keywords: Autoencoders, Hyperspectral unmixing, Clustering, Transfer learning

Improving Autoencoders Performance for Hyperspectral Unmixing using Clustering

Bartosz Grabowski[0000-0002-2364-6547], Przemysław Głomb[0000-0002-0215-4674], Kamil Książek[0000-0002-0201-6220], Krisztián Buza[0000-0002-7111-6452]

bgrabowski@iitis.pl, przemg@iitis.pl, kksiazek@iitis.pl, chrisbuza@yahoo.com

Simplified Title

Improving Autoencoders Performance for Hyperspectral Unmixing using Clustering

Abstract

Hyperspectral cameras acquire images containing information across the electromagnetic spectrum, which convey useful information about the scene. To enable effective analysis of such data, spectral unmixing is often used. It is an important task in hyperspectral imaging, allowing one to obtain the information about spectral endmembers which make up each hyperspectral pixel. This task, traditionally solved with dedicated statistical methods, has recently been explored with deep learning methods. One of the methods well-suited to this task are autoencoders. These neural networks are initialized using multiple random weights, and their initialization often has a significant impact on their efficiency. Because of that, to improve the initialization of autoencoders for the spectral unmixing task, we propose to use the pre-training scheme consisting of clustering-based artificial labeling. We test the approach on two popular hyperspectral datasets, i.e. Samson and Jasper Ridge. Our experiment delivers promising results, improving autoencoders effectiveness in the case of Samson dataset, i.e. for 25-class labeling endmembers' and abundances' errors improve by 0.045 and 0.008, respectively. The worse results in the case of Jasper Ridge dataset (improvement of the endmembers' error by 0.001, and worsening of the abundances' error by 0.006 for 25-classes labeling) show that more research is required to understand when the proposed approach improves the results of the spectral unmixing. The auxiliary experiments that we also conduct allow us to partially answer that question.

I Introduction

In this work, we explore the transfer learning approach to enhance the efficiency of autoencoders in the problem of spectral unmixing. We test our solution on the problem of hyperspectral unmixing because it is an important step in hyperspectral image processing, for which autoencoders are an effective solution. We utilize the clustering algorithm to generate artificial labels for the unlabeled dataset. We then use these labels to pre-train the autoencoder on the classification task. Last, we use the pre-trained model to perform unmixing on the original, unlabeled dataset. We test our pipeline using two different hyperspectral datasets as well as four different weight initialization methods. In the case of one of the datasets, the results show that the proposed approach improves the unmixing quality. In the case of the second dataset, the proposed approach does not bring about improvement. Given that with the subsequent experiments, we partially answer the question of why this is the case.

II State of the Art

The classical methods of hyperspectral unmixing include geometrical, statistical, and sparse regression-based approaches, e.g., SISAL and N-FINDR. Moreover, the hyperspectral unmixing task was solved using autoencoders by numerous authors. The problem of nonoptimal weights initialization in the case of deep learning models was observed in multiple studies, including instability of autoencoder training in the case of hyperspectral unmixing. In this work, we utilize the autoencoders for the spectral unmixing task. However, we also use the pre-training method to pre-train the models and reduce the problem of nonoptimal weight initialization.

III Original Contribution

We propose a way to improve the quality of hyperspectral unmixing in the case of autoencoder neural network, i.e. we utilize self-supervised learning approach, in which classification pretext task based on artificial labels is used to pre-train the model, making the subsequent unmixing easier to perform. We test our approach using two different hyperspectral datasets and four different weight initialization methods. In situations where our approach does not bring about any improvement over the baseline, we conduct additional experiments to find out why this is so.

IV Methodology

In our experimental study, we utilized the autoencoder model from [1] as well as the clustering method from [2] to generate artificial labels for pre-training. The autoencoder is composed of encoder, which transforms the input to the latent space, and the decoder, which transforms the output of the encoder so that it resembles the input to the model. The clustering method works as follows: For a given hyperspectral scene, it is divided into the given number of rectangles. The pixels in a given rectangle belong to the same cluster and are given the same artificial label. This simple labeling technique allows us to prepare the original dataset for the pre-training of the autoencoder. We use the artificial labels to pre-train the encoder part of the autoencoder on the classification task. Then the autoencoder is used to perform unmixing on the original dataset.

V Results

The results of our experiments are mixed. In the case of one of the datasets, the proposed pre-training resulted in lower errors compared to baseline results. However, when considering the second dataset, our approach did not result in statistically significant improvement, and in some cases, even higher errors were obtained compared to the baseline. In summary, the investigated approach has the potential to be a valuable tool for improving the effectiveness of autoencoders in the task of spectral unmixing, but more research needs to be done to learn under what conditions the method can bring improvements. Our analysis of the classification maps, where we observed that more empty maps were present in the case of the second dataset, can be considered a first step in that direction.

VI Evaluation

To test the proposed approach, we performed the experiment using two hyperspectral datasets as well as four weight initialization methods. The experiment was composed of 50 runs per weight initialization. The number of classes for the artificial labels, which is a hyperparameter of the used clustering method, was varied to better test the proposed pipeline. Furthermore, statistical tests were utilized to verify whether the differences between baseline model initializations and pre-training experiments were statistically significant.

VII Conclusions

In this work, we have investigated the usage of an unsupervised pre-training method in the case of autoencoders and the spectral unmixing task. We conducted the experiment using two different hyperspectral datasets and four different weight initialization methods. Moreover, we used a known artificial label generation algorithm based on clustering. The results can be used to improve the effectiveness of autoencoders in the task of spectral unmixing, and in the case where such improvement is not observed, it can provide a possible explanation for this fact.

References

[1] Książek, K., Głomb, P., Romaszewski, M., Cholewa, M., and Grabowski, B. Stable training of autoencoders for hyperspectral unmixing, 2021.

[2] Masarczyk, W., Głomb, P., Grabowski, B., and Ostaszewski, M. Effective training of deep convolutional neural networks for hyperspectral image classification through artificial labeling. *Remote Sensing 12*, 16 (2020).

DOI: https://doi.org/10.1007/978-3-031-21967-2_44

Keywords: Email signature extraction, Hybrid approach.

Meet your email sender - hybrid approach to email signature extraction

Jelena Graovac[0000-0002-9323-4695], Ivana Tomašević[0000-0003-3764-1269], Gordana Pavlović-Lažetić[0000-0002-0665-1053]

jgraovac@matf.bg.ac.rs, ivana@matf.bg.ac.rs, gordana@matf.bg.ac.rs

Simplified Title

Meet your email sender - hybrid approach to email signature extraction

Abstract

Email signature is considered imperative for effective business email communication. Despite the growth of social media, it is still a powerful tool that can be used as a business card in the online world which presents all business information including name, contact number, and address to recipients. Signatures can vary a lot in their structure and content, so it is a great challenge to automatically extract them. In this paper, we present a hybrid approach to automatic signature extraction. The first step is to obtain the original most recently sent message from the entire email thread, cleaned from all disclaimers and sufficient lines, making the signature to be at the bottom of the email. Then we apply the Support Vector Machine (SVM) Machine Learning (ML) technique to classify emails according to whether they contain a signature. To improve obtained results we apply a set of sophisticated Information Extraction (IE) rules. Finally, we extract signatures with great success. We trained and tested our technique on a wide range of different data: Forge dataset, Enron with our own collection of emails, and a large set of emails provided by our native English-speaking friends. We extracted signatures with a precision 99.62% and recall of 93.20%.

I Introduction

Email is one of the most used communication services. Its usage is steadily growing, with more than 4 billion users worldwide in 2021 and about 6.8 billion email accounts – and it continues to grow. Email signatures serve the purpose of business cards or letter pads in this electronic world. They contain the sender's name, organization, location, phone number, company's personal web page URL, social networks addresses and so on. Still, signatures may be highly varying depending on the individuals or based on the company information. Our goal in this paper is to develop a hybrid - machine learning/rule based approach to automatic email signature extraction.

Although some commercial solutions for signature extraction from emails do exist on the market, more sophisticated analysis and use, broader coverage and range of data, specific collections and clients, and the need for an as perfect methodology as possible, combining Machine Learning (ML) and Information Extraction (IE) methods, was the challenge of the project that inspired this work and the paper.

II State of the Art

Signature extraction is envisaged more as an image detection than a Natural Language Processing problem and still, no significant effort in image detection deep learning methods has been put into solving this problem. Some strategies for the component-level analysis of plain text email messages are based on applying different machine learning algorithms to a sequential representation of an email message, represented as a sequence of lines, and each line defined as a set of features. Other methods use a combination of geometrical analysis and linguistics analysis to convert the two-dimensional signature block into one-dimensional reading blocks and to identify functional classes of text in a one-dimensional reading block. These methods are trained and tested on quite a limited set of data.

III Original Contribution

Since machine learning classifiers may make substantial errors, we provided additional Information Extraction technique-based verification of the decision made by the machine learning classification algorithm. For each email, we considered the email address and metadata associated with it (first and last name, if present). We also used three different datasets (collections of emails) for training and testing our hybrid approach, with a high volume of high-quality data, providing high-quality classification results.

Table 1: Result of SVM email classification with 10-cross validation on Forge, Enron and our own email datasets.

Dataset	P	R	ACC	F1
ForgeUnique	0.924152	0.987207	0.945879	0.954639
ForgeUniqueNotBots	0.948665	0.989293	0.962169	0.968553
Enron	0.857342	0.943707	0.867096	0.897685
OurEmails+Enron	0.919156	0.959001	0.936205	0.938490

IV Methodology

Following the strategy described in [1], we performed an experimental study, specifying a specific classification procedure consisting of the following steps:

- **Data preprocessing.** First, we extracted the most recent emails from the threads and applied different preprocessing techniques to ensure that the signatures would be at the bottom of the email if they existed (we deleted all reply lines, disclaimer lines, notification lines, long lines, etc.)
- **SVM email classification and IE refinement.** Then we applied the SVM supervised learning method to classify all emails into the P category (signature-containing emails) or N category. Obtained results are refined using additional IE rules.
- **Signature extraction.** At the end, for all emails that contain the signature (classified into the P category) we extracted signature blocks.

V Results

We have built a robust signature extraction tool. A Python library has been defined for feature rewriting, feature pattern design, feature-based file representation, SVM model training, SVM model testing, elimination of unwanted lines, reply lines extraction, etc. A set of rules for signature extraction has also been developed.

We applied the designed protocol and the tool to three datasets - a set of Forge formatted emails containing P – 1243 emails, N - 1474 emails, Enron email dataset complemented with our own mail collections and mailing lists (P – 400 emails and N - 400 emails), and Dataset-3 – quite a large email dataset in its original format (not in Forge format), consisting of 15264 emails without signatures and 5105 emails with signatures.

VI Evaluation

The results obtained could be evaluated in terms of quality measures Recall, Precision, Accuracy, and F1 measure, for different datasets, in the table 1.

When the SVM algorithm is applied to signature extraction from the emails of the P class, the results obtained on the Dataset-3 dataset are presented in the table 2.

Table 2: Result of signature extractor on Dataset-3 with SVM method only (V2.1) and hybrid method (V2.2).

Dataset-3	P	R	ACC	F1
V2.1	0.964784	0.759530	0.932816	0.849941
V2.2	**0.996231**	**0.932027**	**0.982081**	**0.963060**

VII Conclusions

The results obtained suggest that the developed tool may be successfully applied to many different fields and tasks, such as preprocessing emails for text-to-speech systems, automatic formation of personal address lists, email threading, etc. An application that may have a broad social and economic impact is a development of a network of contacts as a core of a successful customer relationship management system.

References

[1] Carvalho, V. R., and Cohen, W. W. Learning to extract signature and reply lines from email. In *Proceedings of the Conference on Email and Anti-Spam* (2004), vol. 2004.

DOI: https://doi.org/10.1007/978-981-19-8234-7_10

Keywords: Time Series, ML, Resilient Supply Management, FMCG

Data-driven Resilient Supply Management Supported by Demand Forecasting

Marek Grzegorowski[0000-0003-4740-0725], Andrzej Janusz[0000-0002-9763-1399], Jarosław Litwin, Łukasz Marcinowski

M.Grzegorowski@mimuw.edu.pl, {andrzej.janusz, jaroslaw.litwin}@qed.pl, l.marcinowski@fitfoodpoland.pl

Simplified Title

Data-driven Resilient Supply Management and Demand Forecasting

Abstract

The article discusses several challenges related to resilient supply management and demand forecasting. Both of those topics are of great importance for food retailers and producers who aim at reducing the risk of lost sales opportunities and food waste. In the investigated case study of FitBoxY.com, due to the overestimated demand and too large deliveries, historically, even 30% of the products were overdue. The developed ML framework integrated with the supply management system enabled optimization of business costs and reduced food waste from overestimated demand. The experimental evaluation showed that, with the developed solution, it is possible to improve demand forecasting by nearly 50% compared to estimates proposed by human operators.

I Introduction

Providing reliable forecasts of the future demand for fast-moving consumer goods (FMCG) is a big challenge, especially for retailers and producers in the food market. Here, the demand misestimation may lead to lost sales opportunities and extra costs related to food waste. A special case requiring even more attention is sales through vending machines, where the limited capacity causes additional difficulties in optimal supply management (SM). The correct analysis of data from a dispersed points-of-sale (PoS) network causes many difficulties, for instance, related to distributed data sources and their integration or challenges connected to a small number of historical purchase transactions of a particular product in an analyzed PoS. In this regard, researchers indicate an opportunity to build data-driven SM systems.

II State of the Art

Food production is a complex process under high uncertainty resulting in differences between the planned supply and actual demand. Predicting demand may be modeled as the regression of the historically observed sales in the investigated location. Several approaches are often adapted to dealing with this task, including simple statistical methods, multivariate machine learning models, or obtaining near-optimal results with optimization meta-heuristics. Another essential element of data-driven supply management systems is a concise and understandable data representation [2, 3]. Recently, entrepreneurs have also considered applying prescriptive analytics to support decision-making [1]. For such systems to make reliable recommendations, it is critical to employ effective mechanisms of demand forecasting.

III Original Contribution

In the article, we discuss the possibility of integrating the supply management system with a machine learning (ML) framework that allows for predicting the demand for a distributed sales network. The prediction of future sales over a specific period of time basing on historical data can be treated as a standard problem in the field of time series analysis. However, due to the specificity of the FMCG data in the food industry, where products have a short life cycle and are often substituted, this problem is very challenging. In a discussed case study of FitBoxY.com, out of over 200 different products, only 10% are sold regularly over a long period. As an effect, we observe a very big number of relatively short and scattered time series of sales (as a single time series, we consider the sales history of a given product in one PoS). Fitting a prediction model for such data is difficult for many state-of-the-art ML methods. It is, therefore, necessary to properly select and adjust utilized algorithms so that they could effectively predict the sale of new products with a short sales history, thus effectively coping with the so-called cold-start problem [4].

IV Methodology

We deployed the developed framework and confirmed its effectiveness with the analysis of real data collected from the FitBoxY.com. This technological start-up delivers healthy meals to the dispersed network of vending machines of own construction deployed in many office buildings. FitBoxY offers meals prepared with high-quality ingredients and packed under a protective atmosphere, avoiding chemical preservatives. The consequences of such a decision are a slightly higher price and a short expiry date of two weeks. Thus, accurately adjusting supply for the actual demand is critical, otherwise resulting in increased food waste and financial loss. In the article, we present the results of predictions with several univariate auto-regressive methods, as well as multivariate machine learning models operating on multidimensional representations. Data constitute time series extracted from historical daily sales of products, separately for each PoS. However, for each of the investigated approaches (univariate vs. multivariate), the evaluated methods expected different representations.

V Results

The focus of our solution is on the optimization of business costs and the reduction of food waste resulting from overestimated demand at points-of-sale. We achieve this goal by embedding machine learning algorithms into the developed solution architecture. We highlighted challenges related to the specific nature of the data that we collect. In particular, we explained the need for feature extraction to handle the cold start of new products and new locations of points-of-sale. The presented experimental evaluation of several forecasting algorithms showed that by using our approach it is possible to significantly improve (by nearly 50%) over the demand estimates proposed by human operators.

VI Evaluation

Experimental analysis was performed independently on 4 data sets and selected uni- and multivariate machine learning algorithms. The time series were constructed based on historical transactions of FitBoxY products offered at each vending machine (aggregated daily). The analyzed data are limited to the period starting May 2017 (the very beginning of the FitBoxY operations) and ending early 2020. Avoiding the COVID-19 pandemic period was necessary for the proper evaluation of univariate models, for which long gaps in series (e.g., due to lock-downs) would enforce data imputation, resulting in high uncertainty of predictions. Furthermore, this data better reflect the expected post-pandemic reality. There are two different prediction horizons examined. The first of 14 days corresponds to the shelf-life of products. The second of 7 days match the median time between deliveries. In the article, we present the evaluation of the sales prediction with root mean square error (RMSE) and mean absolute percentage error (MAPE) - micro-averaged for all the examined periods and time series.

VII Conclusions

The combination of IoT-driven, low-cost vending machines, data-driven supply management, and the resilient ML framework allows for optimizing costs associated with running a chain of unmanned points-of-sale. Our solution addresses the issue of food waste and the overwhelming amount of garbage produced in the world. This fact places our solution in the frame of the primary mission of modern businesses that opt to be profitable, yet responsible and sustainable.

This research was co-financed by Polish National Centre for Research and Development (NCBiR) grant no. POIR.01.01.01-00-0963/19-00 and by Polish National Science Centre (NCN) grant no. 2018/31/N/ST6/00610.

References

[1] Grzegorowski, M., Janusz, A., Lazewski, S., Swiechowski, M., and Jankowska, M. Prescriptive analytics for optimization of fmcg delivery plans. In *Proceedings of IPMU'22* (2022).

[2] Grzegorowski, M., Litwin, J., Wnuk, M., Pabis, M., and Marcinowski, L. Survival-based feature extraction - application in supply management for dispersed vending machines. *IEEE Transactions on Industrial Informatics* (2022).

[3] Grzegorowski, M., and Ślęzak, D. On resilient feature selection: Computational foundations of r-C-reducts. *Information Sciences 499* (2019), 25–44.

[4] Janusz, A., Grzegorowski, M., Michalak, M., Wróbel, Ł., Sikora, M., and Ślęzak, D. Predicting Seismic Events in Coal Mines Based on Underground Sensor Measurements. *Engineering Applications of Artificial Intelligence 64* (2017), 83–94.

DOI: https://doi.org/10.1007/978-3-031-21743-2_21
Link to repository: https://github.com/UNEP-Economy-Division

Keywords: NLP, Deep-Learning, SDG, United Nations

SDG-Meter : a deep learning based tool for automatic text classification of the Sustainable Development Goals

Jade Eva Guisiano1[0000-0002-9048-3602], Raja Chiky2, Jonathas De Mello

jade.guisiano@etu.sorbonne-universite.fr, raja.chiky@outscale.com, jonathas.demello@un.org

Simplified Title

SDG-Meter : a deep learning based tool for automatic text classification of the Sustainable Development Goals

Abstract

The 17 Sustainable Development Goals (SDGs) are a ”shared blueprint for peace and prosperity for people and the planet, now and into the future”. Since 2015, they help pointing out pathways to solve interlinked challenges being faced globally. The monitoring of SDGs is essential to assess progress and obstacles to realise such shared goals. Streams of SDG-related documents produced by governments, academia, private and public entities are assessed by United Nations teams to measure such progress according to each SDG, requiring labelling to proceed to more in-depth analyses. Such laborious task is usually done by the experts, and rely on personal knowledge of the links between the documents contents and the SDGs. While UNEP has experts in many fields, links to the SDGs that are outside their expertise may be overlooked. In this context, we propose to solve this problem with a multi-label classification of texts using Bidirectional Encoder Representations from Transformers (BERT). Based on this method, we designed the SDG-Meter, an online tool able to indicate to the user in a fully automatic way the SDGs linked to their input text but also to quantify the degree of membership of these SDGs.

I Introduction

As part of the work of the United Nations Environment Programme (UNEP), multiple documents (policy recommendations, toolkits, reports, project submissions, progress reports, etc.) are analysed by SDG experts for properly indexing them according to SDGs. Experts have to read the document and identify which SDGs are mentioned, or are related to it. These mapping exercises are time-consuming and subjective to personal knowledge of the links between the document contents and the SDGs. While UNEP has experts in many fields, links to the SDGs that are outside their expertise may be overlooked. This work, based on the subjectivity of each expert, introduces a bias in the overall mapping work, which therefore does not allow for a comparison of the associations made for each text. This association work (labeling) based on text analysis therefore requires increased knowledge of the terminology of each SDG to achieve an optimal analysis. The improvement of the whole text labeling process consists firstly in adopting automation of this process (time saving for the experts), and secondly, in designing a single intelligence with a solid knowledge of the terminologies of each SDG (bias reduction). Aiming to address these issues, we propose the online tool SDG-Meter which allows to automatically classify a text — without expert intervention — according to the 17 SDGs but also to quantify each link between the SDGs and a text. The SDG-Meter is based on the use of a powerful deep learning algorithm which is currently one of the most powerful and efficient algorithms in the field of Natural Language Processing. Following a series of tests, it was confirmed that our tool is able to perform a classification identical to that of an expert.

II State of the Art

It exists various tools which permit to classify texts according to SDGs such as SDG-Pathfinder[1], OSDG[2] and LinkedSDG[3] However, these tools have some limitations such as the unsupervised methods always require the time-consuming intervention of experts to map the detected features to each SDG. Moreover, these methods are generally based on keyword analysis and do not take into account the context of the texts. In most cases, SDG classification tools and methods do not quantify the link between a text and an SDG, which is useful for establishing a more accurate classification.

[1] https://sdg-pathfinder.org
[2] https://osdg.ai
[3] https://linkedsdg.officialstatistics.org/#/

III Original Contribution

The SDG-meter is an online tool which allow to link text to one or more SDGs that it deals with and also to quantify its degree in percentage of belonging to these different SDGs without any human intervention. The SDG-Meter is based on the use of one the most powerful natural language processing algorithm named BERT(Bidirectional Encoder Representations from Transformers)[1].

IV Methodology

In order to constitute the training base of our SDG Meter we have extracted text labelled by experts from IISD ”SDG Knowledge Hub” website[4] which provides users labelled articles such as ”News”, ”Guest Article” and ”Policy Brief” so three different writing styles. Based on this dataset, the multi-label classification algorithm BERT was trained to identify and quantify the degree to which a text belongs to the 17 SDGs. this experimental approach relies on the sole use of a deep learning algorithm to classify a text according to the SDGs, a method that has never been mentioned in the literature.

V Results

The SDG-Meter permits to automatically classify text — without expert intervention — according to the 17 SDGs but also to quantify each link between the SDGs and a text. The SDG-Meter is available as an online tool[5] that accepts any form of English text limited to a length of 512 words (BERT algorithm limitation).

VI Evaluation

A series of more than 400 tests were performed using new IISD annotated text, in 98%(accuracy) of the cases the SDG Meter is able to retrieve the SDGs identified by the experts with a significant order of importance. This experimental approach relies on the sole use of a deep learning algorithm to classify a text according to the SDGs, a method that has never been mentioned in the literature. The BERT algorithm has been selected for its recognized performance in the field of natural language processing, and no expert intervention is required during the classification process.

VII Conclusions

The SDG-Meter, fully autonomous, performs classifications very similar to those of the experts without ever needing their information in the process. It provides a common annotation base (no more judgement bias between various experts), which can be used by any type of actor, organization or company wishing to know their position regarding the sustainable development objectives. Some improvements can however be made, such as the removal of the word limit per text which would allow the analysis of long text documents but also the expansion of its learning base which would allow to increase its performance via the enrichment of its vocabulary.

References

[1] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. N. Bert: Pre-training of deep bidirectional transformers for language understanding.

[4] `https://sdg.iisd.org/`

[5] `http://62.160.8.100`

DOI: https://doi.org/10.1007/978-3-031-21967-2_34

Keywords: glaucoma detection, domain generalisation, transfer learning, fundus cameras, retinal fundus images

Domain Generalisation for Glaucoma Detection in Retinal Images from Unseen Fundus Cameras

Hansi Gunasinghe[0000-0002-7136-4691], James McKelvie[0000-0002-6460-4175], Abigail Koay[0000-0002-4130-9931], Michael Mayo[0000-0001-8222-3782]

hg99@students.waikato.ac.nz, james@mckelvie.co.nz, a.koay@uq.edu.au, michael.mayo@waikato.ac.nz

Simplified Title

Domain Generalisation for Glaucoma Detection

Abstract

Out-of-distribution data produced by unseen devices can heavily impact the accuracy of machine learning model in glaucoma detection using retinal fundus cameras. To address this issue, we study multiple domain generalisation methods together with multiple data normalisation methods for glaucoma detection using retinal fundus images. RIMONEv2 and REFUGE, both public labelled glaucoma detection datasets that capture fundus camera device information, were included for analysis. Features were extracted from images using the ResNet101V2 ImageNet pretrained neural network and classified using a random forest classifier to detect glaucoma. The experiment was conducted using all possible combinations of training and testing camera devices. Images were preprocessed in five different ways using either single or combination of three different preprocessing methods to see their effect on generalisation. In each combination, images were preprocessed using median filtering, input standardisation and multi-image histogram matching. Standardisation of images led to greater accuracy than other two methods in most of the scenarios with an average of 0.85 area under the receiver operator characteristic curve. However, in certain situations, specific combinations of preprocessing techniques lead to significant improvements in accuracy compared to standardisation. The experimental results indicate that our proposed combination of preprocessing methods can aid domain generalisation and improve glaucoma detection in the context of different and unseen retinal fundus camera devices.

I Introduction

Images taken by retinal fundus camera are useful for glaucoma screening. In this study, we evaluate the accuracy of deep learning-based automated glaucoma detection with respect to the camera model in order to assess generalisation performance.

II State of the Art

Most of the studies are limited to a single clinical setup or images from an individual fundus camera. Such limitations questions the external validity of a model. There is limited variability in data, a small number of images, and sometimes include class-imbalanced data.

III Original Contribution

As far as the authors are aware, this research is the first to consider the retinal fundus camera model variability issue using publicly available data. We demonstrate that simple changes in image preprocessing can generalise machine learning models trained on the retinal fundus images created by different devices. Other research uses a pretrained ResNet model for fine-tuning, but here, it is used as a feature extractor for glaucoma classification. We altered one of the input standardisation methods to improve training accuracy. This research introduced a multi-image histogram matching algorithm to optimise model generalisation on test data.

IV Methodology

REFUGE [2] and RIMONEv2 [1], labelled glaucoma detection datasets drawn from three different models of fundus camera (1655 images in total), were used. The REFUGE dataset contains images from two camera models (we denote them as REFUGED1 and REFUGED2). Images were cropped around the optic nerve head to match the RIMONEv2 images. Median filtering, image standardisation and randomised multi-image histogram matching were applied as preprocessing steps separately on images of each camera in order to compare them. Image features were extracted using ImageNet pretrained RESNET101V2 neural network. Glaucoma detection models based on random forests were trained afterward. The experiment was conducted by assigning images from one camera as the training set and images from another camera as the test set. Finally, we obtained the area under the receiver operating characteristic curve (AUROC).

There were six preprocessing configurations based on the three preprocessing methods. Three experiment setups were followed. The first experiment had one device as training device and another device as testing device. The second experiment was performed by setting up training images from two devices and testing on the third device. The third/baseline experiment was to train and test on the same device using every method.

V Results

The highest AUROC of 0.8870 is achieved when the model is trained on REFUGED1 and tested on REFUGED2. The model gives testing 0.8807 AUROC when tested on REFUGED1 and trained on the other two devices. Standardising the inputs with the median filter prior gives 0.9457 for REFUGED2 and 0.9706 AUROC for RIMONEv2 in the baseline experiment. No preprocessing of the images results in very poor performance where the first experiment achieves 0.6487, the second experiment achieves 0.6708 and the third achieves 0.8800 on average. The values are below the average of preprocessed cases.

VI Evaluation

In summary, standardising the input images is the best configuration. It gives an overall higher average AUROC score compared with the scores returned for other settings by the random forest classifier. Furthermore, the histogram matching method performs better for RIMONEv2 and REFUGED2, as shown by the results in the first experiment.

VII Conclusions

The changed input standardisation image preprocessing method is better together with median filtering in training space. Even though the multi-image histogram matching method did not perform well in the experiments, proper parameter optimisation may lead it to be useful for generalising the deep learning glaucoma detection models when tested across multiple fundus cameras.

Applying data augmentation will potentially improve the results. One such augmentation method is Neural Style Transfer which performs well in related research [3]. We expect that changing the classifier from random forests to other algorithms, such as eXtreme Gradient Boosting (XGBoost) will also lead to better performance in glaucoma classification.

References

[1] Batista, F. J. F., Diaz-Aleman, T., Sigut, J., Alayon, S., Arnay, R., and Angel-Pereira, D. Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis & Stereology 39*, 3 (2020), 161–167.

[2] Orlando, J. I., Fu, H., Breda, J. B., et al. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis 59* (2020), 1361–8415.

[3] Toth, M., and Kiss, A. Retinal blood vessel segmentation on style-augmented images. *Studia Universitatis Babeș-Bolyai Informatica 66*, 1 (2021), 74–85.

DOI: https://doi.org/10.1007/978-3-031-21967-2_35

Keywords: BRDF modeling, anisotropy criterion, hyperspectral BRDF

BRDF Anisotropy Criterion

Michal Haindl[0000-0001-8159-3685], Votěch Havlíček

haindl@utia.cas.cz,havlicek@utia.cas.cz

Simplified Title

A criterion for detection of anisotropic bidirectional reflectance distribution function (BRDF).

Abstract

Visual scene recognition is predominantly based on visual textures representing an object's material properties. However, the single material texture varies in scale and illumination angles due to mapping an object's shape. We present an anisotropy criterion of bidirectional reflectance distribution function (BRDF), which allows deciding if a simpler isotropic BRDF model can be used or if it is necessary to use a more complex anisotropic BRDF model. The criterion simultaneously shows dominant angular orientations for the anisotropic materials. The anisotropic criterion is tested on several isotropic and anisotropic surface materials, with BRDF computed from the measured seven-dimensional Bidirectional Texture Function.

I Introduction

A human observer recognizes a visual scene using shape and material attributes. Unfortunately, the surface material's appearance changes under variable observation conditions [1], negatively affecting its automatic and reliable recognition in numerous artificial intelligence applications. A multidimensional visual texture is an appropriate paradigm for a surface reflectance function model capable of characterizing variable observation conditions. The best measurable representation is the seven-dimensional Bidirectional Texture Function (BTF) [1]. However, an enormous amount of visual BTF data, in the range of terabytes, measured on a single material sample, inevitably requires state-of-the-art storage, compression, modeling, visualization, and quality verification. The Bidirectional Reflectance Distribution function (BRDF) [1] is a simplified model which describes a material reflectance dependence on illumination and viewing angles while neglecting their spatial dependency, among others. We present an automatic anisotropy criterion to select between anisotropic and isotropic materials.

II State of the Art

Various analytic BRDF models were published primarily for isotropic materials, i.e., whose reflection does not depend on the surface's orientation (rotation invariant). Most BRDF models are restricted to isotropic materials, and few models are capable of modeling anisotropic materials. The modeling quality for anisotropic materials is usually significantly worse. Thus there is a need to decide if a simpler isotropic BRDF model can be used or if it is necessary to use a more complex anisotropic BRDF model.

II.1 Bidirectional Reflectance Distribution Function

A physically plausible BRDF must be non-negative (1), and it obeys the symmetry (2), and energy conservation properties (3):

$$\begin{aligned} BRDF(\lambda,\theta_i,\varphi_i,\theta_v,\varphi_v) &> 0 \ , & (1)\\ BRDF(\lambda,\theta_i,\varphi_i,\theta_v,\varphi_v) &= BRDF(\lambda,\theta_v,\varphi_v,\theta_i,\varphi_i) \ , & (2)\\ \int_\Omega BRDF(\lambda,\theta_i,\varphi_i,\theta_v,\varphi_v)\cos\theta_i d\omega_i &\leq 1 \ , & (3) \end{aligned}$$

where θ_i, θ_v are illumination and viewing elevation angles, φ_i, φ_v are illumination and viewing azimuthal angles, $\omega_i = [\theta_i, \varphi_i]$, and λ is the spectral index. A BRDF can be isotropic or anisotropic. The anisotropic BRDF model depends on five variables

$$Y^{BRDF} = BRDF(\lambda,\theta_i,\varphi_i,\theta_v,\varphi_v) \ , \quad (4)$$

while the isotropic, i.e., when the reflected light does not depend on surface orientation, only on four variables

$$Y^{BRDF} = BRDF(\lambda,\theta_i,|\varphi_i-\varphi_v|,\theta_v) \ . \quad (5)$$

III Original Contribution

This paper's contribution is a novel anisotropy criterion, which allows for deciding if a simpler isotropic BRDF model can be used or is necessary to use an anisotropic BRDF model.

IV Methodology

The suggested anisotropy criterion ε (11) depends on the selected range of BRDF measurements and can be applied to any number of spectral bands with a straightforward modification of the equation (11).

$$\begin{aligned}
\vec{\varepsilon}(k) &= \frac{1}{n(k)} \sum_{\forall\theta_i} \sum_{\forall\theta_v} \vec{\alpha}(\theta_i, \theta_v, k) \ , && (6)\\
\vec{\varepsilon} &= \frac{1}{n_k} \sum_{\forall k} \vec{\varepsilon}(k) = \frac{1}{n_k} \sum_{\forall k} \frac{1}{n(k)} \sum_{\forall\theta_i} \sum_{\forall\theta_v} \vec{\alpha}(\theta_i, \theta_v, k) \ , && (7)\\
\vec{\alpha}(\theta_i, \theta_v, k) &= |\vec{f}_{BRDF}(\theta_i, \theta_v, \phi_i, \phi_v) - \vec{\mu}_{BRDF}(\theta_i, \theta_v, k)| \ , && (8)\\
\vec{\mu}_{BRDF}(\theta_i, \theta_v, k) &= \frac{1}{n_{\theta_i,\theta_v}(k)} \sum_{\forall \triangle\phi = k} \vec{f}_{BRDF}(\theta_i, \theta_v, k) \ , && (9)\\
k &= |\phi_i - \phi_v| \ , && (10)\\
\varepsilon &= |\vec{\varepsilon}| = \sqrt{\sum_{\forall\lambda} \varepsilon_\lambda^2} \ , && (11)
\end{aligned}$$

where $n(k)$ is the number of all angular combinations for a specific k, n_k is the number of all possible differences k (i.e., $n_k = 226$ for 81×81 angular format), $\vec{\mu}_{BRDF}(\theta_i, \theta_v, k)$ (9). Spectral curves $f(\alpha(\lambda, \theta_i, \theta_v, k))$ denote anisotropy directions.

V Results

Tab. 1 summarizing anisotropy criterion values $\vec{\varepsilon}$ (7), ε (11) for presented isotropic and anisotropic materials. The largest criterion value, 30, has the most anisotropic spruce wood, while the isotropic stone and green glass values are only 1.76 and 5.48, respectively. The smaller the standard deviation (std) value, the smaller the modeling error can be expected from an isotropic BRDF model.

Table 1: Anisotropy criterion

	glass01	stone0	wood05	wood35	wood45	wood57	wood65
			ayouz	limba	alder	spruce	wenge
$\vec{\varepsilon}$	2.96	1.08	7.76	12.71	8.25	15.06	5.62
	3.19	0.99	7.83	13.99	9.73	17.31	5.39
	3.60	0.97	8.36	15.30	10.02	19.06	5.33
ε	5.48	1.76	13.84	24.32	16.22	30.18	9.44
std	0.32	0.05	0.27	1.06	0.78	1.85	0.13

VI Evaluation

We tested the anisotropy criterion on our extensive UTIA BTF database [2], composed of material images under varying illumination and viewing directions. The anisotropy wood materials were tested on the Wood UTIA BTF Database. All BRDF tables were computed from the BTF measurements.

VII Conclusions

The presented results indicate that the anisotropic criterion can reliably differentiate between isotropic and anisotropic materials and thus can be used to select the appropriate class of BRDF nonlinear models. The criterion can be easily used for high-dynamic or hyperspectral measurements with a straightforward modification to any number of spectral bands.

References

[1] Haindl, M., and Filip, J. *Visual Texture*. Advances in Computer Vision and Pattern Recognition. Springer-Verlag London, London, January 2013.

[2] Haindl, M., Mikeš, S., and Kudo, M. Unsupervised surface reflectance field multi-segmenter. In *Computer Analysis of Images and Patterns*, G. Azzopardi and N. Petkov, Eds., vol. 9256 of *Lecture Notes in Computer Science*. Springer International Publishing, September 2015, pp. 261 – 273.

Link to arXiv: `https://arxiv.org/abs/2211.00709`
DOI: `https://doi.org/10.1007/978-3-031-21967-2_43`
Link to repository: `https://github.com/sgk1997/semantic-pivot-ed`

Keywords: Event detection, Information extraction, Natural language processing, Deep learning

Semantic Pivoting Model for Effective Event Detection

Hao Anran, Hui Siu Cheung, Su Jian

`{S190003,asschui}@ntu.edu.sg,sujian@i2r.a-star.edu.sg`

SIMPLIFIED TITLE

Enhancing Event Detection via Semantic Representation Learning based on Event Type Labels.

ABSTRACT

Event Detection, which aims to identify and classify mentions of event instances from unstructured articles, is an important task in Natural Language Processing (NLP). Existing techniques for event detection only use homogeneous one-hot vectors to represent the event type classes, ignoring the fact that the semantic meaning of the types is important to the task. Such an approach is inefficient and prone to overfitting. In this paper, we propose Semantic Pivoting Model for Effective Event Detection (SPEED), which explicitly incorporates prior information during training and captures more semantically meaningful correlation between input and events. Experimental results show that our proposed model achieves the state-of-the-art performance and outperforms the baselines in multiple settings without using any external resources.

I INTRODUCTION

Event Detection (ED), which is a primary task in Information Extraction, aims to detect event mentions of interest from a text. ED has wide applications in various domains, such as news, business, and healthcare. ED is formulated as identifying *event triggers* which are the words that best indicate mentions of events and classifying these triggers into a pre-defined set of *event types*. For example, in the sentence S1, the underlined words are the triggers of an `Attack` event and an `Injure` event, respectively:

S1: A bomb **went off** near the city hall on Friday, **injuring** 6.

The state-of-the-art ED models are predominantly deep learning methods, which represent words using high dimensional vectors and automatically learn latent features based on training data. However, the limited size and data imbalance of ED benchmarks pose challenges to the performance and robustness of current deep neural models [1]. While recent works on ED [1, 2, 3] continue to push the performance limit, they miss the important fact that the types are semantically meaningful. The models only treat each event type class homogeneously as one-hot vectors and are therefore agnostic to the semantic difference or association of the types.

In this paper, we propose directly incorporating the type semantic information by utilizing the class label words of the event types (e.g., “attack” and “injure”) to guide ED. To this end, we propose a Semantic Pivoting Model for Effective Event Detection (SPEED), which uses the event type label words as auxiliary context to enhance trigger classification through a two-stage network. We highlight the fact that the label words are natural language representations of the meanings of the target types, which allows us to: (1) use them as initial semantic pivots for ED, and (2) encode them in the same manner as the input sentence words and enhance the representations of both via the attention mechanism. We empirically evaluate our SPEED model on the ACE 2005 benchmark and show that it achieves state-of-the-art performance.

II STATE OF THE ART

Existing solutions on ED are feature-based approach and deep learning (i.e., representation-based) approach. The latter is predominant and often achieves state-of-the-art performance. Recent works can be categorized into three major approaches: (i) proposing architectures with more sophisticated inductive bias, (ii) leveraging linguistic tools and knowledge bases, (iii) using external or automatically augmented training data.

III ORIGINAL CONTRIBUTION

We propose the SPEED model, which directly incorporates the type semantic information by utilizing the class label words of the event type. Evaluation on ACE 2005 shows that our model outperforms strong baselines. Without using external resources, our model achieves performance comparable to previous models trained with extra data or using linguistic tools.

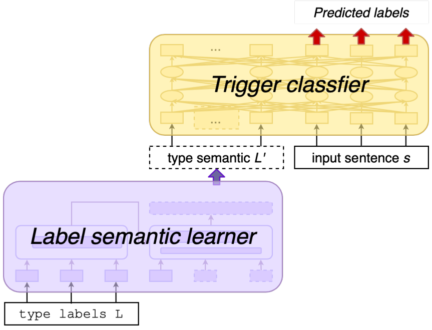


Figure 1: Architecture of the proposed SPEED model.

IV Methodology

We proposed a two-stage Transformer-based model consisting of a Label Semantic Learner and a Trigger Classifier. The Label Semantic Learner obtains semantic representation for the event types. Using such information as a task prior, the subsequent Trigger Extractor learns to detect trigger words from input sentences effectively and classify them into the correct types. To jointly train the Label Semantic Learner and the Trigger Extractor, we introduce the Gumbel-Softmax module in the Label Semantic Learner to resolve the indifferentiality issue.

V Results

In the standard setting, we compare SPEED with state-of-the-art ED models on the ACE2005 benchmark. Without using linguistic tools or external resources, our proposed SPEED model achieves 77.1% in F1, outperforming the baseline models by 0.4%-8.0% in F1. Among all the models, SPEED achieves the highest recall with good precision.

VI Evaluation

To show the data efficiency of our proposed SPEED model, we evaluate on scarce training data in comparison with the baseline models. Our model performs significantly better than the two baselines under the settings. With less training data, the performance of the baselines drops significantly by 3.0%-28.5% in F1, while the performance of our proposed SPEED model only drops by 2.9%-7.4%. With an extremely limited amount (20%) of data, SPEED can still achieve a reasonable F1 performance of 69.7%, whereas the baselines perform much worse.

We further perform analysis on single and multiple event sentences, respectively, and compare our proposed model with the strongest baselines for the multiple-event scenario. Without using linguistic features, including POS tag and dependency, our SPEED model achieves high F1 (76.8%) performance on multiple event sentences, outperforming the baselines by 4.1%-25.9%. This shows that our proposed architecture can effectively model cross-event interaction, benefiting ED on multiple event sentences.

VII Conclusions

In this paper, we propose a novel semantic pivoted Event Detection model that utilizes the pre-defined set of event type labels for event detection. It features event-type semantics learning via a Transformer-based mechanism. The experimental results show that our model outperforms the state-of-the-art event detection methods. In addition, the proposed model demonstrates several other advantages, such as working well for scenarios of scarce training data and multiple event sentences. The method provides an effective solution for general domain event detection.

References

[1] Chen, Y., Liu, S., Zhang, X., Liu, K., and Zhao, J. Automatically labeled data generation for large scale event extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vancouver, Canada, July 2017), Association for Computational Linguistics, pp. 409–419.

[2] Huang, L., and Ji, H. Semi-supervised new event type induction and event detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Online, Nov. 2020), Association for Computational Linguistics, pp. 718–724.

[3] Yan, H., Jin, X., Meng, X., Guo, J., and Cheng, X. Event detection with multi-order graph convolution and aggregated attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 5766–5770.

DOI: https://doi.org/10.1007/978-981-19-8234-7_12

Keywords: Texture classification, Descriptors, Feature extraction, Genetic programming, Late fusion.

Automated late fusion of low level descriptors for feature extraction and texture classification using data augmentation

Mohamed Hazgui[0000-0002-7080-7163], Haythem Ghazouani[0000-0002-6521-5024], Walid Barhoumi[0000-0003-2123-4992]

mohamed.hazgui@fst.utm.tn, haythem.ghazouani@enicar.u-carthage.tn, walid.barhoumi@enicarthage.rnu.tn

Simplified Title

An automated descriptor fusion method for robust texture classification using data augmentation.

Abstract

Feature extraction is an important task for texture image classification. Many descriptors have been proposed in the literature in order to describe textured images locally as well as globally. Researchers' interpretations differ on the effectiveness of these descriptors depending on the field of application, but no one can deny their complementarity. However, fusing different descriptors is not always easy, notably because of their different types (local vs. global, dense vs. sparse ...) and the heterogeneity of the generated features. In this work, we propose to use genetic programming to generate and fuse two different texture classifiers based respectively on HOG and uniform LBP descriptors. Indeed, the proposed method includes a late fusion and data augmentation process in order to combine the classifier's results while using small set of training data. The suggested method benefits from the different information captured by both descriptors while being robust to rotation changes. The performance of the proposed method has been validated on four challenging datasets including different variations. Results show that the proposed method significantly outperforms other low-level methods as well as GP-based methods intended for texture description and classification.

I Introduction

Texture classification can rely heavily on feature extraction to accurately describe an image's content. However, describing texture data can be challenging specially when dealing with image transformations such as scale variation, illumination changes and rotation. Indeed, in order to provide a prominent set of features, it is important first to identify the best regions where information is available. For this purpose, many methods use conventional descriptors such as Local Binary Patterns and Histograms of Oriented Gradients in order to construct a feature vector used lately to perform classification. However, these methods do not benefit from the information that both descriptors can provide simultaneously since they work differently and are hardly combined. In addition, they require human knowledge to configure, which can be costly. In order to automatize this process, some studies focused on Evolutionary Computation (EC) and particularly Genetic Programming (GP) [1, 2]. Indeed, GP is based on an evolutionary trial and error process. Thus, it is known to be very efficient when it comes to finding the best solution among a set of possible outcomes. GP-based methods showed promising results when applied to tasks such as feature extraction and texture classification. However, they have always dealt with texture as a local feature ignoring the fact that it can also be described in a global way. In fact, considering one type of feature can result in a loss of performance when dealing with complex classification problems. This research tries to benefit from GP architecture as well as HOG and LBP description power to perform an early fusion process and evolve robust classifiers that can outperform state-of-the-art methods. Such a process can handle the problem of human expert intervention and image changes by automatically fusing local and global features.

II State of the Art

In order to reduce human intervention and increase performance, many studies tried to benefit from GP and its ability to perform well on image-related tasks such as feature extraction and image classification. They showed good results on challenging tasks, including various image transformations such as change of scale, illumination and rotation. These GP-based methods combined well known image descriptors with GP architecture to extract a prominent set of features using a set of mathematical operators and functions. However, they only considered one type of feature and did not try to combine local and global features to increase performance. Other methods managed to combine different types of descriptors proving their complementarity and benefiting from different ways of capturing information. They also operated automatically, meaning that they managed to reduce human intervention by working only on a reduced search space. However, in order to adapt to the GP architecture, these studies relied only on an early fusion process and did not investigate the effects of a late fusion process that can combine different types of features without losing information generated by one or another.

III Original Contribution

This work tries to take advantage of different features (local and global) by combining image descriptors that capture information differently to perform feature extraction and classification. The developed method benefits from GP in order to evolve classifiers that have been combined lately using a data augmentation and fusion process. This allows the proposed method to train using a limited number of instances. The whole process is fully automated since it does not require human intervention and can deal with challenging transformations such as scale, rotation, and illumination.

IV Methodology

In order to develop the proposed method, an experimental study composed of three phases was conducted. First, a GP architecture was designed in order to optimize the feature selection and extraction processes. This GP architecture uses a set of mathematical operators as well as a fitness function to evolve a classifier based on two descriptors which are the Histogram of Oriented Gradients and Local Binary Patterns. This is called the fusion process and allows our method, during the training phase, to select the best descriptor that is suitable for the given task. During the test phase, a data augmentation and voting process are performed in order to predict the instance label and thus benefit from the features previously extracted to increase the classifier's robustness.

V Results

In order to examine its performance, the proposed method was tested and compared with other state-of-the-art methods on multiple datasets designed for texture classification tasks. It outperformed the majority of conventional descriptors combined with non-GP classifiers as well as similar GP-based classifiers. It also showed better robustness compared to CNN architectures for problems with a limited number of training instances. The accuracy of our method reaches more than 90% in some cases, even when trained on 15 samples and tested on the rest of the unseen data.

VI Evaluation

To evaluate our research, multiple experiments were conducted. The datasets were split into training and testing sets where only a limited number of instances were provided. This evaluation process allowed us to assess the performance of the proposed method when facing real-case scenarios where only a small amount of data is available. The provided datasets contained multiple variations and presented an important similarity between samples. This choice is explained by our motivation to develop a classifier that can handle different image transformations. The results show that the GP architecture presents better robustness when dealing with this kind of issue and can adapt easily to extract important features and, thus, improve the classification task.

VII Conclusions

The proposed method proved to be efficient for texture classification tasks outperforming the well-known state-of-the-art methods. It can operate without human intervention in problems where the number of instances is limited. Its fusion and augmentation processes increase the classifier's robustness when dealing with image variations such as changes of scale, illumination and rotation. The scope of this method can be extended from texture classification to other tasks such as feature extraction, face detection and medical segmentation. Its GP architecture can adapt to many problems and can easily be modified to include additional layers and functions optimizing region detection and extraction as well as feature selection.

References

[1] Bi, Y., Zhang, M., and Xue, B. Genetic programming for automatic global and local feature extraction to image classification. In *2018 IEEE Congress on Evolutionary Computation (CEC)* (2018), pp. 1–8.

[2] Hazgui, M., Ghazouani, H., and Barhoumi, W. Genetic programming-based fusion of hog and lbp features for fully automated texture classification. *The Visual Computer (2021)* (2021).

DOI: https://doi.org/10.1007/978-3-031-21743-2_43
Link to repository: https://github.com/tranhamduong/Vietnamese-Spelling-Correction-testset

Keywords: Vietnamese Spelling Correction, BERT - Transformer

A Combination of BERT and Transformer for Vietnamese Spelling Correction

Hieu Ngo Trung[1], Duong Tran Ham[1], Tin Huynh[1], Kiem Hoang[1]

{ngotrunghieu,tranhamduong,huynhngoctin,hoangkiem}@siu.edu.vn

SIMPLIFIED TITLE

Sequence to sequence model for Vietnamese Spelling Correction

ABSTRACT

Recently, many studies have shown the efficiency of using **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) in various Natural Language Processing (NLP) tasks. Specifically, English spelling correction task that uses Encoder-Decoder architecture and takes advantage of BERT has achieved state-of-the-art result. However, to our knowledge, there is no implementation in Vietnamese yet. Therefore, in this study, a combination of Transformer architecture (state-of-the-art for Encoder-Decoder model) and BERT was proposed to deal with Vietnamese spelling correction. The experiment results have shown that our model outperforms other approaches as well as the Google Docs Spell Checking tool, achieves an 86.24 BLEU score on this task.

I INTRODUCTION

Spelling correction studies that took advantage of the Encoder-Decoder model have attracted much attention and achieved state-of-the-art in the English spelling correction task [1, 2]. Therefore, this paper aims to apply these architectures and techniques to improve the performance of correcting Vietnamese spelling errors.

II STATE OF THE ART

Implementing Language Model in English Spelling Correction was found successful. However, there is still no implementation in Vietnamese that can be used in practice.

III ORIGINAL CONTRIBUTION

- Applying the Transformer architecture and leveraging the pre-trained BERT to solve the Vietnamese spelling correction problem.
- Constructing a large and creditable dataset based on the most common practical Vietnamese spelling errors. The evaluation dataset is published for the Vietnamese NLP community in related works.

IV METHODOLOGY

This study focuses on experimental approach by combining the Transformer with pre-trained BERT to provide a sequence-to-sequence deep learning model to solve the Vietnamese spelling correction problem illustrated in Figure 1.

V RESULTS

For the objective of comparison and practical application, we compare to google docs spellchecking tool. The experiment results have shown that our model and google docs spellchecking tool achieve 86.24 and 0.68 BLEU scores respectively.

Model	BLEU Score
Google Docs spellchecking tool	0.6829
Transformer + vinai/phobert-base	0.8027
Word2Vec	0.8222
Transformer + bert-multi-cased	0.8624

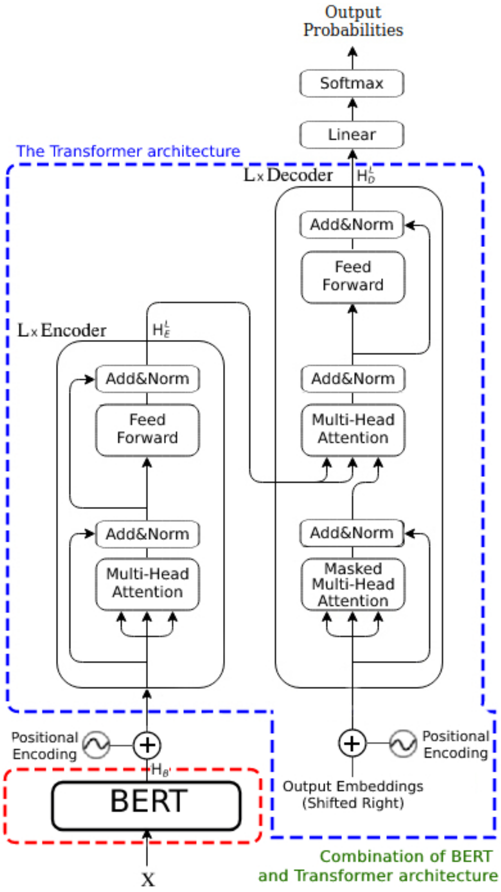


Figure 1: Proposed combination between BERT and Transformer

VI Evaluation

- There are a few patterns that our model gain out performance: telex and edit-distance error types. This happened partially because we designated more of these types of errors to achieve our goal.

- The google docs spellchecking tool has another advantage over our model: the ability to restrict unnecessary correction. Additionally, the emergence of proper nouns also makes our model ineffective. When it comes to proper nouns, especially Vietnamese proper names, our model tends to correct them, which should not be the case.

VII Conclusions

This paper implements a combination of BERT and Transformer architecture for the Vietnamese spell correction task. The experimental results show that our model outperforms other approaches with a 0.86 BLUE score and can be used in real-world applications.

References

[1] Yuan, Z., and Briscoe, T. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (June 2016), Association for Computational Linguistics, pp. 380–386.

[2] Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., and Liu, T. Incorporating BERT into neural machine translation. In *Eighth International Conference on Learning Representations* (2020).

DOI: https://doi.org/10.1007/978-3-031-21967-2_32
Link to repository: https://github.com/minhrongcon2000/mask-recognition

Keywords: face mask detection, convolution neural network, AIoT

A Correct Face Mask Usage Detection Framework by AIoT

Minh Hoang Pham, Sinh Van Nguyen[0000-0003-0424-5542], Tung Le[0000-0002-9900-7047], Huy Tien Nguyen[0000-0002-9948-1048], Tan Duy Le[0000-0001-6597-0209], Bogdan Trawinski[0000-0002-2956-6388]

ldtan@hcmiu.edu.vn

No simplified title given.

Abstract

The COVID-19 pandemic, which affected over 400 million people worldwide and caused nearly 6 million deaths, has become a nightmare. Along with vaccination, self-testing, and physical distancing, wearing a well-fitted mask can help protect people by reducing the chance of spreading the virus. Unfortunately, researchers indicate that most people do not wear masks correctly, with their nose, mouth, or chin uncovered. This issue makes masks a useless tool against the virus. Recent studies have attempted to use deep learning technology to recognize wrong mask usage behavior. However, current solutions either tackle the mask/non-mask classification problem or require heavy computational resources that are infeasible for a computational-limited system. We focus on constructing a deep learning model that achieves high-performance results with low processing time to fill the gap in recent research. As a result, we propose a framework to identify mask behaviors in real-time benchmarked on a low-cost, credit-card-sized embedded system, Raspberry Pi 4. By leveraging transfer learning, with only 4-6 hours of the training session on approximately 5,000 images, we achieve a model with accuracy ranging from 98 to 99% accuracy with the minimum of 0.1 seconds needed to process an image frame. Our proposed framework enables organizations and schools to implement cost-effective correct face mask usage detection on constrained devices.

I Introduction

The year 2021 witnessed one of the most lethal viruses in the human history, the Corona virus and its variants. Despite countless efforts to ease the effects, it is nothing compared to its growing speed. As a result, it is of utmost importance that an automatic system be developed to encourage people to wear mask correctly in crowded area to mitigate the spread of the virus as much as possible. There is a lot of study tackling this problem, and they follow a two-step process: face detection first and mask behavior classification later ([2, 3]). However, the current approach either has fast inference time but low performance (as in [2]) or high performance but low inference time (as in [3]). Our approach tries to obtain the best of both worlds by following the same approach but redefining the problem into a binary classification problem between correct and incorrect mask behavior. In this paper, we also propose general guidance on deploying a medium-sized deep learning model on an AIoT system using three different hardware architectures: edge computing, fog computing, and cloud computing.

II State of the Art

Most studies tackle the problem of recognizing correct mask behavior by extracting the face from an image captured by a media device and determining the behavior for each extracted face afterward. For example, the work of Ejaz et al. [2] uses the Viola-Jones algorithm (Viola and Jones [4]) to extract face from a camera and principal component analysis (PCA) to group mask behavior into clusters, based on which the classification is performed. Despite good inference properties, this approach is not scalable as PCA has a high sensitivity to shifting variants. The study of Fan and Jiang [3] leverages deep learning for both face recognition and mask usage recognition tasks by utilizing Context Attention Module (Woo et al. [5]), which is not good for real-time use cases due to its large size. The approach of this paper aims to solve the problem of incorrect mask usage behavior with high accuracy while providing a model with a low processing time that can be deployed to an AIoT system.

III Original Contribution

The contribution of this paper is three-fold. Firstly, this paper provides a lightweight model with 99% accuracy on the MaskedFace-net dataset [1] and an average frame rate of approximately 7 FPS on a Raspberry Pi 4 model B. Secondly, we provide a benchmark on the effectiveness of deploying an AI application on two different architectures: edge computing and fog computing. Last but not least, the author develops an AIoT (Artificial Intelligence of Things) framework that combines face detection models and face mask classification models to act as a gate to control the mask usage behavior before accessing highly-sensitive areas.

IV Methodology

For system design, we use a system with a combination of a receiver (handling simple image processing and result display), and an external computation (handling predictions and model inference). To decrease the latency, we suggest the use of edge computation, where the receiver is placed in the same building as the external computation. In addition, we use a convolutional neural network to effectively recognize the mask usage behavior at scale.

V Results

We use transfer learning with various backbones and obtain good results on MobileNet with 99% accuracy and 0.7 hours of testing time (evaluated on approximately 90,000 images).

Table 1: Model performance on MaskedFace-net dataset

Model name	**Precision**	**Recall**	**F1-score**	**Training time (hours)**	**Testing time (hours)**
ResNet50	50.07%	50.07%	50.07%	6.1814	0.7505
MobileNet	**99.39%**	**99.39%**	**99.39%**	4.7844	0.5503
VGG16	98.73%	98.73%	98.73%	4.8856	0.5353

VI Conclusions

In this study, we investigate the performance of three different models for the problem of recognizing mask usage behavior. We also utilize transfer learning to accomplish the task more efficiently by retaining the configuration trained on the ImageNet dataset and adding custom, fully-connected layers that contain parameters suitable for our problems. The results show that MobileNet is the best solution for accuracy and processing time.

We propose a general and extensible framework for deploying the model to an AIoT system with external computation. This crucial part allows the system to execute deep learning models. In addition, we also demonstrate some examples of system success and failure, where most of the failure falls into the safe case of mask equipment. Such failures can also be alleviated by collecting more data on such cases and using different weighting techniques to penalize the model's wrong prediction for the case.

Even though our system has high accuracy, there is still room for improvement. With the development of face recognition, we can further utilize facial landmarks to identify each case of incorrect mask usage. Model optimization is another direction in which model architecture should be minimized to lower the inference time. Lastly, bringing GPU to the embedded system is also attractive as it allows the deep learning model to operate smoothly on the edge without needing external computation.

References

[1] Cabani, A., Hammoudi, K., Benhabiles, H., and Melkemi, M. Maskedface-net–a dataset of correctly/incorrectly masked face images in the context of covid-19. *Smart Health 19* (2021), 100144.

[2] Ejaz, M. S., Islam, M. R., Sifatullah, M., and Sarker, A. Implementation of principal component analysis on masked and non-masked face recognition. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)* (2019), pp. 1–5.

[3] Fan, X., and Jiang, M. Retinafacemask: A single stage face mask detector for assisting control of the covid-19 pandemic. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2021), IEEE, pp. 832–837.

[4] Viola, P., and Jones, M. Robust real-time object detection. In *International Journal of Computer Vision* (2001).

[5] Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 3–19.

DOI: https://doi.org/10.1007/978-3-031-21967-2_54

Keywords: Heterogeneous cluster, Resource allocation, Multi-criteria decision , Big Data , Hadoop

Effective Resource Utilization in Heterogeneous Hadoop environment through a Dynamic inter-cluster and intra-cluster Load Balancing

Emna Hosni[0000-0003-3430-2966], Wided Chaari, Nader Kolsi, Khaled Ghedira

emna.hosni@ensi-uma.tn, wided.chaari@ensi-uma.tn, nader.kolsi@esct.uma.tn, Khaled.ghedira@uik.ens.tn

SIMPLIFIED TITLE

Efficient resource utilization in a heterogeneous environment based on dynamic load balancing between and within clusters.

ABSTRACT

The Hadoop cluster hosts multiple parallel workloads requiring various resource usage (CPU, RAM, etc.). In practice, in heterogeneous Hadoop environments, resource-intensive tasks may be allocated to the lower performing nodes, causing load imbalance between and within clusters and and high data transfer cost. These weaknesses lead to performance deterioration of the Hadoop system and delays the completion of all submitted jobs. To overcome these challenges, this paper proposes an efficient and dynamic load balancing policy in a heterogeneous Hadoop YARN cluster. This novel load balancing model is based on clustering nodes into subgroups of nodes similar in performance, and then allocating different jobs in these subgroups using a multi-criteria ranking. This policy ensures the most accurate match between resource demands and available resources in real time, which decreases the data transfer in the cluster. The experimental results show that the introduced approach allows reducing noticeably the completion time s by 42% and 11% compared with the H-fair and a load balancing approach respectively. The obtained finding also reveal that our approach optimizes the use of the available resources and avoids cluster over-load in real time.

I INTRODUCTION

In a heterogeneous Hadoop environment, over-allocation of resources for some jobs and under-utilization of cluster resources can occur [1]. In this situation, it may be necessary to move jobs from low-performance nodes to high-performance nodes in real time to finish fast. Effective load balancing is crucial to avoid the overhead of data transfer inter-cluster and intra-cluster, as well as optimize resource utilization. These issues are more common with different node and task characteristics. Moreover, if jobs cannot be executed effectively in heterogeneous clusters, data transfer overheads may be incurred because available resources are not used efficiently to meet resource requirements. Therefore, to balance workloads according to the resources available in the cluster, data must be distributed accurately and efficiently to reduce data transfer costs. Practically, in a heterogeneous Hadoop cluster the CPU cores, memory size and storage speed, etc, are not similar. Due to this heterogeneity, a load imbalance can be generated when task requirements do not match the resources available in the Node Managers. The proposed model is designed for the Yarn architecture. First, the system profiles the available nodes by grouping them into clusters with similar capabilities. Then, the Node Manager uses a multi-criteria decision to rank each job according to its resource demand using the generated node groups. Finally, a dynamic allocation of jobs is achieved to the most appropriate resources in the cluster.

II STATE OF THE ART

In recent years, the heterogeneity of computing systems in clusters has become an important research area, especially in the Hadoop Yarn system, thanks to the development of various load-balancing strategies and scheduling approaches. In fact, several studies have been conducted to deal with the heterogeneity problems in the Hadoop system, mainly load imbalance and resource wastage. The heterogeneous environment may cause an imbalance in resource utilization between over-loaded and under-loaded hosts, which degrades resource usage. In the literature, some works have improved the scheduling decision of Hadoop in order to improve load balancing. Others only take into account the heterogeneity of the jobs and neglect the capacities of the nodes, which affects the use of resources and lead to poor cluster performance. In addition to the use of prediction models [3] and cosine similarity [2] in Yarn to optimize the execution time and resource utilization. However, failure to consider multi-user and to monitor the overload and under-load of nodes during task execution leads to inefficient and inaccurate resource utilization between high-performance and low-performance nodes. Therefore, optimal node selection must be performed to achieve a good match between resource requirements and real-time node capabilities.

III Original Contribution

The proposed contribution achieves the best load balancing that reduces the number of remote jobs, and improves the locality rate and resource utilization. It also avoids under-loading and over-loading in heterogeneous Hadoop environments.

IV Methodology

the purpose of our work is to achieve the best match between available resources and resource requirements in Hadoop to improve resource utilization and subsequently reduce inter-cluster and intra-cluster data transfer. The presented hybrid load balancing system performs iterative clustering of nodes using the efficient k-means algorithm, which creates groups of nodes with similar performance. After that, the system labels the jobs using the Analytic Hierarchy Process AHP ranking while considering their resource demand. The jobs are sorted according to their overall weights. Then, the system combines the node groups and job scoring to achieve an accurate match between the required and available resources. The node with the highest utilization of available resources has the highest priority for job execution. In this article we present an experimental work.

V Results

Our experimental environment was deployed in two physical machines (10 VM). We implemented our load-balancing algorithm in all nodes. The k-means algorithm generates three groups of nodes that have an intra-group similarity in terms of resource availability. The multi-criteria job scoring process dynamically provided the ranking of 10 jobs submitted at the same time. Different types of jobs such as WordCount and TeraSort are used to carry out the experiment. Our load-balancing algorithm noticeably reduced the competition time of 10 jobs by 23% , 37% compared with H-fair and LB approach [1], respectively. This allows to rapid release of the containers for the next job, which enhance the overall performance of the distributed computing systems.

VI Evaluation

The theory of multi-criteria decision-making is adopted since it integrates several criteria and their order of preference to select the best option among many alternatives according to the expected result in real time i.e. a better load balancing in a heterogeneous environment. Our approach achieved the minimum time spent while processing the job in the high-performance group and in the medium-performance group. It assign the resource-intensive jobs to the group of high-performance nodes, and inside each group, the nodes are sorted in descending order. Thus, job will be dynamically allocated based on the load balancing constraints and the current capacity range of each node to avoid over-loading in the heterogeneous cluster.

VII Conclusions

An efficient hybrid load-balancing approach applied in a heterogeneous Hadoop cluster was proposed. It combines iterative clustering of nodes and a dynamic multi-criteria decision that scores jobs according to the required resource. Our main objectives were to improve resource utilization, reduce the job competition time and avoid load imbalance in a heterogeneous cluster. The proposed research work can be used in many domains where the use of massive data in real-time requires an optimized use of resources (embedded system, shared resources between computing devices (Iot), permanent availability of resources, etc.). In future work, we will improve our approach with a higher scale while evaluating the energy consumption of Iot heterogeneous devices in real-time.

References

[1] Bawankule, K. L., Dewang, R. K., and Singh, A. K. Load balancing approach for a mapreduce job running on a heterogeneous hadoop cluster. In *International Conference on Distributed Computing and Internet Technology* (2021), Springer, pp. 289–298.

[2] Postoaca, A. V., Pop, F., and Prodan, R. h-fair: asymptotic scheduling of heavy workloads in heterogeneous data centers. In *2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)* (2018), IEEE, pp. 366–369.

[3] Wang, M., Wu, C. Q., Cao, H., Liu, Y., Wang, Y., and Hou, A. On mapreduce scheduling in hadoop yarn on heterogeneous clusters. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications* (2018), IEEE, pp. 1747–1754.

DOI: https://doi.org/10.1007/978-981-19-8234-7_13
Link to repository: https://github.com/huetran1611/DD-SSD-problem

Keywords: Drone, Healthcare, Routing, Synchronization

Door-to-door sampling service with drone

Tran Thi Hue[1,2], Nguyen Quang Anh[2], Tran Van Thanh[2], Pham Phu Manh[2], Huynh Thi Thanh Binh[2], Nguyen Khanh Phuong[2]

huett@hvnh.edu.vn, anh.nq183476@sis.hust.edu.vn, thanhvantran99@gmail.com, phumanh1998@gmail.com, binhht@soict.hust.edu.vn, phuongnk@soict.hust.edu.vn

Simplified Title

Door-to-door sampling service with drone

Abstract

Minimizing customer waiting time during service is the key to success of door-to-door service companies. The paper introduces the problem of addressing the integration of a drone into the existing sampling service system in which a set of technicians go to customers' locations to get samples and bring them back to the laboratory. We propose mathematical modeling and two meta-heuristics to solve the problem. Experiments are conducted to compare the MILP solutions with those obtained from meta-heuristics. The numerical results demonstrate the significant gain when implementing the proposed drone integration compared to the conventional technician-alone sampling service system.

I Introduction

Drones could be used to transport packages of different sizes and limited weight in postal service. In this context, deploying drones with trucks could improve not only the service time and quality but also the operating costs and thus contribute to safeguarding the environment [1]. Moreover, drones also have been used to deliver health-related items. The goal of this paper is to contribute to the investigation of possible integration. Hence it requires synchronization between vehicles, making the problem much more challenging.

We formally introduce and define the *Door-to-door sampling service with drone (DD-SSD)*, addressing the integration of a drone into the existing sampling service system in which a set of technicians go to customers' locations to get samples and bring them back to the laboratory. The first mathematical model and meta-heuristic algorithms for the DD-SSD are our second contribution. Computational results are discussed to qualify search strategies on the quality of the meta-heuristics and the management of the Door-to-door sampling service system.

II State of the Art

When considering the hybrid mode, research avenues include two major variants: 1) only drones perform the delivery [2]; 2) both drones and trucks perform the delivery, some examples are in [3]. However, most of the works described in the literature assume a single delivery per trip for the drones, and each drone coordinates with a fixed truck [1].

III Original Contribution

The sampling service system in the DD-SSD is composed of a laboratory where a set of technicians and a drone are based. A number of locations where customers are required to get samples are available. The route planning consists of two parts: 1) each technician only performs one trip that departs from the laboratory, gets samples from one or several customers, then goes back to the laboratory either with or without taken samples; 2) while the drone does either one trip or a sequence of trips, each trip starts from the laboratory to visit one or several technicians at customer locations to get samples and bring them back to the laboratory, and must not be longer than the maximum flight time. Thus, samples could be brought to the laboratory by either technicians or the drone.

All the technicians must leave the laboratory from time 0, and the last sample must be brought to the laboratory at the latest time $L^{'}$. The waiting time to be tested for each customer's sample is the difference between its arrival time at the laboratory and the time at which the sample is taken. The aim of the problem is to minimize the total waiting time of all customers' samples.

IV Methodology

To solve this problem, this study provides a MILP mathematical model, two meta-heuristics algorithm.

The first meta-heuristics algorithm is called Bi-level genetic algorithm, each solution to the problem is encoded into two chromosomes: the Technician chromosome represents the routes of technicians and the Drone chromosome encodes the drone trips.

The second is called Tabu search algorithm (TS), with an initial feasible solution z; at each iteration of the TS, one neighborhood is selected probabilistically based on the current value of r, then the selected neighborhood is explored, and the best move is chosen. This move must not be tabu, unless it improves the current best solution z_{best}. The search is terminated when the maximum number of iterations IT_{max} is reached or after IT_{imp} iterations without improvement on the best solution.

V Results

Table 1 shows the Performance comparison between algorithms.

Table 1: Performance comparison between algorithms

Number of	GUROBI		Bi-level GA			Tabu search			
customers	Best Avg	Sol.status	Best Avg	Std.	Time(s)	Best Avg	Std.	Time(s)	GAP to GA(%)
6	35.33	Optimal	35.33	0.1	68	35.33	0.01	0.5	0.0
10	71.91	Feasible	67.18	1.83	356	65.07	0.01	3	-3.14
12	79.38	Feasible	70.62	0.77	498	70.62	0.05	4	0.0
20	-	Unknown	120.58	3.15	723	116.04	0.05	27	-3.77
50	-	Unknown	1129.86	56.37	1374	697.91	0.06	486	-38.23
100	-	Unknown	3293.88	152.71	2851	1543.4	0.21	2076	-53.14
Avg	-	-	786.24	35.82	979	421.40	0.065	432.75	-16.38

Table2 describes three scenarios: 1) using $|K|$ technicians; 2) using $|K|+1$ technicians; 3) drone+$|K|$, using one drone and $|K|$ technicians.

Table 2: Comparison of with and without drone

Number of	drone+$\|K\|$	$\|K\|$	$\|K\|+1$
customers	Best	GAP(%)	GAP (%)
6	35.33	264.81	104.65
10	65.07	301.86	118.81
12	70.62	88.57	59.15
20	116.04	116.88	84.38
50	697.91	92.89	70.12
100	1543.40	59.27	48.65
Average	421.40	154.05	80.96

VI Evaluation

In Table 1, the rightmost column *GAP to GA(%)* displays the average gaps to the best Bi-level GA solutions of the solutions obtained by the TS. With large-scale customer instances, TS performs better computing experimental. Table 2 displays the superior performance of the proposed drone integration.

VII Conclusions

We introduced the DD-SSD problem by addressing the integration of a drone into the existing sampling service system. Experimental results clearly illustrated the superior performance of the proposed drone integration compared to the conventional technician-alone sampling service system.

References

[1] Macrina, G., Di Puglia Pugliese, L., Guerriero, F., and Laporte, G. Drone-aided routing: A literature review. *Transportation Research Part C: Emerging Technologies 120* (2020), 102762.

[2] Poikonen, S., and Golden, B. Multi-visit drone routing problem. *Computers & Operations Research 113* (2020), 104802.

[3] Raj, R., and Murray, C. The multiple flying sidekicks traveling salesman problem with variable drone speeds. *Transportation Research Part C: Emerging Technologies 120* (2020), 102813.

DOI: https://doi.org/10.1007/978-981-19-8234-7_14

Keywords: Natural Language Inference, Error Analysis, BERTology

Error Investigation of Pre-trained BERTology Models on Vietnamese Natural Language Inference

Tin Van Huynh[0000-0003-4990-2868], Huy Quoc To[0000-0002-0936-245X], Kiet Van Nguyen[0000-0002-8456-2742], Ngan Luu-Thuy Nguyen[0000-0003-3931-849X]

tinhv@uit.edu.vn, huytq@uit.edu.vn, kietnv@uit.edu.vn, ngannlt@uit.edu.vn

Simplified Title

Error Investigation of Pre-trained BERTology Models on Vietnamese NLI

Abstract

Natural Language Inference tasks have emerged in recent years and attracted significant attention from the natural language processing research community. There has been much success in this task with many quality datasets in English and Chinese for research and demonstrating the impressive performance of machine learning models. Pre-trained models play a crucial role, which is reflected in their superior performance compared to other models. However, they are still far from perfect and have many obstacles to the characteristics of the data. Especially in Vietnamese, we have just seen the emergence of the ViNLI benchmark dataset to serve the research community. In this paper, we experiment and analyze how the characteristics in the ViNLI benchmark dataset affect the performance of the pre-trained BETology-based models. In addition, the data parameters of ViNLI are also measured and analyzed on the accuracy of these models to see if it has any impact on the accuracy of the model.

I Introduction

The original NLI task, known as Recognizing textual entailment, required the machine learning model to capture the semantics of a given pair of premise and hypothesis sentences. The remarkable point in this task is the presence of many high-quality large datasets such as English, Chinese, Korean, Indonesian, and Persian. As a low-resource language, Vietnamese still has many limitations for outstanding research in this NLI task. However, recently, the research community has witnessed the launch of the ViNLI dataset, which Huynh et al. [2] developed for Vietnamese. This dataset has yielded some positive research results, so it is hoped to promote more and better research outcomes in the future.

In this paper, we try to investigate the behavior of the pre-trained BERT [1] language model and variant models of BERT through the lens of the Vietmanses NLI task. Vietnamese is an interesting language, but not much research has been done. From the current research results from the ViNLI dataset [2], we focused on setting up experiments in this paper. We deeply analyzed the features contained in ViNLI to see what affects the pre-trained model performance. This study can help us better understand pre-trained models as well as the ViNLI dataset. We hope these analyses point to potential future studies to improve the Vietnamese NLI task outcomes further. These findings may suggest new approaches in the data construction process or other techniques to enhance the accuracy of the machine learning models, especially BERTology models.

II State of the Art

In NLP tasks in general and NLI in particular, models based on Transformer architecture achieve very high performance. However, we find that there isn't usually much focus on deep analysis of models, as well as how special features of the data affect BERTology models so that we can make recommendations to improve NLI tasks more and more accurately. In this paper, we have an in-depth analysis of 4 model transformers, SBERT, mBERT, PhoBERT, and XLM-R, according to data characteristics in the process of building data VINLI such as writing rules in guidelines along with routines, the behavior of annotators. The results of the in-depth analysis show that the influence of data characteristics on the models is so many.

III Original Contribution

Our two main contributions are described as follows. (1) We re-implement SBERT, which is one of the state-of-the-art NLI models on the Vietnamese. (2) Together with SBERT, we analyze the impacts of other BERTology models such as mBERT, PhoBERT, and XLM-R according to Vietnamese concepts contained in the ViNLI dataset to understand the capabilities of each model better.

IV Methodology

Our study experiments with multilingual pre-trained models on the ViNLI dataset. We use the accuracy measures and F1-score to evaluate the performance of those models. The ViNLI benchmark dataset is used for experiments on pre-trained models.

Besides experiments with pre-trained models, including multilingual BERT, PhoBERT, and XLM-R established on ViNLI by Huynh et al. [2], we also experimented with another pre-trained model, SBERT [3]. The SBERT model is pre-trained in many different languages, including Vietnamese. We use these pretrained models provided by HugggingFace's library in our experiments. The parameters in the SBERT model are set up as follows: learning_rate = 1e-05, batch_size = 16, max_length = 256, in addition, we set epoch = 10.

V Results

Compared with the experimental results of Huynh et al. we found that the performance of the SBERT model is lower with the accuracy on the dev and test sets of 59.29% and 58.17%, respectively. Besides, the experimental results on SBERT have a rather large gap compared with other pre-trained models, especially when compared with the XLM-R_large model (83.02% on the dev set and 81.36% on the test set). This difference in accuracy is more than 23% on both the dev set and test set.

As mentioned above, in this study, we do not focus on improving the accuracy of models on ViNLI but on analyzing the result of pre-trained models. We try to investigate what affects the performance of these models. Specifically, we explore how the characteristics of the ViNLI dataset affect the performance of these pre-trained models. The issues in ViNLI that we are interested in analyzing include the influence of the annotation rule, word overlap, sentence length on performance, the ability to capture annotation artifacts of pre-trained models, and error analysis by confusion matrixes.

Some of the highlights we found when analyzing the results are as follows: On the entailment rules, all four models, SBERT, mBERT, PhoBERT, and XLM-R have the worst performance on pairs of sentences generated from the rule ”Turn adjectives into relative clauses”. Both the mBERT and XLM-R models have the highest accuracy on the rule ”Turn the object into relative clauses.” Regarding contradiction rules, four pre-trained models have the best predictive ability on pairs of ”Use negative words” rule with high accuracy, especially mBERT and XLM-R models achieve nearly 90%. The analysis results also show that the SBERT model has the worst performance on the hypothesis sentences generated from the ”Replace words with antonyms” rule with only 32.87%. Besides, most of the performance of pre-trained models decreases remarkably as the new word rate increase from 0 to more than 80%. Moreover, annotation artifacts leave clues on the hypothesis sentence that help language inference models correctly predict the label.

VI Conclusions

There are many interesting findings relating to data characteristics in ViNLI and the accuracy of models. In particular, most models have relatively low accuracy on the sentences entailment hypothesis generated from the rules ”Turn adjectives into relative clauses” and ”Create conditional sentences”. The contradiction hypothesis generated from the ”Use negative words” rule is straightforward for the models to predict correctly. In addition, when multiple rules are combined to create a contradiction hypothesis, the prediction models are more accurate. Word overlap or premise and hypothesis length also significantly affect the model's performance. Pre-trained models can make predictions thanks to the clues of the annotation artifacts, although the accuracy is not too high.

References

[1] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, Minnesota, June 2019), Association for Computational Linguistics, pp. 4171–4186.

[2] Huynh, T. V., Nguyen, K. V., and Nguyen, N. L.-T. ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics* (Gyeongju, Republic of Korea, Oct. 2022), International Committee on Computational Linguistics, pp. 3858–3872.

[3] Reimers, N., and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 3982–3992.

DOI: https://doi.org/10.1007/978-3-031-21967-2_22

Keywords: Graph Classification, Graph Mining, Graph Embedding

Graph Classification via Graph Structure Learning

Tu Huynh, Tuyen Thanh Thi Ho, Bac Le

huynhtu9910@gmail.com, tuyenhtt@ueh.edu.vn, lhbac@fit.hcmus.edu.vn

Simplified Title

GC-GSL: Graph Classification via Graph Structure Learning

Abstract

With the ability of representing structures and complex relationships between data, graph learning is widely applied in many fields. The problem of graph classification is important in graph analysis and learning. There are many popular graph classification methods based on substructures such as graph kernels or ones based on frequent subgraph mining. Graph kernels use handcraft features, hence it is so poor generalization. The process of frequent subgraph mining is NP-complete because we need to test isomorphism subgraph, so methods based on frequent subgraph mining are ineffective. To address this limitation, in this work, we proposed novel graph classification via graph structure learning, which automatically learns hidden representations of substructures. Inspired by doc2vec, a successful and efficient model in Natural Language Processing, graph embedding uses rooted subgraph and topological features to learn representations of graphs. Then, we can easily build a Machine Learning model to classify them. We demonstrate our method on several benchmark datasets in comparison with state-of-the-art baselines and show its advantages for classification tasks.

I Introduction

Graph classification is a significant problem as graph data has become increasingly popular and widely applied in many fields. However, there are still limitations in previous works, such as poor generalization or high computational cost. In this paper, inspired by the success of doc2vec [1] and Weisfeiler-Lehman graph kernel [2], we proposed a new method, which automatically learns the embedding of each graph using node information and substructures, to improve the graph.

II State of the Art

There are many popular substructure-based graph classification methods, such as graph kernels or frequent subgraph mining. For the former, graph kernels usually work on graph elements such as walk or path. However, these methods are difficult to find a suitable kernel function that captures the semantics of the structure while being computationally tractable. For the second group of mining frequent subgraphs in graphs, it is time-consuming because of the high cost of the subgraph mining step.

III Original Contribution

There are two main contributions in this paper. First, we proposed a graph-embedding neural network, which automatically learns the graph embedding corresponding to each graph. The learned embedding not only reflects the characteristics of the graph itself but also contains the relationship between the graphs. Second, through our experiments on several benchmark datasets, we demonstrate that our method is comparable with the graph kernels and other methods based on feature vector construction.

IV Methodology

Inspired by doc2vec [1] to learn document embedding, we extend the idea to graph embeddings. Doc2Vec exploits the way in which words/word sequences compose documents to learn their embedding. Similar to doc2vec, our method **GC-GSL** views a graph as a document and the rooted subgraphs in the graph as words. To mine the rooted subgraph ”vocabulary” set, we apply Weisfeiler-Lehman relabeling method [2] to avoid the subgraph isomorphism test - an NP-complete problem, and help our method be more effective. Besides, in **GC-GSL**, the topological attribute vector is added to the model to learn the general information of the graph. Specifically, the architecture of the graph embedding neural network of **GC-GSL** includes three layers: an input layer, a hidden layer, and an output layer. The input layer gets a one-hot vector of graphs whose length is equal to the number of graphs in the dataset. Next, there is only one hidden layer whose number of neurons equals the expected dimensionality of feature vectors. The embedding matrix between the input layer and the hidden is the embedding of the graphs we need to train. Finally, the output layer consists of two parts: the first part is the topological attributes vector with 16 dimensions corresponding to 16 features of the graph, and the second part of the output layer is taken from the subgraph "vocabulary" set. After training the graph embedding, a conventional Machine Learning algorithm can be applied for doing classification tasks. This graph embedding training is an unsupervised learning method, and it only uses the information and structures extracted from graphs, including the topological attributes vectors, the rooted subgraphs in graphs, and graphs themselves to be used for training. Therefore, it does not depend on graph labels, and only learns embedding through substructures and information of graphs. Moreover, this graph embedding model automatically learns the corresponding embedding for each graph, and the embedding that we get after training not only reflects the components of the graph itself but also reflects information about relationships between graphs.

V Results

Table 1: Average Accuracy (± std dev.) for our method **GC-GSL** and state-of-the-art baselines on benchmark datasets. Bold font marks the best performance in a column.

Datasets	**MUTAG**	**PROTEINS**	**NCI1**	**NCI109**	**PTC_MR**	**IMDB-B**	**IMDB-M**
WL	80.72 ± 3.00	72.92 ± 0.56	80.13 ± 0.50	80.22 ± 0.34	56.97 ± 2.01	-	-
Deep WL	82.94 ± 2.68	73.30 ± 0.82	80.31 ± 0.46	80.32 ± 0.33	59.17 ± 1.56	-	-
DDGK	**91.58 ± 6.74**	-	68.10 ± 2.30	-	**63.14 ± 6.57**	-	-
AWE	87.87 ± 9.76	70.01 ± 2.52	62.72 ± 1.67	63.21 ± 1.42	59.14 ± 1.83	**74.45 ± 5.83**	**51.54 ± 3.61**
FSG	81.58 ± 0.08	71.61 ± 0.03	77.01 ± 0.03	74.58 ± 0.02	60.29 ± 0.05	64.40 ± 0.05	46.53 ± 0.04
GC-GSL	83.86 ± 2.16	**76.55 ± 1.02**	**82.04 ± 0.45**	**81.86 ± 0.33**	60.11 ± 1.17	68.46 ± 1.12	46.39 ± 0.44

VI Evaluation

To evaluate the performance of the graph classification task, we compute the accuracy of each method. The experimental result shows that our method has higher accuracy than other methods. Furthermore, the performance of our method in 3 datasets of MUTAG, NCI1, and NCI109 have greater than 80% accuracy.

VII Conclusions

Due to the efficiency of our proposed method, it can be applied to solve many graph classification problems, such as mutated protein prediction, unknown compound finding, etc. Another application of our strategy is for many other graph-level tasks, such as graph clustering or community detection.

References

[1] Le, Q., and Mikolov, T. Distributed representations of sentences and documents. In *International conference on machine learning* (2014), PMLR, pp. 1188–1196.

[2] Shervashidze, N., Schweitzer, P., Van Leeuwen, E. J., Mehlhorn, K., and Borgwardt, K. M. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research 12*, 9 (2011).

DOI: https://doi.org/10.1007/978-981-19-8234-7_15

Keywords: Natural Language Processing, 2D Analytic Geometry, Machine Learning, Support Vector Machine, Python

Supervised learning use to acquire knowledge from 2D analytic geometry problems

Anca-Elena Iordan[0000-0001-9853-7102]

anca.iordan@cs.utcluj.ro

Simplified Title

Support Vector Machine - 2D analytic geometry problems

Abstract

Understanding 2D analytic geometry problems described in natural language is an important and difficult open research task. In this paper it explores the theme of identifying geometric elements in unstructured documents and their classification. The proposed solution is based on the automatic recognition of geometric elements, achieved with the help of supervised learning. The chosen system is based on a model resulting from an automatic learning, in a supervised manner, made with the help of the support vector machine. A significant point for increasing the performance level is to provide a training data set as balanced as possible, a target attained with the help of the subsampling technique. The originality retrieves in the preprocessing phase of the data, both in the case of identifying the geometric elements characteristics, and in the case of finding a solution for balancing the data set used.

I Introduction

Currently, there are several software systems used in the automatic solution of geometry problems, but they receive the hypothesis and the conclusion in a specific format. Improving them would mean that the automatic solution would start from the statement of the geometry problem in natural language. Automatic identification of geometric elements from geometry problems expressed in natural language is a well-known challenge in the field of natural language processing. This paper represent the steps towards an automatic identification of geometric elements found in problems of 2D analytic geometry in English. The approach is based on an automatic identification of geometric elements using support vector support.

II State of the Art

The general objectives of this research topic can be considered in three phases: data preparation, use of the classifier for training and testing, and then analysis and validation of the system. However, various challenges have arisen throughout the development process, the resolution of which has been of significant importance in achieving final performance. Experiments show that the existence of a balanced data set is a key factor and an important step for achieving the best possible performance of the system based on supervised learning [1].

III Original Contribution

In order to extract the hypothesis and the conclusion from the statement of a geometry problem expressed in natural language, it is necessary to define a formalized structure for identifying the geometric concepts. Thus, the geometric elements [3] can be categorized into four major clusters: abstract geometric elements (point, line, segment, ellipse, hyperbola, parabola, circle, conic), relational geometric elements (intersection(line, conic), tangent-line(conic, point), normal-line(conic, point)), quantitative geometric elements (length(segment), area(ellipse), measure(angle)), and logical geometric elements (parallel(line, line), tangent(conic, line), collinear(point, point, point), perpendicular(line, line)).

To enlighten the process of obtaining the geometric concepts and their classification in the four major clusters previously presented, SVM supervised machine learning strategy will be used. SVM grants overfitting protectorship, which does not rely on the number of characteristics specified to the classifier, thus it has the prospective to handle a very extensive characteristic space. During the development of the system, a problem was encountered related to the report between the terms that did not represent the interest and the geometric notions existing in the collections of geometry problems. Being large enough to influence the classification in a negative way, the classifier became biased towards the majority cluster, which represented the negative examples, and the system tended to classify the geometric elements as being of this type. To solve the problem, the support vector classifier was chosen, which offers the advantage of avoiding the overfitting problem [2], regardless of the number of features used. ”Overfitting” is a system modelling error that occurs when a function is mapped too precisely on a data set, and thus its behaviour when entering new data will not be desired, failing to classify correctly.

IV Methodology

The weight of the clusters provides information about the importance of that cluster, so the higher it is, the better the classifier will tend to identify that cluster. Following analysis, it was chosen to set the cost to 1 and the termination tolerance to 0.001. Also, because the custer containing the negative examples is not of much interest, we reduced its weight to 0.2, with the other clusters having a weight of 1. In this way, the meaning of terms that do not represent any kind of geometric concepts is reduced. The excellent results obtained by the developed geometric software acquired a mean F-score identical with 0.8452, if it takes into account the rank of negative examples, which demonstrated to be the optimal classification. The mean F-score computed precisely for geometric concepts has a value of 0.7961. If this weighted metric is computed, the valuables are acquired: 0.9107, if the “none” cluster is taken into account and 0.8017 if just the geometric concepts are analyzed.

V Results

The first analysis of the data was performed to verify whether the features, generated with the help of the word lemma, with the completion of the training set, bring or not improvements to the system, taking into account that they could introduce the phenomenon of over-learning. A initial appraisal was made after the subsampling of preponderante cluster (none) based on the investigation of the speech side used in the geometric concepts. The outcomes are shown in Table 1. The concept of prior sampling is observed too in this situation. This subsampling established a huge advantage to the performance, the increases being also at the step of a few percent for each cluster. In the second situation, the recall metric becomes bigger than the precision metric, which affordess appreciable input for software validation. For a higher survey of results, Table 1 presented the F-score values for every cluster used in the classification: abstract geometric element, relational geometric element, quantitative geometric element, logical geometric element and negative examples (none).

Table 1: Precision, Recall, and F-score.

	Subsampling based on speech part analysis		Subsampling based on superficial parsing		System performances
	Precision	Recall	Precision	Recall	F-score
Abstract geometric element	72.82	64.97	81.17	81.92	79.63
Relational geometric element	72.03	63.29	79.72	75.16	77.95
Quantitative geometric element	69.92	62.65	77.03	77.06	74.79
Logical geometric element	67.42	61.23	75.36	73.56	73.79
None(negative examples)					88.92

VI Conclusions

This research work covered the subject of identifying geometric concepts and classifying them in four clusters: abstract geometric components, relational geometric components, quantitative geometric components, and logical geometric components. The chosen approach consists of automatic recognition of elements, reached with the assistance of supervised learning. The first stage of software design was to examine the available experimental proof for this work, in order to better comprehend the component function in the procedure of solving. It has been presented that a fundamental point for growing the performance level is to deliver a collection of training data as balanced as possible, a scope reached through subsampling strategy.

References

[1] Gan, W., Yu, X., and Wang, M. Automatic understanding and formalization of plane geometry proving problems in natural language: A supervised approach. *International Journal on Artificial Intelligence Tools 28* (2019).

[2] Poon, H., Yap, W., Tee, Y., Lee, W., and Goi, B. Hierarchical gated recurrent network with adversarial and virtual adversarial training on text classification. *Neural Networks 119* (2019), 299–312.

[3] Quaresma, P., Santos, V., Graziani, P., and Baeta, N. Taxonomies of geometric problems. *Journal of Symbolic Computation 97* (2020), 31–55.

Link to arXiv: https://arXiv.org/abs/2104.12473

DOI: 10.1007/978-3-030-73280-6

Keywords: Granger causality; Sentiment analysis; Forecast; Cryptocurrencies; Twitter; Dogecoin

Forecasting cryptocurrency price fluctuations with Granger causality analysis

David L. John[0000-0002-4797-0915], Bela Stantic[0000-0003-0475-7951]

David.John2@griffithuni.edu.au, B.Stantic@griffith.edu.au

SIMPLIFIED TITLE

Forecast cryptocurrency price fluctuations

ABSTRACT

Forecasting various economic indicators has been a primary interest in economics and has attracted the attention of many researchers. Granger causality analysis has become quite popular in the econometrics literature and it aims to determine whether one time series is useful in forecasting another. In this work through the use of Granger causality analysis we investigate whether Twitter sentiment, expressed in large scale collections of daily tweets, can be correlated or even predictive of future prices of cryptocurrencies. The proposed framework considers tweets that mention the cryptocurrency “Dogecoin” and analyses the textual content of each of these tweets using a modified version of the lexicon-based sentiment polarity analysis method, VADER. The generated, Twitter sentiment time series is then compared to a time series of the closing prices of Dogecoin for each day. Granger causality analysis showed a unidirectional relationship between Twitter sentiment and cryptocurrency prices for day lags ranging from 2 to 4 days (with a 3-day lag having the lowest statistical significance value). This was also accompanied by a Pearson correlation coefficient of *r* = 0.6940 and a clear visual correlation between the two time series (with this 3-day lag). Findings indicate that Twitter sentiment is directly correlated and can be predictive of the future prices of cryptocurrencies.

I INTRODUCTION

Statistical analysis is a vital research tool used by businesses, researchers, governments, and other organizations to draw valid predictions about the future. This involves investigating patterns, trends, and relationships using quantitative data to improve data organisation and future projections. The primary purpose of this research is to investigate whether the innovative and unique analysis technique for predicting future values of a time series, introduced by the econometrician Sir Clive Granger called Granger causality, can be used to predict equity markets. In particular, a time series of public sentiment, expressed as the quantification of a large scale collection of daily tweets is used to predict the price of the cryptocurrency Dogecoin. Given the recent events that have occurred with GameStop, the sentiment surrounding it on Reddit, and the current excitement on Twitter around the cryptocurrency Dogecoin, this seems to be a suitable time to investigate this topic [1]. This work endeavours to answer the question, “Can Granger causality be used to predict future prices of Dogecoin using Twitter sentiment analysis?”.

II STATE OF THE ART

Granger causality analysis has been applied to a plethora of studies that aim to identify a predictive or causative relationship between two-time series. These studies have been mainly confined to the field of economics and used to relate variables such as financial development, tourism development, and economic growth. Other works use this technique to relate various economic indicators to predict stock market fluctuations. These works have all claimed a strong correlation between many of these variables, and some have demonstrated, by using Granger causality analysis, that many of these variables can be accurately predicted, highlighting the power of Granger causality as a practical analysis technique. Many existing solutions that aim to forecast economic indicators' prices mainly identified relationships between financial news media or social media and the stock market. The data analysis techniques employed in these works (specifically text analysis and sentiment analysis) prove to be very effective in predicting equity markets. Many methods of analysing news media were used in these studies; examples of these include text mining, feature extraction, feature selection, and machine learning methods.

III Original Contribution

We looked into harnessing Granger causality analysis, a statistical hypothesis test, to investigate whether one variable can forecast another variable. This technique is applied to Twitter sentiment and Dogecoin price to determine if Granger causality be used to predict future prices of Dogecoin.

IV Methodology

This theoretical study makes use of a modified version of a comprehensive, simple rule-based model for general sentiment analysis called VADER (for Valence Aware Dictionary for sEntiment Reasoning) to quantify Twitter sentiment. The tweets analysed were extracted worldwide through the use of the public Twitter API, using code developed in the Big Data and Smart Analytics lab at Griffith University. A collection of 5,331,040 public tweets were extracted and recorded between 5 May, 2021 and 31 May, 2021 and only references to the cryptocurrency Dogecoin; this filtering included tweets that had mention of ”dogecoin”, ”dogearmy”, ”dogecoinRise”, ”dogeEurope” or ”dogecoins”. The closing prices of Dogecoin for each day in the same period (between 5 May, 2021 and 31 May, 2021) were extracted from Yahoo! Finance[1] by using web scraping techniques. This method was developed based on code provided at GitHub[2]. Granger causality is then used to find the correlation between social media sentiment and Dogecoin where, in order to maintain a common scale for comparison, the sentiment and Dogecoin prices are converted to z-scores.

V Results

A statistically significant, unidirectional Granger causality relation for Twitter sentiment and Dogecoin prices was found for lags ranging from 2 to 4 days, with a 3-day lag having the lowest p-value of 0.0063.

This means that changes in sentiment, according to the results and fundamentals of Granger causality, prompt changes in the price of Dogecoin 2 to 4 days later. Additionally, this can also be seen by a visual comparison of the two time series, which show that both frequently overlap and trend in the same direction at many time points. This is also reflected in a Pearson correlation coefficient for these two data sets (with a 3-day lag) of $r = 0.6940$. This value, as well as the statistically significant results of Granger causality analysis accompanied by a clear visual correlation between the two time series, all provide strong evidence that changes in the past values of Twitter sentiment (at time t) predict a similar rise or fall in the price of Dogecoin three days later (at $t+3$). Therefore, Twitter sentiment about Dogecoin, as calculated by a method developed in the Big Data and Smart Analytics lab at Griffith University, has predictive value with regard to the price of Dogecoin.

VI Evaluation

The reason that Granger causality analysis was used rather than powerful machine learning techniques, such as Deep learning, which have given excellent results in previous works, is because the desired outcome of this analysis is simply a confirmation on whether Twitter sentiment is correlated to Dogecoin price to a statistical significance. Essentially a yes or no answer is required, which is precisely what Granger causality analysis will provide, given its predictive power when applied to time series data. The main assumption of this research is that if Twitter sentiment is shown to Granger-cause changes in the price of Dogecoin over the selected time period, then this can be related to any time period. Based on this assumption, the conclusion that can be drawn from these results is that Twitter sentiment about Dogecoin, as calculated by a method developed in the Big Data and Smart Analytics lab at Griffith University, has predictive value with regard to the price of Dogecoin.

VII Conclusions

The statistical analysis technique, Granger causality, has been shown to have many useful applications, particularly in the field of economics. Identifying relationships between various economic indicators in multiple countries highlights the usefulness and effectiveness of this technique. This paper has shown how Granger causality analysis can be used to predict future prices of cryptocurrencies by using Twitter sentiment data analytics.

References

[1] You, W., Guo, Y., and Peng, C. Twitter’s daily happiness sentiment and the predictability of stock returns. *Finance Research Letters 23* (2017), 58–64.

[1] https://finance.yahoo.com/quote/DOGE-USD/history

[2] https://github.com/Gunjan933/stock-market-scraper

Link to arXiv: https://arXiv.org/abs/2104.12473

DOI: 10.1007/978-3-030-73280-6

Keywords: Sentiment analysis; Social media; Natural language processing; Machine learning

Machine learning or lexicon based sentiment analysis techniques on social media posts

David L. John[0000-0002-4797-0915], Bela Stantic[0000-0003-0475-7951]
David.John2@griffithuni.edu.au, B.Stantic@griffith.edu.au

SIMPLIFIED TITLE

Sentiment analysis techniques on social media

ABSTRACT

Social media provides an accessible and effective platform for individuals to offer thoughts and opinions across a wide range of interest areas. It also provides a great opportunity for researchers and businesses to understand and analyse a large volume of online data for decision-making purposes. Opinions on social media platforms, such as Twitter, can be very important for many industries due to the wide variety of topics and large volume of data available. However, extracting and analysing this data can prove to be very challenging due to its diversity and complexity. Recent methods of sentiment analysis of social media content rely on Natural Language Processing techniques on a fundamental sentiment lexicon, as well as machine learning oriented techniques. In this work, we evaluate representatives of different sentiment analysis methods, make recommendations and discuss advantages and disadvantages. Specifically we look into: 1) variation of VADER, a lexicon based method; 2) a machine learning neural network based method; and 3) a Sentiment Classifier using Word Sense Disambiguation, Maximum Entropy and Naive Bayes Classifiers. The results indicate that there is a significant correlation among all three sentiment analysis methods, which demonstrates their ability to accurately determine the sentiment of social media posts. Additionally, the modified version of VADER, a lexicon based method, is considered to be the most accurate and most appropriate method to use for the semantic analysis of social media posts, based on its strong correlation and low computational time. Obtained findings and recommendations can be valuable for researchers working on sentiment analysis techniques for large data sets.

I INTRODUCTION

Sentiment analysis, also referred to as opinion mining, is the field of study which focuses on the analysis and quantification of people's sentiments, opinions, attitudes, emotions and appraisals. The primary purpose of this research, in order to give guidelines to researchers, is to identify and evaluate typical representatives of different sentiment analysis methods, test their performance both with regard to accuracy and time complexity, make recommendations, and discuss advantages and disadvantages. The accuracy of three different sentiment analysis techniques are investigated. A comparison of: 1) a sentiment analysis tool calculated using propriety methodology from the Big Data and Smart Analytics lab at Griffith University, which is built on top of the sentiment analysis tool VADER; 2) a machine learning technique using a neural network as implemented in [1]; and 3) a Sentiment Classifier using Word Sense Disambiguation, Maximum Entropy, and Naive Bayes Classifiers, is carried out.

II STATE OF THE ART

Sentiment analysis techniques have been widely used to analyse social networks from determining the public's opinions towards specific topics, issues or events to aiding businesses and organisations in improving their services, and they have proven to be an important and valuable tool. To date, the vast majority of sentiment analysis has been carried out on written text, with the intent of predicting the sentiment of a given written statement. In this context, sentiment analysis can be considered the primary research field of Natural Language Processing (NPL), which, with the aid of machine learning techniques, aims to identify and extract certain insights from written text. Many existing solutions which aim to quantify sentiment are based on a variety of different approaches, such as lexicon-based methods, machine learning methods and deep learning methods.

III Original Contribution

We plan to determine which of the three analysis techniques is considered to be the most appropriate method to use for the semantic analysis of social media posts by evaluating both the accuracy and efficiency based on strong correlations and low computational time. Obtained findings and recommendations can be valuable for researchers working on sentiment analysis techniques for large data sets.

IV Methodology

This theoretical study makes use of the public Twitter API to collect data. A collection of 10,000 original public tweets (not including retweets) were extracted to time complexity, which is related to the cryptocurrency Dogecoin; and that have a length of text larger than 150 characters to avoid short messages that usually do not have significant semantic content. As mentioned above, the three methods used to comparatively analyse the sentiment included: a modified version of VADER; a machine learning technique using a neural network, and a Sentiment Classifier. To enable a comparison of the sentiment values calculated, a sentiment values are normalised to 1 to ensure that the sentiment is always between -1 (the lower limit for negative sentiment) and +1 (the upper limit for positive sentiment). To quantitatively determine the relations between the three sentiment analysis methods, multiple regression analysis is used to test the correlation of the three-time series.

V Results

Using multiple regression analysis among the three sentiment analysis time series, results indicate a statistically significant correlation among the sentiment values of all three sentiment analysis methods. The most significantly correlated of these is the modified version of VADER and the machine learning technique using a neural network with a p-value of 7.301 E-53. These statistically significant results of multiple regression analysis ($p \ll 0.001$) accompanied by a clear visual correlation among the three plots show that all sentiment analysis methods are able to calculate similar results, thus providing strong evidence that these three methods can calculate an accurate sentiment value for textual data. To determine the most beneficial method to use for this analysis, the time taken to calculate sentiment values was also considered. Given that the modified version of VADER has the strongest correlation with both the other two methods used (MV vs. NN p = 7.301 E-53 and MV vs. SC p = 7.996 E-11) and has the shortest computational time, it is reasonable to suggest that it would be the most appropriate of the three methods to use for this type of analysis.

VI Evaluation

The criteria that have been adopted in selecting the methodologies involved included a variety of different types of sentiment analysis methods. The main assumption of this research is that a significant correlation among all three sentiment analysis methods demonstrates their ability to accurately determine the sentiment of social media posts, even though there is no ‘true’ sentiment value to compare them to. The method with the strongest correlation to the other two methods (along with the consideration of computational time) is considered to be the most accurate and most appropriate method to use for the semantic analysis of social media posts. Based on these assumptions, the conclusion that can be drawn from these results is that the modified version of VADER, a lexicon-based method, is considered to be the most accurate and most appropriate method to use.

VII Conclusions

Social media sentiment analysis has been shown to be valuable for researchers and businesses to understand and analyse a large volume of online data. In this paper, a comparison of three different sentiment analysis techniques for short text messages was carried out. The results of the relations between the three sentiment analysis methods, obtained by multiple regression analysis, provide evidence that these methods are able to have a statistically significant correlation and accuracy in their quantification of sentiment in Twitter posts. Based on a strong correlation and low time complexity, it was concluded that the modified version of VADER is considered to be the most accurate and most appropriate method to use. These findings and recommendations can be valuable for researchers that need sentiment polarity analysis for short text messages and large data sets.

References

[1] Stantic, B., Mandal, R., and Chen, E. Target sentiment and target analysis. *Report to the National Environmental Science Program. Reef and Rainforest Research Centre Limited, Cairns* (2020).

DOI: https://doi.org/10.1007/978-3-031-21743-2_8

Keywords: Financial data, portfolio, classification

Portfolio Investments in the Forex Market

Przemysław Juszczuk[0000-0001-7893-5410], Jan Kozak[0000-0002-2128-6998]

przemyslaw.juszczuk@@ue.katowice.pl, jan.kozak@@ue.katowice.pl

SIMPLIFIED TITLE

Portfolio Investments in the Forex Market

ABSTRACT

Investing in the forex market seems to be an especially challenging task due to the large variety of dependencies related to instruments. Among the crucial aspects that should be considered is the correlation between the currency pairs. In this article, we derive a general investing schema considering the signal generation based on the well-known classification methods and verify the quality of these signals with the idea of portfolio building. To do so, we derive a two-stage process, where the first stage is devoted to deriving the classifier capable of generating the trading signals on the forex market. We use the set of the most popular market indicators, and the decision about the potential buy (or sell) signal is dependent on the values of these indicators. Eventually, we derive the classifier in which quality is measured on the basis of accuracy, recall, and precision. Further, we use signals generated by the classifier to adjust the account balance of the decision-maker and estimate the relation between the quality of classification and the final account balance. Experiments are performed using the trading system implemented by the authors on the real-world data covering several years.

I INTRODUCTION

In this article, we try to fill the gap related to the portfolio problem on the forex market and move towards the investing process based on the set of instruments simultaneously, rather than consider a set of signals independently. To do so, we present the investing approach involving the classification methods used to generate signals. Further, these signals are considered simultaneously, leading to the situation that the decision-maker portfolio could include several currency pairs for a given time. Profit/loss from these currency pairs is adjusted to the decision-maker portfolio. However, we assume that no additional information about the correlation among instruments is considered.

II ORIGINAL CONTRIBUTION

Our main goal is to investigate whether the quality of signals classification performed on the data is directly related to the profits achieved by the decision-maker at the end of the investing period. Therefore, we implement a trading system including the signals classification module and investing module to verify that. Furthermore, we compare the classification quality performed on the well-known algorithms with the final account balance measured in dollars.

III METHODOLOGY

In our system we use the following notations and concepts:

- t will be a discrete moment of time (reading) in which the instrument (currency pair) value and the market indicator values are calculated;
- T will be a time period, for which the whole investing process occurs. T consists of large number of t;
- CP will be a full set of currency pairs available in the system with cp_i as i-th currency pair;
- time period – will be a time, which has to pass between two successive readings;
- I – will be a set of indicators available in the system;
- PT – will be the portfolio of the decision-maker. This set is initially empty, however, the currency pairs cp are added to the portfolio, while the signals are observed on the market;
- c – will be the counter indicating the number of readings, for which the given cp is already in the portfolio;

- *max* – is the maximal number of readings, for which the cp_i can be present in the portfolio;
- *p* – is the number of readings which must pass to evaluate the decision for a given signal.

Algorithm 1: The investing process

```
begin
    Create the empty set $PT = \emptyset$
    Select set of currency pairs $CP$ and number of readings $T$
    Set additional parameters (number of decision classess and counter $c$)
    for each reading $t$ in $T$ do
        for each $cp_i$ in $CP$ do
            Set decision class for $cp_i$ according to formula (2)
            if decision for $cp_i$ is (BUY OR STRONG BUY) AND $cp_i \notin PT$ then
                add $cp_i$ to $PT$
                set counter $c = 0$ for $cp_i$
            if decision for $cp_i$ is (SELL OR STRONG SELL) AND $cp_i \in PT$ then
                remove $cp_i$ from $PT$
                update the value of account in $
        for each $cp_i$ in $PT$ do
            increase counter $c$
            if $c = max$ for $cp_i$ then
                remove $cp_i$ from $PT$
                update the value of account in $
    close all opened positions
    derive the final account balance to the decision-maker
```

IV Results

We performed the classification with the use of the CART algorithm. In actual results, we focused on deriving the decision about the general instrument value direction; however, from the point of view of the decision-maker, it could be essential to know the strength of the movement as well. Moreover, the number of decision classes was set arbitrarily. Still, it is possible to include more decision classes corresponding to the trend's strength (in such an example, the decision class depends directly on the range of the instrument value movement). In border cases, it is even possible to move towards the regression task, where instead of predicting the decision class, we rather expect the exact instrument value.

V Conclusions

Despite the diversity of data used in the experiments, results were ambiguous, and no direct relation between the quality of classification and the account balance was observed. It concludes that these two problems (data classification and portfolio management on the forex market) should be considered independently. Thus the relatively good classification of the data does not imply the overall positive account balance at the end of the investing period.

The analyzed classification problem could be easily transformed into a regression case. Instead of predicting the decision classes, we rather focus on deriving the exact value of the instrument. The common assumption in the investing field is to know the general direction of the instrument value; however, for ongoing transactions learning the exact instrument value range could be vital.

Link to arXiv: http://arxiv.org/abs/2211.08799
DOI: https://doi.org/10.1007/978-3-031-21967-2_14

Keywords: recommendation system, association rules, apriori algorithm, frequent pattern growth algorithm, factorization machines, prediction's time, quality of recommendations

Speeding Up Recommender Systems Using Association Rules

Eyad Kannout[0000-0001-7543-774X], Hung Son Nguyen[0000-0002-3236-5456], Marek Grzegorowski[0000-0002-3236-5456]

eyad.kannout@mimuw.edu.pl, son@mimuw.edu.pl, m.grzegorowski@mimuw.edu.pl

Simplified Title

Improving Recommendation Speed Using Frequent Pattern Mining and Association Rules.

Abstract

Recommender systems are considered one of the most rapidly growing branches of Artificial Intelligence. The demand for finding more efficient techniques to generate recommendations becomes urgent. However, many recommendations become useless if there is a delay in generating and showing them to the user. Therefore, we focus on improving the speed of recommendation systems without impacting the accuracy. In this paper, we suggest a novel recommender system based on Factorization Machines and Association Rules (FMAR). We introduce an approach to generate association rules using two algorithms: (i) apriori and (ii) frequent pattern (FP) growth. These association rules will be utilized to reduce the number of items passed to the factorization machines recommendation model. We show that FMAR has significantly decreased the number of new items that the recommender system has to predict and hence, decreased the required time for generating the recommendations. On the other hand, while building the FMAR tool, we concentrate on making a balance between prediction time and accuracy of generated recommendations to ensure that the accuracy is not significantly impacted compared to the accuracy of using factorization machines without association rules.

I Introduction

Recommendation systems can be built using different techniques to generate more relevant recommendations for the users [2]. However, these recommendations might become useless if the recommendation engine does not produce them in a proper time frame. In this paper, we work on finding a novel approach that incorporates association rules in generating the recommendations using the factorization machines algorithm to improve the efficiency of recommendation systems. It is worth noting that the factorization machine model is used to evaluate our method and compare the latency of FM before and after using the association rules. However, in practice, our method can be combined with any other recommendation engine to speed up its recommendations.

II State of the Art

Many online services are trying to boost their sales by implementing recommendation systems that estimate users' preferences or ratings to generate personalized offers and thus recommend items that are interesting for the users [1]. Recommendation systems can be built using different techniques which leverage the rating history and possibly some other information, such as users' demographics and items' characteristics [2]. The goal is to generate more relevant recommendations. However, these recommendations might become useless if the recommendation engine does not produce them in a proper time frame.

III Original Contribution

The main contributions of this paper are as follows: 1) proposing a method that uses the apriori algorithm or frequent pattern growth (FP-growth) algorithm to generate association rules which suggest items for every user based on the rating history of all users; 2) utilizing these association rules to create short-listed set of items that we need to generate predictions for them; 3) employing factorization machines model to predict missing user preferences for the short-listed set of items and evaluate the top-N produced predictions.

IV Methodology

In this section, we introduce the FMAR framework, which proposes two hybrid models that utilize factorization machines and frequent pattern mining algorithms to speed up the process of generating the recommendations.

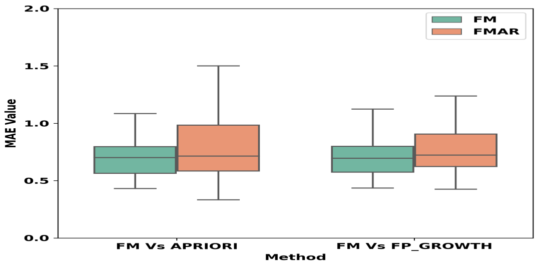


Figure 1: MAE Comparison

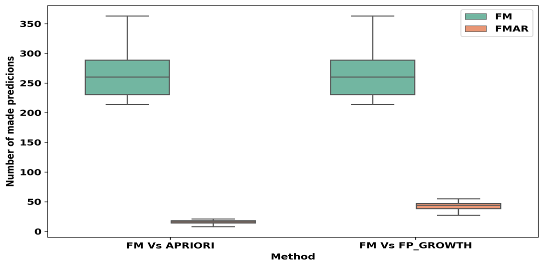


Figure 2: Comparison of the speed of methods

IV.1 Factorization Machine Apriori Based Model

In this model, we utilize factorization machines and apriori algorithms to speed up the process of generating the recommendations. Firstly, we use the apriori algorithm to create a set of association rules based on the rating history of users. Secondly, we use these rules to create users' profile which recommends a set of items for every user. Then, when we need to generate recommendations for a user, we find all products that are not rated before by this user, and instead of generating predictions for all of them, we filter them using the items in the users' profile. Finally, we pass the short-listed set of items to a recommender system to estimate their ratings using FM model.

IV.2 Factorization Machine FP-Growth based Model

In this model, we introduce the second version of FMAR where FP-growth algorithm has been employed to generate the association rules. In general, FP-growth algorithm is considered as an improved version of apriori method. However, what makes FP-growth algorithm different from apriori algorithm is the fact that in FP-growth no candidate generation is required. This is achieved by using FP-tree (frequent pattern tree) data structure which stores all data in a concise and compact way. Moreover, once the FP-tree is constructed, we can directly use a recursive divide-and-conquer approach to efficiently mine the frequent itemsets without any need to scan the database over and over again. After finding the frequent itemsets, we generate the association rules and users' profiles in the same way as in FM Apriori-based model.

V Results

We conduct comprehensive experiments using MovieLens 100K dataset to evaluate the performance of the FMAR recommender system. In the first experiment, we calculate the mean absolute error (MAE) generated in both recommendation engines. The main goal of this approach is to show that the quality of recommendations is not significantly impacted after filtering the items in the testing set using the association rules. In the second experiment, we evaluate FMAR by comparing its recommendation with FM using Normalized Discounted Cumulative Gain (NDCG). In the third evaluation method, we run Wilcoxon Rank-Sum test on the results of previous experiments. In the last experiment, we compare FM and FMAR in terms of the speed of their operation, measured as the number of predictions performed by the factorization machines model.

VI Evaluation

The experimental results show that FMAR has improved the efficiency of recommender systems. Furthermore, the experiments also indicate that the accuracy of FMAR is very close to the results produced by the standard recommender system (cf. Fig 1 and Fig 2).

VII Conclusions

This article introduces FMAR, a novel recommender system, which methodically incorporates the association rules in generating the recommendations using the factorization machines model. Our study evaluates two approaches to creating association rules based on the users' rating history, namely: the apriori and frequent pattern growth algorithms. These rules are used to decrease the number of items passed to the model to estimate the ratings, reducing the latency of the recommender system prediction.

References

[1] Kannout, E. Context clustering-based recommender systems. In *Proceedings of the FedCSIS 2020* (2020), M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds., vol. 21 of *ACSIS*, pp. 85–91.

[2] Kannout, E., Grodzki, M., and Grzegorowski, M. Utilizing frequent pattern mining for solving cold-start problem in recommender systems. In *Proceedings of the FedCSIS 2022* (2022), M. Ganzha, L. A. Maciaszek, M. Paprzycki, and D. Slezak, Eds., vol. 30 of *ACSIS*, pp. 217–226.

DOI: https://doi.org/10.1007/978-3-031-21967-2_47

Keywords: forecast, trend, MusicBrainz, music industry

Music industry trend forecasting based on MusicBrainz metadata

Marek Kopel[0000-0002-2273-684X], Damian Kreisich

marek.kopel@pwr.edu.pl,238028@student.pwr.edu.pl

SIMPLIFIED TITLE

Music industry trend forecasting.

ABSTRACT

In this paper forecast analysis for music industry is performed. The trends for years 2020-2024 are calculated based on forecasting for time-series metadata from online MusicBrainz dataset. The analysis takes on music releases throughout the years from different perspectives, e.g. the release format, type of music releases or release length (playback time). In the end, all the results are discussed before final conclusions are drawn.

I INTRODUCTION

The dataset used in this research is metadata for music releases, artists, recordings, and other entities connected to information about the music industry. This paper's theme is data analysis within the music industry – focusing on trends within the data. The analysis topics focus on music releases, which carry information on what was released, how it was released, and what is featured within the releases among other. The metadata source is an open database MusicBrainz. The specific goal is to gather knowledge on trends and changes in the music industry and prepare a solution to perform similar analyses on demand.

II STATE OF THE ART

MusicBrainz is a community-maintained database on music releases. It was used previously e.g., to analyze artist mutual influence [1], a gender gap and gender differences in the music industry, music sentiment and genre classification. The authors are unaware of any existing trend analysis for MusicBrainz metadata. On the other hand, trend forecasting is another field on its own. Trends can be analyzed for any kind of time-related data. Just to name a few examples: forecasting a price of a cryptocurrency, fashion trends, oil prices, and seasonal climate conditions.

III ORIGINAL CONTRIBUTION

The chosen forecasting method is Prophet. It's model has been applied to analyze (1980-2019) and predict (2019-2024) the trends on:

- music recordings released as single vs album,
- medium for release: LP (vinyl), MC (cassette), CD, digital (mp3, etc.),
- album playback time and the number of tracks it contains.

IV METHODOLOGY

Prophet (presented in [2]) is chosen as the forecasting method. The components of the Prophet model are linear regression with changepoints, seasonality and holidays/events. Every parameter for Prophet model can be set by hand using analyst in the loop approach – visualize, adapt the model and repeat. Regarding the parameters, the Prophet has capabilities to automate the process when the hyper parameters are set properly. The parameters that are sought after during the tuning process are: *changepoint_prior_scale*, which directly affects the amount of changepoints set in the model, *seasonality_prior_scale* and *seasonality_prior_mode*, which are both responsible for the seasonality component itself.

V Results

As a result of forecasting for album releases there is a simple conclusion, that most likely number of albums released within few next years will most likely be falling down. When it comes to singles, the situation feels straightforward, the only possible outcome from the forecast is rise in the number of released singles.

The digital release trend is almost linear, rising from nearly 0% in the early '00s. By the year 2024, we can expect 94%, even up to 99%, of the music to be released in digital format. This means that the physical format forecast is all down to 0%: CD had the highest releases up to 80% in '90s and '00s, MC - dropped drastically throughout '80s from 90% and LP stayed statically low at a few % the whole analyzed time.

The average number of tracks per album rose from slightly below 10 in the mid-'80s to a peak of a little above 12 by the end of the millennium and then dropped back to 10.5 today and is expected to go again below 10 in the next 5 years. The forecast for album length is a derivative of the number of tracks. There are also no big magnitude changes. In the analyzed scope the average album playback time stays between 40 and 50 minutes with a decreasing trend. New album releases should aim at around 40 minutes in length.

VI Evaluation

Using Prophet requires specific column names for the data, which are *ds* and *y*, where *ds* stands for the time and *y* is the value of a variable at the time. Data with the labels specified properly can be directly fed into the Prophet model, which results in setting up all the parameters in the model, taking into account all the parameters set by the analyst. In the end, the model can produce a forecast with uncertainty, in correlation to the dataset, with built-in Python visualisation tools. It makes it the best solution to use for music metadata-based forecasting, as it provides all the desired tools, and ways to visualise the forecast along with counting the errors and accuracy for the model, based on cross-validation.

VII Conclusions

The outcomes should be interesting for musicians and anybody wanting to re- lease music in the next few years. The timeline covered in the analysis is years 1980 – 2019, with forecasts reaching up to the year 2024. Trending is connected to the changes in society and technology. This is what drives the music industry to new grounds every few years. In terms of media, it was e.g. the CDs in the mid-80's or the digital era in music that formed in 2004 in the form of online music stores and streaming services. Today no musicians or labels can afford to avoid the internet as the leading media to release music, socialize with fans, and target and market their products.

References

[1] Kopel, M. Analyzing music metadata on artist influence. In *Asian Conference on Intelligent Information and Database Systems* (2015), Springer, pp. 56–65.

[2] Taylor, S., and Letham, B. Forecasting at scale. peerj preprints, 2017.

DOI: 10.1007/978-981-19-8234-7_18
Keywords: Non-small cell lung cancer, metastases, survival analysis

Predicting metastasis-free survival using clinical data in non-small cell lung cancer

Emilia Kozłowska[0000-0002-3069-3085], Monika Giglok[0000-0002-8196-5191], Iwona Dębosz-Suwińska[0000-0002-4554-8905], Rafał Suwiński[0000-0002-3895-7938], Andrzej Świerniak[0000-0002-5698-5721]

emilia.kozlowska@polsl.pl, andrzej.swierniak@polsl.pl

SIMPLIFIED TITLE

Prediction of time to metastasis based on clinical information

ABSTRACT

Lung cancer is the most common and the deadliest type of cancer with 5-year overall survival equal to 15%. One of the main reasons for the high mortality of lung cancer is the development of local and distant metastases. Lung cancer patients mostly die because of distant metastases rather than the primary tumor. Thus, here we tackle the problem of predicting when a patient relapse with a distant metastatic tumor. This information is relevant not only to assess a patient's prognosis but also to guide the first-line treatment. Here, we applied clinical data from over 400 patients to predict the time to metastatic relapse which is also called metastasis-free survival (MFS). Using Cox regression, we have got a fairly good prediction with a c-index =0.63 for a model with three clinical covariates. In addition, we created also a nomogram that could be applied to predicting the probability of metastases in newly diagnosed patients. In conclusion, solely based on clinical data, it is possible to predict the time to metastasis.

I INTRODUCTION

The main goal is to predict the time elapsed from cancer diagnosis until the emergence of metastasis, i.e., distant tumor. The intermediate goals to achieve this are the following: 1) acquisition and processing of clinical data, 2) identification of significant clinical features that could be applied for prediction of time to metastasis, and 3) development of a statistical model that could predict the risk of metastasis.

II STATE OF THE ART

Currently, there is no single biomarker that could be applied for the prediction of time to metastasis in lung cancer. Researchers are still working on finding biomarkers to detect if metastasis will appear sooner or later. We are even far from understanding the mechanism of metastasis emergence.

III ORIGINAL CONTRIBUTION

There are available statistical models that could predict the emergence of brain metastases of lung cancer [1]. However, there is no model that could indicate if patient will develop metastasis. Here we aim to fill the gap.

IV METHODOLOGY

We utilized a cohort of non-small cell lung cancer patients' clinical data for metastasis prediction. We applied survival analysis, i.e., Kaplan-Meier estimator and Cox regression, for metastasis prediction.

V RESULTS

We developed a clinical data signature that could be applied to predict the time to metastasis. The signature is based on three clinical features: a total dose of radiotherapy, platelet level at the diagnosis, and N that describe local tumor dissemination to lymph nodes.

VI EVALUATION

We applied classical survival analysis that aims at predicting time-to-event that could range from predicting patient death to time-to-patient relapse. The Kaplan-Meier model was applied to estimate the probability of not developing metastasis in a given time interval. Cox regression, however, predicts so-called hazard ratio.

As a metric of a Cox regression model predictive power, we applied a concordance index that is between 0 and 1. A value above 0.5 means that the model performs well.

VII CONCLUSIONS

The developed statistical model could be applied in predicting the risk of metastasis at the time of diagnosis by oncologists.

REFERENCES

[1] KAMER, I., STEUERMAN, Y., DANIEL-MESHULAM, I., PERRY, G., IZRAELI, S., PERELMAN, M., GOLAN, N., SIMANSKY, D., BARSHACK, I., NUN, A. B., GOTTFRIED, T., ONN, A., GAT-VIKS, I., AND BAR, J. Predicting brain metastasis in early stage non-small cell lung cancer patients by gene expression profiling. *Translational Lung Cancer Research 9*, 3 (jun 2020), 682

DOI: https://doi.org/10.1007/978-3-031-21743-2_56

Keywords: convolutional neural network, deep learning, knowledge discovery, performance evaluation metrics

CORDIS Partner Matching Algorithm for Recommender Systems

Dariusz Król[0000-0002-2715-6000], Zuzanna Zborowska, Paweł Ropa, Łukasz Kincel

dariusz.krol@pwr.edu.pl, p.ropa@nomtek.com, l.kincel@nomtek.com

SIMPLIFIED TITLE

Developing a recommendation method for matching scientific and business partners from the Community Research and Development Information Service (CORDIS).

ABSTRACT

The purpose of this paper is developing a method for recommending business and scientific partners matchmaking with use of a deep learning model based on historical data on previously completed European Union projects. The paper starts with an introduction to recommender systems, followed by a systematic literature review on the subject. The next part describes the course of the research and its implementation of two deep learning approaches: (1) the entity embeddings of organizations and (2) the embedding space of keywords. The paper ends with a summary of the entity embedding-based recommendation characterized by coverage, average accuracy, low Gini index and the entropy measure.

I INTRODUCTION

The subject of the work lies within the field of machine learning, in particular deep learning, determining the closeness of business and scientific relationships, as well as defining a measure of matchmaking success. The work focuses on the construction of a learned classifier that performs the task of selecting the most likely partnership and evaluating its effectiveness. The learned model will be tested experimentally, taking into consideration measures of recommender system evaluation presented in literature. The result of the project will be the creation of a prototype recommending a ready-made solution, i.e. it will identify partners who have similar business/scientific goals and thus their relationship will potentially show the greatest similarity. Potential recipients may be organizations looking for institutions to collaborate with as well as the organizers of conferences or fairs.

II STATE OF THE ART

The recommender systems can be divided into three categories.

- *Content-based recommender systems* - attributes of items are compared to find similar entities. The main advantage of content-based recommendation is that the items can be recommended even though they are not yet rated by any user. However the method may lack novelty as it has a high probability of recommending items from the same category.
- *Collaborative recommender systems* - recommendation is based on the preferences of other users (user-based) or based on the rating a user has given to similar items (item-based). The first subtype bases on the assumption that users, who made similar decisions in the past, are likely to still act in a similar manner. Among the techniques used within that type is Vector Space Model and Term Frequency-Inverse Document Frequency (TF-IDF).
- *Hybrid recommender systems* - a technique integrating two or more methods to benefit from their advantages and avoid their flaws.

III ORIGINAL CONTRIBUTION

The problem of recommendation in the partner matchmaking process is neither pure content-based nor collaborative in nature. Therefore the approach that needs to be followed must be of a hybrid nature.

IV Methodology

Before the data can be fitted into neural network it must undergo several preparation procedures. First, it has to be fetched from the data source, which is not always straightforward. Secondly, the data has to be analyzed and cleaned. We have to make sure the data has the most accurate entries. The next step is the preparation of descriptions. Descriptions have to be tokenized, we have to remove stop words and the resulting set must be simplified to its lemmatized form. This can be achieved with Python based tools and libraries. After the data preparation we were left with **2384** organizations with lemmatized meaningful words that could be further processed to eventually be fitted into the neural network model.

V Results

Two neural network models and four different recommendation list creation methods were proposed. The proposed neural network models base on the connections between the organizations that took part in the same projects in the past. If a new organization appears in the database or the description of the organization is updated both of the models would have to be trained again. The small portion of the data that we experimented on makes training the model time consuming only to a small extent but bigger amounts could slow down remarkably the recommendation process. On the other hand, the proposed method takes into consideration the attributes of organization and their relation to the others which makes it more resistant to scarcity of data in one of the fields.

VI Evaluation

In our case we need to choose evaluation metrics to determine which recommendation list among those created is the best. There are 4 metrics proposed: *accuracy*, *coverage*, *Gini index* and *entropy*. Each metric determines a characteristic of a distribution and we asses obtained lists in terms of the desired value of a given metric. The best accuracy is achieved in the case of a recommendation list based on ranks. Entity and keywords lists reach similar values (0.5) for all the chosen lengths. The worst in terms of accuracy is the list based on averages. Its values do not exceed 0.2. The second metric that we chose to evaluate the recommendation lists was coverage. Coverage is the percentage of items which appear at least once in the recommendation lists [1]. Besides the accuracy of recommended items and the coverage of the organizations in the lists we decided to check fairness of the distribution of recommended items. *Gini index* is a measure that takes into account how uniformly items appear in the recommendation lists [1]. The last measure was entropy. Entropy is the measure of the uniformity of a distribution. Uniform distributions have higher information gain and therefore are more desired when higher diversity of recommended items is needed.

The results of training and the evaluation of obtained recommendation lists do not lead to a strong conclusion that any of the proposed list creation methods is significantly better than the others. One recommendation list is always significantly better than the other but it differs depending on which metric is taken into consideration. However, for a majority of metrics, the entity embedding-based list gives good results. It is characterized by high coverage, average accuracy, low Gini index and the highest entropy. If any method were to be chosen to be proposed, the method based purely on the entity embedding-based is the first to recommend.

VII Conclusions

The main goal of the research was to perform experiments on the available data with a use of different deep learning techniques. After a detailed systematic literature review there were two deep learning approaches proposed. The first focuses on the entity embeddings of organisations. During the learning process it tries to place similar organisations closer in the embedding space and thus provides a possibility to determine which partners may be interested in a collaboration. The second, on the other hand, focuses on the keywords describing each organisation and creates an embedding space of those keywords. Based on the keywords' proximity the model tries to predict whether two organisations characterized with a given set of keywords may have collaborated. The two models give us a measure of similarity which in the end lets us create four recommendation lists. Using common evaluation measures we evaluated which of the created lists gives the best recommendation. In the course of the research it became clear that recommendation problems are complex and require a lot of experimentation to adjust all the parameters to fit the given data. The evaluation part showed that even effective prediction models may not be the most effective when it comes to create a diverse, uniformly distributed and accurate recommendation lists. Because the problem is so complex there is no simple solution to it problem and the answer differs depending for example on the length of the list that we want to obtain or whether we want to create more accurate or more diverse lists.

References

[1] Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., and Burke, R. Fairmatch: A graph-based approach for improving aggregate diversity in recommender systems. pp. 154–162.

DOI: `https://doi.org/10.1007/978-981-19-8234-7_19`
Keywords: Parallel programming, TPL, Performance of processing, Data parallelism

Impact of Design Decisions on Performance of Embarrassingly Parallel .NET Database Application

Piotr Karwaczyński, Marcin Sitko, Sylwia Pietras, Bogdan Marczuk,
Jan Kwiatkowski[0000-0003-3145-0947], and Mariusz Fraś[0000-0001-5534-3009]

{pkarwaczynski,msitko,spietras,bmarczuk}@sygnity.pl
{mariusz.fras,jan.kwiatkowski}@pwr.edu.pl

Simplified Title

The impact of design decisions on performance of such a .net database application that can be easily decomposed to a set of parallel tasks

Abstract

The implementation of parallel applications is always a challenge. It embraces many distinctive design decisions that are to be taken. The paper presents issues of parallel processing with use of .NET applications and popular database management systems. In the paper, three design dilemmas are addressed: how efficient is the auto-parallelism implemented in the .NET TPL library, how do popular DBMSes differ in serving parallel requests, and what is the optimal size of data chunks in the data parallelism scheme. All of them are analyzed in the context of the typical and practical business case originated from IT solutions which are dedicated for the energy market participants. The paper presents the results of experiments conducted in a controlled, on-premises environment. The experiments allowed to compare the performance of the TPL auto-parallelism with a wide range of manually-set numbers of worker threads. They also helped to evaluate 4 DBMSes: Oracle, MySQL, PostgreSQL, and MSSQL in the scenario of serving parallel queries. Finally, they showed the impact of data chunk sizes on the overall performance.

I Introduction

The primary purpose of the research was to answer three questions: (1) How efficient is the auto-parallelism implemented in .NET TPL? (2) How do popular DBMSes differ in serving parallel requests? (3) What is the impact of the size of data chunks on the performance of tested service? These questions arose during the development of Meter Data Management (MDM) system, supporting storage and management of data in the energy industry. A typical pattern for data processing in MDM is to perform the same operation on each element of a large collection of measurement data. To improve performance, the parallel processing of data is used with the objective to distribute the load between computing units of a multi-core CPU of the application server.

One of a few key design issues was to control the number of threads allocated to a parallel algorithm. In the TPL, auto-parallelism (automatic parallelization of the code) is implemented – it automatically picks a number of parallel threads and partitions the computations in a loop. However, it was not known how efficient is the auto-parallelism implemented in TPL. Another issue was the choice of a relational DBMS. It was decided to check to what extent the choice of the DBMS impacts the overall performance of the business scenario in which many concurrent, high-volume reads are performed, as in MDM. Finally, selection of the optimal size of data chunks turned out to be disputable. On one extreme, no fragmentation means sequential execution. On the other side, too tiny data chunks impose large communication overheads.

II State of the Art

The Task Parallel Library was introduced in .NET Framework 4.0 in 2010. It is a well-established and convenient way of achieving parallelism in a C# code. It is able to dynamically scale the degree of concurrency in order to use available computing units efficiently. However, there are few publications analyzing its performance and none evaluating its auto-parallelism methods.

Interesting insights on speedup achieved through parallelization of code for computing π using TPL and OpenMP are given in [1]. The authors show that even for the embarrassingly parallel problem the libraries do not achieve linear speedup. The problem considered in that paper varies from ours as it completely focuses on pure calculations. TPL performance is measured in [2] to compare the library with another concurrent computation model – the actor model. The authors ran tests with different numbers of threads but the results were published as averages. Consequently, it was impossible to compare our and their trends. Handling concurrency in selected databases is addressed in [3]. Its authors focus on testing the implementations of isolation levels. To some extent, this aspect may influence our results, however it is not directly aligned with our research questions.

III ORIGINAL CONTRIBUTION

The original contribution of the reseach are the experimental evaluations of: (1) the auto-parallelism implemented in .NET TPL, (2) the impact of the DBMS on the overall performance of the business scenario in which many concurrent, high-volume reads are performed, and (3) the impact of data chunk sizes on the performance of the tested parallel operations.

IV METHODOLOGY

The conducted experiments were based on the business case that consists in reading the daily profiles of electricity consumption with a 15-minute resolution from the measurement database for the purpose of their validation, transformation, and transfer to the appropriate recipient. It is one of the basic services of the MDM. It significantly strains computational resources in production deployments as well as it is very susceptible to parallelization. The experiments were performed on two 16-core servers with hyper-threading.

In the first experiment, the efficiency of TPL auto-parallelism method was measured and compared to eight degrees of parallelization manually set to: 1, 2, 4, 8, 16, 32, 64, and 128 threads. The tests were performed for each of the four considered DBMSes. In the second experiment, different sizes of data chunks were used in order to measure their impact on the performance of the parallel application. To this end, the daily consumption profiles for 100k measurement points (9,6m of data records) were partitioned into separate data chunks of various sizes, then read and processed in parallel. The tests were performed for each of the 4 considered DBMSes, with the TPL auto-parallelism method enabled.

V RESULTS

In short, the degree of parallelization automatically determined by the TPL reached data reading rates very close to those achieved by the manually set number of concurrent threads equaled the number of processor cores. Regarding DBMSes, the reading rates were the highest and very similar for MSSQL and PostgreSQL. Oracle coped slightly worse, especially for more threads. MySQL seemed to make the least use of parallelization and got the poorest results. Interestingly, it was observed that the higher the reading rate, the wider the confidence intervals for the reading rate means. Using such a data chunk size that each thread from the pool has exactly one task to be performed gave maximum read rates. Smaller sizes increased the time spent on communication between the application and database servers. Larger sizes reduced the possibility of parallel execution of the task.

VI EVALUATION

If the performance is a top priority, the manually set degree of parallelization backed with a few experiments would be a better choice than the TPL auto-parallelism method. However, if the target application runtime environment is unknown or may be subject to changes, opting for the latter will be a rational approach, setting the degree of parallelization several percent worse than optimal.
The meaningful differences in handling high-volume, concurrent reads by considered DBMSes were observed. The experimental results show what performance levels may be expected for each of them in a function of a number of threads requesting data.
Finally, if the optimal number of threads for a given problem and an execution environment is known, then it is rational to consider such distribution of data among threads where each thread does exactly one iteration.

VII CONCLUSIONS

The achieved results have been allowing us to make the right design decisions, based on knowledge instead of intuition. The outcomes of the conducted research may be directly applied in any database applications with parallel data processing, built on .NET framework and TPL library.

REFERENCES

[1] VIŠTICA, M., HASELJIĆ, H., MAKSUMIĆ, A., NOSOVIĆ, N.: *Comparison of speedups for computing π using .NET TPL and OpenMP parallelization techonologies*. In: X Int. Symp. Telecommunications (BIHTEL), 2014.

[2] ZMARANDA D., POP-FELE L.-L., GYŐRÖDI R., GYŐRÖDI C.: *Actor Model versus TPL for applications development*. In: Proc. 16th Int. Conf. Engineering of Modern Electric Systems (EMES), 2021.

[3] LIAROKAPIS, D., O'NEIL, E., O'NEIL, P.: *Testing Concurrency in Databases still Matters*. In: Int. Conf. Information Technologies (InfoTech), 2020.

DOI: `https://doi.org/10.1007/978-3-031-21743-2_5`
Link to repository: `https://github.com/hduc-le/SimCPSR`

Keywords: paper submission recommendation, contrastive learning, sentence embedding, recommendation system

SimCPSR: Simple Contrastive Learning for Paper Submission Recommendation System

Duc H. Le[†], Tram T. Doan[†], Son T. Huynh, Binh T. Nguyen

`{hduc.lee,tramdoan93,huynhthanh1234vn}@gmail.com,ngtbinh@hcmus.edu.vn`

SIMPLIFIED TITLE

SimCPSR: Simple Contrastive Learning for Paper Submission Recommendation System

ABSTRACT

The recommendation system plays a vital role in many areas, especially in academic fields, to support researchers in submitting and increasing the acceptance of their work through the conference or journal selection process. This study proposes a transformer-based model using transfer learning as an efficient approach for the paper submission recommendation system. By combining essential information (such as the title, the abstract, and the list of keywords) with the aims & scopes of journals, the model can recommend the Top K journals that maximize the acceptance of the paper. Our model had developed through two stages: (i) Fine-tuning the pre-trained language model (LM) with a simple contrastive learning framework. We utilized a simple supervised contrastive objective to fine-tune all parameters, encouraging the LM to learn the document representation effectively. (ii) The fine-tuned LM was then trained on different combinations of the features for the downstream task. This study suggests a more advanced method for enhancing the efficiency of the paper submission recommendation system compared to previous approaches when we respectively achieve 0.5173, 0.8097, 0.8862, 0.9496 for Top 1, 3, 5, and 10 accuracies on the test set for combining the title, abstract, and keywords as input features. Incorporating the journals' aims and scopes, our model shows an exciting result by getting 0.5194, 0.8112, 0.8866, and 0.9496, respecting Top 1, 3, 5, and 10. We provide the implementation and datasets for further reference at https://github.com/hduc-le/SimCPSR.

I INTRODUCTION

Choosing a suitable journal for submitting new work is difficult for most researchers, including young and experienced people. Wang and coworkers proposed the recommendation system for computer science publications early to support the researchers in selecting a relevant journal to increase their work acceptance opportunities. The problem has been studied by many other researchers later.

In this work, we investigate the paper submission recommendation problem using general paper information and the aims and scopes of the journals as sufficient attributes in our method. We aim to deeply extract the semantic relationship between the paper submission and the journal through those available features. We tackle the problem by applying the transformer architecture [3] as an encoder to extract the input feature effectively and utilize the contrastive learning framework [1] to enhance the model's robustness in the downstream task.

II STATE OF THE ART

The paper recommendation system had developed by using statistical frameworks such as Chi-square, the term frequency-inverse document frequency (TF-IDF) for feature selection from the paper attributes, and linear logistic regression (LLR) or Multi-layer Perceptrons (MLP) model to classify relevant journals. To improve the model efficiency in feature extraction, researchers address the problem by using GloVe and FastText, combining Convolutional Neural Network (CNN) and LSTM, then show a significant improvement. However, using pre-trained word embedding models like GloVe or FastText could draw some limitations since the embedding vectors are not learnable. Furthermore, because of the complexity of natural language, the CNNs model could be less efficient in capturing the language semantics.

[†] Master students at the University of Science, Vietnam National University in Ho Chi Minh City

III Original Contribution

The contribution of our work can be described as follows: (1) We propose a new framework for the paper submission recommendation problem using the transformer architecture, which shows a significant advance to tackle. Moreover, the experimental results show that our approach has a competitive performance compared to the previous works. (2) Leveraging contrastive learning, the powerful method of sentence embedding, we enhance the framework's robustness in learning semantic relationships among documents or sentences. (3) Our method provides a basic framework that can be extended further by applying other transformer models to achieve better performance in the paper submission recommendation problem.

IV Methodology

To learn a sufficient sentence representation for the input document, we present a simple contrastive learning framework whose contrastive loss function can be used to fine-tune model parameters. For a mini-batch of N input pairs, let $\mathbf{h}_i$ and $\mathbf{h}_i^+$ denote the embedding or latent representation of x_i and x_i^+; the contrastive objective was defined as:

$$\ell_i = -\log \frac{e^{sim(\mathbf{h}_i,\mathbf{h}_i^+)/\tau}}{\sum_{i=1}^{N} e^{sim(\mathbf{h}_i,\mathbf{h}_j^+)/\tau}}, \tag{1}$$

where τ is a temperature hyper-parameter and $sim(\mathbf{h}_1, \mathbf{h}_2)$ is the cosine similarity.

After the fine-tuning process, the fine-tuned LM can be applied to different feature combinations for the downstream task on groups of attributes. Furthermore, during the classification stage, for a particular embedding vector representing a paper, we compute the cosine similarity between it and all of the embedding vectors of journals to obtain the cosine feature vector, which serves as a contributed feature. After that, we concatenate them together for feeding to a final Softmax layer to classify Top K labels.

V Results

Our experiments, which involve various combinations of paper attributes, demonstrate that, in contrast to earlier approaches, the proposed model can help to increase performance when incorporating additional information regarding the "aims & scope" of journals. Furthermore, our result demonstrates the significance of the abstract feature when a model using input data that includes it performs significantly. For example, the best performance of the proposed method in the Accuracy@K (K=1, 3, 5, 10) is 0.5194, 0.8112, 0.8866, and 0.9496 when using combinations of paper attributes and "aims & scope" of the journal.

VI Evaluation

We use the set of paired paper attributes with the corresponding aims and scope of the journal to perform fine-tuning on the Distil-RoBERTa (a distilled version of RoBERTa [2]) model. Then we train it on different combinations of features to solve the classification problem as the downstream task. We then use the AdamW to optimize the Cross-Entropy loss. To get the Top K values—which represent the probability that an input can match the K labels—we finally apply the Softmax layer and compute the accuracy at each K value. Finally, we utilize the Accuracy@K metric, defined as the ratio between the number of accurate items and the number of viewed items at each K, to evaluate the performance of the proposed model, where K = 1, 3, 5, and 10.

VII Conclusions

The experimental results show that the proposed approach has competitive performance and is an advanced method for enhancing the efficiency of the paper submission recommendation system. Our study especially gives a practical approach or a basic framework to utilize the transformer model to obtain efficient sentence embedding. Based on this method's idea, one can apply it to other studies in learning effective sentence representation in recommendation systems separately and natural language processing in general.

References

[1] Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings, 2021.

[2] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR abs/1907.11692* (2019).

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017.

DOI: https://doi.org/10.1007/978-3-031-21743-2_19

Keywords: Knowledge Graph Embedding, Link Prediction, Dynamic Convolution, Attention over Convolution Kernel

Embedding Model with Attention over Convolution Kernels and Dynamic Mapping Matrix for Link Prediction

Thanh Le[0000-0002-2180-4222], Nam Le[0000-0002-2273-5089], Bac Le[0000-0002-4306-6945]

{lnthanh,lhbac}@fit.hcmus.edu.vn,18120061@student.hcmus.edu.vn

Simplified Title

A link prediction model for knowledge graph based on deep learning and dynamic mapping matrix approaches

Abstract

Knowledge Graph Completion, especially its sub-task link prediction attracts the attention of the research community and industry because of its applicability as a premise for developing several potential applications. Knowledge graph embedding (KGE) shows promising results to solve this problem. This paper focuses on the neural networks-based approach for KGE, which can extract features from the graphs better than other groups of embedding methods. The ConvE model is the first work using 2D convolution over embeddings and stacking multiple nonlinear feature layers to model knowledge graphs. However, its computation is inefficient and does not preserve translation between entity and relation embedding. Therefore, dynamic convolution was designed to solve limited representation capability issues and show the promised performance. This work introduces a mixture model that incorporates attention into performing the convolutional operation on projection embeddings. The TransD idea is used to project entity embedding from entity space to relation space. Then, it is stacked with relation embedding to perform dynamic convolution over stacked embedding without reshaping, following the idea that comes from Conv-TransE. So the translational property between the entity and the relation is preserved, and their diversity is considered. We experimented on benchmark datasets and showed how our proposed model is better than baseline models in terms of MR, MRR, and Hits@K.

I Introduction

Knowledge graphs are knowledge bases that have a graph structure and represent real-world facts as triplets (head entity, relation, tail entity). There are numerous techniques for creating this data structure, but they are typically divided into two categories: manually and semi-automatically constructed strategies. as a result, it is easy to overlook information in the data. The fundamental problem to solve is link prediction. Predict the missing object in triplets to solve it. Most of the techniques use knowledge graph embedding, which can be helpful in link prediction. It encodes entities and relations into a continuous low-dimensional vector space and employs these embeddings for task prediction.

II State of the Art

There are several models have been proposed, and we can categorize them into: (i) Translation distance-based models; (ii) Semantic matching-based models; (iii) Graph network-based models; (iv) Neural network-based models.

- Translation distance-based models: to model object interactions, a variety of geometric transformations can be employed. TransE, TransH, and TrasnD are a few examples.
- Semantic matching-based models: to capture the semantic similarity between objects, use multiplication and matrix decomposition. DistMult and ComplEx are two methods.
- Neural network-based models: to capture deep interaction features between objects, neural network techniques can be used such as convolution and capsule networks. ConvE, ConvKB, and InteractE are a few examples.
- Graph network-based models: based on message-passing to send information and encode features for the whole graph and then use other models to decode and make predictions. RGCN, SACN, and CompGCN are a few models.

III Original Contribution

The ConvAP model is presented in this work, which fuses the TransD and Conv-TransE ideas, improving calculation efficiency and taking into account the diversity of objects in the knowledge base while preserving the translation characteristics between entities and relations. Furthermore, dynamic convolution was used to improve the representation power. We evaluate our model on the standard FB15k-237, WN18RR, kinship, and UMLS datasets and find that it outperforms previous models in terms of MR (Mean Rank), MRR (Mean Reciprocal Rank), and Hits@K (K = 1, 3, 10).

IV Methodology

Our model's main idea is to define two vectors for each entity and relation. The first vector, as in the TransD [2] model, represents the meaning of an entity or relation, and the second vector is the projection vector. The projection vector represents how a real embedding is projected into relation vector space. The score is then determined by a dynamic convolution [1] over 2D embedding with no reshaping after stacking the projected and relation embedding. For each entity embedding, the mapping matrix $\mathbf{M}_r$ is defined to project entity from entity space to relation space:

$$\mathbf{M}_r = \mathbf{r}_p \mathbf{e}_p^\top + \mathbf{I}^{dt\times d} \tag{1}$$

where $\mathbf{r}_p$, $\mathbf{e}_p$ are relation and entity projection vectors, respectively and $\mathbf{I}$ is the identity matrix. With the mapping matrices, the projected vector is defined as $\mathbf{e}_{sp} = \mathbf{M}_r \mathbf{e}_s$. The stacked of $\mathbf{e}_{sp}$ and $\mathbf{r}_r \in \mathbb{R}^d$ is used as the input for a dynamic convolution layer and procedure a predicted score:

$$\psi_r(\mathbf{e}_s, \mathbf{e}_o) = f(vec(f(DynamicConv([\mathbf{e}_{sp}, \mathbf{r}_r]))\mathbf{W}))\mathbf{e}_o \tag{2}$$

where $\mathbf{r}_r \in \mathbb{R}^d$ is a relation parameter depending on r and f is a nonlinear function. For training our model parameters, the logistic sigmoid function $\sigma(.)$ is used like ConvE to the scoring function: $p(\mathbf{e}_s, \mathbf{e}_r, \mathbf{e}_o) = \sigma(\psi_r(\mathbf{e}_s, \mathbf{e}_o))$, minimizing the the binary cross-entropy loss function

V Results

ConvAP and ConvE have comparable performance on FB15k-237. In particular, when compared to ConvE, our model improved 2.041% on Hits@10 and significantly when compared to DistMult or ComplEx. Unfortunately, when compared to Conv-TransE, our model does not improve substantially; the results are nearly identical due to our model's inability to use context information outside of the local receptacle. We obtained similar results on the WN18RR to the FB15k-237 experiment. It has continued to outperform the ConvE, ComplEx, and R-GCN across all evaluation metrics. Futhermore, a significantly improved results also found on UMLS and Kinship datasets.

VI Evaluation

To evaluate the ability of the proposed model, we use standard metrics for link prediciton. The proportion of correct entities ranked in top k, where k is 1, 3 and 10 (Hits@1, Hits@3, Hits@10), and the mean reciprocal rank (MRR) are reported. Let r_h is rank in the set that was generated by replacing the head entity and r_t is rank in the set that generated by replacing tail entity and the number of test set is n_t. Hits@K and Mean Reciprocal Rank (MRR). To avoid some corrupted triples, we use the filtered setting to filter out all valid triples before ranking: the lower the MR score, the higher MRR or, the higher Hits@K, the better performance.

VII Conclusions

This paper introduced a mixture model that employs dynamic convolution to improve computation performance. Furthermore, the model considers multiple types of entities and keeps the model complexity balanced by utilizing the dynamic mapping matrix resulting from project entity embedding to relation space. The performance of the ConvAP model has improved when compared to state-of-the-art models such as ConvE. ConvAP has performed admirably in terms of convergence time, achieving relatively high performance in a few epochs. We hope that this will help many applications such as recommendation systems, question answering systems, and so on.

References

[1] Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., and Liu, Z. Dynamic convolution: Attention over convolution kernels. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 11027–11036.

[2] Ji, G., He, S., Xu, L., Liu, K., and Zhao, J. Knowledge graph embedding via dynamic mapping matrix. In *ACL* (2015).

DOI: https://doi.org/10.1007/978-3-031-21743-2_34

Keywords: Knowledge Graph Embedding, Link Prediction, Multi-relational Graph, Mixed-curvature Spaces

Mixed Multi-relational Representation Learning for Low-Dimensional Knowledge Graph Embedding

Thanh Le[0000-0002-2180-4222], Chi Tran[0000-0001-6345-8658], Bac Le[0000-0002-4306-6945]

{lnthanh, lhbac}@fit.hcmus.edu.vn, tdchi18@clc.fitus.edu.vn

Simplified Title

MuREL: A mixed multi-relational representation approach for knowledge graph embedding

Abstract

Hyperbolic embeddings have recently received attention in machine learning because of their better ability to handle hierarchical data than Euclidean embeddings. Moreover, Hyperbolic models are also being developed for multi-relational knowledge graphs which contain multiple hierarchical relationships and have achieved promising results, such as MuRP, HyperKG, and ATTH. However, not all data is hierarchical. We also found that most of the geometry models were trained to attain good results on high-dimensional embeddings and low-dimensional embeddings are often of little interest. Besides, neural networks and graph networks models have had impressive performance. However, they also require a relatively high-dimensional embedding to achieve good results, making them limited to use in large-scale knowledge graphs. To address these issues, in this paper, we introduce a new model named **MuREL** (**Mu**lti-**R**elational **E**uclidean **L**orentzian), which learns a mixed embedding between two spaces, Euclidean and Lorentzian. They are not only suitable for a variety of data types but also work well on low-dimensional embeddings. Experiments on standard benchmark datasets from the task of link prediction show that our model outperforms existing Euclidean and Hyperbolic models, especially at lower dimensionality.

I Introduction

KG can be seen as a data structure describing real-world objects and their relationships. Among the problems on KGs, we are focusing our attention on the link prediction problem. The goal of it is to predict new links between nodes in KGs. However, real-world KGs are highly complex, and this task has many challenges.

To tackle this challenge, graph embedding is introduced as an effective approach. The main idea of it is to map entities/relations to a low-dimensional embedding space while preserving the inherent structures of KGs.

In this paper, we propose MuREL, which embeds a multi-relational KG in a mixture space between Euclidean and Lorentzian to handle a wide variety of data types well and get good performance even on low-dimensional embeddings. Experiments show that our model has competitive results compared to state-of-the-art geometry-based models, particularly at a low-dimensional embedding.

II State of the Art

For Euclidean space, it is simple and needs a small number of parameters to be used on large KGs. However, translations make this space fail to capture critical relational components such as symmetry/anti-symmetric, inversion and composition.

For Complex space, it is able to overcome weaknesses in Euclidean space well. However, almost models in this space often have bad results when running on low-dimensional embeddings and require a high memory cost to run on high-dimensional embeddings to attain good results.

For Hyperbolic space, it handles hierarchical data well. However, this space often uses a fixed curvature and can lead to suboptimal results. In addition, the Riemannian optimization function in this space often has a slower convergence speed than the usual optimization functions such as Adam or Adagrad.

III Original Contribution

Our main contributions are as follows:

- To our knowledge, we are first to combine Lorentz and Euclidean distances to solve the LP problem.
- We provide some mathematical hypotheses to demonstrate MuREL's ability on hierarchical datasets.
- Showing the effectiveness of MuREL on three standard benchmark datasets compared to other geometry-based models, giving a detailed analysis of the performance on different embedded dimensions.

IV Methodology

IV.1 MuREL

In Euclidean space, the scoring function is defined as follow:

$$\phi_{\mathrm{E}}(\mathrm{e_h},\mathrm{e_r},\mathrm{e_t}) = -\mathrm{d_E}(\exp_0(\mathbf{R}\log_0(\mathbf{e}_\mathrm{h})),\mathbf{e}_\mathrm{t}+\mathbf{e}_\mathrm{r})^2+\mathrm{b_h}+\mathrm{b_t} \tag{1}$$

where $\mathbf{e}_\mathrm{h}$, $\mathbf{e}_\mathrm{t} \in \mathbb{E}^\mathrm{d}$ are head and tail Euclidean embeddings, $\mathrm{R} \in \mathbb{R}^{\mathrm{d}\times\mathrm{d}}$ is a diagonal relation matrix, $\mathbf{e}_\mathrm{r} \in \mathbb{E}^\mathrm{d}$ is a translation vector of relation $\mathrm{e_r}$, $\mathrm{d_E}$ is the Euclidean distance, $\mathrm{b_h}$, $\mathrm{b_t} \in \mathbb{R}$ are scalar biases of head and tail entities.

In Lorentizian space, the scoring function is defined as follow:

$$\phi_{\mathrm{L}}(\mathbf{l}_\mathrm{h},\mathbf{l}_\mathrm{r},\mathbf{l}_\mathrm{t}) = -\mathrm{d_L}(\exp_0(\mathbf{R}\log_0(\mathbf{l}_\mathrm{h})),\mathbf{l}_\mathrm{t}+\mathbf{l}_\mathrm{r})^2+\mathrm{b_h}+\mathrm{b_t} \tag{2}$$

where $\mathbf{l}_\mathrm{h}$, $\mathbf{l}_\mathrm{t} \in \mathbb{L}^\mathrm{d}$ are Lorentzian embeddings of head and tail entities and $\mathbf{l}_\mathrm{r} \in \mathbb{L}^\mathrm{d}$ is a translation vector of relation $\mathrm{l_r}$ and $\mathrm{d_L}$ is the squared Lorentzian distance.

Finally, we combine Eq. 1 and Eq. 2 to produce mixed embeddings for MuREL, and the scoring function of it is designed as follows:

$$\phi(\mathrm{h},\mathrm{r},\mathrm{t}) = -(\mathrm{d_E}(\exp_0(\mathbf{R}\log_0(\mathbf{e}_\mathrm{h})),\mathbf{e}_\mathrm{t}+\mathbf{e}_\mathrm{r})^2+\mathrm{d_L}(\exp_0(\mathbf{R}\log_0(\mathbf{l}_\mathrm{h})),\mathbf{l}_\mathrm{t}+\mathbf{l}_\mathrm{r})^2)+\mathrm{b_h}+\mathrm{b_t} \tag{3}$$

To get the predicted probabilities of true triples, a non-linearity activation function such as a logistic sigmoid is applied to the score function, i.e. $\sigma(\phi(\mathrm{h},\mathrm{r},\mathrm{t}))$.

IV.2 Training

To train MuREL, we use the reciprocal negative sampling technique by adding (t, r^{-1}, h) for every triple (h, r, t). Then, k negative samples are created for each true triple (h, r, t) by corrupting either the tail (h, r, t′) or the head (t, r^{-1}, h′) entity with a randomly chosen new entity from the training set $\mathscr{E}$. Finally, the model is trained to minimize the Bernoulli negative log-likelihood loss.

V Results

Model	WN18RR				FB15k-237			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
RotatE	0.387	0.330	0.417	0.491	0.290	0.208	0.316	0.458
MuRE	0.458	0.421	0.471	0.525	0.313	0.226	0.340	0.489
ComplEx-N3	0.420	0.390	0.420	0.460	0.294	0.211	0.322	0.463
MuRP	0.465	0.420	0.484	0.544	0.323	0.235	0.353	0.501
ROTE	0.463	0.426	0.477	0.529	0.307	0.220	0.337	0.482
ROTH	**0.472**	0.428	**0.490**	**0.553**	0.314	0.223	0.346	0.497
MuREL	**0.472**	**0.433**	0.483	0.552	**0.333**	**0.243**	**0.366**	**0.512**

Table 1: Link prediction results on WN18RR and FB15k-237 for a low-dimensional embedding (d = 32). Best results in bold and second best in underlined. Except for our model, all results are taken from [1].

VI Evaluation

For evaluation metrics, we use two ranking-based metrics, Mean Reciprocal Rank (MRR) and Hits@k (k ∈ {1, 3, 10}), to evaluate our model. The filtered setting [1] is also applied as a standard evaluation method.

From Table 1 we can see that in WN18RR, MuREL obtains the best performance on most metrics, except Hits@3 and Hits@10. Compared to the baseline model MuRE, MuREL outperforms MuRE for all metrics, the improvement ranges from 2.548% to 5.143%. In FB15k-237, MuREL is better than the rest of the models for all metrics. Compared with MuRE, MuREL has slightly better results, ranging from 4.703% to 7.647%.

VII Conclusions

In this paper, we introduced MuREL, a mixed model that utilizes Euclidean and squared Lorentzian distances to handle hierarchical datasets for both low and high-dimensional embeddings better. Through experiments, our model has shown quite promising results compared to the previous geometry-based models and implies that the combination of two geometric spaces has successfully handled a variety of data types.

References

[1] Chami, I., Wolf, A., Juan, D.-C., Sala, F., Ravi, S., and Ré, C. Low-dimensional hyperbolic knowledge graph embeddings. *arXiv preprint arXiv:2005.00545* (2020).

DOI: https://doi.org/10.1007/978-981-19-8234-7_21

Keywords: Knowledge Graph Embedding, Link Prediction, Graph Convolutional Networks, Dynamic Graph Attention Networks.

A Novel Integrating Approach Between Graph Neural Network and Complex Representation for Link Prediction in Knowledge Graph

Thanh Le [0000-0002-2180-4222], Loc Tran [0000-0002-0108-503X], Bac Le [0000-0002-4306-6945]

{lnthanh, lhbac}@fit.hcmus.edu.vn, txloc18@clc.fitus.edu.vn

Simplified Title

Applying graph structure information to complex space for improving the performance of the link prediction on Knowledge Graph

Abstract

Deep learning brings high results in many problems, including Link Prediction on Knowledge Graphs (KGs). Although there are many techniques to implement deep learning into KGs, Graph Neural Networks (GNNs) have recently emerged as a promising direction for representing the structure of KGs as input for a decoder. With this structural information, GNNs can help to retain more information from the original graph than conventional embeddings like TransE, TransH, RESCAL. As a result, the learning model achieves higher accuracy in predicting missing links between entities in the KG. Meanwhile, several studies have successfully demonstrated the intrinsic properties of the embedding process in complex space while keeping many binary relations (symmetric and asymmetric). Thus, this paper proposes deploying GNNs into complex space to increase the model's predictive capability. Another issue with GNNs is that they are susceptible to over-squashing when a large amount of information propagating between nodes is compressed down to a fixed representation space. As a result, we utilize a dynamic attention mechanism to minimize the adverse effects of these factors, and experiments on benchmark datasets have indicated that our proposal achieves a significant improvement compared to baseline models on almost all standard metrics.

I Introduction

Knowledge Graphs (KGs) are becoming a widely used term in the field of artificial intelligence. Some of its outstanding applications are in a number of areas, such as medicine and e-commerce. Among the KG-related challenges, we are putting our efforts into tackling the link prediction problem, whose objective is discovering the missed links in KG to accomplish itself. In fact, data in the KG is regularly gathered from various sources, including manually. Moreover, identifying the relations between data aids in order to complete KGs. Furthermore, we could forecast potential relations between entities in the future. Although numerous rule-based and probability-based methods for addressing this problem were initially proposed, these models suffer the exploding computational cost when performing on large KGs. As a result, the embedding approaches emerged and gained widespread attention in recent years because of their above mentioned characteristics.

II State of the Art

Currently, KGE is classified into two categories: Translational and Semantic matching models. At the Translational based models, the models projecting entities and relations into vector or matrix representation with simplicity, and scalable characteristics, TransE and its extensions such as TransH, TransR, and TransD are the top striking models of translational research. The semantic-matching models include bilinear models and neural network-based models. Bilinear methods such as the RESCAL model, DistMult, and ComplEx can mine the intrinsic properties of KGs by using tensor decomposition, characterized by the less time-consuming and effective computation except for RESCAL.

In neural network-based models, researchers apply the success of Convolutional neural networks (CNNs) in KGs and archive potential results such as ConvE, ConvKB, but it suffers to time-consuming, increases model complexity, and other related CNN's problems. Another variant of CNNs is Graph Convolutional Networks, levering the same convolutional operation but performing on graph data while considering graph features during the embedding phase. Some recent models such as RGCN, VR-GCN, and TransGCN on the graph neural networks branches. Furthermore, incorporating additional information into the embedding process, such as literals, textual descriptions, entity, and relation types, has been shown to yield higher quality embedding and is especially effective in many downstream tasks.

III Original Contribution

We propose implementing Graph Convolutional Networks (GCNs) in complex space and conducting experiments to demonstrate the effectiveness of this approach. Moreover, the ComplEx model is employed as a decoder to utilize the graph features in the complex space because of its simplicity and scalability over large graphs. Additionally, we also realize that GCNs mainly solve problems in real space, which is the main weakness leading to marginal performance growth in recent years, thus we apply GCNs to the complex space to prove our conclusion. Besides that, the reason we integrate the GCNs and Dynamic Graph Attention Networks (GATv2[2]) is to tackle the bottleneck problem of GNNs.

IV Methodology

Our proposed framework includes the encoder and decoder. For the encoder, entities and relations are embedded into the complex space to create two main components, including real component and imaginary component for each entity and relation. The Graph Aggregation Layer takes the entities embedding as input, which consists GCN and GATv2 layers. We utilize GCN layer as the first layer to propagate information between neighbors because it is scalable on large KGs and can operate on local graph neighbors. Next, we expand the interaction between entities and assess the neighbor's contributions by adding a Graph Attention Layer (GAL) to assign the different weights for the importance of the neighbor's features in the entity's neighborhood. We use GATv2 in our framework to reduce the amount of information passed through an activation function and alleviate the bottleneck problem of GNN because of the exponential growing information into fixed-size vectors[1].

For the decoder, the real and imaginary parts of the entity and relation are passed into the scoring function; we utilize the ComplEx scoring function to compute the plausibility of forming triplets. Next, a sigmoid function is applied to get the prediction results.

V Results

Complex-GNN improved its baseline model (ComplEx[3]) in all of the metrics on FB15K-237 and WN18RR. Compared with models using Graph Convolutional Networks like TransE-GCN, VR-GCN, R-GCN, our model shows its potential ability when integrating graph information into complex space and then enhancing the ability to handle binary relations. Besides that, the convergence of our model shows that it is fast and requires at least 300 epochs for WN18RR and 1000 epochs for FB15k-237.

VI Evaluation

We use two widely used datasets to evaluate the link prediction task, including FB15K-237 and WN18RR. We also use two common benchmarks, including Hit@k, and mean reciprocal rank (MRR), to evaluate the performance of the link prediction task.

We fine-tune the hyperparameters {learning rate, embedding dim, hidden dim, and dropout rate} by running grid search. Through experiment, the following hyperparameters gave the best result on FB15k-237: {0.001, 300, 400, 0.3}; on WN18RR: {0.003, 200, 400, 0.4} and the other configuration based on the framework of ConvE. We use PyTorch version 1.9.1 and run on NVIDIA Tesla V100-DGXS-32GB.

VII Conclusions

In this paper, we have proposed ComplEx-GNN model based on integrating graph neural networks and complex representation for the link prediction problem. The gap between our model with ComplEx and LiteralE and GNNs models proves the importance of structural information, indicating the power of complex space in representing entities and relations.

References

[1] Alon, U., and Yahav, E. On the bottleneck of graph neural networks and its practical implications. *arXiv preprint arXiv:2006.05205* (2020).

[2] Brody, S., Alon, U., and Yahav, E. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491* (2021).

[3] Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. Complex embeddings for simple link prediction. In *International conference on machine learning* (2016), PMLR, pp. 2071–2080.

DOI: https://doi.org/10.1007/978-3-031-21743-2_32

Keywords: Knowledge Graph Embedding, Link Prediction, Literals, Convolution Neural Network

Embedding and Integrating Literals to the HypER Model for Link Prediction on Knowledge Graphs

Thanh Le [0000-0002-2180-4222], Tuan Tran [0000-0002-6185-7353], Bac Le [0000-0002-4306-6945]

{lnthanh, lhbac}@fit.hcmus.edu.vn, tttuan18@clc.fitus.edu.vn

Simplified Title

Embedding and Integrating Literals to the HypER Model

Abstract

The link prediction on knowledge graphs is now one of the challenges that are gaining a lot of interest from the academic community. The leading solution for this problem is based on graph embedding. Recently, in embedding approaches, convolutional neural networks (CNN) have produced promising results, especially the HypER model. The HypER model outperforms the preceding approaches to maximize the quantity of information from the source entities and relations. However, HypER and other CNN-based methods only focus on retaining information (i.e., structure) of knowledge graphs in low dimension embedding spaces while ignoring literals of the entities. However, the literals can also have a significant impact on relation construction. As a result, this paper proposes an improved model called HypERLit, which is based on the HypER model and incorporates literals. Experiments prove that the role of literals significantly influences the accuracy of the prediction model on the benchmark datasets, including FB15k, FB15k-237, and YAGO3-10. Furthermore, our model outperforms the HypER and other CNN-based models on almost standard metrics.

I Introduction

Knowledge is incredibly significant in today's society and is the accumulation of experience during human evolution. Consequently, that has resulted in a large amount of information. A structure called the ”Knowledge Graphs” (KGs) is utilized to store them efficiently. One problem is that the data in KGs may be incomplete, even with large KGs. Therefore, researchers in KGs field have devised the link prediction (LP) problem. Its main task is to predict relations between the source and target entities in the triples. One of the typical research branches of LP problem is knowledge graph embedding (KGE), which has recently achieved good results. The survey found that current embedding models focus on keeping information on entities and relationships when converting from a high embedding dimension to a low embedding dimension. It means that literals in the knowledge graph are often ignored during embedding, even though they strongly impact creating relationships between entities. With that motivation, we decide to find ways to improve current LP models, especially the HypER model, by integrating literals into entities.

II State of the Art

II.1 ConvE

ConvE can be said to be the pioneer model in applying CNN to KGE. ConvE model performs convolution on a stacked 2D matrix of entity and relation embeddings. As a result, they increase the interaction between source entities and relations as well as make the expressiveness of the model more efficient. However, the limitation also lies in the concatenation of source and relation embeddings. It makes the features extracted at the convolution step not completely possess the properties of both source and relation embeddings.

II.2 ConvR

ConvR is a model that uses relation embeddings to create filters instead of taking them into forming a 2D matrix for the convolution step like ConvE. Consequently, they increased the representation for embeddings. However, ConvR generates filters directly from the relation embeddings without considering the possibility that they can also significantly affect the result of the convolution.

III Original Contribution

The paper proposes a HyperLit - a LP model that has filters generated from relations through the hypernetwork. It helps entities that have literals integration, and relations are better represented. Through the positive results of HyperLit model it has contributed to enriching the information of the KGs.

IV Methodology

First of all, we outline some significant underlying models, including HypER [1] and LiteralE [3]. HypER is a convolutional network model that recently produced good results. Unlike ConvE employs randomly initialized filters, and ConvR uses filters formed directly from reshaping embedded relational vectors, HypER's filters are generated by a hypernetwork [2]. When performing convolution, this hypernetwork has the effect of tweaking the filters to extract the features of source embeddings efficiently. As a result, HypER may optimize the quantity of information gathered from source entities and represent them fully. Otherwise, the LiteralE model efficiently inserts literals into source entities. This efficiency is proved when substituting the source embedding in ConvE with the results of the LiteralE model. Building on the strengths of these two models, we designed a model named HypERLit by integrating literals into the filters and finding a way to attach literal information to entity embeddings. Hence, the model has more information to predict relations between entities.

The HypERLit model aggregates information from the source entity with literals by passing them through a Gate and does the same for all entities through an entity matrix whose each row is an entity vector in KGs. The related information is then transmitted via the hypernetwork in order to generate filters for convolution with the information accumulated through the previous Gate. Hence, we obtain the features retrieved from the source entity, literals, and relation. These features are projected into a d-dimensional space, and then a nonlinear function is applied to them. The acquired results are multiplied by the matrix of entities integrated with literals in the first stage and sent through the sigmoid function to generate the score of each triple in this matrix corresponding to each target entity.

V Results

For the FB15k dataset, the HypERLit model is higher than other models, about $0.001 - 0.622$ on the Hits@K measure and $0.004 - 0.583$ on the MRR measure. Despite the fact that the FB15k dataset is no longer routinely employed in assessing later models due to data leakage issues, the comparison reveals that HypERLit is capable of doing well in this dataset. Similarly, on the FB15k-237 dataset, our model reaches a higher Hit@K and MRR result of about $0.001 - 0.009$. Because this dataset does not contain inverse triples, it is more suitable for the triple training process as ConvKB model. However, for the other models, HypERLit is still better. On YAGO3-10 dataset, the result of HypERLit is about $0.005 - 0.098$ higher. The information about the entities of YAGO3-10 becomes more, and the predictions can easily achieve equivalent accuracy even without the literal involved.

VI Evaluation

From the experimenting results of HyperLit model, it can be included that HypER is the best CNN-based LP model for integrating literals to entities. HyperLit model can use maximize the ability of HyPER model in handling the amount of information that is combined between literals and entities. In addition, the experimental results also show the efficiency when using the gate structure of the LiteralE model to be able to integrate literals into entities, which is also a central criterion of the HyperLit model.

VII Conclusions

Experiments on three standard datasets, including FB15k, FB15k-237, and YAGO3-10, have shown that the proposed model achieves better results than other CNN-based models on metrics MRR, Hits@K (1, 3, 10). It demonstrates that the inclusion of literals provides additional information for deciding which relation is most likely.

References

[1] Balažević, I., Allen, C., and Hospedales, T. M. Hypernetwork knowledge graph embeddings. In *International Conference on Artificial Neural Networks* (2019), Springer, pp. 553–565.

[2] Ha, D., Dai, A., and Le, Q. V. Hypernetworks. *arXiv preprint arXiv:1609.09106* (2016).

[3] Kristiadi, A., Khan, M. A., Lukovnikov, D., Lehmann, J., and Fischer, A. Incorporating literals into knowledge graph embeddings. In *International Semantic Web Conference* (2019), Springer, pp. 347–363.

DOI: https://doi.org/10.1007/978-3-031-21743-2_23
Link to repository: https://github.com/aliotopal/HRadversImgs/blob/main/original-advers.md

Keywords: High resolution adversarial images, Convolutional Neural Networks, Black-box attack, Evolutionary Algorithm

Strategy and Feasibility Study for the Construction of High Resolution Images Adversarial against Convolutional Neural Networks

Franck Leprévost[0000-0001-8808-2730], Ali Osman Topal[0000-0003-0141-4742], Elmir Avdusinovic[0000-0002-8292-8747], Raluca Chitic[0000-0003-1113-2343]

Franck.Leprevost@uni.lu, Aliosman.Topal@uni.lu, Elmir.Avdusinovic.001@student.uni.lu, Raluca.Chitic@uni.lu

Simplified Title

Construction of high-resolution adversarial images against convolutional neural networks

Abstract

Convolutional Neural Networks, that perform image recognition, assess images by first resizing them to their fitting input size. In particular, high resolution images are scaled down, say to 224 × 224 for CNNs trained on ImageNet. So far, existing attacks, that aim at creating an adversarial image that a CNN would misclassify while a human would not notice any difference between the modified and the unmodified image, actually work in the 224 × 224 resized domain and not in the high resolution domain. Indeed, attacking high resolution images directly leads to complex challenges in terms of speed, adversity and visual quality, that make these attacks infeasible in practice. We design an indirect strategy that addresses effectively this issue. It lifts to the high resolution domain any existing attack that works in the CNN's input size domain. The adversarial noise is of the same size as the original image. We apply this strategy to construct efficiently high resolution adversarial images of good visual quality that fool VGG-16 trained on ImageNet.

I Introduction

The profusion of images in our modern-day society and the need to analyze quickly the information they contain for a large series of applications (self-driving cars, face recognition and security controls, etc) has led to the emergence of tools to automatically process and sort this type of data. Trained convolutional neural networks (CNNs) are among the dominant and most accurate tools for automatic object recognition and classification. Nevertheless, CNNs can be led to erroneous classifications by specifically designed adversarial images. For instance, starting with an original image classified by a CNN in a given category, the target scenario essentially consists in choosing a target category, different from the original one, and in creating a variant of the original image that the CNN will classify in the target category, although a human would classify this adversarial image still in the original category, or would be unable to notice any difference between the original and the adversarial image.

Such attacks are classified according to the knowledge about the CNN at the disposal of the attacker. In this hierarchy, black-box attacks are the most challenging ones, since no knowledge about the architecture of the CNN (number and type of layers, weights, etc.) is assumed. While addressing images of moderate size is done routinely nowadays, high-resolution images were not handled efficiently so far. To the best of our knowledge, our contribution is the first attempt that succeeds in doing so.

II State of the Art

So far, all attacks — black-box or not — addressed images of moderate size, ranging from 32 × 32 (typically for CNNs trained on CIFAR-10) up to 224 × 224 (typically for CNNs trained on ImageNet), or resized to such values that the CNNs handle natively, what is called here the ”low resolution” domain. The construction of adversarial images is then achieved by adding some carefully designed adversarial noise to the potentially resized original image. In particular, the adversarial noise created by all these attacks is in the ”low resolution” domain handled natively by the CNNs, so that the obtained adversarial images are as large as the CNN's input size. This means that these attacks explore a search space of size that does not depend on the size of the original image, but that coincides with the size of the CNN input. Creating adversarial images of large size leads to three challenges in terms of speed, adversity and visual quality. Firstly, the complexity of the problem increases quadratically with the size of the images. Secondly, the noise introduced in the ”high resolution” domain should be assessed as adversarial in the ”low resolution” domain: It should ”survive” the resizing process to fit the CNN. Thirdly, the noise introduced in the ”high resolution” domain should be imperceptible to a human eye looking at the images at their native size, and not merely once they are reduced to fit the ”low resolution” domain.

III Original Contribution

This paper is a first step towards the creation of adversarial noise of size of the original image, whatever this size may be. Our contribution is essentially twofold. Firstly, we describe an indirect strategy that leads to the construction of images in the ”high resolution” domain adversarial for the target scenario performed on a trained CNN. The conceptual design of the indirect strategy, that furthermore lists indicators relevant to the problem, is flexible enough to lift to the ”high resolution” domain attacks considered as efficient in the ”low resolution” domain. Secondly, we perform a feasibility study of this indirect strategy with 10 explicit HR images and on one CNN, namely VGG-16 trained on ImageNet. We lift to the ”high resolution” domain a black-box attack based on an evolutionary algorithm. We prove experimentally that our strategy is highly efficient in terms of speed and of adversity, and is reasonably efficient in terms of visual quality.

IV Methodology

Our study is both theoretical (for the strategy part, applying to any attack), and experimental (for the case study of a black-box attack on one trained CNN). Our methodology relies on a thorough lifting strategy. Given a clean image in the ”high resolution” domain, we downsize it with a degrading function to the ”low resolution” domain handled by the CNN to deceive. Then this reduced clean image is attacked (the specific attack is of no importance), leading to a ”low resolution” adversarial image. Then we lift this ”low resolution” adversarial image to the ”high resolution” domain with an enlarging function. We check whether reducing this ”high resolution” adversarial leads to a ”low resolution” adversarial. This is not automatic since the enlarging and degrading functions are not reciprocal functions one from the other. We introduce indicators (the Loss function in particular) that measure how adversity and visibility are affected by our strategy, and by the choice of the enlarging and degrading functions.

V Results

The indirect strategy is applied to a series of 10 very diverse high resolution images, of sizes 600 × 641 up to 2171 × 1740. The CNN to deceive is VGG-16 trained on ImageNet, and the considered attack is a black-box evolutionary algorithm-based targeted attack. We set as objective that the obtained high-resolution adversarial images, once reduced to the size handled naturaly by the CNN, should be classified in the pre-defined target categories with a confidence exceeding 55%. This is achieved within 55 minutes while a direct attack (aiming at creating the adversarial noise directly in the high-resolution domain without following the indirect path of our strategy) did not give any indication of success after 40 hours.

VI Evaluation

The performed feasibility study on one CNN, on 10 high resolution images, and one attack not only showed that our indirect strategy, aiming at lifting any existing attack, is efficient in practice, at least for one instantiation of the problem. It also helped to select appropriate degrading and enlarging functions, and to validate the pertinence of the indicators we used (especially the Loss function that we created).

VII Conclusions

Following up on the present conference paper, we expanded our study to challenge the lifting of the evolutionary-algorithm based attack on more CNNs [1]. More recently, we designed a new refined strategy aiming at the creation of an appropriate adversarial noise directly in the high-resolution domain. On-going experiments indicate that this new direction is promising for the creation in reasonable time of adversarial images natively in the high-resolution domain with a substantially increased visual quality.

References

[1] Leprévost, F., Topal, A. O., Avdusinovic, E., and Chitic, R. A strategy creating high-resolution adversarial images against convolutional neural networks and a feasibility study on 10 cnns. *Journal of Information and Telecommunication* (2022).

DOI: https://doi.org/10.1007/978-3-031-21967-2_29

Keywords: Quality of Experience (QoE), Quality of Service (QoS), metrics, evaluation, performance, Computer Vision (CV), Video Quality Indicators (VQI), Key Performance Indicators (KPI), User-Generated Content (UGC), in-the-wild content

User-Generated Content (UGC)/In-The-Wild Video Content Recognition

Mikołaj Leszczuk[0000-0001-9123-1039], Lucjan Janowski[0000-0000-0000-0000], Jakub Nawała[0000-0002-5671-3726], Michał Grega[0000-0001-7633-8663]

qoe@agh.edu.pl

Simplified Title

Recognition of User-Generated Content

Abstract

According to Cisco, we are facing a three-fold increase in IP traffic in five years, ranging from 2017 to 2022. IP video traffic generated by users is largely related to user-generated content (UGC). Although at the beginning of UGC creation, this content was often characterized by amateur acquisition conditions and unprofessional processing, the development of widely available knowledge and affordable equipment allows one to create UGC of a quality practically indistinguishable from professional content. Since some UGC content is indistinguishable from professional content, we are not interested in all UGC content, but only in the quality that clearly differs from the professional. For this content, we use the term "in the wild" as a concept closely related to the concept of UGC, which is its special case. In this paper, we show that it is possible to deliver the new concept of an objective "in-the-wild" video content recognition model. The value of the F measure in our model is 0.988. The resulting model is trained and tested with the use of video sequence databases containing professional and "in the wild" content. These modeling results are obtained when the random forest learning method is used. However, it should be noted that the use of the more explainable decision tree learning method does not cause a significant decrease in the value of measure F (an F-measure of 0.973)

I Introduction

According to Cisco, we are facing a three-fold increase in IP traffic in five years, ranging from 2017 to 2022. IP video traffic generated by users is largely related to user-generated content (UGC). Although at the beginning of UGC creation, this content was often characterized by amateur acquisition conditions and unprofessional processing, the development of widely available knowledge and affordable equipment allows one to create UGC of a quality practically indistinguishable from professional content. Since some UGC content is indistinguishable from professional content, we are not interested in all UGC content, but only in the quality that clearly differs from the professional. For this content, we use the term "in the wild" as a concept closely related to the concept of UGC, which is its special case.

II State of the Art

It should be noted that a review of the literature on the subject indicates that the problem of recognizing UGC content has already been considered, but not with respect to "in-the-wild" content. In addition, some references are related to UGC content, but not necessarily to multimedia content. For example, [3] is about classifying the text. However, even the methods for classifying multimedia content disclosed during the literature review are not necessarily based on video data. For example, the model described in [1] uses audio features, while the model presented in [2] uses video, however, while classifying it is supported by tags and metadata such as shot length.

III Original Contribution

In this paper, we show that it is possible to deliver the new concept of an objective "in-the-wild" video content recognition model.

Reproduced with permission from Springer Nature

IV Methodology

To make our considerations about "in-the-wild" content more concrete, we used "in-the-wild" video databases. Our further research used 3 publicly available in-the-wild video databases: (i) CVD-2014, (ii) LIVE-Qualcomm, and (iii) KoNViD-1k. The database of video sequences has been supplemented with a "counterweight" in the form of professional quality video sequences. For this purpose, the "NTIA simulated news" database was used.

The number of video frames in all databases used is hundreds of thousands, but in reality, the frames belonging to one shot are quite similar to each other. Therefore, in a further analysis, we operate at the level of recognizing the entire shot. As the frames belonging to one shot are quite similar to each other, they have similar values of the video indicators. Consequently, the experiment operates on averaged video shots. It applies a set of video indicators and outputs a vector of results (one for each video indicator). The results will be combined later with the ground truth (content "in the wild" versus professional content). Together, they constitute input data for modeling.

In total, we used 10 video indicators. They come from our AGH Video Quality (VQ) team.

The possession of the data sets ("in-the-wild" content and professional content) allows us to construct a model that detects "in-the-wild" video content.

To build the model, we start with the indicator transformation, which is a typical step in a data analysis. Then, we assume an "in-the-wild" video content by classifying into two classes (the video shot is "in-the-wild" content, the video shot is professional content). Further modeling is done with Scikit-learn – a free software machine learning library for the Python programming language. Decision tree and random forest algorithms and modeling tools are tested for the detection of "in-the-wild" video content.

The resulting model is trained and tested with the use of video sequence databases containing professional and "in the wild" content. Finally, modeling using a random forest turns out to be more effective, while a decision tree gives a more explainable model.

V Results

These modeling results are obtained when the random forest learning method is used. The value of the F measure in our model is 0.988. However, it should be noted that the use of the more explainable decision tree learning method does not cause a significant decrease in the value of measure F (an F-measure of 0.973).

VI Evaluation

For both the decision tree and the random forest, we make 5000 attempts to train and test the model, each time randomly dividing the division into training and test sets and reporting the results obtained on the test set. The reported results are always average results.

VII Conclusions

In this paper, we show that it is possible to deliver the new "in-the-wild" concept of an objective video content recognition model. The value of the measured accuracy of a model (the parameter of the F measure) achieved is 0.988.

These modeling results are obtained when the random forest learning method is used. However, it should be noted that the use of the more explainable decision tree learning method does not cause a significant decrease in prediction accuracy (measure F of 0.973).

The results presented are work in progress. Although the current results are highly promising, they still require additional validation since the training and test datasets are relatively limited (especially for professional content). Therefore, additional selected video sequences from the database of 6000+ professional YouTube news clips should be used. These video sequences are currently being reviewed and any "in-the-wild" shots are being eliminated from them (adding them to the "in-the-wild" part of the set being prepared). This allows us in the future to make a more precise validation of the model or also its correction.

Finally, since we only compare the results of the implementation of the random forest and the decision tree, the implementation of other machine learning techniques will be useful for future research.

References

[1] Guo, J., and Gurrin, C. Short user-generated videos classification using accompanied audio categories. In *Proceedings of the 2012 ACM international workshop on Audio and multimedia methods for large-scale video analysis* (2012), pp. 15–20.

[2] Guo, J., Gurrin, C., and Lao, S. Who produced this video, amateur or professional? In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval* (2013), pp. 271–278.

[3] Marc Egger, A., and Schoder, D. Who are we listening to? detecting user-generated content (ugc) on the web. *ECIS 2015 Completed Research Papers* (2015).

DOI: https://doi.org/10.1007/978-3-031-21967-2_17

MLP-Mixer approach for corn leaf diseases classification

Li-Hua Li, Radius Tanone

lhli@gm.cyut.edu.tw,s11014903@gm.cyut.edu.tw

Simplified Title

MLP-Mixer approach for corn leaf diseases classification

Abstract

Corn is one of the staple foods in Indonesia. However, corn leaf disease poses a threat to corn farmers in increasing production. Farmers find it difficult to identify the type of corn leaf that is affected by the disease. Seeing the development of corn that continues to increase, prevention of common corn leaf disease needs to be prevented to increase production. By using open dataset, the modern MLP-Mixer model is used to train the smaller size of datasets for further used in predicting the classification of diseases that attack corn leaves. This experiment uses an MLP-Mixer with a basic Multi-Layer Perceptron which is repeatedly applied in feature channels. This makes the MLP-Mixer model more resource efficient in carrying out the process to classify corn leaf disease. In this research, a well-designed method ranging from data preparation related to corn leaf disease images to pre-training and model evaluation are proposed. Our model shows the performance of 98.09 % test accuracy. This result is certainly a new trend in image classification, so that it can be a solution in handling computer vision problems in general. Furthermore, the high precision achieved in this experiment can be applied to small devices such as smartphones, drones, or embedded systems. Based on the images obtained, these results can undoubtedly be a solution for corn farmers in recognizing the types of leaf diseases in order to achieve smart farming in Indonesia.

I Introduction

Agriculture is very popular in Indonesia, where 10 million people work as farmers, including corn farmers, out of a population of over 273 million. In fact, common rust, leaf blight, and gray leaf spot are all diseases that affect corn leaves. Farmers may have difficulty identifying the types of diseases that attack corn leaves due to the large size of the land and the potentially varying conditions of corn leaves. Given the difficulties in classifying the types of diseases on corn leaves, information technology may be one solution for assisting farmers in recognizing the types of diseases on corn leaves. Since Google Research, Brain Team introduced the MLP-Mixer for vision [1], this research on corn leaf diseases image classification has become the main target in implementing this model. This research focuses on how to develop a modern MLP-Mixer pre-trained model with few datasets, trained and then able to predict the classification of a disease in corn leaves. The MLP-Mixer model uses fewer resources than other algorithms so that in future developments it can be used on devices with smaller sizes. The purpose of this research is how to help farmers on agricultural land to be able to work more quickly and efficiently in recognizing diseases in corn leaves.

II State of the Art

Since the introduction of Transformer as part of DL, many models have been created that use patches [2] as input images. This is undoubtedly an alternative to the state-of-the-art Convolutional Neural Network (CNN) [3] for problem solving in the field of computer vision. Since Google Brain released the MLP-Mixer model for image classification, it has provided several benefits, one of which is low computational resources. In this experiment, the MLP-Mixer model was used to classify disease recognition in corn leaves. The MLP-Mixer model was created from the ground up with the goal of performing image classification. Due to the limited scope of the study, the performance of this experiment was not compared to other deep learning models such as CNN, Vision Transformer, etc. This study focuses solely on how the performance of MLP-Mixer to perform problem solving on the introduction of disease classification on corn leaves.

III Original Contribution

According to the original contribution, this experiment employs a basic Transformer to classify disease recognition in corn leaves. This differs from image classification research in general, which employs image classification techniques such as CNN. According to the experimental results, MLP-Mixer can classify images with an accuracy of more than 95%. This will undoubtedly have a positive impact on the field of computer vision research. Furthermore, proper training time and parameter settings can aid in improving accuracy. Furthermore, MLP-Mixer employs a relatively simple low computational system to speed up the training process and produce good results.

IV Methodology

This experiment's method consists of several stages of simulating the data requirements in which the appropriate data requirements will be seen to carry out the classification process. The following step is data collection, in which datasets will be retrieved from open datasets on Kaggle. Following that, data will be prepared for use in the stages of the MLP-Mixer-based deep learning training process. After completing the stages of dataset preparation, the next step is to create a model from scratch. At this point, the parameters will be configured., as well as the model will be trained and evaluated. Model evaluation will be carried out until the model produces the best performance and achieves the expected accuracy.

V Results

The experimental results show good results regarding corn leaf classification using the MLP-Mixer model. This experiment has a top accuracy of more than 98%. This means that the MLP-Mixer is an excellent alternative deep learning model for image classification in the field of computer vision. Some of the advantages gained from using this MLP-Mixer model are as described in the original paper. The MLP-Mixer also has a few advantages that simplify its architecture, such as identical layer sizes, each layer consisting of only two MLP blocks and accepting input of the same size. Another significant point is that all image patches are projected linearly using the same projection matrix. This model also has a small number of parameters, 218,123 in total. This will undoubtedly help to reduce the cost and speed of the computational process in image classification. Furthermore, accuracy results can be improved by using suitable parameters and a longer training time.

VI Evaluation

This study employs an experimental method, which requires enough time to prepare the dataset in order to build and test a model from scratch. Experiments are carried out until the maximum output is obtained. Evaluation metrics used in this experiment are Precision, Recall, Accuracy and F1-Score. However, based on the results presented in this paper, setting parameters and training the model for a longer period of time will certainly produce better results.

VII Conclusions

In conclusion, the model developed from this experiment yields promising results with an accuracy greater than 98%. This can certainly assist application developers in being able to use this model so that it can be implemented in the agricultural field on a practical level. The model developed into a lite version based on MLP-Mixer, for example, can be used on smartphones, drones, embedded devices, and other devices. Finally, disease recognition on corn leaves can be implemented in various devices and used directly in agricultural areas.

References

[1] I. Tolstikhin *et al.*, “MLP-Mixer: An all-MLP architecture for vision.” [Online]. Available: https://github.com/google-research/vision_transformer.

[2] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale.” [Online]. Available: https://github.com/.

[3] P. Ouppaphan, “Corn disease identification from leaf images using Convolutional Neural Networks,” *ICSEC 2017 - 21st Int. Comput. Sci. Eng. Conf. 2017, Proceeding*, pp. 233–238, Aug. 2018, doi: 10.1109/ICSEC.2017.8443919.

DOI: https://doi.org/10.1007/978-3-031-21967-2_46

Keywords: Diffie-Hellman key agreement, non-commutative cryptography

Extensions of the Diffie-Hellman key agreement protocol based on exponential and logarithmic functions

Z. Lipiński [0000-0001-6722-4402], J. Mizera-Pietraszko [0000-0002-2298-5037]

zlipinski@uni.opole.pl, jolanta.mizera-pietraszko@awl.edu.pl

Simplified Title

Extensions of the Diffie-Hellman key agreement protocol

Abstract

We propose a method of constructing cryptographic key exchange protocols of the Diffie-Hellman type based on the exponential and logarithmic functions over the multiplicative group of integers modulo n. The security of the proposed protocols is based on the computational difficulty of solving a set of congruence equations containing a discrete logarithm. For the multiplicative group of integers modulo n we define the non-commutative group of their automorphisms. On the defined group we construct non-commutative key exchange protocol similar to the Anshel-Anshel-Goldfeld key exchange scheme.

I Introduction

The Diffie-Hellman key exchange is one of the most utilized key agreement protocols in the secure network communication. There are symmetric and asymmetric (public) key exchange variants of the protocol. In case of the symmetric Diffie-Hellman key exchange protocol the users A and B agree the modulus p being the prime number and the primitive root r from the multiplicative group $(\mathrm{Z}/(p))^*$ of integers modulo p. The user A selects the number a and the user B selects the number b from $(\mathrm{Z}/(p))^*$. Both communicating parties exchange the numbers $r^a \bmod p$ and $r^b \bmod p$. The common key is $k_{a,b} = (r^b)^a = (r^a)^b \bmod p$. Security of the protocol is based on computational difficulty in determining the value of discrete logarithm $\log_r(r^a) \bmod \varphi(p)$. The standard X9.42 defines the Diffie-Hellman public key exchange scheme in which the private-public keys is the pair $(a, g^a \bmod p)$, $a, g \in (\mathrm{Z}/(p))^*$. According to the standard, the modulo parameter should be a prime number of the form $p = jq+1$, where q is a large prime and $j \geq 2$. The base g of the public key $g^a \bmod p$ is of the form $g = h^j \bmod p$, where h is any integer with $1 < h < p-1$ such that $h^j \bmod p > 1$. The base g does not have to be a generator of the cyclic group $(\mathrm{Z}/(p))^*$. For the symmetric Diffie-Hellman key exchange protocol the crucial problem is determination of the primitive root elements from the group $(\mathrm{Z}/(p))^*$.

In this article we discus the problem of finding the primitive root elements in cyclic groups. We identify the set of primitive roots in the group $(\mathrm{Z}/(p))^*$ as the common part of the complement of the sets of p-residues. This definition allows in an efficient way to construct new algorithm for searching the primitive root elements in cyclic groups.

The purpose of this article is to present several generalization of the Diffie-Hellman key exchange protocol based on the functions which permute the elements of the group $(\mathrm{Z}/(p))^*$. The security of the discussed protocols is based on the computational difficulty of solving a set of congruence equations containing the discrete logarithm. We define four invertible functions on $(\mathrm{Z}/(p))^*$ determined by elements of the group. These are the exponential function $R(x) = r^x \bmod p$ determined by the primitive roots r of the group $(\mathrm{Z}/(p))^*$, the monomial function $E(x) = x^e \bmod p$ determined by the elements e from $(\mathrm{Z}/\varphi(p))^*$, the automorphism function $M(x) = m \cdot x \bmod p$ and the Chebyshev polynomials of the first kind $C_n(x) \bmod p$, where $\gcd(n, p^2-1) = 1$ and $\varphi(p)$ is the Euler's totient function. With each of the function there is related a group which permute the elements of $(\mathrm{Z}/(p))^*$. The group $G_p^{(r)}$ determined by the exponential functions $R(x)$ is non-abelian. The groups $G_p^{(e)}$, $G_p^{(m)}$, $G_p^{(C)}$ generated by the functions $E(x)$, $M(x)$ and $C_n(x)$ are abelian subgroups of $G_p^{(r)}$ for prime $p > 11$. We use the $G_p^{(r)}$ group to construct new symmetric key agreement protocol based on the idea of the Anshel-Anshel-Goldfeld key agreement scheme.

II State of the Art

There are several generalization of the Diffie-Hellman key exchange protocol. For example, J. Partala in 2018 proposed a generalization of the Diffie–Hellman scheme called Algebraic generalization of Diffe-Hellman (AGDH). Its security is based on the hardness of a solution of the homomorphic image problem, which requires to compute the image of a given element under an unknown homomorphism of two algebras selected as a encryption platform. A. G. Chefranov and A. Y. Mahmoud in 2013 proposed a matrix-based Diffie-Hellman key exchange protocol. The security of the proposed protocol is based on exploiting of a non-invertible public matrix in the key generating process. In 2003 J. H. Cheon and B. Jun proposed a polynomial time algorithm to solve the Diffie-Hellman conjugacy problem in braid groups. In 2012 M. Eftekhari proposed a Diffie-Hellman like key exchange procedure which security is based on the difficulty of computing discrete logarithms in a group matrices over noncommutative ring. In the algorithm the exponentiation is hidden by a matrix conjugation which ensures it security. In 2008 D. Cash, E. Kiltz and V. Shoup proposed new computational problem called the twin Diffie–Hellman problem. The twin DH protocol allows to avoid the problem of an attack on the public keys exchanged in the standard Diffie–Hellman scheme. I. F. Blake and T. Garefalakis analyzed the complexity of the discrete logarithm problem, the Diffie–Hellman and the decision DH problem. The authors showed that if the decision DH problem is hard then computing the two most significant bits of the DH function is hard.

III Original Contribution

In the article we showed how to construct a Diffie–Hellman like cryptographic protocols by means of functions which permutate the elements of the cyclic group $(\mathrm{Z}/(p))^*$. For example in the DH$_{r,e}$ algorithm, which utilises the exponential and monomial functions $R(x)$, $E(x)$, the common cryptographic key has the form $k_{a,b} = r^{(ab)^e} \bmod p$, where the primitve root r and the number $e \in (\mathrm{Z}/\varphi(p))^*$ are known and the numbers $a,b \in (\mathrm{Z}/(p))^*$ are secret. In the DH$_{2r}$ algorithm, which utilises the exponential function $R(x)$, the common cryptographic key has the form $k_{a,b} = r_1^{r_2^{a+b}} \bmod p$, where the primitve roots r_1, r_2 are known and the numbers $a,b \in (\mathrm{Z}/(p))^*$ are secret. In the DH$_{e,\log}$ algorithm, the third modification of the Diffie-Hellman protocol, the common cryptographic key has the form $k(r_a, r_b) = (\log_{r_b}(r_a))^e \bmod \varphi(p)$, where the primitve root r and the number e are known and the numbers $a,b \in (\mathrm{Z}/(p))^*$ are secret. We showed that by composition of functions which permutate the elements of the cyclic group $(\mathrm{Z}/(p))^*$ it is possible to construct an infinite series of cryptographic algorithms.

With each function which permutate the elements of the cyclic group $(\mathrm{Z}/(p))^*$ one can associate a finite, in general non-abelian group. In this article we proposed a new symmetric cryptographic protocol on that group motivated by the Anshel-Anshel-Goldfeld key exchange scheme.

IV Methodology

In the article we discuss the problem of searching generators in finite, cyclic groups. We discuss the problem of generalization of the Diffie–Hellman cryptographic protocol by means of functions which permutate the elements of the cyclic group $(\mathrm{Z}/(p))^*$. We also discuss constructing cryptographic systems over non-commutative finite groups and discuss security of proposed algorithms.

V Conclusions

We proposed three cryptographic key exchange protocols of the Diffie-Hellman type based on the exponential and logarithmic functions over the multiplicative group of integers modulo prime number p. The security of the proposed protocols is based on the computational complexity in solving a set of congruence equations containing the discrete logarithm. For the multiplicative group of integer numbers modulo p we constructed the non-commutative group $G_p^{(r)}$ of their automorphisms. On the defined group we constructed a non-commutative key exchange protocol similar to the Anshel-Anshel-Goldfeld key exchange scheme. The security of the proposed protocols is based on the difficulty of finding path in a defined cipher graph G_{2N}^{cipher} build of $2N$ nodes and solution of the set of certain matrix equations in $G_p^{(r)}$. Proposed cryptographic algorithms can be used in information systems as a public key agreement protocols, the symmetric version of the proposed cryptographic algorithms can be used to assure confidentiality of data transmitted over public network.

DOI: https://doi.org/10.1007/978-3-031-21967-2_52

Keywords: non-commutative cryptography, Cracovian quasigroup

Symmetric and asymmetric cryptography on the special linear Cracovian quasigroup

Z. Lipiński [0000-0001-6722-4402], J. Mizera-Pietraszko [0000-0002-2298-5037]

zlipinski@uni.opole.pl, jolanta.mizera-pietraszko@awl.edu.pl

Simplified Title

Symmetric and asymmetric cryptography on the Cracovian quasigroup

Abstract

We propose symmetric and public key cryptographic protocols on the non-associative special linear Cracovian quasigroup. The $\mathrm{SL_n(Z)}$ Cracovian quasigroup is the set of $n \times n$ matrices with determinant one and entries in the ring of integers Z with the non-associative multiplication. The strength of the proposed symmetric cipher is based on the NP-hardness of the non-negative matrix factorization problem. We define the totient function for a given matrix from the special linear group over the ring Z/mZ of integers modulo m as the order of a cyclic subgroup generated by this matrix. Defined in this way totient function allow us to construct the public key cryptographic protocol on the special linear group $\mathrm{SL_n(Z/mZ)}$ and generalize it to the Cracovian quasigroup.

I Introduction

In the article we propose the symmetric and asymmetric cryptographic protocol defined on the special linear group of matrices over the ring of integers $\mathrm{SL_n(Z)}$. The non-commutativity of the group allows us to construct simply symmetric cryptographic protocol which strength is based on the NP-hardness of the non-negative matrix factorization problem. The asymmetric cryptographic protocol is a generalization of the RSA cryptosystem to the special linear group $\mathrm{SL_n(Z/mZ)}$ over the ring on integers modulo m. Each element A from $\mathrm{SL_n(Z/mZ)}$ generates a cyclic group $\langle A \rangle_m$. We define the totient function $\lambda(m,A)$ in $\mathrm{SL_n(Z/mZ)}$ as the order of the cyclic group generated by A, i.e.,

$$\lambda(m,A) = |\langle A \rangle_m|.$$

All matrices B from the group $\langle A \rangle_m$ have the following property

$$B^{\lambda(m,A)} = I \bmod m.$$

To define a pair of public-private cryptographic keys in $\mathrm{SL_n(Z/mZ)}$ we determine two integers e,d such that $e \cdot d = 1 \bmod \lambda(m,A)$, where $gcd(e,\lambda(m,A)) = 1$. A given sequence $\overline{\alpha}$ of $n(n-1)$ integers to be encrypted we write as a matrix $U_{\overline{\alpha}}$ from $\mathrm{SL_n(Z/mZ)}$. In general the totient $\lambda(m,U_{\overline{\alpha}})$ of the matrix $U_{\overline{\alpha}}$ is unknown. When we conjugate it with a given matrix A

$$U_{A,\overline{\alpha}} = U_{\overline{\alpha}} A U_{\overline{\alpha}}^{-1} \bmod m, \tag{1}$$

then the two matrices A and $U_{A,\overline{\alpha}}$ have the same totient, i.e., $\lambda(m,A) = \lambda(m,U_{A,\overline{\alpha}})$. The sets $\{A, U_{\overline{\alpha}}, h_{\overline{\alpha}}, e, m\}$ and $\{A, U_{\overline{\alpha}}, h_{\overline{\alpha}}, d, m\}$ define a pair of public-private cryptographic keys in $\mathrm{SL_n(Z/mZ)}$. The hash value $h_{\overline{\alpha}}$ of the string $\overline{\alpha}$ is calculated by the sender before encryption to allow the receiver uniquely recover it in the process of decryption.

In this article we define a new non-commutative and non-associative algebraic structure called Cracovian quasigroup. A quasigroup $\{Q,*\}$ is a set Q with a binary multiplication operation $* : Q \times Q \to Q$ in which specification of any two elements $A,B \in Q$ in the equation $A * B = C$ determines the third $C \in Q$ uniquely. Special linear Cracovian quasigroup $(\mathrm{KSL_n(Z)},*)$ is the set of matrices from $\mathrm{SL_n(Z)}$ with the matrix multiplication $A * B := B^T A$, where B^T means matrix transposition and on the right hand side of the definition it is an ordinary matrix multiplication. The elements of the quasigroup $(\mathrm{KSL_n(Z)},*)$ have an inverse $\mathrm{Inv(A)} := (\mathrm{A}^{-1})^{\mathrm{T}}$, i.e., $A * B = B^T A = I$ and $B * A = A^T B = I$, if $B = A^{-T}$, where I is the identity matrix. Note, that I is the right identity element of $\mathrm{KSL_n(Z)}$, i.e., $A * I = A$ but $I * A = A^T$. The Cracovians are the non-associative quasigroups, which means that the associator, $[A,B,C] = (A*B)*C - A*(B*C)$, of the three matrices A,B,C from $\mathrm{KSL_n(Z)}$ equals to $[A,B,C] = (C^T B^T - B^T C)A$ and in general it is non-zero.

The main purpose of the article is to generalize the constructed cryptographic protocols in $\mathrm{SL_n(Z)}$ and $\mathrm{SL_n(Z/mZ)}$ to the non-associative Cracovian quasigroup $\mathrm{KSL_n(Z)}$ and $\mathrm{KSL_n(Z/mZ)}$ respectively. The advantage of using the non-associative algebraic structures in cryptography is that the multiplication of three or more elements depends on the order their multiplication. This means, that passing from the matrix group to the non-associative matrix quasigroup we increase the complexity of matrix factorization problem. For a product of $(N+1)$ elements from the non-associative quasigroup we can obtain $C_N = \frac{(2N)!}{(N+1)!N!}$ results depending on the order of multiplication. The formula C_N counts the number of possible ways to insert the parentheses into a product of $(N+1)$ elements.

II State of the Art

The idea of using the non-abelian groups and quasigroups in cryptography has its origin in the solutions of three famous problems in combinatorial group theory proposed by M. Dehn in 1911. These are the word problem, the conjugacy problem and the isomorphism problem for finitely presented groups. A presentation $\{S,D\}$ of a group G is a set S of group generators and the collection of words D on the elements of S and their inverses that define G. For finitely presented group G the word problem is the algorithmic problem of deciding for arbitrary words g of G whether or not $g = I$ in G. The conjugacy problem asks whether there exists $g \in G$ such that $g_2 = g^{-1}g_1g$ in G, i.e., whether two elements g_1 and g_2 from G are conjugate. For the isomorphism problem it must be decided whether two given presentations define the isomorphic groups. In 1954 Novikov constucted a finitely presented group for which the conjugacy problem is unsolvable. In 1955, Novikov and Boone independently showed that there are finite group presentations whose word problem is undecidable. Using the Novikov's results, Adian and Rabin showed that the isomorphism problem for finitely presented groups is unsolvable. In 1985 Wagner and Magyarik devised the first public-key protocol based on the unsolvability of the word problem for finitely presented groups. A non-deterministic public-key cryptosystem based on the conjugacy problem on braid group, similar to the Diffie-Hellman key exchange system, was proposed by K.H. Ko et al. in 2000. In the Anshel-Anshel-Goldfeld key agreement system and the public-key cryptosystem the authors used the braid groups where the word problem is easy to solve but the conjugacy problem is intractable. This is due to the fact that on braid groups the best known algorithm to solve the conjugacy problem requires at least exponential running time. The earliest quasigroup-based public-key cryptosystem (PKC) was proposed by Koscielny and Mullen in 1999. A. Kalka generalized the Anshel-Anshel-Goldfeld PKC to the non-associative algebraic structures, called the left self-distributive systems.

III Original Contribution

We define and discuss the properties of the non-associative and non-commutative Cracovian quasigroup $\mathrm{KSL_n(Z)}$ and $\mathrm{KSL_n(Z/mZ)}$. We develop notation which allow in a easy way to perform basic algebraic operations over elements of the Cracovians. A basic properties of the totient function $\lambda(m,A)$ defined in $\mathrm{SL_n(Z/mZ)}$ and $\mathrm{KSL_n(Z/mZ)}$ quasigroup is presented. Two new cryptographic protocols over special linear group $\mathrm{SL_n(Z/mZ)}$ and Cracovian quasigroup $\mathrm{KSL_n(Z/mZ)}$ are constructed and example are discussed.

IV Methodology

The article presents a theoretical aspects of building cryptographic systems over non-commutative and non-associative algebraic structures. We develop notation which allow to perform basic algebraic operations in the Cracovian quasigroup $\mathrm{KSL_n(Z/mZ)}$. Defined cryptographic protocols in $\mathrm{KSL_n(Z)}$ and $\mathrm{KSL_n(Z/mZ)}$ are analyzed and the case study are presented.

V Conclusions

We proposed new symmetric and asymmetric cryptographic protocols on the special linear group $\mathrm{SL_n}$ and generalized them to the non-associative Cracovian quasigroup $\mathrm{KSL_n}$. The encryption with the symmetric cipher in $\mathrm{SL_n(Z)}$ we defined as the multiplication of given sequence of matrices, each of which represent a letter of the encrypted plain text. The strength of the proposed symmetric cipher is based on the NP-hardness of the non-negative matrix factorization problem. Further, we increased the decryption complexity of the symmetric cipher by passing from the group $\mathrm{SL_n(Z)}$ to $\mathrm{KSL_n(Z)}$ in which the decryption of a cipher text requires solution of the matrix factorization problem on the non-associative quasigroup.

Proposed the asymmetric cipher on the $\mathrm{SL_n(Z/mZ)}$ group was based on the definition of the totient function $\lambda(m,A)$ for arbitrary matrix A from $\mathrm{SL_n(Z/mZ)}$. The pair of encryption keys was defined as the set of the following elements $\{K,U_{\overline{\alpha}},h_{\overline{\alpha}},e,m\}$, $\{K,U_{\overline{\alpha}},h_{\overline{\alpha}},d,m\}$ where the encryption of the plain text encoded in the matrix $U_{\overline{\alpha}}$ was just taking the power $(U_{\overline{\alpha}}KU_{\overline{\alpha}}^{-1})^e \bmod m$, with known the totient function $\lambda(m,K)$ of the matrix K. The decryption was given by the formula $((U_{\overline{\alpha}}KU_{\overline{\alpha}}^{-1})^e)^d \bmod m$, where $e \cdot d = 1 \bmod \lambda(m,K)$. The strength of the proposed asymmetric cipher was based on the difficulty of finding the value $\lambda(m,K)$ and the numbers e, d. Also, to recover of the plain text from the matrix $U_{\overline{\alpha}}KU_{\overline{\alpha}}^{-1}$ requires solution of the non-trivial matrix factorization problem on the group $\mathrm{SL_n(Z/mZ)}$ or on the Cracovian quasigroup $\mathrm{KSL_n(Z/mZ)}$.

DOI: https://doi.org/10.1007/978-3-031-21743-2_31
Link to repository: https://hol.compute.dtu.dk/Scratch.thy

Keywords: Logic, Automated Reasoning, Isabelle Proof Assistant

On Verified Automated Reasoning in Propositional Logic

Simon Tobias Lund, Jørgen Villadsen[0000-0003-3624-1159]

jovi@dtu.dk

Simplified Title

Using the Isabelle Proof Assistant to Formally Verify a Prover for Propositional Logic

Abstract

As the complexity of software systems is ever increasing, so is the need for practical tools for formal verification. Among these are automatic theorem provers, capable of solving various reasoning problems automatically, and proof assistants, capable of deriving more complex results when guided by a mathematician/programmer. In this paper we consider using the latter to build the former. In the proof assistant Isabelle/HOL we combine functional programming and logical program verification to build a theorem prover for propositional logic. Finally, we consider how such a prover can be used to solve a reasoning task without much mental labor.

I Introduction

Using the Isabelle proof assistant [2], we present work towards a system translating a text from a fragment of natural language to formulas in propositional logic, use our logical approach to solve the problem in the text, and next translate back to a conclusion expressed in natural language.

II State of the Art

We find that the few existing formally verified provers for propositional logic in Isabelle/HOL [1, 3] should be simplified and/or evaluated on case studies.

III Original Contribution

We present a novel and verified prover for propositional logic based on implication ($\rightarrow$) and falsity ($\bot$), and showcase its use. The main contribution is presenting modern automated reasoning methodology and tools to a broader audience.

IV Methodology

This is a case study in automated reasoning. We show how tools from the intersection of computer science and logic can be used for rapid, high-assurance development and alleviating the mental labor involved in certain reasoning tasks. To this purpose, we use the proof assistant Isabelle/HOL (Higher-Order Logic) to develop and verify a prover for propositional logic. We then showcase this verified prover by using it to solve a (fairly complicated) riddle by Raymond Smullyan.

V Results

We obtain a prover for checking if formulas of classical propositional logic based on $\rightarrow$ and $\bot$ are tautologies (true for all interpretations). For ease of use, we extended the logic with abbreviations for other propositional operators. The verification of the prover consisted of showing termination and giving a proof for soundness and completeness. Both are verified by the Isabelle proof assistant. The proof of soundness and completeness shows that our prover returns valid/true for a formula exactly when the formula is true for any interpretation of the propositional symbols. Thus, the prover will be able to analyse any formula we give it, and whatever result it returns will be correct.

We show how, once we have access to a propositional prover (which we can trust to a high degree as a product of our verification), we can use it to avoid having to perform difficult reasoning tasks manually. This showcase consists of translating a riddle to propositional logic, and then running the prover on the formula obtained by taking the implication from the assumptions of the riddle to one of the possible outcomes. The prover then returns valid/true if it is sound to conclude the given outcome. We would liken this use of tautology checkers to quickly conclude results about problems which can be reduced to propositional logic to the use of calculators to calculate results about problems which can be reduced to arithmetic expressions. For this, it is obviously not necessary to use a self-developed propositional solver, and one of the many efficient SAT-solvers are better suited.

VI Evaluation

We use the formally verified prover to solve the following riddle by Raymond Smullyan:

> *You find yourself on a desert island inhabited by knights and knaves. Knights only ever tell the truth and knaves only ever tell lies. In a clearing, you come upon three islanders: Ann, Bob, and Cat. You ask Bob how many knights there are among them. Bob answers something in a foreign language. You then ask Ann what Bob said. Ann answers: "He said there is one knight among us." Cat then reacts: "Don't listen to Ann, for she is a knave!" Who, if any, among them are knights?*

The execution time in Isabelle/HOL for the file is around a second. This includes the formal verification of the prover as well as the solution to the riddle. The only complex proof method is *(simp_all, blast, meson, fast)*. The alternatives *(simp_all, blast+)* and *(simp_all, blast 7, blast 5, blast 6)* are around ten times slower but lower arguments to the proof method *blast* do not succeed within ten seconds.

The standard proof method *auto* fails. It is identical to *(auto 4 2)* but *(auto 5 2)* succeeds and it takes only around twice the time of the relatively quick *(simp_all, blast, meson, fast)* so it can be considered a very robust proof whereas in principle the proofs starting with *simp_all* can fail in future releases of Isabelle/HOL due to changes in the simplifier. The use of arguments to *auto* is rare but not as such problematic.

VII Conclusions

Our paper presents a method for development where formally verified programs can be produced without much added work. This is done by combining functional programming and logic in Isabelle/HOL. We also show how another facet of automated reasoning – fully automated tautology checkers – can be used for other reasoning tasks by using our verified prover to find the solution to a riddle. Automated reasoning can, in our opinion, be beneficially used by anyone in the computer science field (and many other fields as well) for decreasing both labour and miscalculations.

References

[1] Michaelis, J., and Nipkow, T. Formalized Proof Systems for Propositional Logic. In *23rd International Conference on Types for Proofs and Programs (TYPES 2017)* (2018), A. Abel, F. N. Forsberg, and A. Kaposi, Eds., vol. 104 of *LIPIcs*, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, pp. 6:1–6:16.

[2] Nipkow, T., Paulson, L. C., and Wenzel, M. *Isabelle/HOL - A Proof Assistant for Higher-Order Logic*, vol. 2283 of *Lecture Notes in Computer Science*. Springer, 2002.

[3] Villadsen, J. Tautology checkers in Isabelle and Haskell. In *Proceedings of the 35th Edition of the Italian Conference on Computational Logic (CILC 2020)* (2020), F. Calimeri, S. Perri, and E. Zumpano, Eds., vol. 2710, CEUR-WS.org, pp. 327–341.

DOI: `https://doi.org/10.1007/978-3-031-21743-2_10`

Keywords: localization, positioning, interpolation, Wi-Fi

Impact of Radiomap Interpolation on Accuracy of Fingerprinting Algorithms

Juraj Machaj[0000-0002-7544-8796], Peter Brida[0000-0002-5442-9246]

juraj.machaj@feit.uniza.sk, peter.brida@feit.uniza.sk

Simplified Title

Effect of received signal strength interpolation on accuracy of positioning algorithms.

Abstract

The positioning using Wi-Fi signals and fingerprinting algorithms achieved a lot of attention lately. The main drawback of the fingerprinting localization is the process of radiomap creation, which is labour intensive and time-consuming procedure. Therefore, some solutions for crowdsourcing and dynamic map creation were proposed. However, the problem of these is that users usually don't move regularly thru all the areas, which leads to undersampling of certain parts of the localization area. In this paper interpolation algorithms are used to increase radiomap density and thus reduce the problem with undersampling. To evaluate the impact of interpolation on the performance of fingerprinting algorithms the NN, KNN and WKNN algorithms were tested on dynamic radiomap without interpolation as well as with radiomaps created using linear, inverse distance weight and Kriging interpolations. The paper shows the results achieved in the real-world scenario and shows the impact of the interpolation algorithms on the performance of the above-mentioned localization algorithms.

I Introduction

Fingerprinting positioning algorithms require a Received Signal Strength (RSS) database, i.e. radiomap, to calculate position. The radiomap can be created by measurements at predefined positions. Recently a lot of work focused on crowdsourcing radiomap data, reducing time complexity. The accuracy of positioning algorithms depends on the density of the radiomap, distribution of reference points in the area may have an impact on performance. The paper is focused on the use of interpolation algorithms to increase the density of radiomap and improve localization accuracy in areas with low density of radiomap.

II State of the Art

II.1 Fingerprinting localization

Fingerprinting localization is based on a comparison of data reported by a mobile device with radiomap. The Euclidean distance is widely used for the comparison of RSS samples in deterministic algorithms. There are three basic modifications of deterministic algorithm implementation. In case when only one point from radiomap is selected, the algorithm is called NN (Nearest Neighbour); when multiple points with the smallest distances are selected, algorithms are called KNN (K-Nearest Neighbour) and WKNN (Weighted K-Nearest Neighbour). KNN algorithm calculates position as an average of selected points, while WKNN uses inverse distances as weights [2].

II.2 Dynamic radiomap

Dynamic radiomap is created using IMU and Wi-Fi samples from the smartphone. IMU data is used in particle filter-based pedestrian dead reckoning algorithm together with information about the environment. Output from the algorithm can be used to estimate positions where RSS samples were taken. Generated reference points can be merged with existing ones when similarity conditions are met [1].

II.3 Interpolation algorithms

In the data processing part, three interpolation algorithms were implemented. Linear interpolation calculates RSS values at positions using tessellation triangles [3]. Inverse distance weighting interpolation takes into account the impact of distance on interpolated RSS value. While Kriging interpolation is based on finding a mathematical function that describes the spatial structure of interpolated data.

III Original Contribution

The original contribution of the paper is the implementation of multiple interpolation techniques on crowdsourced RSS measurements and the evaluation of the impact on localization algorithms.

IV Methodology

Results were achieved based on data measured in real world conditions using off-the-shelf smartphones. The data were processed, visualised and evaluated using scripts prepared in Matlab environment. Therefore, the study is experimental.

V Results

Achieved results show that the lowest mean error was achieved by a simple NN algorithm and IDW interpolation. However, when compared to NN without interpolation, the improvement is not significant. Results achieved by the WKNN algorithm with linear interpolation are the best among WKNN and KNN algorithms. Therefore, linear interpolation seems to provide the best results when algorithms use multiple reference points to calculate position.

VI Evaluation

The impact of interpolation algorithms was evaluated using experimental measurements in real world scenario, in order to provide a reliable outcome. From the achieved results, it is possible to conclude that interpolation has a limited impact on the accuracy of used localization algorithms.

VII Conclusions

Multiple interpolation techniques for received signal strength were tested to improve the accuracy of localization algorithms. The data was collected in real world environment. The interpolation had minimum effect on the simplest algorithm; on the other hand, the impact of interpolation was higher for more complicated algorithms. Achieved results show that interpolation can increase accuracy by approximately 10% in the case of KNN and WKNN algorithms.

References

[1] Brida, P., Machaj, J., Racko, J., and Krejcar, O. Algorithm for dynamic fingerprinting radio map creation using imu measurements. *Sensors 21*, 7 (2021).

[2] Honkavirta, V., Perala, T., Ali-Loytty, S., and Piche, R. A comparative survey of wlan location fingerprinting methods. In *2009 6th Workshop on Positioning, Navigation and Communication* (2009), pp. 243–251.

[3] Talvitie, J., Renfors, M., and Lohan, E. S. Distance-based interpolation and extrapolation methods for rss-based localization with indoor wireless signals. *IEEE Transactions on Vehicular Technology 64*, 4 (2015), 1340–1353.

DOI: https://doi.org/10.1007/978-3-031-21967-2_24

Keywords: Whitelisting, random forest, software security, semi-supervised learning, decision support system

Random Forest in Whitelist-based ATM Security

Michal Maliszewski, Urszula Boryczka

michal.maliszewski@dieboldnixdorf.com, urszula.boryczka@us.edu.pl

Simplified Title

Whitelist-based Protection of the Automated Teller Machines Against Logical Attacks Using Random Forest Algorithm

Abstract

Accelerated by the COVID-19 pandemic, the trend of highly-sophisticated logical attacks on automated teller machines (ATMs) is ever-increasing nowadays. Due to the nature of attacks, it is common to use zero-day protection for the devices. The most secure solutions available use whitelist-based policies, which are extremely hard to configure. This article presents the concept of a semi-supervised decision support system based on the Random forest algorithm for generating a whitelist-based security policy using the ATM usage data. The obtained results confirm that the Random forest algorithm is effective in such scenarios and can be used to increase the security of the ATMs.

I Introduction

In 2020, the number of automatic teller machines exceeded 3 million around the globe. Millions of people use ATMs daily, which is why the security of cash systems is of utmost importance. ATMs are exposed to two types of attacks: physical and logical, with the latter becoming more frequent in the last years.

A large part of logical attacks is prepared for a specific type of device, considering the installed software stack and its potential security issues. As the threat is yet unknown, standard protection mechanisms such as intrusion detection systems and anti-viruses are not the preferred form of ATM protection. To ensure the security of the devices against yet unknown threats, the whitelist approach is used.

Alas, creating whitelist-based protection is currently a manual process that may generate numerous errors resulting in an insufficient protection level. The person responsible for the device security must determine how the software installed on the ATM operates and classify it to the predefined security rules to ensure proper working. For example, program *p* must get permission to write to directory *r*, where it can store its logs.

In these studies, we decided to automate the ATM protection process using the Random forest algorithm to mitigate the problem. We expected that a properly implemented solution would significantly reduce the number of errors and the time needed to protect the device properly.

II State of the Art

As for ATM security itself, there are not many publications on the subject. Most of them are based on an analysis of the threats or the current situation. However, as part of our work on ATM security, we published articles on the use of grouping algorithms to protect these systems from logical attacks using the sandbox mechanism [1, 2]. The solutions we proposed are highly effective in limiting the spread of viruses; however, they do not guarantee such comprehensive protection as the whitelisting approach discussed in these studies.

III Original Contribution

According to our knowledge, in our research, for the first time, a Random forest algorithm was used to protect ATMs against logical attacks. Moreover, it is the first proposal to automate the process of protecting the ATMs against logical attacks, based on whitelists.

IV Methodology

To achieve the results, a number of pre-processing methods were used, including, among others: the removal of noises, duplicates, modification of some attributes (e.g., paths to programs or required by them resources), or methods of selective attribute selection depending on their information gain ratio. Moreover, the classifiers used and the method of their creation was also adapted to the presented problem. This applied primarily to determining the size of the Random forest and the selection of attributes used in the construction of individual decision trees. Due to the lack of literature in the field of ATM security and the possibility of using real data from physical devices in Diebold Nixdorf laboratories, our research was experimental.

V Results

During the studies, we are able to detect, using a discrete event system and our pre-processing methods, a set of programs running in the operating system. That observation allows us to create whitelist-based security protection. The Random forest algorithm used for this purpose provides excellent classification results, with correctness exceeding 90% in every applied quality criterion.

It is worth noting that using the results of previous supervised classifications in the training set, a significant increase in the quality of the classification was noted. Depending on the measurement method, the quality improvement ranged from 1.71% to 5.29%.

VI Evaluation

The results achieved in these studies show that machine learning techniques, especially the Random forest algorithm can be successfully used to protect ATMs.

To assess the quality of the classification, we adopted standard and well-known metrics, usually used in classification problems: Accuracy, Precision, Recall and F-score. Additionally, we relied on expert knowledge to define the ideal classification to verify the results.

Please be aware that our goal was not to find the best classification algorithm in the research area. Instead, our goal was to find a well-described algorithm accepted by the banking sector, which we could easily extend in our future research depending on the requirements set by the industry.

VII Conclusions

Even though it is challenging to configure whitelist solutions, it is safe to say that this approach is the future of cybersecurity for single-purpose systems or for the Internet of Things (IoT). The automatic creation of security configurations is tough to achieve, as there are still security goals that require expert knowledge and human supervision. However, it is much easier for specialized systems to provide complex program classification from observations than having someone do them manually.

The results of these studies can be used to better secure ATMs as well as other systems with a fixed software stack, e.g., cashier dispensers, medical devices, vehicle charging stations, server solutions, and many others.

As the classification system is independent of the technology used, it can be used together with many whitelist security solutions available on the market.

References

[1] Maliszewski, M., and Boryczka, U. Using MajorClust algorithm for sandbox-based ATM security. In *2021 IEEE Congress on Evolutionary Computation (CEC)* (2021), pp. 1054–1061.

[2] Maliszewski, M., Pristerjahn, S., and Boryczka, U. DBSCAN algorithm as a means to protect the ATM systems. In *2018 Innovations in Intelligent Systems and Applications (INISTA)* (07 2018), pp. 1–6.

DOI: https://doi.org/10.1007/978-981-19-8234-7_23
Link to repository: https://github.com/nurassyl7/python/blob/main/python.txt

Keywords: Smart technologies in motor transport, simulation event-probabilistic model, smart traffic light, IoT device

Traffic management in Smart City

Mansurova M.E.[0000-0002-9680-2758], Belgibaev B.A.[0000-0002-6857-3775], Zhamangarin D.S.[0000-0002-2526-6492],
Zholdas N.A.[0000-0002-4941-8198]

Madina.Mansurova@kaznu.edu.kz, bbelgibaev@list.ru, dus_man89@mail.ru, Zholdas.Nurassyl@kaznu.edu.kz

Simplified Title

Traffic management in Smart City.

Abstract

The article discusses new smart technologies for monitoring and controlling road traffic in the city of Almaty. It is shown that the megapolis has specific features in automating the processes of controlling the movement of vehicles. They are connected with the polycentricity of the city and a number of mutually intersecting traffic flows in the central part of Almaty. The algorithms and applications presented in the article for the AnyLogic PLE environment open up practical opportunities to predict, based on an event-probabilistic simulation model of traffic, the dynamics of the development of congestion processes with a few hours ahead of time.

I Introduction

The development of modern digital information exchange technologies on a global scale and the use of intelligent control systems in the automotive industry have significantly changed the theoretical, methodological, and practical approaches to the organization of optimal traffic in densely populated cities. Current traffic flow management mathematical models are deterministic in nature, enabling the empirical formulation of contradictory initial and boundary conditions related to identifying the type of flow and the capacity of the transport artery at maximum load. This results in one of the vehicle lanes having insufficient capacity during ”peak hours” and significantly underutilizing the lanes for oncoming traffic. The city of Almaty has a number of densely inhabited residential communities and suburbs, whose residents work in Almaty’s downtown [1]. Due to the mutual intersection of the traffic flows, congestion develops during peak hours. Modern sophisticated automated control systems face a pressing challenge in trying to find efficient ways to lighten the burden on the city’s transportation network. The authors of this article have developed a conceptual scheme for adaptive traffic light control and a module for calculating traffic flow parameters [3]. In the development environments of VMware Workstation Pro, TIA Portal V13, Logo! Soft Comfort there are developments on the creation of adaptive self-adjusting digital traffic light controllers based on Siemens microcontrollers in a network version [2].

II State of the Art

System analysis using simulation and simulation modeling in AnyLogic, VMware Workstation Pro, TIA Portal V13, Logo! Soft Comfort environments has shown that the dynamic change in the time phases of traffic light objects, as collective intelligent IoT devices, allows to change and actively manage the dynamics of traffic flows on the most loaded sections of highways. The AnyLogic environment’s local application of simulation modeling enables the landscape connection of roadways to the anticipated flow of vehicles on major highways at the entrance/exit to highly populated cities. Traffic signal cyclograms are chosen semi-empirically and modern advanced experimental studies of the dynamics of traffic flows dependent on the time of day are conducted in the AnyLogic environment. The difficulty of this system’s integration with the traffic management system is a drawback.

III Original Contribution

RC 2 does not have a special interface device with microcontrollers and graphical reprogrammable interfaces for displaying current traffic information. The proposed program codes for this controller are ”closed” and do not allow them to be reprogrammed to meet the requirements of Smart City. The proposed design technical solution makes it possible to raise the level of traffic management in the country to Industry 4.0 technologies.

IV Methodology

The scientific foundations of the traffic management in Almaty require experimental studies using the LO@RA WAN network technology. The collected data from LO@RA mobile sensors allows for a system analysis of the array of dislocation of cars in the form of a closed information and mathematical model of a Smart City using artificial intelligence. This approach has a streaming and economic component. It is advantageous for the traffic police of the city to create their own computer network and save money resources.

V Results

In accordance with the technology of building traffic simulation models in the AnyLogic PLE environment, it is necessary to use the primitives of the Traffic library. The design of a section of the road network begins with the construction of the main route and the creation of an intersection at the points of closure of the routes. The given modeling technology was used to build simulation models of the development of the traffic situation in the different areas of Almaty city. Calculations and variation of the values of the sources of cars show weak influence of multi-level car bridges on reducing the level of congestion. The main reason for this is the same as in the morning hours of the highways of the western direction. All the cars tried to go to the central part of the city. The absence of a ring multi-lane bypass road leads to the appearance of this phenomenon.

VI Evaluation

The simulation experiments showed the dependence of the appearance of congestion processes on the intensity of traffic on Gagarin Street. The increase in traffic intensity along Gagarin Street from 1000 to 1200 cars per hour leads to the emergence of congestion processes that develop into congestion after 4 cycles of traffic lights.

VII Conclusions

It is shown that the replacement of the analog actuator in the traffic light controller RC 2 with a modern microcontroller with artificial intelligence elements allows to obtain a number of important additional functionality, such as: accounting for astronomical, weekly and daytime illumination of highways and their workload on days of the week. The improvement of the RC 2 traffic light controller to an IoT device also enables an interactive way to control traffic lights and road signs from a mobile or stationary LED screen. The calculation schemes and formulas obtained in the AnyLogic environment for optimizing the cyclograms of the operation of the ”smart” traffic light, allowing the monitoring mode to predict and prevent effectively manage the flows of vehicles, resolve conflict and emergency situations. It is shown that simulation modeling in the AnyLogic PLE environment is a promising direction for the development of algorithms for time forecasting the development of congestion processes and traffic regulation in large cities of the country. Simulation experiments, taking into account the improved algorithms for changing the phases of traffic lights, allow to dynamically control the traffic situation at complex intersections. The calculated data indicate the need to take administrative measures to reduce oncoming traffic flows that perform a logistical function for the wholesale markets of the country’s megacities. This work was carried out and sponsored within the framework of the scientific project AR09261344 ”Development of methods for automatic extraction of geospatial objects from heterogeneous sources for information support of geographic information systems”.

References

[1] *Almaty Sustainable Transport Strategy for 2013-2023.* http://almatyinvest.kz.

[2] Mansurova, M., Akhmetov, B., Zhamangarin, D., and Smilov, N. *Mobile intelligent road sign.* RK Patent No. 5097 for the utility model, 2021.

[3] Mansurova, M., Belgibaev, B., Ixanov, S., Karimsakova, D., and Zhamangarin, D. *Interaction of Adjacent Smart Traffic Lights During Traffic Jams at an Intersection.* Applied Mathematics and Information Sciences, 2021.

DOI: https://doi.org/10.1007/978-3-031-21743-2_26

Keywords: Arrhythmia, Recurrence Plot, Convolutional neural network, Electrocardiography, Deep Learning

ECG Signal Classification using Recurrence Plot-Based Approach and Deep Learning for Arrhythmia Prediction

Niken Prasasti Martono[0000-0003-0204-9230], Toru Nishiguchi, Hayato Ohwada

niken@rs.tus.ac.jp, 7422537@ed.tus.ac.jp, ohwada@rs.tus.ac.jp

SIMPLIFIED TITLE

ECG Classification using Machine Learning to Predict Arrhythmia

ABSTRACT

Automatic electrocardiogram (ECG) analysis is crucial in diagnosing heart arrhythmia but is limited by the performance of existing models owing to the high complexity of time series data analysis. Arrhythmia is a heart condition in which the rate or rhythm of the heartbeat is abnormal. The heartbeat may be excessively fast or slow or may have an irregular pattern. Research has shown that the use of deep Convolutional Neural Networks (CNNs) for time-series classification has several advantages over other methods. They are highly noise-resistant models and can very informatively extract deep features that are independent of time. Five classes of heartbeat types in the MIT-BIH arrhythmia database were classified using the resilient and efficient deep CNNs proposed in this study. The proposed method achieved the best score (95.8% accuracy) for arrhythmia detection using the deep learning classification method.

I INTRODUCTION

Over the years, a variety of automatic ECG categorization algorithms, based on signal processing techniques, have been presented. Commonly used methods include the wavelet transform, frequency analysis, support vector machines (SVMs), artificial neural networks (ANNs), decision trees, linear discriminant analysis, and Bayesian classifiers.

II STATE OF THE ART

The implementation of deep-learning algorithms has recently emerged as the most popular strategy. Deep convolutional neural networks (CNNs) have recently yielded impressive results in tasks such as image classification and speech recognition, and this technology can greatly improve healthcare and clinical practice. To date, the most effective systems have used supervised learning to automate examination diagnoses [2, 1].

III ORIGINAL CONTRIBUTION

This research proposes a CNN-based ECG arrhythmia classification model, with recurrence plot-based data used in the training, to maximize the ability of the classifier to analyze time-series data. We identify and learn ECG data with the goal of improving the accuracy by constructing and optimizing neural networks. The ECG dataset was organized into five categories in our study to provide a general assessment of heart health, as well as an important and dependable reference for future diagnosis by the doctor.

IV METHODOLOGY

The overall process implemented in this study is as follows. First, data preprocessing is conducted, including signal labeling, splitting, normalization, and implementing the recurrence plot method on the ECG signal data. Then, we transform and preprocess the ECG data using recurrence plots. We implemented CNNs and an activation function to introduce nonlinear characteristics into the neural network model. A rectified linear unit (ReLU) is one of the most commonly used activation functions in CNNs.

V RESULTS

The results indicate that our classifier could classify most of the classes well, with classes N and V exhibiting the best performance. For classes, N, S, V, and Q, the precision was above 70.0%, with 97.4% being the highest score. However, for class Q, owing to the very limited number of samples compared with the other classes, the performance result seems to be very low. Using 5-fold cross validation, the classifier has 93.3% accuracy, 70.0% precision, 66.4% recall, and 72.0% F-1 score. The low results for the last three evaluations could be attributable to the low prediction score of class Q.

VI Evaluation

we proposed a CNN-based ECG arrhythmia classification algorithm. This study used a data reconstruction method; in this case, recurrence plots were used to prepare the training data. Using this method reduced the need for complex feature processing and calculation. The ECG records in the MIT-BIH arrhythmia database were processed to obtain 2-D outputs and used as model input data. Finally, the trained model classified the ECG signal into five classes: normal beat, supraventricular ectopic, ventricular ectopic, fusion, and unknown beats.

VII Conclusions

The optimized CNNs model uses an RELU activation function, dropout, and other technologies to create a network architecture. Our proposed method performs well in the classification, with an overall average accuracy rate of 93.3% and best accuracy of 95.8% lead ECG data, which is superior to previous works on arrythmia prediction using deep CNNs and recurrence plot. It could accurately categorize ECG signals according to the results, and should be useful to physician in predicting arrhythmia.

References

[1] Martis, R. J., Acharya, U. R., and Min, L. C. Ecg beat classification using pca, lda, ica and discrete wavelet transform. *Biomedical Signal Processing and Control 8* (2013), 437–448.

[2] Varatharajan, R., Manogaran, G., and Priyan, M. K. A big data classification approach using lda with an enhanced svm method for ecg signals in cloud computing. *Multimedia Tools and Applications 77* (4 2018), 10195–10215.

Link to arXiv: https://arxiv.org/abs/2207.09869
DOI: 10.1007/978-3-030-73280-6
Keywords: Semi-pseudo-labeling, 3D data augmentation, Neural network training, 3D object detection, Machine learning

A Novel Neural Network Training Method for Autonomous Driving Using Semi-Pseudo-Labels and 3D Data Augmentations

Tamás Matuszka, Dániel Kozma
tamas.matuszka@aimotive.com, daniel.kozma@aimotive.com

Simplified Title

A novel neural network training method for autonomous driving

Abstract

Training neural networks to perform 3D object detection for autonomous driving requires a large amount of diverse annotated data. However, obtaining training data with sufficient quality and quantity is expensive and sometimes impossible due to human and sensor constraints. Therefore, a novel solution is needed for extending current training methods to overcome this limitation and enable accurate 3D object detection. Our solution for the abovementioned problem combines semi-pseudo-labeling and novel 3D augmentations. For demonstrating the applicability of the proposed method, we have designed a convolutional neural network for 3D object detection which can significantly increase the detection range in comparison with the training data distribution.

I Introduction

The task of 3D object detection [1] requires high-quality annotations, which are hard to acquire and expensive to label, especially for distant objects. Consequently, several datasets do not include annotations at distances of over 100 meters, while an Advanced Driver-Assistance System must be able to perceive objects beyond this limit. The paper describes a solution for overcoming this limitation in training datasets.

II State of the Art

Several approaches exist to facilitate neural network training. One of the most popular solutions is transfer learning, where a neural network is trained on a particular dataset and then fine-tuned on another dataset. Self-supervised learning has resulted in breakthroughs in language modeling. Pseudo-labeling is a simple solution that uses the same model's predictions as true labels during the training. However, none of these solutions helps the model to produce predictions that are not part of the training distribution.

III Original Contribution

The first contribution of the paper is the definition of 3D augmentations. Most 2D data augmentations are easy to generalize in three dimensions. However, zooming in or out changes the image scale, altering the position and egocentric orientation of annotated objects in 3D space too, which must be handled during the augmentation.

To provide a supervision signal for image-visible objects without 3D annotation, we introduced semi-pseudo-labeling (SPL) as a method where pseudo-labels are generated by a neural network trained on a simpler task and utilized during the training of another network performing a more complex task.

IV Methodology

We developed a virtual camera-based solution for enabling 3D augmentations. The method includes a zoom in/zoom-out step, which makes it appear that objects are moving closer to or further from the ego car while the camera matrix is modified, ensuring the transformation in image space is consistent with the real-world space. The principal challenge is dealing with unannotated objects that appear in the image. Fig 1 describes the effects of zoom augmentation where objects with and without annotations are depicted with grey and red rectangles, respectively.

In this example, the object marked with the red rectangle is located beyond the limit of the annotation range; hence there is no information even on the presence of the object, but it is still within the required operating range and visible on the image. Therefore, it cannot be used to supervise the loss. We used SPL for this case, the output of a regular 2D object detector (i.e., semi-pseudo-labels) can be used while training a 3D detector. This method allowed us to utilize image-visible objects without 3D annotations during the training phase (see Fig 2) and thus significantly extend the detection range.

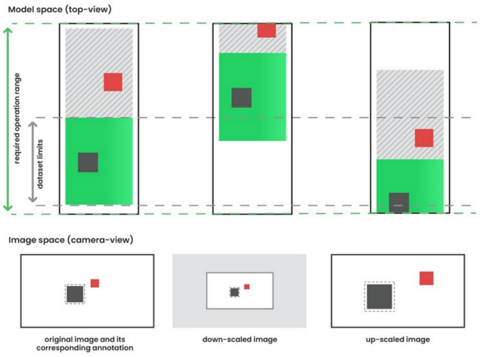


Figure 1: Effects of zoom augmentation on image and model space data.

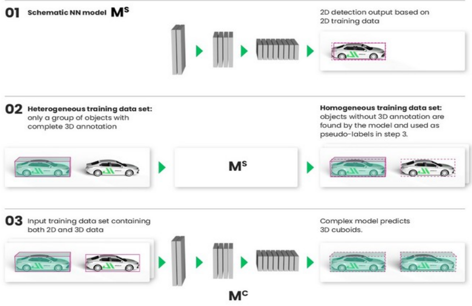

01 Schematic NN model M^S
2D detection output based on 2D training data
02 Heterogeneous training data set: only a group of objects with complete 3D annotation
M^S
Homogeneous training data set: objects without 3D annotation are found by the model and used as pseudo-labels in step 3.
03 Input training data set containing both 2D and 3D data
M^C
Complex model predicts 3D cuboids.

Figure 2: SPL applied in 3D object detection using 2D detection as a simple task.

V RESULTS

We developed a single-stage object detector based on the YOLOv3 [3] architecture which utilizes our SPL method and 3D data augmentations during the training.

VI EVALUATION

We benchmarked the performance of the trained neural network on Argoverse [2] and an in-house dataset. Overall, the model trained with our method has better performance in bird's-eye-view space as well as in image space than the baseline model. Furthermore, the detection area is extended significantly (from 120m to 200m).

VII CONCLUSIONS

We applied our proposed method for 3D object detection using a YOLO-variant. However, one could use it with different model architectures too, or in other fields (for example, object segmentation) by reproducing the proposed training method.

REFERENCES

[1] ARNOLD, E., AL-JARRAH, O. Y., DIANATI, M., FALLAH, S., OXTOBY, D., AND MOUZAKITIS, A. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems 20*, 10 (2019), 3782–3795.

[2] CHANG, M.-F., LAMBERT, J., SANGKLOY, P., SINGH, J., BAK, S., HARTNETT, A., WANG, D., CARR, P., LUCEY , S., RAMANAN, D., ET AL. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 8748–8757.

[3] REDMON, J., AND FARHADI, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

DOI: https://doi.org/10.1007/978-981-19-8234-7_24

Keywords: Temporal Data Mining, Returns, E-Commerce

Identifying non-intuitive Relationships within Returns Data of a Furniture Online-Shop using Temporal Data Mining

Katherina Meißner[0000-0002-1538-4129], Anthony Boyd Stevenson[0000-0002-1402-8560] and Julia Rieck[0000-0002-6569-0239]

{meissner, stevenson, rieck}@bwl.uni-hildesheim.de

Simplified Title

Automatically Finding Interesting or Unknown Conditions in Returns Data of a Furniture Online-Shop.

Abstract

Along with the growth in online retailing, there is a simultaneous increase in the number of returned products that have to be processed. These returns depend on various and often changing conditions, such as the product presentation in an online-shop, the quality of the product or of the logistics partner. Measures for improving the returns process and reducing the corresponding costs can thus not only be implemented on the basis of a static analysis. Rather, the temporal context of returns characteristics should be studied for latent relationships that may have been previously neglected for managing the returns process. For providing insights into such non-intuitive relationships, we propose a data mining framework combining frequent itemset mining, time series clustering, forecasting, and scoring in order to detect returns characteristics with an interesting temporal behavior.

I Introduction

Nowadays, a lot of people buy their goods online. Online bought goods are mainly purchased because of the way they are presented. These goods cannot be physically examined before purchase, must be processed in a warehouse, and must be delivered by a shipping provider. Something going wrong in any of these process steps can lead to the customer returning the purchased goods to the retailer. This is frustrating for both the retailer and the customer. The retailer has costs for shipping and taking back the product, while the customer is angry that the order was unsatisfactory. To reduce the amount of financial damage a return can cause, the retailer is motivated to process the returns in the best way possible. Information on the order process, the customer, the shopping cart, and the product is recorded for each return. This information is divided into different features. If specific features frequently come together, that is in the same return, this may be used as a basis for improving returns management. Changes in how often they are co-occurring are especially interesting. For this reason, this study looks at how feature combinations that are either interesting or unknown can be automatically found.

II State of the Art

If goods are ordered online, a paper slip, a so-called return bill, is usually put in the package. With this paper slip, a return can be initialized and documented for processing purposes by manually being filled out by the customer. These packages and the paper slip must also be checked manually and processed by the retailer. After processing, the retailer knows which ordered items were returned and why. The different reasons can be collected and analyzed. As a result, it is possible to plan and take actions to make the returns process more efficient. In order to better understand the customers' causes of returns and to improve the processing of returns in the warehouse, it is a good idea to automatically find out conditions for returns that are not yet known and non-intuitive for the retailer. It is especially important in this context to examine the changes in these conditions over time. Only then can the retailer react quickly to changing conditions in the returns process, for example to the growing number of orders and returns during the corona pandemic.

III Original Contribution

In order to help retailers to quickly find unknown and changing conditions in the returns process, we have built a data mining framework consisting of different steps. This framework takes not only the information directly related to returns processing into account, such as the manually completed returns bill and the condition of the returned goods. It also looks at other metadata associated to the order and returns process, such as order and shipping times, the customer's shopping cart, and detailed product information. All this data is combined and looked at over time to improve returns processing in the warehouse, for example, by adjusting employee schedules. The newly gained knowledge can be used to make profitable decisions for improvement.

IV Methodology

In this study, we apply the data mining framework presented by [1] to the returns data of an online retailer in a case study. We want to find unknown conditions for returns within the data automatically. To do so, the framework begins with frequent itemset mining. Here, we count the number of returns for each month with a certain condition, called a frequent itemset (step 1). These results in a time series showing the frequency of the given returns conditions. Note that we have not only one condition, but many since we can combine all information collected for a return. Therefore, many time series are obtained that must be analyzed in order to find interesting and unknown conditions. For this analysis, we first use time series clustering (step 2) to groups of similar time series. For each group of time series, we then find a suitable forecasting method (step 3). By forecasting the time series, we can see how many returns with the condition of a certain return we will have in the future. For planning actions to improve the returns process, it is not enough to just look into the past and present but also to take future directions into account. The forecast is therefore used in a scoring procedure (step 4) that automatically finds interesting and unknown returns conditions. The parameters of the scoring procedure may be changed to the need of the returns manager. The time series with the highest score values are then displayed in a dashboard (step 5) and analyzed by returns managers, so that actions can be planned and taken.

V Results

Based on the returns data of our cooperation partner, a retailer selling furniture online, we can find good parameters for all methods used in the framework steps. With about 300,000 returns from 2014 to 2020, we have about 5,500 different return conditions after applying frequent itemset mining. For each, we build a time series showing the number of corresponding returns. By using 5 clusters, or time series groups, we receive a good representation of the given data structure. One forecasting method is chosen for every time series group, which works well for all the time series within the group because the time series within one cluster are all similar. Based on the case study's findings, we adjusted the scoring procedure and the options for the returns managers to adapt the procedure to their needs. This way only the most relevant aspects can extracted and presented to the decision-makers.

VI Evaluation

The results of the data mining framework were presented to the cooperation partner, a furniture company in the e-commerce sector. To evaluate the findings from the framework, we used returns data from previous years. Our assumption was that the framework would find unknown return conditions in the data, as it had already done in various other data sets. The case study revealed that the framework could present new and unknown return conditions to the cooperation partner. For example, the framework found that the number of returns is higher when the number of days between ordering and shipping a product is large. This is because the customers buy the product elsewhere if the shipping takes too long. In this case, they then send the product back. The cooperation partner used this new knowledge to take measures to speed up the shipping process.

VII Conclusions

We presented and applied a multi-step framework for analyzing returns data. The results of the framework showed that interesting and unknown returns conditions can be identified. These can then be used as indicators for possible improvements in returns management within a furniture company. Furthermore, we were able to quantify the monetary impact of the findings together with the company. We also learned that our framework can be deployed to optimize preventive returns management. For example, if it is found that certain commodity groups seem to be more return prone in recent months, customers can be offered different prevention measures, such as higher discounts. The framework should then be adapted in the first step and use association rules instead of frequent itemsets in order to obtain connections that can easily be transferred into business decisions. Returns data from retailing sectors other than furniture can also be analyzed with the framework at hand if the attributes to be included are adapted to the retailer's needs. Moreover, the framework is not bound to returns data. It could also be used, for example, for analyzing sales data.

References

[1] Meissner, K., and Rieck, J. Strategic planning support for road safety measures based on accident data mining. *IATSS Research 46*, 3 (2022), 427–440.

DOI: `https://doi.org/10.1007/978-3-031-21967-2_51`

Keywords: IoT, Cyberattacks

Vulnerability Analysis of IoT Devices to Cyberattacks Based on Naïve Bayes Classifier

Jolanta Mizera-Pietraszko[0000-0002-2298-5037], Jolanta Tańcula [0000-0001-9416-2171]

jolanta.mizera-pietraszko@awl.edu,pl,jtancula@uni.opole.pl

SIMPLIFIED TITLE

Vulnerability analysis of IT system devices based on statistical methods

ABSTRACT

IoT or Smart World, as a global technology, is a rapidly growing concept of ICT systems interoperability covering many areas of life. Increasing the speed of data transmission, increasing the number of devices per square meter, reducing delays - all this is guaranteed by modern technologies in combination with the 5G standard. However, the key role is played by the aspect of protection and security of network infrastructure and the network itself. No matter what functions are to be performed by IoT, all devices included in such a system are connected by networks. IoT does not create a uniform environment, hence its vulnerability in the context of cybersecurity. This paper deals with the selection of a method aimed to classify software vulnerabilities to cyber-attacks and threats in the network. The classifier will be created based on the Naive Bayes method. However, the quality analysis of the classifier, i.e., checking whether it classifies vulnerabilities correctly, was performed by plotting the ROC curve and analyzing the Area Under the Curve (AUC).

I INTRODUCTION

The specific nature of IoT technology shows many vulnerabilities to cyberattacks and security gaps, hence nowadays IoT security is a priority issue. This paper deals with the selection of a method to classify software vulnerabilities to cyber-attacks and threats in the network. The specific nature of IoT technology shows many vulnerabilities and security gaps, hence nowadays IoT security is a priority issue. The threat posed by the very structure of IoT is primarily its distributed organization. Due to the distributed management system, ensuring the substantial security level and monitoring the devices poses a challenge. The classifier will be created based on the Naive Bayes method. However, the quality analysis of the classifier, i.e., checking whether it classifies vulnerabilities correctly, was performed by plotting the ROC curve and analyzing the Area Under the Curve (AUC)

II STATE OF THE ART

There are many classification methods in the literature, they include: classification by induction of decision trees, Bayesian classifiers, neural networks, statistical analysis, metaheuristics, rough sets and many other statistical methods. The concept of ROC curves has long been known and has been used to detect signals. Currently, they are used mainly in econometrics, statistics and medicine. In this paper we chose the Bayesian classifier because it is a simple probabilistic classifier and it is not complicated to implement.

III ORIGINAL CONTRIBUTION

In this paper, we deal with the problem assigning the vulnerability of a network to one of two groups: existence of vulnerability or not. The classifier of vulnerability will be created based on the Naive Bayes method and the quality analysis of the classifier was performed by plotting the ROC curve and analyzing the Area Under the Curve. The attributes in the database represent different types of vulnerabilities in the table. After determining the attributes, we extract the training set. The data have been encoded and the attributes have been replaced with parameters. The classifier defined was used to analyze susceptibility whether it is high (H) or low (L), in the proposed model. In the last column represents the response of the system, where all cases classified as positive are marked, i.e. there is a high level of vulnerability of the system to the network attacks. In order to evaluate quality of the classifier, a number of positively and negatively identified scenarios was analyzed, additionally we analyzed the cases that were determined to be positive, but yielded a negative result and as well as those identified as negative but which turned out to be positive. All of the scenarios' average values were calculated and presented in confusion matrix. At this stage, we obtain the following cut-off values and we plot ROC curves that is the relationship between SE and 1-SP. The resulting graph and the area under the curve will allow to determine if the Bayesian classifier is correctly selected.

IV METHODOLOGY

Our study is experimental. An IoT network model was proposed. Then, we performed the network scanning. The network scans have been labeled with the corresponding ID number in database. The last column that is system vulnerability is the response of the system to the database queries about the occurrence of attributes in the system. After the computation and construction confusion matrix has been done, the results were collected. In the following part of the simulation, the execution of the algorithm was described by individual steps. ROC plot was constructed, AUC field analysis was described and final classifier quality was analyzed.

V RESULTS

Naïve Bayes classifier has been constructed correctly for the vulnerability study. After plotting the ROC curve, it took a form of concave. It means, that the classifier works well because it correctly classifies the selected parameters. The area under the AUC curve is much larger than 0.5 which is interpreted that the diagnostic power of the test is high.

VI EVALUATION

Our fundamental assumptions of the method chosen for evaluating our research was to find a way to determine whether the vulnerability of the IoT system is threatening its functioning or not. However, the results obtained are not perfect. This is indicated by the final plot of the ROC curve. The curve is not smooth enough, so more parameters or more detailed would presumably better describe the characteristics of the system vulnerability. Perhaps scanning a larger area or more complex networks in real time would produce better results.

VII CONCLUSIONS

The article shows how to select a system classifier, so that the vulnerability level allows to take appropriate steps to make the system fully secure. This paper is just about exploring IoT system vulnerabilities to improve its security. The use of known classification methods, including Bayes algorithm, as well as the use of ROC curves for vulnerability analysis, allow to determine whether the level of security of such a network is sufficient or not. We show how to select a system classifier based on CVSS 3.1, so that the vulnerability level allows to take appropriate steps to make the system fully secure. The analysis of ROC curves and AUC field allows to determine whether the selected method gives good results. According to the selected parameters, the selection of the training set, the classification of the system responses and the plotting of the curve showing good classification quality, this task has been accomplished. The authors hope that the method described in this paper will allow to determine the level of susceptibility of the system both in a small area and in a slightly larger one, like a region.

REFERENCES

1. H. Mahmoud, M. Thabet, M. H. Khafagy and F. A. Omara, Multiobjective Task Scheduling in Cloud Environment Using Decision Tree Algorithm, *IEEE Access*, vol. 10, pp. 36140-36151, (2022).
2. S. Alex, K. J. Dhanaraj and P. P. Deepthi, Private and Energy-Efficient Decision Tree-Based Disease Detection for Resource-Constrained Medical Users in Mobile Healthcare Network, in *IEEE Access*, vol. 10, pp. 17098-17112, (2022).
3. A. G. C. Menezes, M. M. Araujo, O. M. Almeida, F. R. Barbosa and A. P. S. Braga, Induction of Decision Trees to Diagnose Incipient Faults in Power Transformers, *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 29, no. 1, pp. 279-286, Feb. (2022).

Link to arXiv: https://arxiv.org/abs/2207.12086
DOI: https://doi.org/10.1007/978-3-031-21743-2_3
Link to repository: https://github.com/nphdang/CCRAL

Keywords: Data Augmentation, Classification, Counterfactual Reasoning, Active Learning, Tabular Data

Efficient Classification with Counterfactual Reasoning and Active Learning

Azhar Mohammed[0000-0002-1363-0576], Dang Nguyen[0000-0002-0401-988X], Bao Duong[0000-0001-9850-0270], Thin Nguyen[0000-0003-3467-8963]

mohammedaz@deakin.edu.au, d.nguyen@deakin.edu.au, duongng@deakin.edu.au, thin.nguyeng@deakin.edu.au

Simplified Title

Improving the classification performance of a machine learning model on the tabular data with Counterfactual reasoning and Active learning.

Abstract

Data augmentation is one of the most successful techniques to improve the classification accuracy of machine learning models in computer vision. However, applying data augmentation to tabular data is a challenging problem since it is hard to generate synthetic samples with labels. In this paper, we propose an efficient classifier with a novel data augmentation technique for tabular data. Our method called CCRAL combines causal reasoning to learn counterfactual samples for the original training samples and active learning to select useful counterfactual samples based on a region of uncertainty. By doing this, our method can maximize our model's generalization on the unseen testing data. We validate our method analytically and compare it with the standard baselines. Our experimental results highlight that CCRAL achieves significantly better performance than those of the baselines across several real-world tabular datasets in terms of accuracy and AUC.

I Introduction

In spite of great success in computer vision, applying data augmentation to tabular data is challenging. There are three main reasons. First, an image is typically invariant to a minor modification, e.g., flip, zoom, or rotation. In contrast, a slight change in a record in tabular data can result in a totally different outcome. All features (i.e., pixels) in images are i.i.d (independent and identical distributed), whereas each feature in tabular data (e.g., Sex or Age) has different ranges of values. Finally, one transformation operator can be applied to all features in images, whereas each feature in tabular data often requires a relevant transformation operator depending on the feature type (continuous, discrete, or categorical). We propose an efficient classification method with a new data augmentation technique for tabular data. Our approach has two main steps. First, we use causal reasoning to learn counterfactual samples for the original training samples [2]. Each counterfactual sample is a variant of an original sample whose all feature values are the same except the intervened feature. Since the counterfactual samples may have different outcomes from the original ones, we obtain their labels via a matching method. Second, we augment counterfactual samples with real samples to create a new training set for the classifier. Since not all counterfactual samples are useful, we select the meaningful ones that potentially improve the classification performance using an active learning-based method [3]. Our active learning is an uncertainty-based approach. It determines difficult to predict samples, then obtains their counterfactual version to enrich the training data. Using both real and counterfactual samples, our classifier improves its generalization, resulting in better accuracy on unseen testing samples.

II State of the Art

As our method augment the new data to existing data, we compare our method (CCRAL) with two strong baselines: (1) Standard: this method uses available training samples to train a classifier, and (2) Counterfactual: this method uses the counterfactual samples of all original training samples in the training process. For a fair comparison, we measure the accuracy and AUC of each method on the same hold-out test set. We also use the same classifier for all methods, namely the Support Vector Machine (SVM) with the linear kernel and C = 1 for the regularization. Note that other machine learning classifiers can be used with our method. We evaluate methods on each dataset for five times with different train-test data splits and report the averaged accuracy and AUC.

III Original Contribution

We demonstrate the efficacy of CCRAL on five standard real-world tabular datasets. The obtained results show that CCRAL generalizes better and is more robust towards unseen testing samples, where it significantly outperforms other methods. Our novel approach combines counterfactual reasoning with active learning for selecting counterfactual samples. Since the region margin α has values being in the range of [0, 0.5], we use a grid search (or Bayesian optimization [1]) to find the best α that derives the best classifier measured on a validation set.

IV Methodology

In this experimental study, our main aim is to improve classification accuracy. So, instead of using only given training samples in the given data (i.e., training), we try to obtain more training samples, which is very helpful in improving the classifier's generalization. When the classifier observes more training samples, it is more robust and its classification accuracy is often improved on unseen test samples. This process is often called data augmentation, which has become the state-of-the-art method to improve the performance of deep learning models in computer vision. Our approach, called Classifier with Causal Reasoning and Active Learning (CCRAL), has two main steps: (1) learning counterfactual samples using causal reasoning and (2) training a classifier with both real and counterfactual samples using active learning.

V Results

We conduct extensive experiments on five real-world tabular datasets to evaluate the classification performance (accuracy and AUC) of our method CCRAL, comparing it with two strong baselines as mentioned above. Our method outperforms the two strong baselines on five datasets in terms of accuracy and AUC. The improvement in accuracy and AUC of our method compared to the standard model is significant, i.e., nearly 12.15%* increase in accuracy along with 7.91%* increase in AUC performance. On the other hand, CCRAL outperforms the second baseline Counterfactual model with 2.6%* more accuracy and 7.32%* more AUC performance over five datasets.

*The percentage calculated is averaged over the five datasets used in the experiments.

VI Evaluation

For the evaluation of this experimental study we tested our method on five real-world datasets against two strong baselines. As our aim is to improve classification performance on the tabular data, we use accuracy and AUC performance as metrics for evaluation. As the treatment feature needed to be binary for obtaining counterfactual samples, we select the datasets that has at least one binary feature, which is considered as treatment feature. From the results mentioned above, we can conclude that, with the augmentation of selective counterfactual samples using active learning, the standard machine learning model generalizes better over unseen test data with improved accuracy and AUC performance.

VII Conclusions

In this paper, we have introduced an efficient classifier (named CCRAL) with a novel data augmentation technique for tabular datasets. We generate counterfactual data by flipping the binary value of the treatment feature of original training samples and obtaining their labels using a matching method. We use active learning to select useful counterfactual samples based on a region of uncertainty depending on the predicted scores of the original training samples. This method can be widely used over all kinds of tabular datasets where the given treatment feature is binary, especially when there are limited training samples.

References

[1] Nguyen, D., Gupta, S., Rana, S., Shilton, A., and Venkatesh, S. Bayesian optimization for categorical and category-specific continuous inputs. In *AAAI* (2020), vol. 34, pp. 5256–5263.

[2] Pearl, J. *Causality: models, reasoning, and inference.* Cambridge University Press, 2009.

[3] Settles, B. *Active Learning.* Morgan & Claypool Publishers, 2012.

DOI: https://doi.org/10.1007/978-981-19-8234-7_26

Keywords: knowledge dissemination, fuzzy aggregation norms

Analysis of Dynamics of Emergence and Decline of Scientific Ideas Based on Optimistic and Pessimistic Fuzzy Aggregation Norms

Aleksandra Mrela[0000-0002-2059-864X], Oleksandr Sokolov[0000-0002-6531-2203], Veslava Osinska[0000-0002-1306-7832], Wlodzislaw Duch[0000-0001-7882-4729]

a.mrela@ukw.edu.pl, osokolov@fizyka.umk.pl, wieo@umk.pl, duch@fizyka.umk.pl

Simplified Title

Analysis of Dynamics of Emergence and Decline of Scientific Ideas

Abstract

Scientists develop new concepts, methods, and techniques to solve practical and theoretical problems. These ideas are disseminated among scientists representing the same discipline; some are known even among non-scientists (artificial intelligence, machine learning, or fractals). The spreading of new ideas in the scientific community is going with various intensities; some quickly emerge and die, and others exist for a long time. Scientists, like other people, would like to disseminate their concepts. They use different methods to spread them, such as publishing papers, writing books, giving lectures, taking part in conferences, being a member of interdisciplinary teams, and so on. The paper presents the two different methods of analyzing the knowledge dissemination of scientific ideas based on fuzzy logic. One of these methods involves fuzzy aggregation norms and calculating and analyzing knowledge dissemination coefficients. The other considers the article's citations and extrapolates results to citations of papers consisting of the scientific field. Moreover, the multi-agent model of emergence and decline of scientific ideas among the scientific community is coded in Netlogo and presented to show the results of the multi-agent model of spreading the scientific concept.

I Introduction

Nowadays, there are available enormous bibliometric databases collecting information about the achievements of scientists, so simulations showing the development of disciplines or research areas based on different types of data, such as numbers of publications, citations, and categories of publications according to disciplines, are helpful. The main goal of these simulations is to search for groups of researchers considering their scientific interests and achievements.

II State of the Art

Modeling scientific knowledge space has many benefits for researchers, science managers, and philosophers; such visualizations and simulations can help scientists search for researchers whose interests are similar or select the set of papers to read because they are similar or not in the proximity space of the previously studied one. Hence, scientists look for appropriate applications, including visualization methods, to analyze how some disciplines and researchers' achievements develop depending on input data. The spread and dissemination of scientific ideas can be simulated and visualized by completing data vectors.

III Original Contribution

We propose the new knowledge dissemination and forgetting coefficient based on fuzzy logic and analyze its values to visualize some papers' citations' impact on two research fields: fuzzy logic and Gestalt psychology. We use the simulation of the multi-agented model of knowledge and confidence dissemination. This model simplifies situations of receiving and spreading or forgetting some scientific ideas; however, changing parameters of knowledge dissemination and forgetting can observe scientific concepts' behaviors. Models based on fuzzy logic give many opportunities to simulate bibliometric measures.

IV Methodology

Since many articles can be related to different disciplines, fuzzy logic is a better choice than classical one. The optimistic and pessimistic fuzzy aggregation norms are applied to aggregate bibliographic data to calculate the knowledge dissemination coefficients. The spread of the presented ideas is presented by choosing the highly cited articles based on the number of papers that cite them. Collecting data from some research areas and calculating these coefficients are the foundation of the simulations and visualization of ideas dissemination.

The study is mainly theoretical. However, computer simulation was developed.

V Results

The paper presents two methods of analyzing knowledge dissemination among the scientific community. One way uses the fuzzy logic based on scientometric data downloaded from the Web of Science, which enables us to look through the number of articles, and their importance measured by the number of citations per year. The time range of the knowledge dissemination coefficient shows how vital the specified article might be for its concept of knowledge dissemination.

The second method involves extrapolating empirical data to the normal distribution and comparing their standard deviations. We observe that the citation numbers follow the probability density function of the normal distribution for selected papers. Moreover, we state that the results observed for articles can be extrapolated on the behavior of concepts, as scientific concepts and publications are developed for some time and based on citations.

It can be expected that for every research field, there are articles that play the role of seeds of growth in this area. Two influential articles were chosen, and their impact on their discipline could be observed by analyzing the citations (and knowledge dissemination coefficients). Thus, the impact of the predominant authors of the development of the considered field, like, for example, L. Zadeh, can be analyzed by considering the number of papers and the dynamics of their papers' citations.

VI Evaluation

The presented method shows that it is possible to identify the most influential articles in the considered research area. Moreover, a dynamics model of knowledge dissemination of ideas in two research areas was prepared, showing that scientific theories emerge and decline. The developed model is also helpful in simulating future dynamical changes using citation patterns. Experts in the bibliometric sciences estimated the usefulness of these methods and the simulation results.

VII Conclusions

The Diss Fuzzy application, was developed based on this methods. The simulation was carried out using a computer program written by one of the authors in a multi-agent programmable modeling environment, NetLogo. Based on this simulation, it can be observed how the individual highly cited article influences scientists' knowledge dissemination. Moreover, aggregated values of knowledge dissemination coefficients can show the growth and decline of scientific ideas that might be noticed in the history of science. Moreover, it is interesting how the levels of trust in the scientific concept are spread out or forgotten.

In the future, the dynamic behavior of citations of different articles to predict the citing range using this method can be observed. Therefore, using our method, it is possible to predict new pivotal papers or scientific schools, which will be fundamental in field development.

References

[1] Darko, A., Chan, A.P.C., Adabre M.A., Edwards, D.J., Hosseini, M.R., Ameyaw, E.E.: Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities, Automation in Construction, vol. 112, (2020). doi: 10.1016/j.autcon.2020.103081

[2] Sokolov, O., Osinska, W., Mrela, A., Duch, W.: Modeling of Scientific Publications Disciplinary Collocation Based on Optimistic Fuzzy Aggregation Norms, Advances in Intelligent Systems and Computing, Vol. 853 Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology - ISAT 2019 Part II, ISBN 978-3-319-99995-1, 145–156, (2019).doi: 10.1007/978-3-319-99996-8.

[3] Sokolov, O., Osinska, V., Mrela, A., Duch, W., Burak, M.: Scientists' Contribution to Science and Methods of Its Visualization, Proceedings of 40th Anniversary International Conference on Information Systems Architecture and Technology - ISAT 2019, Swiatek, J., Borzemski, L., Wilimowska, Z. (ed.), Advances in Intelligent Systems and Computing, Part II, pp. 159–168, (2020). doi: 10.1007/978-3-030-30604-5-14.

[4] Zadeh, L.A.: Fuzzy sets, Information and Control, 8, 338—353, (1965). doi: 10.1016/s0019-9958(65)90241-x

DOI: https://doi.org/10.1007/978-3-031-21743-2_24

Keywords: Deep learning; Auto-encoders; Anomalies detection; Unsupervised learning

Using deep learning to detect anomalies in traffic flow

Manuel Méndez, Alfredo Ibias, Manuel Núñez [0000-0001-9808-6401]

manumend@ucm.es, a.ibias@sanoscience.org, manuelnu@ucm.es

SIMPLIFIED TITLE

Using deep learning to detect anomalies in traffic flow

ABSTRACT

Uncertainty is an ever present challenge in data analysis. In particular, it is important to detect, as precisely as possible, unforeseen phenomena. In this paper we study the usefulness of two deep learning based methods (CNN auto-encoder and BiLSTM auto-encoder) to detect anomalies in situations that can be defined in terms of time series. In order to evaluate our approaches, we consider traffic flow data and perform experiments in two orthogonal scenarios: a guided scenario (training only with data considered as 'normal' after a naive labelling) and a basic scenario. Our results show that if we train the models using only the considered 'normal' data, the obtained models do not achieve good results because none of them are able to detect all type of abnormal data correctly. In contrast, both models can detect all type of time series anomalies when we consider the basic scenario.

I INTRODUCTION

Anomaly detection [2] is a process that consists in detecting different abnormal behaviours, called *outliers*, in a dataset and provide useful information about their causes.

The anomaly detection in traffic flow presents a main issue that does not appear in other cases: the lack of knowledge about what an anomaly is. Therefore, this technique requires the application of unsupervised models. At a first glance, we might be tempted to think that anomalies are the same as extreme values. Nevertheless, ideally, a well-performed model should be able to detect not just the extreme values but the values that imply, for example, the failure of a given pattern or a sudden increase or decrease from the previous value. Summarising, the definition of either a "normal" or an "abnormal" data should be a time-dependent quantity.

II STATE OF THE ART

Several deep learning methods have been proposed to identify abnormal data in time series. Recurrent neural networks (RNN), mainly long short-term memory (LSTM) networks based models, are the primary option. RNN are applied in two different scopes. On the one hand, RNN aims to predict the future values and, subsequently, compare them with the actual values or with predefined thresholds by determining whether the corresponding data is or not an abnormal value. On the other hand, auto-encoders (AE) or variational auto-encoders (VAE) based on RNNs (specifically, on LSTMs) are developed. Their aim is to restore the original values and compare the actual and the reconstructed values. Then, a threshold is established. If the discrepancy between the actual and the reconstructed value is greater than the threshold, then the corresponding data will be considered as abnormal.

Anomaly detection in traffic flow by employing auto-encoders is an unexplored research line. There are some recent approaches dealing with this task by using other techniques. A method known as *local outlier factor* was employed in the city of Odense. By leveraging the availability of video surveillance, an algorithm based on fuzzy theory was developed to detect anomalies in road traffic flow. An original and novel method, which does not take into account the historical data but uses expert feedback to deal with the fluid definition of anomaly, was developed in Vienna and was able to provide high accuracy. To detect anomalies in Zagreb traffic flow, a tensor based method was proposed. An adapted k-nearest neighbours method was used with the same purpose in Beijing.

III ORIGINAL CONTRIBUTION

In this work, we present two models to detect anomalies in time series data by using previous data as input. We apply both models in two different scenarios.

IV Methodology

An auto-encoder is an algorithm composed by an encoder and a decoder. The idea is that the encoder will translate the input, X, from its high-dimensional space to a lower-dimensional space. Then, the decoder will take that lower-dimensional representation of the input, what we usually called the *latent vector*, and will try to reconstruct the original input in the high-dimensional space, producing $\hat{X}$. Due to its nature, the decoder usually has a structure that mimics the one of the encoder, being its goal to obtain a value $\hat{X}$ as similar as possible to the original input, X. The dissimilarity between the original and reconstructed data is defined by a loss function. Therefore, the goal of the model is to reduce, as much as possible, this value during the training process.

We apply auto-encoder models to traffic flow data obtained in the city of Madrid. First, we develop a CNN [3] auto-encoder. In addition to the dependent variable, this model will use another eight independent variables related to traffic flow, that is, we will be working within a multivariate time series. The second model is a novel Bidirectional LSTM [1] auto-encoder model. This model will only use as input the traffic flow data (dependent variable), that is, we will work within a univariate time series. These models will be applied in both a basic scenario and a guided scenario (training only with data considered as 'normal' after a naive labelling).

V Results

Our results show that if we train the models using only the considered 'normal' data, the obtained models do not achieve good results in guided scenario because none of them are able to detect all type of abnormal data correctly. In contrast, both models can detect all type of time series anomalies when we consider the basic scenario.

VI Evaluation

Both models have been applied in a guided and a basic scenarios with the goal of detecting abnormal data. In the guided scenario, we break up values higher than an established quantile into an abnormal set.

VII Conclusions

Results indicate that the CNN auto-encoder, although can detect anomalies of the three types, cannot detect a significant number of them while BiLSTM auto-encoder can only detect anomalies of one type (global anomalies). Therefore, we conclude that the usefulness of these models in this scenario is limited. However, in the basic scenario, we appreciate that both models can virtually detect all anomalies of the three types (global anomalies, contextual anomalies and collective anomalies). These anomalies can be appreciated both graphically and by date.

References

[1] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural Computation 9*, 8 (1997), 1735–1780.

[2] Kotu, V., and Deshpande, B. Anomaly detection. In *Data Science*, second ed. Morgan Kaufmann, 2019, ch. 13, pp. 447–465.

[3] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Computation 1*, 4 (1989), 541–551.

Link to arXiv: https://arXiv.org/abs/2104.12473
DOI: https://doi.org/10.1007/978-3-031-21743-2_14

Keywords: Air quality prediction, Deep Learning, Transformer Networks.

Using Deep Transformer based Models to predict ozone levels

Manuel Méndez, Carlos Montero, Manuel Núñez [0000-0001-9808-6401]

manumend@ucm.es,cmonte09@ucm.es,manuelnu@ucm.es

Simplified Title

Using Deep Transformer based Models to predict ozone levels

Abstract

Ozone (*O*3) is an air pollutant that has harmful effects in human health when its concentration exceeds a certain level. Therefore, it is important to advance in methods that can appropriately predict *O*3 levels. In this paper we present a new model to estimate 4h, 12h, 24h, 48h and 72h ahead *O*3 concentration levels. We rely on Deep Transformer Networks. Interestingly enough, these models were originally developed to be used in Natural Language Processing applications but we show that they can be successfully used in classification problems. In order to evaluate the usefulness of our model, we applied it to predict *O*3 levels in the centre of Madrid. We compare the results of our model with four baseline models: two LSTMs and two MLPs. Accuracy (*Acc*) and Balanced Accuracy (*BAC*) are the metrics employed to evaluate the goodness of all the models. The results clearly show that our Deep Transformer based Network obtains the best results.

I Introduction

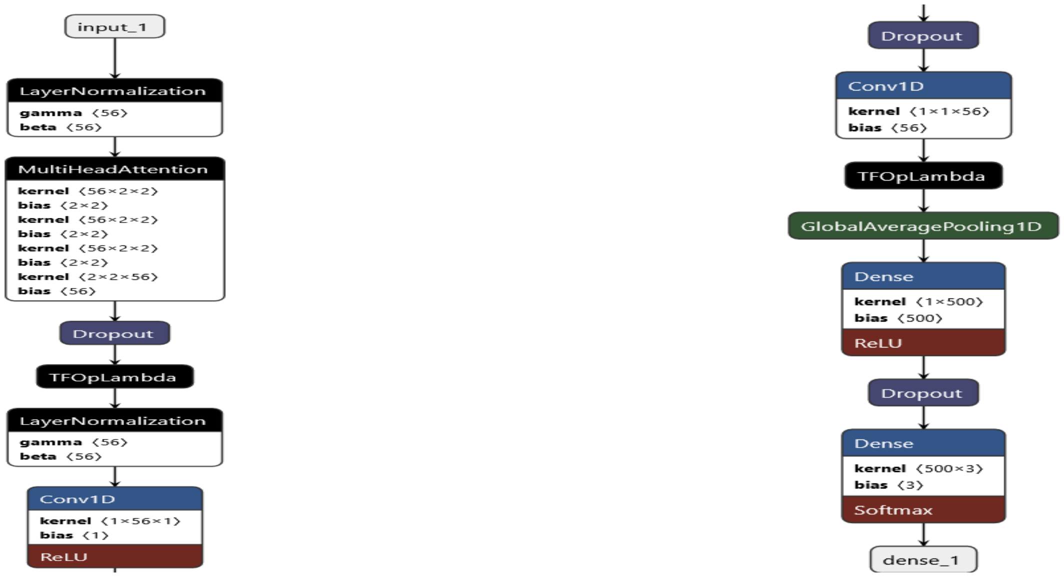
Air pollution is one of the major problems currently faced by humanity. Pollutants like ozone (*O*3), nitrogen dioxide (*NO*2), sulphur dioxide (*SO*2), carbon monoxide (*CO*) and particulates matter (*PM*2.5, *PM*10) are some of the most common air pollutants. They are also the pollutant included in the air quality index measurement. In this paper we focus on *O*3 (ozone). Ozone is a colourless gas located in the atmosphere. It is one of the most common existing pollutants and, as such, it is one of the pollutants taken into account to determine the air quality index. The exposure to this pollutant has severe effects in human health such as eyes and nose irritation and inflammation, lung function reduction, exacerbation of respiratory diseases and increased susceptibility to diseases infection, among others. Ozone concentration levels depend on complex processes that happen in the atmosphere. Precursor gases, such as nitrogen oxides, are chemically transformed into ozone when they are exposed to solar light. Since sunlight is needed for ozone formation, its concentration highly depends on the time of the day and on the current meteorological conditions. Moreover, human emission of these precursor pollutants by fabrics and traffic have affected the increase of ozone concentration in the last decades. Most major cities around the world have specific protocols to deal with high concentrations of pollutants. In the case of Madrid, used in this paper as case study, when the *O*3 concentration exceeds a certain level, the authorities take measures to reduce the *O*3 effects in population. These measures range from recommendations to reduce physical exercise outdoors for vulnerable people to prohibitions of outdoor activities, specially sport activities. Therefore, it is very important to be able to estimate future *O*3 levels in order to alert the population of possible future recommendations or restrictions.

II State of the Art

Several studies have used either statistical machine learning techniques or deep learning models to forecast pollutant concentrations.

Our approach is based on Transformer Networks. They were developed as a Natural Language Processing tool that improves classical LSTM networks and Recurrent Neural Networks (RNN) in tasks such as text classification and translation. Its potential is based on self-attention layers, which estimate the attention weights between input variables. In recent years, Transformers have been used to solve tasks in other fields such as image recognition, multi-class classification and time series prediction.

To the best of our knowledge, Deep Transformer based Models have not been used to analyse air pollutants. Although machine learning techniques have been used to forecast ozone concentration values (that is, as part of a regression model), we are not aware of their use to predict ozone levels as such (that is, as part of a classification model).



(a) Part a.

(b) Part b.

Figure 1: Deep Transformer based Model structure

III Original Contribution

We present a new model to predict the $O3$ concentration level. Our approach applies a novel Transformers-based model to predict the $O3$ concentration level. Unlike classical time series forecasting models, Transformers do not process the data ordered. On the contrary, they process the entire sequence and use self-attention techniques to find dependencies between variables. We considered an optimised Transformers-based model and our preliminary experiments revealed that it was a good candidate to overcome the accuracy and the balanced accuracy of usual time series classification networks such as MLP or LSTM.

IV Methodology

We present a new model based on Deep Transformer Networks [1] to predict $O3$ levels concentrations. The main characteristics of our model can be found in Figure 1. We compare the results provided by our model with four neural network baselines models. We also make an analysis of the variation of the model accuracy and balanced accuracy depending on the modification of three hyperparameters in short-term and in long-term cases. For this, we use as case of study the task of predicting $O3$ levels in the centre of Madrid. We consider a total of fourteen predictors variables.

V Results

Our results show that our proposal is better than the baseline models based on the evaluation metrics. On average, the balanced accuracy of our proposal is 4.3% better than the one corresponding to the best baseline model.

VI Evaluation

We apply the proposed model and the baseline models to predict the categories in the next 4, 12, 24, 48 and 72 hours. Comparing the model with the previously defined MLP and LSTM networks, we appreciate that, in general, our developed model obtains better results, particularly in long-term cases.

VII Conclusions

Our experiments show that our model overcomes, in general terms, the accuracy and the balanced accuracy of four baselines neural network models. The capability of the proposed model to detect minority class observations, particularly, in 24 hours in advance case is one of the main advantages of proposed model respect to the baseline methods. The hyperparameters analysis suggests us that the behaviour of balanced accuracy depending on hyperparameters modifications differs between long term and short term cases.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *31st Conf. on Neural Information Processing Systems, NIPS'17* (2017), pp. 1–11.

DOI: https://doi.org/10.1007/978-981-19-8234-7_27

Keywords: Deep features, Support vector machines, Transfer learning, Memetic algorithm

Fusing Deep Learning with Support Vector Machines to Detect COVID-19 in X-Ray Images

Jakub Nalepa[0000-0002-4026-1569], Piotr Bosowski, Wojciech Dudzik[0000-0003-4080-3644], Michal Kawulok[0000-0002-3669-5110]

{jnalepa,michal.kawulok}@ieee.org

Simplified Title

Using support vector machines to classify the features extracted with a deep network for detecting COVID-19 from X-ray images

Abstract

Deep neural networks are powerful learning machines that have laid foundations for most of the recent advancements in data analysis. Their most important advantage lies in learning how to extract the features from raw data, and these deep features are later classified with fully-connected layers. Although there exist more effective classifiers, including support vector machines, their high computational complexity is a serious obstacle in using them for classifying highly-dimensional and often huge datasets of deep features. We introduce a new framework which allows us to classify the deep features with evolutionarily-optimized support vector machines and we apply it to a real-life problem of detecting COVID-19 from X-ray images. We demonstrate that the proposed approach is highly effective and it outperforms well-established transfer learning strategies, thus improving the potential of existing pre-trained deep models. It can be particularly beneficial in cases when the amount and quality of labeled data is insufficient for performing full training of a network, but still too large for training a regular support vector machine.

I Introduction

Deep learning is highly effective in data analysis and processing, and it has allowed for achieving important advancements in many fields. Deep neural networks learn at the same time how to extract the valuable features from the input data and how to classify these features. This is achieved by combining the convolutional layers that extract the features from input images with fully-connected layers that classify the extracted feature vectors. However, to train such neural networks, large amounts of data are needed. Otherwise, the network may learn how to classify the data presented during training, but it later fails when it receives the data it has not seen during training.

II State of the Art

There have been many attempts to deal with the problem of insufficient amounts of labeled data and their poor quality. A simple, but highly effective approach, termed *transfer learning* [2], is to train the network with some other data, before its top layers are trained using the data it is intended to classify. This reduces the required amounts of data from the considered domain—the network can learn how to extract the features from other large datasets. Later, it is trained to classify the features using small amounts of the data that belong to the considered domain.

The classifiers trained during transfer learning are the fully-connected layers. Although they are not considered the best classifiers, their main advantage is that they can be trained together with the convolutional layers that extract the deep features. On the other hand, support vector machines (SVMs) are known to be very effective in solving difficult data classification tasks. Their main disadvantage is that the time and memory complexity of SVM training is very high, which means that even in transfer learning, the training data may be too large to train an SVM. To deal with that problem, the methods for selecting the SVM training data were developed. They include evolutionary algorithms, in which a population of training sets is evolved to obtain the best solution. Our simultaneously-evolved SVM (SE-SVM) [1] optimizes several elements of SVMs, including the training set, features used for classification, and the method's settings.

III Original Contribution

In the reported research, we used our SE-SVM technique to select the training samples and features extracted using a deep network that was earlier trained with large amounts of data. This allows for using an SVM as a classifier in transfer learning. In this work, we demonstrate that the features extracted from X-ray images using deep networks can be classified with SE-SVM to diagnose COVID-19 disease. In this way, we combine the strengths of deep networks (concerned with feature extraction) with the high classification skills of SVMs. This is a general scheme that may potentially be applied to any real-life classification problem.

IV Methodology

An outline of the proposed technique is presented in Figure 1. In our study, any earlier-trained convolutional neural network (CNN) can be used for extracting the deep features. The most valuable features are selected using SE-SVM along with the training samples—this results in a final set of picked features together with the SVM model trained over these features. To classify any new data sample, the very same trained network is employed to extract the selected features which are later classified using the trained SVM model.

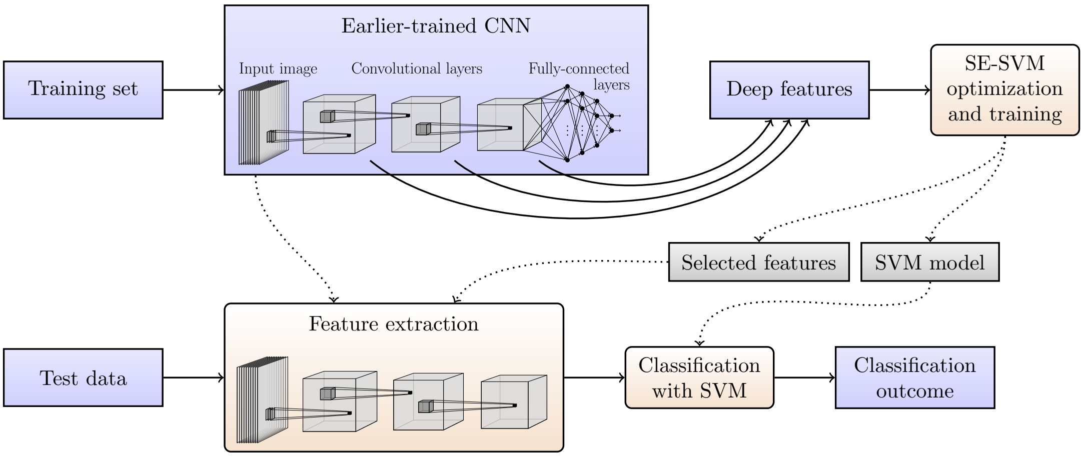


Figure 1: Outline of the proposed solution. Blue blocks indicate the data, and the light orange blocks show the actions. The outcome of SE-SVM optimization procedure is shown with gray blocks.

V Evaluation

In our experimental study, we used four different network architectures of a varied number of parameters, and in all cases, their convolutional layers were trained using the same large ImageNet dataset. We have compared two approaches to transfer learning—using the original fully-connected layers and using an SVM optimized with our SE-SVM technique. Also, we performed full training of the networks with the COVID-19 dataset, and we classified the test data using both the trained network and, after training, an SVM with the deep features extracted by the fully-trained network. The obtained results showed that for the transfer learning, SE-SVM outperforms the classifier based on fully-connected layers in all cases, and for the fully-trained networks, it offers an advantage for smaller models, while for bigger ones, the results are similar in both cases.

VI Conclusions

The proposed approach to classify the deep features with an SVM optimized using an evolutionary algorithm offers better results for transfer learning and makes us less dependent on the selected network architecture—even for smaller models, SE-SVM obtains high classification scores, close to those obtained with larger models. In our ongoing work, we are applying the introduced framework to other computer vision and pattern recognition tasks.

Acknowledgements

This work was supported by the National Science Centre under Grant DEC-2017/25/B/ST6/00474.

References

[1] Dudzik, W., Kawulok, M., and Nalepa, J. Optimizing training data and hyperparameters of support vector machines using a memetic algorithm. In *Proc. ICMMI* (2019), pp. 229–238.

[2] Huh, M., Agrawal, P., and Efros, A. A. What makes ImageNet good for transfer learning? *arXiv preprint arXiv:1608.08614* (2016).

Link to arXiv: https://hal.inria.fr/hal-03737859/
DOI: https://doi.org/10.1007/978-3-031-21743-2_25
Link to repository: https://github.com/NohaNekamiche/GANS_Workshop_SOAI_Algiers

Keywords: Image generation, Generative Adversarial Network, Deep Convolutional Generative Adversarial Network, Batch Normalization.

A Deep Convolution Generative Adversarial Network for the Production of Images of Human Faces

Noha Nekamiche, Chahnez Zakaria, Sarra Bouchareb, Kamel Smaïli

hn_nekamiche@esi.dz, c_zakaria@esi.dz, sa.bouchareb@univ-bouira.dz, smaili@loria.fr

Simplified Title

DCGAN for image generation

Abstract

Generative models get huge attention by researchers in different topics of artificial intelligence applications, especially generative adversarial networks (GANs) which have demonstrated good performance in data generation. In this paper, we would like to explore the potential of this class of models in producing human faces images. For that, we will use Deep Convolutional Generative Adversarial Network (DCGAN). Since that, the evaluation of GANs is still difficult even with the existing metrics like Inception Scores (IS), Mode Score (MS), Kernel Inception Distance (KID), Fréchet Inception Distance (FID), Multi-Scale Structural Similarity (MS-SSIM), etc. Thus, the best possible evaluation remains that carried out by human evaluators. This is why we propose a new hybrid measure combining qualitative and quantitative evaluation, we called this measure: Measuring the Quality of the Features of an Image (MEQFI). The images produced with the DCGAN method were trained on three well known datasets from the literature and the results were evaluated with MEQFI.

I Introduction

Generative adversarial networks (GANs) are a class of generative models proposed by Ian J. Goodfellow et al. [1], which have demonstrated good performance in data generation.

In our research work, we have explored the potential of this class of models. Thus we have focused on human face generation tasks using Deep Convolutional Generative Adversarial Networks (DCGAN) [2].

Since that, the evaluation of GANs is still difficult even with the existing metrics like Inception Scores (IS), Fréchet Inception Distance (FID), Multi-Scale Structural Similarity (MS-SSIM), etc. Thus, the right way to evaluate them is by human evaluators. This problem inspired us to introduce a hybrid subjective human evaluation metric that calculates the score of each generated image based on human feedback. We called this measure: Measuring the Quality of the Features of an Image (MEQFI).

II State of the Art

Generative adversarial networks (GANs) are widely used in data generation, they get huge success since their introduction by Ian J. Goodfellow.

They consist of two different neural networks a generator that generates data and discriminator that classifies the generated and real data.

In our work we used a deep convolution generative adversarial network to generate human faces. One of the hardest tasks how to find a good metric that allows to evaluate the images achieved by a process based on GANs algorithm.

The existing solutions categorized into two types of evaluation:

II.1 Qualitative evaluation

Qualitative evaluation, where we will evaluate the model based on the human subjective evaluation [3] [16] and according to the feedback, one can get an idea about the quality of the produced images. Which is done in several research areas in artificial intelligence.

II.2 Quantitative evaluation

Quantitative evaluation, where a specific numerical score is calculated indicating the quality of the produced images. In this topic, we identified 24 quantitative metrics to evaluate GANs

III ORIGINAL CONTRIBUTION

Inspired by the two categories of evaluation : Qualitative evaluation and Quantitative evaluation.

We have proposed a hybrid measure based on a qualitative evaluation for each present feature and which is summarized by a quantitative score. This score will be named Measuring the Quality of the Features of an Image (MEQFI). For that, we define a list of F features that will be used to calculate the score of each generated image. For each feature i, we attribute a mark based on human feedback to estimate its probability P_i. We multiply this probability by an activation coefficient. α_i. This score is given by:

$$Score(m_k) = \frac{1}{F-1} \sum_{i=1}^{F} \alpha_i * P_i \tag{1}$$

After calculating the score of each generated image, we calculate the average score over all the generated images.

IV METHODOLOGY

Our research is a case study in which we have used DCGAN model and trained it with three well-known datasets (CelebA, CelebA-HQ, LFW).

For the evaluation, we judge our results based on the BCE loss function and the quality of the generated images and we compare those results with the result of our hybrid subjective evaluation metric MEQFI.

V RESULTS

We tested our model DCGAN by using 12800 images from three widely used dataset of images: CelebA, CelebA-HQ (only female faces) and LFW and we have used several values of epochs: 100, 150 and 200.
According to the generator's and discrimibator's loss function, we find that the D loss decrease faster than the G loss, because the generation task is harder than the classification task. We also notice that our generator learns to fool the Discriminator by creating more or less real images. But as the training progresses the generation process improves and the synthesized images is more accurate. And finally, observing the results of our experiments, we find that it is preferable to use only 100 epochs, because increasing the number does not help to improve the training.

After having our model's results, we used our evaluation metric MEQFI to evaluate the quality of the generated images.

So, we fixed a group of features Nose, Eyes, Eyebrows, Mouth, Ears, Skin tone to calculate the score of each image by using MEQFI. For that, we ask a group of evaluators to attribute a mark from one to five for each feature of each image.

As a results, the CelebA gets a total score on the 64 generated images of 59.18%, the CelebA-HQ-female recieves a total score of 60.74% and the LFW gets a total score of 49.12%.

VI EVALUATION

The evaluation criteria of our model performance are based on how to reduce the noise in the training process of the generator and the discriminator, in other words how to reduce the loss function of both generator (G) and discriminator (D) and produce high quality generated images of a human face.

In our work, we evaluated the results of each dataset based on the variation of the loss function of both G and D and the score achieved by our hybrid human evaluation metric MEQFI.

VII CONCLUSIONS

We have developed a new measure to evaluate the quality of the produced images by our model DCGAN. This is the measure MEQFI that allows to take into account the evaluation of each annotator and combines them into a single score. The scores achieved by this measure are correlated to what we observed with the loss function during the learning step.

REFERENCES

[1] GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A., AND BENGIO, Y. Generative adversarial nets. *Advances in neural information processing systems 27* (2014).

[2] RADFORD, A., METZ, L., AND CHINTALA, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

Link to arXiv: https://arxiv.org/abs/2208.02432
DOI: https://doi.org/10.1007/978-981-19-8234-7_28
Link to repository: https://github.com/anhduy0911/ACIIDS2022-PIKA

Keywords: Pill Recognition, Knowledge Graph, Graph Embedding

Image-based Contextual Pill Recognition with Medical Knowledge Graph Assistance

Anh Duy Nguyen, Thuy Dung Nguyen, Huy Hieu Pham, Thanh Hung Nguyen, Phi Le Nguyen

{duy.na184249,dung.nt184244}@sis.hust.edu.vn,hieu.ph@vinuni.edu.vn,{hungnt,lenp}@soict.hust.edu.vn

Simplified Title

Graph assisted framework for contextual Pill Recognition

Abstract

In many healthcare applications, identifying pills given their captured images under various conditions and backgrounds has been becoming more and more essential. Several efforts have been devoted to utilizing the deep learning-based approach to tackle the pill recognition problem in the literature. However, due to the high similarity between pills' appearance, misrecognition often occurs, leaving pill recognition a challenge. To this end, in this paper, we introduce a novel approach named PIKA that leverages external knowledge to enhance pill recognition accuracy. Specifically, we address a practical scenario (which we call contextual pill recognition), aiming to identify pills in a picture of a patient's pill intake. Firstly, we propose a novel method for modeling the implicit association between pills in the presence of an external data source, in this case, prescriptions. Secondly, we present a walk-based graph embedding model that transforms from the graph space to vector space and extracts condensed relational features of the pills. Thirdly, a final framework is provided that leverages both image-based visual and graph-based relational features to accomplish the pill identification task. Within this framework, the visual representation of each pill is mapped to the graph embedding space, which is then used to execute *attention* over the graph representation, resulting in a semantically-rich context vector that aids in the final classification. To our knowledge, this is the first study to use external prescription data to establish associations between medicines and to classify them using this aiding information. The architecture of PIKA is lightweight and has the flexibility to incorporate into any recognition backbones. The experimental results show that by leveraging the external knowledge graph, PIKA can improve the recognition accuracy from 4.8% to 34.1% in terms of *F1*-score, compared to baselines.

I Introduction

Medicines are used to cure diseases and improve patients' health. Medication mistakes, however, may have serious consequences, including diminishing the efficacy of the treatment, causing adverse effects, or even leading to death. According to a WHO report, one-third of all mortality is caused by the misuse of drugs, not by disease. Moreover, according to Yaniv *et al.* [3], medication errors claim the lives of about six to eight thousand people every year. To emphasize the significance of taking medication correctly, WHO has chosen the subject Medication Without Harm for World Patient Safety Day 2022.

Medication errors may fall into many categories, one of which is incorrect pill intake, which occurs when the drugs taken differ from the prescription. This is due to the difficulty in manually distinguishing pills owing to the wide variety of drugs and similarities in pill colors and shapes. In such a context, we dedicates this work for tackling a practical application that recognizes pills in the patient's pill intake picture.

II State of the Art

Machine learning (ML) is now an effective method for object classification. Many studies have employed machine learning in the pill recognition challenge (eg. [2]). Some common techniques such as convolutional neural networks (CNN) and Graph Neural Networks (GNN) are often used. Furthermore, numerous efforts have strived to improve pill recognition accuracy by incorporating handcrafted features such as color, shape, and imprint. As an example, Ling et al. [1] investigated the problem of few-shot pill detection. The authors proposed a Multi-Stream (MS) deep learning model that combines information from four streams: RGB, Texture, Contour, and Imprinted Text. In addition, they offered a two-stage training technique to solve the data scarcity constraint; the first stage is to train with all samples, while the second concentrates only on the hard examples.

Despite numerous efforts, pill recognition remains problematic. Especially, pill misidentification often occurs with tablets that look substantially similar.

III Original Contribution

To summarize, our main contributions are as follows:

- We are the first to address a so-called *contextual pill recognition* problem, which recognizes pills in a picture of a patient's pill intake, by building a dataset containing pill images taken in unconstrained conditions and a corresponding prescription collection.
- We propose a novel method that leverages external knowledge to increase the accuracy and, in particular, to tackle the misclassification of similar pills. Moreover, there are new loss functions and a training strategy to enhance the classification accuracy.
- We conduct thorough experiments on a dataset of drugs taken in real-world settings and compare the performance of the proposed solution to existing methods. The experimental findings indicate that our proposed model outperforms significantly the baselines.

IV Methodology

Unlike the existing works, we focus on a practical application that recognizes pills in the patient's pill intake picture. We propose a novel deep learning-based approach to solve the contextual pill recognition problem by leveraging external prescription information. Our main idea is that by using such external knowledge, we can learn the relationship between the drugs, such as the co-occurrence likelihood of the pills. This knowledge will be utilized to improve the pill recognition model's accuracy.

Specifically, we design a method to construct a prescription-based knowledge graph representing the relationship between pills. We then present a graph embedding network to extract pills' relational features. Finally, we design a framework to fuse the graph-based relational information with the image-based visual features to make the final classification decision.

V Results

Extensive trials in numerous scenarios have led us to the conclusion that the introduction of external knowledge - in this case, the co-occurrence relationship between pills - can improve the performance of the model in recognizing pills. We are actively advancing this project by collecting the additional pill and prescription datasets necessary to validate the proposed method and demonstrate its applicability in other clinical contexts.

VI Evaluation

We perform several experiments on our custom pill images captured with mobile phones under unconstraint environments. The results showed that the proposed framework outperforms the baselines by a significant margin, ranging from 4.8% to 34.1% in terms of F1 -score. We also analyzed the effects of the prescription-based medical knowledge graph on pill recognition performance and discovered that the graph's accuracy is critical in boosting the overall system's performance.

VII Conclusions

Having seen the positive result that our proposed framework achieved, we are actively developing this study by gathering more pill and prescription datasets required to verify the suggested technique and prove its usefulness in different clinical settings. We believe that leveraging the external knowledge will improve the accuracy of pill identification significantly.

References

[1] Ling, et al. Few-shot pill recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020).

[2] Ting, H.-W., et al. A drug identification model developed using deep learning technologies: experience of a medical center in taiwan. *BMC Health Services Research 20* (2020).

[3] Yaniv, et al. The national library of medicine pill image recognition challenge: An initial report. In *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)* (2016), pp. 1–9.

DOI: `https://doi.org/10.1007/978-3-031-21743-2_17`
Link to repository: `https://github.com/baolongnguyenmac/fl-ml`

Keywords: Federated learning, non-IID, Meta-learning, Personalization layer

Meta-learning and Personalization layer in Federated learning

Bao-Long Nguyen[0000-0002-6411-8943], Cuong Tat Cao[0000-0003-1803-843X], Bac Le[0000-0002-4306-6945]

`{18120201,18120296}@student.hcmus.edu.vn,lhbac@fit.hcmus.edu.vn`

Simplified Title

Meta-learning and Personalization layer in Collaborative learning

Abstract

Federated learning systems are confronted with two challenges: systemic and statistical. Non-IID data is acknowledged to be a primary component in causing statistical challenges. To address the federated learning system's substantial performance loss on non-IID data, we offer the `FedMeta-Per` algorithm (which combines meta-learning methods and personalization layer approaches into a federated learning system). In terms of performance and personalization, `FedMeta-Per` has been shown in experiments to outperform typical federated learning algorithm, algorithms using personalization layer techniques and algorithms using meta-learning in system optimization.

I Introduction

Conventional machine learning approaches require users to transmit data from their devices (which might contain a lot of sensitive information) to a server for training, seriously affecting the privacy of the data. Edge computing [3] was born to perform computing and storing data right on edge devices. Therefore, the exchange of data process is no longer performed, thus ensuring data privacy to a certain level. Based on the idea of edge computing, federated learning (with FedAvg algorithm) aims to train machine learning model on separated datasets distributed on edge devices, helping to protect user privacy better.

However, federated learning systems also have their own problems. Specifically, in Horizontal federated learning, imbalanced data between devices (also known as non-IID data) can negatively affect the training performance [1]. Moreover, `FedAvg` can not give the model a high degree of personalization for each user. Our research aims to solve the statistical problem as well as the personalization problem.

In this work, we assume that (1) the system has ensured the security as well as well maintained the privacy of users and (2) the non-IID scenario is labeled imbalanced between clients.

II State of the Art

Meta-learning-based approach. Meta-learning is a learning method that can provide model the ability to fast adapt on a new dataset by letting the meta model learn a task distribution. If we consider each user as a task, then training the global model on a set of users (federated learning approach) is equivalent to training the machine model on a task distribution. A disadvantage of this approach is that it treats local clients (who have been in the system for a long time) and new clients (who have just gotten into the system) with the same initialization.

Personalization layer-based approach. Some studies proposed maintaining a part (called *personalization layers*) of neural network at each client to capture the features of them, while the rest of the network (called *based layers*) is co-trained by all clients. A downside of this approach is that the based layers cannot quickly adapt to new datasets.

III Original Contribution

Our method combines meta-learning algorithms and personalization layer techniques into a federated learning system and calls it `FedMeta-Per`. Experiments show that `FedMeta-Per` easily achieved higher convergence as well as better personalization than the aforementioned works.

IV Methodology

Data preparation. The dataset of client i is divided into two parts: a support set ($\mathscr{D}_i^{support}$) containing 20% of data and a query set ($\mathscr{D}_i^{query}$) containing 80% of data.

Training phase. Based on the personalization layer approach [2], we divided the neural network into two part: (1) based layers, co-trained by all clients; (2) personalization layers, maintained by each client. We used `MAML` and `Meta-SGD` in the proposed method. Our algorithm is illustrated using `MAML` as follows:

- Client side: Combine the based layer (receiving from the server) and the personalization layer to form the full parameter suit and perform local training:
  - Train: $\hat{w}_i^{t+1} \leftarrow w_i - \alpha \nabla_{w_i} f_{local}\left(w_i^t, \mathscr{D}_i^{support}\right)$
  - Meta-train: $w_i^{t+1} \leftarrow w_i^t - \beta \nabla_{w_i} f_{local}\left(\hat{w}_i^{t+1}, \mathscr{D}_i^{query}\right)$
  - Resolve w_i^{t+1} to form $w_{B(i)}^{t+1}$ and $w_{P(i)}^{t+1}$. Send based layers to server and store personalization layer.
- Server side: Aggregate the based layers:

$$\texttt{MAML}: w_B^{t+1} = \sum_{i=0}^{m} \frac{n_i}{N_m} w_{B(i)}^{t+1}$$

Inference phase. Clients are divided into two types: local clients and new clients. Local clients are clients who have already been in the system for a long time and have already successfully built their own personalization layers. New clients are clients who have just gotten into the system. Their personalization layers are not really good, but they can build a good one over time.

V Results

Convergence ability. Meta-learning provides the system with fast adaptability on new clients, leading to our algorithm converging much faster than `FedAvg`, `FedPer` (an algorithm that uses personalization technique) and reaching the same degree convergence of `FedMeta` (an algorithm that only uses meta-learning in optimization) but higher accuracy.

Personalization ability. Observation of the standard deviation shows that personalization layers provide the system with the high personalization ability that our standard deviation values are smaller than the ones outputted by `FedAvg`, `FedPer` and `FedMeta`.

VI Evaluation

We used acc_{micro} (accuracy w.r.t. all data points) to evaluate the fast convergence ability because we can easily visualize acc_{micro} to observe the convergence process; and acc_{macro} (accuracy w.r.t. all clients) to evaluate personalization ability because it represents the average deviation of clients' accuracy. $F1_{macro}$ is also used to check for imbalanced data processing.

VII Conclusions

Our algorithm can be used for those who want to optimize their Horizontal federated learning system in meta-learning approach and personalization layer-based approach. They can perform more experiments on meta-learning algorithms such as `FO-MAML`, `iMAML`, or other personalization techniques.

Although the proposed method has been experimentally proven to have fast convergence ability and high personalization, it has not been proven mathematically. In the future, further studies are needed to fill this gap.

References

[1] Chen, F., Luo, M., Dong, Z., Li, Z., and He, X. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876* (2018).

[2] Liang, P. P., Liu, T., Ziyin, L., Allen, N. B., Auerbach, R. P., Brent, D., Salakhutdinov, R., and Morency, L.-P. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523* (2020).

[3] Xia, Q., Ye, W., Tao, Z., Wu, J., and Li, Q. A survey of federated learning for edge computing: Research problems and solutions. *High-Confidence Computing 1*, 1 (2021), 100008.

DOI: https://doi.org/10.1007/978-981-19-8234-7_29
Link to repository: https://github.com/nguyenlab/statute_law_IR

Keywords: Legal information retrieval system, Statute law, Statute law information retrieval, COLIEE

A Legal Information Retrieval System for Statute Law

Chau Nguyen[0000-0003-0068-0387] , Nguyen-Khang Le[0000-0001-6585-5470] , Dieu-Hien Nguyen[0000-0001-5238-733X] , Phuong Nguyen[0000-0002-3752-8699], Le-Minh Nguyen[0000-0002-2265-1010]

chau.nguyen@jaist.ac.jp, lnkhang@jaist.ac.jp, ndhien@jaist.ac.jp, phuongnm@jaist.ac.jp, nguyenml@jaist.ac.jp

Simplified Title

A system to retrieve relevant legal articles for statute law

Abstract

The information retrieval task for statute law requires a system to retrieve the relevant legal articles given a legal bar exam query. The Transformer-based approaches have demonstrated robustness over traditional machine learning and information retrieval methods for legal documents. However, those approaches are mainly domain adaptation without attempting to tackle the challenges in the characteristics of the legal queries and the legal documents. This paper specifies two challenges related to the characteristics of the two legal materials and proposes methods to tackle them effectively. Specifically, the challenge of different language used (while the articles use abstract language, the queries may use the language to describe a specific scenario) is addressed by a specialized model. Besides, another specialized model can overcome the challenge of long articles and queries. As shown in the experimental results, our proposed system achieved a state-of-the-art F2 score of 76.87%, with an improvement of 3.85% compared to the previous best system.

I Introduction

COLIEE is a renowned international competition in legal text processing. In 2021, COLIEE included five tasks on case law and statute law. Our work aims to address Task 3: information retrieval on statute law (statute law is the written law that is passed by a body of legislature). There are two characteristics of the legal texts for the statute law information retrieval task that should be considered. While legal articles have an abstract nature, a legal bar query may either be written in an abstract manner or describe a very specific scenario. Hence, it is non-trivial for a sole model to generalize the relationship between the abstract legal articles and the legal bar queries. We propose to determine the specific-scenario queries, then employ a Transformer-based model to address only those queries. Moreover, since a portion of legal articles as well as legal bar queries are relatively long, we propose to employ models specialized in processing long sequences[1].

II State of the Art

LegalBERT is a family of BERT models to assist natural language processing research in the legal domain, computational law, and legal technology applications. It is an adaption of BERT in the legal domain where pre-training is carried out on several fields of English legal text. This particular BERT model has been performing better than the original version of BERT on legal domain-specific tasks. Longformer was proposed as a Transformer-based model with an attention mechanism that scales linearly with sequence length and can process documents of more than thousands of tokens. The ability to deal with long sequences of Longformer makes it a suitable candidate for the statute law retrieval task where the concatenation of a query and an article can be longer than 512 tokens.

III Original Contribution

In this paper, our contributions are three-fold:

- Propose to determine specific-scenario queries to facilitate Transformer-based models to learn better the specific relationship patterns between legal articles and legal bar queries.
- Propose to employ Longformer to address the long articles and long queries.
- Analyze the contribution of the two proposed ideas for the system comprehensively.

IV Methodology

We formulate the retrieval task as a sentence pair classification task, where the input is the pair of query and article, and the label is 1/0, corresponding to whether or not there is an entailment relationship between the sentences. Fine-tuning models requires positive and negative samples. Positive samples are provided in the dataset where most queries have only 1 or 2 relevant articles. For negative samples, we represent the query and all articles in TF-IDF vectors and compute their similarity scores. We consider the top K articles with the highest TF-IDF similarity scores, except for the articles labeled as relevant, to be negative samples.

We propose to filter the specific-scenario queries and manipulate them separately. We design rules to determine if the uppercase in a query implies a legal person or object. We separately handle the specific-scenario queries and their relevant articles following the same procedure for regular queries.

To address long articles and queries, we propose to adopt Longformer[1] as an additional model for capturing the patterns of the long inputs besides the normal-length inputs. We follow the same procedure of fine-tuning LegalBERT[2] to fine-tune the pre-trained Longformer model.

V Results

Our model outperforms the best system in the competition [3] by 3.85%. It can be seen that our model demonstrates superior recall (note that this retrieval task emphasizes recall) while achieving fair precision. The ablation study shows that the specific-scenario models help improve the macro average F2 score significantly, especially when the specific-scenario model can retrieve relevant articles but the other model cannot retrieve any of them.

VI Evaluation

The experiments are conducted on the COLIEE Task 3 dataset, which consists of 748 training queries, 58 validation queries, and 87 test queries. The performance of models is measured using Precision, Recall, and F2 score.

VII Conclusions

This paper specifies two challenges related to two characteristics of the legal bar exam queries and legal articles, which are neglected by previous approaches. The first challenge is that the language used in the legal bar exam queries may be either abstract or scenario-specific. We proposed to determine the specific-scenario queries and tackle them separately. The second challenge is that the long document may limit the performance of many available pre-trained models. We propose to leverage the capability of handling long documents of Longformer to tackle this challenge. Our system can achieve the best F2 score on the dataset of Task 3 (COLIEE 2021). This study encourages further research to observe and tackle particular characteristics of documents in the legal domain.

References

[1] Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer. *arXiv:2004.05150* (2020).

[2] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (Online, Nov. 2020), Association for Computational Linguistics, pp. 2898–2904.

[3] Wehnert, S., Sudhi, V., Dureja, S., Kutty, L., Shahania, S., and De Luca, E. W. Legal norm retrieval with variations of the bert model combined with tf-idf vectorization. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (New York, NY, USA, 2021), ICAIL '21, Association for Computing Machinery, p. 285–294.

DOI: https://doi.org/10.1007/978-3-031-21967-2_33
Link to repository: https://github.com/ngdangha/Vietnamese-Mean

Keywords: 3D face, face reconstruction, morphable model

A new 3D face model for Vietnamese based on Basel Face Model

Dang-Ha Nguyen, Khanh-An Han Tien, Thi-Chau Ma, Hoang-Anh Nguyen The

ndha.uet.vnu@gmail.com, chaumt@vnu.edu.vn

Simplified Title

A base 3D face model for Vietnamese.

Abstract

In recent years, many 3D face reconstruction from image methods have been introduced and most of them have shown incredible results. However, methods such as photogrammetry require images from multiple views and can be very time-consuming. Deep learning based methods, on the other hand, are faster and more efficient but heavily rely on the base face models and training datasets. Meanwhile, most base face models lack Asian facial features, and high-quality Vietnamese facial image databases are still not available yet. In this paper, we propose an approach that increases the accuracy of Vietnamese 3D faces generated from a single image by creating a new mean face shape and training a convolution neural network with our dataset. This method is compact and can improve the quality of 3D face reconstruction using facial image data with specific geographical and race characteristics.

I Introduction

Reconstructing a 3D face from a single image refers to retrieving a 3D face surface model of a person given only one input face image. This is a classical and challenging computer vision task with a wide range of applications. To do so, many researchers have come up with the idea of using deep learning networks to change the morphable face model. These frameworks have shown great results but heavily rely on the generative face model and training data. Thus, these methods can not perform well with face images from different countries, including Vietnam. Our main scope is to create a new framework that helps researchers to improve the quality of the 3D reconstructed face. We assume that we can make the face more realistic with enough training data and a base face model possessing facial attributes and traits of specific locations. That is why we collect more face images and try creating a new face model for Vietnamese in this paper.

II State of the Art

To solve the research problems, there are several methods using deep learning networks to reconstruct the 3D model from a single image. With the output coefficients from the network, they can change the parameter of the base 3D face model to transform the model into the face of the person from the input image. These methods have shown very promising results; however, when testing with Vietnamese people, the generated faces are not convincing enough. This problem might be because of the lack of training data and facial attributes of Vietnamese.

III Original Contribution

Our main contribution in this paper is proposing a framework to increase the accuracy of the 3D face reconstruction from a single image process for Vietnamese by generating a new mean shape with our high-quality face image dataset. After creating the new face model, we test the effectiveness of our method using face images of Vietnamese and compare it with the previous techniques to see if there is any improvement.

IV Methodology

To achieve the presented results, we first created our face image dataset by taking pictures of nearly 200 Vietnamese. The capturing system we used contains 27 DSLR cameras and ten lighting devices. Volunteers are required to perform seven facial expressions in 11 lighting conditions. Then, by utilizing the Deep3D framework [1] and the 2009 Basel Face Model [2], we reconstructed the 3D face models of the volunteers and generated the new mean shape. Finally, we used the pre-trained model of Deng *et al.* [1] and continued training on our dataset so the output coefficients could be fitted with our new face model. Our study is an experiment to see whether our assumption is correct or not and it has shown some positive results.

V Results

From our framework, we have generated a new mean shape called V-Mean which can replace the old shape from the BFM-2009 [2] to construct a new generative face model with Vietnamese facial features (in Figure 1). With the new base face model, we can create 3D faces with more traits of people from Vietnam.

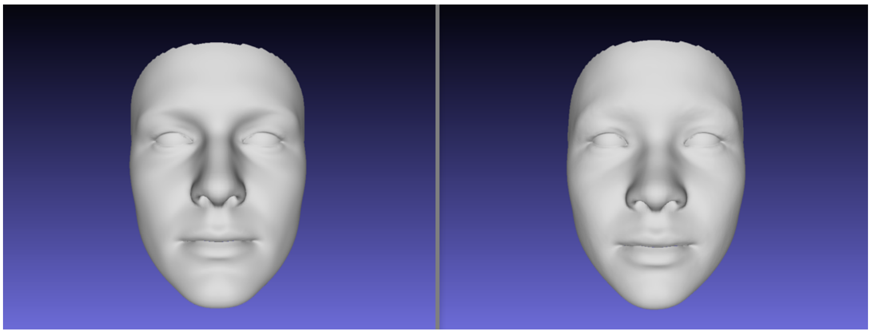


Figure 1: Comparing mean shape from BFM (*left*) with V-Mean (*right*)

VI Evaluation

Due to the lack of ground truth 3D face models, we used two types of measurements named photometric error and cosine distance to test the effectiveness of our framework on Vietnamese facial images. Besides, we think that it is important the method should improve the results of Vietnamese without affecting the reconstructed models of the original faces. Thus, we also used mean square error to verify our approach on a few 3D-scanned non-Vietnamese face models. The results from all mentioned evaluation metrics have shown that our assumption is accurate but we still need more images of people in various conditions to make the final 3D face model more truthful.

VII Conclusions

We have introduced a new mean shape model for Vietnamese, which helps improve the accuracy of the 3D face reconstruction process. Our new model showed better results than the original when tested on both of our custom test sets. However, the truthfulness of the final 3D face model can still be improved by training the network with more Vietnamese face images. Researchers can simply increase the quality of the reconstructed 3D face by using facial images in specific countries.

References

[1] Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., and Tong, X. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2019), pp. 285–295.

[2] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance* (2009), pp. 296–301.

DOI: https://doi.org/10.1007/978-981-19-8234-7_30
Link to repository: https://github.com/lekhang4497/qasper-retriever-reader

Keywords: Question-Answering, Retriever-Reader, Long Sequences

Exploring Retriever-Reader Approaches in Question-Answering on Scientific Documents

Dieu-Hien Nguyen[0000-0001-5238-733X], Nguyen-Khang Le[0000-0001-6585-5470], Minh Le Nguyen

ndhien@jaist.ac.jp, lnkhang@jaist.ac.jp, nguyenml@jaist.ac.jp

Simplified Title

A two-step approach in applying deep learning to answer questions from scientific documents

Abstract

As readers of scientific articles often read to answer specific questions, the task of Question-Answering (QA) in academic papers was proposed to evaluate the ability of intelligent systems to answer questions in long scientific documents. Due to the large contexts in the questions, this task poses many challenges to state-of-the-art QA models. This paper explores the retriever-reader approaches widely used in open-domain QA and their impact when adapting to QA on long scientific documents. By treating one scientific article as the corpus for retrieval, we propose a retriever-reader method to extract the answer from the relevant parts of the document and an effective sliding window technique that improves the pipeline by splitting the articles into disjoint text blocks of fixed size. Experiments on QASPER, a dataset for QA in Natural Language Processing papers, showed that our method outperforms all state-of-the-art models and establishes a new state-of-the-art in the extractive questions subset with 30.43% F1

I Introduction

The task of Question-Answering (QA) in academic papers was proposed to evaluate the ability of intelligent systems to answer questions in long scientific documents. QASPER[2] is a dataset for QA on NLP papers. In QASPER, given an academic article and a question about the article, the QA system must find the answer to the question. Models specialized in processing long sequences [1] may not efficiently capture the semantics of all the words compared to other models processing a smaller number of words. In open-domain QA, only the questions are provided, and QA systems have to find information from a large corpus to answer the questions. The problem of finding a piece of information from a long scientific paper in QASPER can resemble that of the open-domain QA task. No existing works explore the impact of open-domain QA retrieval methods on the QASPER dataset. This paper proposes a novel method that adapts retriever-reader approaches to long scientific papers in QASPER by treating a single paper as a large corpus for retrieval. Treating each scientific document as a corpus for retrieval enables us to utilize open-domain QA techniques and develop a retriever-reader pipeline that achieves the new state-of-the-art performance on QASPER

II State of the Art

DocHopper is a model that iteratively attends to different parts of long, hierarchically structured documents to answer complex questions. QASPER-LED is an encoder-decoder model based on Longformer. As QASPER-LED can support sequence lengths up to about 16,000 tokens, 99% of the articles in the QASPER dataset can be processed without truncation. ETC reader is one of the most efficient models for processing long sequences, with a maximum of 4096 tokens capability. However, models specialized in processing long sequences may not efficiently capture the semantics of all the words compared to other models processing a smaller number of words

III Original Contribution

The main contributions of this paper are:

1. We propose a novel method that adapts retriever-reader approaches from Open-domain QA to the problem of QA in long scientific documents and establishes new state-of-the-art results in the QASPER dataset.

2. We conduct experiments on retrievers and re-rankers to explore the impact of different passage-splitting techniques on retrieval results in QASPER. We find that splitting articles using a sliding window improves the retrieval performance, and the window side of 150 produces the best result.

IV METHODOLOGY

This study is experimental. We consider article A as the corpus for retrieval. A retriever is leveraged to retrieve m passages $P = [P_1, ..., P_i, ..., P_m]$ from article A for a given question $Q = (q_1, ..., q_{|Q|})$, where $P_i = (p_i^1, p_i^2 ..., p_i^{|p_i|})$ is the i-th passage, $P_i \in A$, and $q_k \in Q$ and $p_i^j \in P_i$ are corresponding words. We split each article into multiple, disjoint text blocks of a fixed number of words. These text blocks are referred to as passages, and we consider these passages as our basic retrieval units. In the answer extraction problem, we only consider the questions in QASPER that have extractive answers. A reader is employed to extract the answer span from the relevant passages. Given a context passage $C = (c_1, c_2, ..., c_n)$ and a question $Q = (q_1, ..., q_{|Q|})$. The reader aims to find a text span $(c_i, c_{i+1}, ...c_j)$ from the context that answers the question Q. We employ Transformer-based models for the reader component.

V RESULTS

Our method outperforms all state-of-the-art models including ETC, QASPER-LED[2], and DocHopper[3], which are designed for processing long sequences. The ablation study shows that a performance gain can be observed when applying a sliding window.

VI EVALUATION

All experiments are conducted on the QASPER dataset[2], which consists of 5,049 questions over 1,585 NLP papers. We experiment with the subset of extractive questions, which account for 51.8% of the dataset. We compare our method with previous state-of-the-art models in QASPER, DocHopper [3] and QASPER-LED [2], and other competitive baselines. The performance is compared using the official QASPER evaluation metric F1 score. Our analysis finds that Cross-encoder Re-rankers perform better than Sparse Retrievers on scientific articles and that a window size of 150 words produces the most effective results.

VII CONCLUSIONS

We propose a method to adapt the retriever-reader approach in Open-domain QA to the problem of QA in long scientific documents. Our method allows us to employ efficient Transformer-based readers which overcome the previous limitations posed by long sequences. This study also paves the way for future research on efficient retrievers and readers to be applied in QA on long scientific documents without the obstacle of processing long sequences.

REFERENCES

[1] BELTAGY, I., PETERS, M. E., AND COHAN, A. Longformer: The long-document transformer. *CoRR* (2020).

[2] DASIGI, P., LO, K., BELTAGY, I., COHAN, A., SMITH, N. A., AND GARDNER, M. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Online, June 2021), Association for Computational Linguistics, pp. 4599–4610.

[3] SUN, H., COHEN, W. W., AND SALAKHUTDINOV, R. Iterative hierarchical attention for answering complex questions over long documents, 2021.

DOI: https://doi.org/10.1007/978-981-19-8234-7_31

Keywords: Data mining, High-utility sequential pattern mining, Frequent closed sequence mining, Multi-core processors, Parallel mining.

P-FCloHUS: A Parallel Approach For Mining Frequent Closed High-Utility Sequences On Multi-core Processors

Hong-Phat Nguyen, Bac Le

nhphat1997@gmail.com, lhbac@fit.hcmus.edu.vn

Simplified Title

Mining compact subsets of highly beneficial patterns in sequential data in a parallel manner.

Abstract

Frequent closed high-utility (FCHU) sequences are preferable to frequent closed sequences. Not only because of their utility-based nature that considerately contributes to taking decisive business actions, FCHU sequences also preserve necessary information for re-constructing frequent high-utility sequences. Despite of their vital role, mining FCHU sequences is a time consuming task when facing with large-scale datasets, or especially when the input thresholds are relatively small. To contend with these difficulties, this paper proposes a parallel algorithm named P-FCloHUS for fast mining FCHU sequences by making good use of multi-core processors. By relying on a novel Single scan synchronization strategy that is facilitated by an efficiently Partitioned result space structure, P-FCloHUS successfully alleviates the communication cost between mining tasks and hence speeds up the parallel mining process. Experiments on both dense and sparse datasets show that P-FCloHUS outperforms the state-of-the-art FMaxCloHUSM in terms of runtime performance.

I Introduction

Discovering highly beneficial patterns in sequential data has been always of great interest to researchers. In particular, FCHU patterns [3] are significantly valuable due to their compact nature. That being said, mining FCHU patterns is a challenging task in situations where provided input thresholds are rather small and/or the amount of data is huge. This work aims to alleviate the mining effort in these situations, and thus provides a more efficient mining approach.

II State of the Art

Several studies on mining either high-utility sequential patterns [2] or frequent closed sequences [1] have been proposed recently, but only FMaxCloHUSM [3] was the state-of-the-art algorithm on exploring FCHU patterns. Even though FMaxCloHUSM's performance is not very efficient for large datasets and/or input thresholds are small.

III Original Contribution

A pioneering algorithm for mining FCHU patterns in parallel that outperforms the state-of-the-art FMaxCloHUSM [3] in terms of runtime performance on both dense and sparse datasets.

IV Methodology

Many up-to-date studies on mining beneficial patterns in sequential data have been taken into account. Then with the emphasis on extracting FCHU patterns, we re-implemented FMaxCloHUSM in C++ as a starting point for our research. From there, we continued to design and develop P-FCloHUS by making good use of multi-core processors. Finally, various experiments have been conducted on both approaches.

V Results

Telling from the obtained results from multiple experiments on both dense and sparse dataets, P-FCloHUS takes at least 5 times less than the amount of execution duration required for the mining process as compared to FMax-CloHUSM given the smallest input thresholds in each experiment.

VI Evaluation

All experiments within this work were conducted on 5 real-life datasets, including both dense and sparse ones. For each experiment, the minimum support (or the minimum utility) was fixed while the other was decreased to stretch the performance of both algorithms. Also, a scaling experiment was carried out on several CPU setups to see how that would affect the runtime performance.

VII Conclusions

With these enhancements of P-FCloHUS, businesses and researchers who have been using FMaxCloHUSM or other algorithms for mining FCHU patterns can now speed up their mining process significantly, therefore saving time and effort on the task of FCHU patterns data mining.

References

[1] Bac L, Hai D, T. T. *FCloSM, FGenSM: Two efficient algorithms for mining frequent closed and generator sequences using the local pruning strategy*. International Journal of Knowledge and Information Systems, 2018.

[2] Bac L, U. H., and D, D.-T. *A pure array structure and parallel strategy for high-utility sequential pattern mining*. Expert Systems with Applications, 2018.

[3] Tin T, H. D., and Bac L, F.-V. P. *FMaxCloHUSM: An efficient algorithm for mining frequent closed and maximal high utility sequences*. Engineering Applications of Artificial Intelligence, 2019.

DOI: https://doi.org/10.1007/978-3-031-21967-2_16
Link to repository: https://github.com

Keywords: Concept drift, Data stream, Ensemble learning, Bayesian optimization

Parameter Distribution Ensemble Learning for Sudden Concept Drift Detection

Khanh-Tung Nguyen, Trung Tran, Anh-Duc Nguyen, Xuan-Hieu Phan, Quang-Thuy Ha

tungnk@epu.edu.vn, trungt.cs.edu@gmail.com, 20021336@vnu.edu.vn, hieupx@vnu.edu.vn, thuyhq@vnu.edu.vn

Simplified Title

Parameter Distribution Ensemble Learning for Sudden Concept Drift Detection

Abstract

Concept drift is a big challenge in data stream mining (including process mining) since it seriously decreases the accuracy of a model in online learning problems. Model adaptation to changes in data distribution before making new predictions is very necessary. This paper proposes a novel ensemble method called E-ERICS, which combines multiple Bayesian-optimized ERICS models into one model and uses a voting mechanism to determine whether each instance of a data stream is a concept drift point or not. The experimental results on the synthetic and classic real-world streaming datasets showed that the proposed method is more precise and more sensitive (shown in precision, recall, and F1-score) than the original ERICS models in detecting concept drift, especially for sudden drifts.

I Introduction

Recently, along with the development of technologies, data stream mining has become a hot topic for researchers. One of the most important properties of streaming data is that its distribution may change over time (i.e., concept drift). Concept drift detection is essential in data stream mining since if we cannot point out the drift, predictions made by the model that is based on past data will not be adaptive to new data.

We find that ERICS (Effective and Robust Identifications of Concept Shift) [1], a framework that detects only real concept drifts by using a predictive model, satisfies two properties: model-awareness and explainability. Our study aims to improve this approach to achieve better detecting results.

II State of the Art

Several methods were introduced to deal with the problem of concept drift. Early approaches (such as DDM and EDDM), work by measuring the change in the error rate, while some others (ADWIN, KSWIN, etc.) use the windowing mechanism. Moreover, more complex algorithms like unsupervised or automatic learning are also used in LD3, Meta-ADD, etc. The main drawback of these methods is that they do not achieve high precision and recall in detecting concept drift.

III Original Contribution

In this study, our main contributions are summarized as follows:

- We introduce a Bayesian optimization pipeline to find hyperparameter values that increase the precision, recall, and F1-score of ERICS [1] models.

- We propose a novel framework E-ERICS (Ensemble of ERICS models), an improved version of ERICS by ensembling ERICS models with the high performance achieved in the Bayesian Optimization phase, to detect concept drifts (especially sudden drifts) more precisely and sensitively than the original ERICS model does.

IV Methodology

In general, our study is experimental. We introduce a new method called E-ERICS, a concept drift detection model that uses ensemble learning with a voting mechanism in which the base learners are ERICS models with values of hyperparameters adjusted using Bayesian optimization.

Our E-ERICS method is divided into two phases:

- BO-ERICS Phase (Bayes Optimization on ERICS): Perform Bayesian Optimization on three ERICS hyperparameters: moving average size, sliding window size, and update rate of the threshold. We set F1-score as the acquisition function and the Gaussian process as the surrogate model of the optimization procedure.

- Ensemble Phase: Ensemble selected models which have been found in BO-ERICS. We take the best four base-learners (with the highest F1-scores) found in the BO-ERICS phase as input and use ensemble learning on them to improve the performance of our model by a voting mechanism for every single data point. In our poll, the weight of the base learner with the highest F1-score (M_0) is twice as much as the vote of others (M_1, M_2, M_3), and:

+ Drift points detected by only M_0 but not in M_1, M_2, M_3 will be eliminated.

+ Drift points not detected by M_0 but detected by all M_1, M_2, M_3 will be added to the drift point sets.

V Results

For our experiments, we use six datasets containing sudden and incremental drift points: two synthetic ones (SEA and Hyperplane) produced by scikit-multiflow framework and four real-world datasets from various fields: spams (Spambase), online games (Dota), network intrusion (KDD) and personal income (Adult).

The experiment results show that the E-ERICS model achieves better results compared with the original ERICS model (with the average Precision, Recall, and F1-score: 0.9583, 0.6700, and 0.7509, respectively, while the original ERICS model achieves 0.4867, 0.4077, and 0.3200, and the model with the best performance in BO-ERICS Phase gets 0.9583, 0.6515, and 0.7446).

VI Evaluation

We evaluate our method and the original ERICS using prequential evaluation (interleaved test-then-train) and three classical metrics: Precision, Recall, and F1-score.

In Precision, a drift detection is counted as true positive if it is less than 50 batches after an actual drift point; otherwise, it is a false positive point; while in Recall, a false negative is considered if after 50 batches from the actual drift point, the model does not detect any drift point at all. Any actual drift detected in the range of 50 batches is counted as true positive.

VII Conclusions

In this article, we have presented a new method called E-ERICS (using Bayesian optimization combined with ensemble learning). E-ERICS can be used to detect concept drift (especially sudden drifts) in data streams of different types and fields more accurately than the previous approaches.

References

[1] Haug, J., and Kasneci, G. Learning parameter distributions to detect concept drift in data streams. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), pp. 9452–9459.

Link to arXiv: https://arxiv.org/abs/2204.07002
DOI: https://doi.org/10.1007/978-3-031-21743-2_30

Keywords: Question Answering, Transformer, BERT, XLM-R, Transfer Learning, Machine Reading Comprehension

XLMRQA: Open-Domain Question Answering on Vietnamese Wikipedia-based Textual Knowledge Source

Kiet Van Nguyen[0000-0002-8456-2742], Phong Nguyen-Thuan Do, Nhat Duy Nguyen, Tin Van Huynh[0000-0003-4990-2868], Anh Gia-Tuan Nguyen, Ngan Luu-Thuy Nguyen

kietnv@uit.edu.vn, 18520126@gm.uit.edu.vn, 18520118@gm.uit.edu.vn, tinhv@uit.edu.vn, ngannlt@uit.edu.vn, anhngt@uit.edu.vn

Simplified Title

Retriever-Reader Question Answering System for Vietnamese

Abstract

This paper presents XLMRQA, the first Vietnamese QA system using a supervised transformer-based reader on the Wikipedia-based textual knowledge source (using the UIT-ViQuAD corpus), outperforming the two robust QA systems using deep neural network models: DrQA and BERTserini with 24.46% and 6.28%, respectively. From the results obtained on the three systems, we analyze the influence of question types on the performance of the Vietnamese QA systems.

I Introduction

In recent years, the rapid development of social media has led to an explosion of data and information. People need to find information and knowledge through the support of machine question-answering applications like Google, Siri, and Alexa. QA systems assist people in accessing information and knowledge faster without wasting much time and effort. QA-based tasks are of interest to the Vietnamese natural language processing and computational linguistics community. Machine reading comprehension-based QA systems have gained much attention in recent years. Although several machine reading comprehension corpora have been released for developing QA systems, such as UIT-ViQuAD, UIT-ViWikiQA, and UIT-ViNewsQA, there is no reader-based QA system for Vietnamese yet.

Along with the strong development of machine learning, QA systems have been explored in various corpora and methods. In recent years, QA systems have followed two Retriever-Reader QA systems, such as DrQA and BERTserini. In this paper, we proposed XLMRQA, the first Vietnamese QA system using a supervised transformer-based reader on the Wikipedia-based textual knowledge source (using the UIT-ViQuAD corpus [3]), outperforming the two robust QA systems using deep neural network models: DrQA and BERTserini.

II State of the Art

Unlike previous QA systems without readers, DrQA is a full QA system combining a bigram hash-based TF-IDF retriever with a multi-layer iterative neural network reader trained to predict answers in the passage. BERTserini is a QA system that combines the BERT-based reader and the open-source Anserini toolkit for text retrievers. The system receives a small set of documents as input. In an end-to-end approach, the system combines best practices from document retrieval with a BERT-based reader to determine answers from a large-scale corpus of English Wikipedia articles. For the Vietnamese language, there are still not any QA systems based on the Retriever-Reader mechanism, mainly focusing on the traditional QA system. Therefore, we would like to develop this system as a starting point for the mechanism for Vietnamese QA, outperforming DrQA and BERTserini.

III Original Contribution

Our three main contributions are described as follows. (1) We propose XLMRQA, a retriever-reader-selector QA system for the Vietnamese language, outperforming the F1-score and exact match (EM) of two other SOTA systems: DrQA and BERTserini. (2) We re-implement state-of-the-art QA systems on the Vietnamese Wikipedia knowledge texts: DrQA and BERTserini as baseline systems. The first experiments were performed on the retriever-reader-based QA model on Vietnamese texts. (3) We analyze the impacts of question types on the proposed QA system XLMRQA for Vietnamese, which helps researchers improve the performance of the QA systems.

IV Methodology

Inspired by the DrQA system [1], we present an overview of the QA architecture using a supervised reader for the Vietnamese language. XMRQA is a QA system with three components: text retriever, text reader, and answer selector. The text retriever finds texts or documents and passes them to the text reader to find the candidate's answers. The answer selector finds the answer that best matches the question from the candidate answers predicted by the text reader. In particular, we describe the QA system and its components as follows.

IV.1 Text Retriever

We apply a basic retriever to find k passages that answer the input question, using the question as a bag-of-words question. The text retriever finds passages or documents related to the question from a set of 5,109 passages extracted from the ViQuAD corpus. This corpus was built by aggregating all the passages from the Train, Dev, and Test sets of the ViQuAD corpus consisting of 4,101, 515, and 493 passages, respectively. This paper assesses two different text retrievers, including TF-IDF and the Anserini. To optimize the performance of QA systems, we apply word segmentation to the text retrievers.

IV.2 Text Reader

The retrieved passages are passed to the reader to extract the candidate answers. XLM-RoBERTa (XLM-R) [2] is a multilingual language model trained on a large-scale dataset with 100 languages. XLM-R is used as a pre-trained language model for many tasks such as natural language inference and machine reading comprehension, which achieves state-of-the-art performances. In this paper, we use XLM-R to build a reader as the main component of the XLMRQA system to extract candidate answers before transferring them into the answer selector.

IV.3 Answer Selector

The candidate answer list, the reading score list, and the retrieving score list are fed into this component. Each score in the two score lists corresponds to each answer in the answer list. We then combine the reading score with the retrieving score through linear interpolation to estimate the score for each answer and find the answer with the highest score.

V Results

The three systems in ascending order of results are DrQA, BERTserini, and XLMRQA. The XLMRQA QA system achieves the highest performance with EM and F1 set to the maximum at k=5 with an F1 of 64.99% and an EM of 51.94% on the Test set. Similar to XLMRQA, the BERTserini system achieves the best performance at k=5 with an F1 of 58.30% and an EM of 39.46% on the ViQUAD Test set. The DrQA system achieves the best performance with $k \geq 10$, and these are equal. Nevertheless, we have chosen k=10 as the official value for the DrQA system because it achieves good performance in terms of time. At k=10, the DrQA system achieved an F1 of 37.86% and EM of 18.42% on the Test set.

VI Conclusions

In this paper, we introduced XLMRQA, a QA system based on the retriever-reader-selector mechanism for Vietnamese open-domain texts, which outperformed two state-of-the-art question-answering systems, DrQA, and BERTserini, on the ViQuAD corpus. For assessing the performance of three QA systems on the ViQuAD corpus, we achieved the highest performance with the XLMRQA system: EM of 51.94% and F1 of 64.99%. Analysis of the performance of QA systems was performed on different types of questions. In the future, several future directions are recommended: (1) integrating diverse question words as linguistic features into the QA systems can boost their performances; (2) finding out a method to leverage the power of monolingual and multilingual BERTology-based language models; and (3) expanding our QA system to other low-resource languages.

References

[1] Chen, D., Fisch, A., Weston, J., and Bordes, A. Reading Wikipedia to answer open-domain questions. In *Proceedings of ACL 2017 (Volume 1: Long Papers)* (Vancouver, Canada, July 2017), Association for Computational Linguistics, pp. 1870–1879.

[2] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020* (Online, July 2020), Association for Computational Linguistics, pp. 8440–8451.

[3] Nguyen, K. V., Nguyen, D.-V., Nguyen, A. G.-T., and Nguyen, N. L.-T. A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of COLING 2020* (Barcelona, Spain (Online), Dec. 2020), ICCL, pp. 2595–2605.

DOI: https://doi.org/10.1007/978-3-031-21967-2_3
Link to repository: https://github.com/ngminhhieu/gafs_framework

Keywords: Time series prediction, Feature selection, Genetic Algorithm.

A Lightweight and Efficient GA-based Model-Agnostic Feature Selection Scheme for Time Series Forecasting

Minh Hieu Nguyen[0000-0003-1518-8977], Viet Huy Nguyen[0000-0002-5669-8837], Thanh Trung Huynh[0000-0003-2027-5362], Thanh Hung Nguyen[0000-0001-6290-2841], Quoc Viet Hung Nguyen[0000-0002-9687-1315], Phi Le Nguyen[0000-0001-6547-7641]

hieu.nm2052511m@sis.hust.edu.vn, huy.nv184120@sis.hust.edu.vn, hungnt@soict.hust.edu.vn, lenp@soict.hust.edu.vn, h.thanhtrung@griffith.edu.au, henry.nguyen@griffith.edu.au

SIMPLIFIED TITLE

A Lightweight and Efficient Feature Selection Framework using Genetic Algorithm for Time Series Forecasting.

ABSTRACT

Time series prediction, which obtains historical data of multiple features to predict values of features of interest in the future, is widely used in many fields. One of the critical issues in dealing with the time series prediction task is how to choose appropriate input features. This paper proposes a novel approach to select a sub-optimal feature combination automatically. Our proposed method is model-agnostic that can be integrated with any prediction model. The basic idea is to use a Genetic Algorithm to discover a near-optimal feature combination; the fitness of a solution is calculated based on the accuracy obtained from the prediction model. In addition, to reduce the time complexity, we introduce a strategy to generate training data used in the fitness calculation. The proposed strategy aims to satisfy at the same time two objectives: minimizing the amount of training data, thereby saving the model's training time, and ensuring the diversity of the data to guarantee the prediction accuracy. The experimental results show that our proposed GA-based feature selection method can improve the prediction accuracy by an average of 28.32% compared to other existing approaches. Moreover, by using the proposed training data generation strategy we can shorten the time complexity by 25.67% to 85.34%, while the prediction accuracy is degraded by only 2.97% on average.

I INTRODUCTION

Time series prediction is a critical problem applied in many fields such as environmental indicators prediction and stock forecasting. The performance of a solution is decided by two main factors, namely the input features and the prediction model. Although many works have been devoted to constructing prediction models, that is still rare for feature selection. In practice, the number of features in a prediction problem is usually very large. Unrelated input features may confuse the model, thus degrading the prediction accuracy. Motivated by the aforementioned observations, we aim to propose a novel feature selection framework that allows us to quickly select an optimal feature combination and train the prediction model using the selected features. We are given a dataset consisting of l supporting features, and k features of interest, denoted as $X_1,...,X_l$ and $X_{l+1},...,X_{l+k}$, respectively. We focus on a time series prediction task that obtains the information of the features in m previous timesteps and produces the values concerning the features of interest in the following n timesteps, which can be represented as follows.

Input: $x_i, x_{i+1}, \ldots, x_{i+m-1}$

Output: $\tilde{y}_{i+m}, \tilde{y}_{i+m+1}, ..., \tilde{y}_{i+m+n-1} = \underset{y_{i+m}, y_{i+m+1}, ..., y_{i+m+n-1}}{\text{argmax}} \quad p(y_{i+m}, y_{i+m+1}, ..., y_{i+m+n-1} | x_i, x_{i+1}, \ldots, x_{i+m-1})$,

where x_{i+j} $(j = 0, ..., m-1)$ is the input vector with respect to the j-th timestep; y_{i+m+k} $(k = 0, ..., n-1)$ is a vector represents the value of the features of interest at the $(i+m+k)$-th timestep. Suppose that f is the prediction model. Our aim is to determine l' features among l supporting features to feed into f to achieve sub-optimal accuracy.

II STATE OF THE ART

In [1], the authors proposed a novel approach that utilizes a GA-based wrapper method for feature selection. To reduce the computing complexity, only the top k most important features are retained before going through the GA-based wrapper method. However, we hardly define which k should be, especially when the number of features is very large. In [2], to identify an optimum feature subset for the credit risk assessment problem, the authors used a filter method that reduces the number of input features before putting them in the GA-based wrapper phase. Nevertheless, removing redundant features before processing the wrapper method could possibly miss the combinations of the valuable features Although some existing works exploit GA to select input features, most of them focus on classification problems. Moreover, they all require training the model with the full data to calculate the fitness, leading to huge time complexity.

III ORIGINAL CONTRIBUTION

The main contributions of our paper are as follows: (1) We propose a model-agnostic feature selection framework that can dynamically determine input features for time series prediction problems; (2) We present a GA-based method to search for a near-optimal feature combination; (3) We design a training data generation strategy that helps to reduce the training time while ensuring the accuracy of the prediction model; and (4) We perform extensive experiments on a real dataset to evaluate the effectiveness of the proposed method.

IV METHODOLOGY

Our proposed method comprises a GA-based feature selector, a training data generator, and a prediction model.
GA-based feature selector and prediction model. We use each individual in the GA-based feature selector to represent a feature combination. Therefore, the accuracy of the prediction model when utilizing the corresponding feature combination determines an individual's goodness. In each generation, the individuals with better fitness remain while the others are removed from the population. By repeating these processes, the fitness values of the population are improved over the generations. Finally, the individual with the best fitness value is selected.
Training data generator. We denote by S the length of the original training data set, and s the length of a sub-dataset which is used to calculate the fitness of one individual. Note that s is a hyper-parameter that tradeoffs between the training time and the prediction model's accuracy. In general, the greater s, the higher the accuracy but the larger the training time. The main purpose of the training data generator is to reduce the running time of the framework while ensuring high prediction accuracy. For each individual A, we store all sub-datasets that have been used to calculate the fitness of A. Specifically, let us denote by $\mathscr{D}_A^m$ the set of all datasets that have been used to train the model for calculating the fitness of A until the m-th generation. Suppose $\mathscr{D}_A^m = \{d_1, ..., d_m\}$, where d_i is the sub-dataset used in the i-th generation. Now, we select the sub-dataset for A in the $(m+1)$-th generation as follows. For each timestep t, i.e., t ranges from 1 to $S - s + 1$, we determine a sub-dataset of length s starting from timesteps t, and denote as d_{m+1}^t. For each d_{m+1}^t, we define its overlap degree with $\mathscr{D}_A^m$ as the total length of segments overlapped by d_{m+1}^t and d_i $(i = 1, ..., m)$. Finally, we choose the sub-dataset d_{m+1}^t whose overlap degree with $\mathscr{D}_A^m$ is the minimum.

V RESULTS

We took PM2.5 prediction as the case study and evaluated the proposed approach regarding prediction accuracy and time complexity. The experimental results showed that our approach improved the accuracy by at least 6.9% up to 66.88% with an average of 28.32% compared to the existing ones. Moreover, by reducing the data size used in calculating the fitness, we can shorten the time complexity by 25.67% to 85.34%. while the prediction accuracy declines by only 2.97% on average. As a result, the framework could effectively work in any time series forecasting problem, especially when the number of features is large.

VI EVALUATION

We evaluate the effectiveness of our proposed method concerning two metrics: training time and prediction accuracy. The prediction accuracy is evaluated by MAE (Mean Absolute Error); the lower the MAE, the more accurate the prediction result. We aim to answer the following two research questions: (1) *How much does our GA-based feature selection method improve the prediction accuracy compared to existing approaches?* and (2) *How much does our training data generation strategy can shorten the training time while assuring prediction accuracy?*

VII CONCLUSIONS

This paper proposed a novel feature selection framework for time series prediction problems. The framework can be adapted for problems that require dimension reduction and feature selection. Besides, when it comes to explainable AI, our framework provides significant insights, which are guaranteed by the efficiency of genetic algorithms, within a short amount of time.

REFERENCES

[1] JADHAV, S., HE, H., AND JENKINS, K. Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing 69* (2018), 541–553.

[2] ORESKI, S., AND ORESKI, G. Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert Systems with Applications* (2014), 2052–2064.

DOI: https://doi.org/10.1007/978-3-031-21743-2_35

Keywords: GDPR, semantic parsing, logic representation

Learning to map the GDPR to Logic Representation on DAPRECO-KB

Minh-Phuong Nguyen, Thi-Thu-Trang Nguyen, Vu Tran, Ha-Thanh Nguyen, Le-Minh Nguyen, Ken Satoh

{phuongnm,trangttn,nguyenml}@jaist.ac.jp,vutran@ism.ac.jp,{nguyenhathanh,ksatoh}@nii.ac.jp

SIMPLIFIED TITLE

Construct a Semantic Parser that map the GDPR content to Logic Representation on DAPRECO Knowledge Base.

ABSTRACT

General Data Protection Regulation (GDPR) is an important framework for data protection that applies to all European Union countries. Recently, DAPRECO knowledge base (KB) which is a repository of if-then rules written in LegalRuleML as a formal logic representation of GDPR has been introduced to assist compliance checking. DAPRECO KB is, however, constructed manually and the current version does not cover all the articles in GDPR. Looking for an automated method, we present our machine translation approach to obtain a semantic parser translating the regulations in GDPR to their logic representation on DAPRECO KB. We also propose a new version of GDPR Semantic Parsing data by splitting each complex regulation into simple subparagraph-like units and re-annotating them based on published data from DAPRECO project. Besides, to improve the performance of our semantic parser, we propose two mechanisms: *Sub-expression intersection* and *PRESEG*. The former deals with the problem of duplicate sub-expressions while the latter distills knowledge from pre-trained language model BERT. Using these mechanisms, our semantic parser obtained a performance of 60.49% F1 in sub-expression level, which outperforms the baseline model by 5.68%.

I INTRODUCTION

General Data Protection Regulation[1] is the regulation on protecting EU citizens regarding processing personal data. The DAPRECO knowledge base is a repository of if-then rules representing the regulations in GDPR has been introduced. DAPRECO KB uses the Privacy Ontology (PrOnto), which models legal concepts in GDPR and provides additional concepts needed to represent the semantics of the legal rules in GDPR. The challenge of constructing a semantic parser for logic representation on DAPRECO KB comes from its constraints in the legal domain. To approach the challenging task, we split a complex GDPR statement into simple legal rules and then build a model to generate a logical representation of these simpler rules, inspired by the research on Semantic Parsing and Question Answering dealing with complex sentences.

II STATE OF THE ART

To our best knowledge, we are the first to construct a semantic parser mapping GDPR points to its logic representation on DAPRECO KB. It is difficult to directly map a complex GDPR rule into its original logic expression consisting of multiple logic formulae. To deal with the task of mapping a GDPR statement into its logic representation on DAPRECO KB, we apply the solution of the Semantic Parsing task in Natural Language Processing (NLP). With the approach using Intent Classification and Slot Filling, each logic representation is considered a semantic frame with a defined set of intent and slot information. This method requires annotated data to contain the label of slot information and intent type for each sample, which is difficult to extract from the GDPR data. A more flexible approach is using Neural Machine Translation (NMT); the semantic parser can be adapted to any logic representation syntax.

III ORIGINAL CONTRIBUTION

We constructed two versions of the GDRP Semantic Parsing dataset. The first version of the dataset (**Original data**) consisting of 275 samples is constructed from the current version of the DAPRECO KB. One sample is a pair of GDPR statements and their logic expression. To assist in solving the task of mapping a complex GDPR rule into its logic expression, we constructed a second version of the dataset called **Relaxation data**. In this version, a

[1] https://gdpr-info.eu/

complex sample is split into simple subparagraph-like units. We propose a Sub-expression intersection mechanism to avoid duplicate sub-logic expressions in its logic formulae. We propose PRESEG (*i.e.*, **P**redicate **RE**trieval & **S**ub-**E**xpression **G**eneration) mechanism, which consists of two steps. First, we utilize the power of the pre-trained language model BERT to retrieve well-relevant predicates. After that, we apply a Transformer-based NMT model to generate sub-expressions for each predicate instead of generating the logic representation for the whole GDPR statement, which results in a more correct syntax of logic representation.

IV Methodology

We use Transformer architecture as our strong baseline model. In our proposed mechanism, PRESEG, we utilize the power of the pre-trained language model BERT to support the expression parsing process. The parsing process is split into two steps: (1) *Predicate Retrieval*. This step uses a BERT retrieval model to generate a set of predicates related to an input GDPR statement. In detail, we construct a vocabulary of predicates ($\mathcal{V}^{predicate}$) from the training data, then fine-tune the pre-trained BERT model to predict the relation between text input and each predicate; (2) *Sub-expression Generation*. With each predicate generated from the previous step, we concatenate it with the GDPR statement to generate corresponding sub-expressions using the NMT model. After that, all generated sub-expressions are combined to present the final expression.

V Results

Our *Sub-expression intersection* mechanism effectively deals with the duplicate sub-expressions of DAPRECO KB logic. By using this mechanism, the performance of the baseline model using the Transformer increases on both two versions, Original and Relaxation data. By using the *PRESEG* mechanism on Relaxation data, our proposed model achieved the highest score on the Relaxation data. It boosts the performance on full logic representation to 60.23%, F1 increased by 2.47% compared to the end-to-end baseline model using Transformer.

VI Evaluation

To evaluate our proposed method, we use the F1 score in the sub-expression level which is computed based on the precision and recall score of the output triples compared with the gold triples. Based on our observation, we found that if the model can learn well the constraints between variables in each sub-logic expression according to each predicate, the performance of our model can increase a lot. Therefore the problem of generating correct constraints between variable and predicate requires important future work.

VII Conclusions

In this paper, we propose an effective semantic parser for mapping GDPR to the corresponding logic representation on DAPRECO KB. Firstly, we create Relaxation data for this task by splitting and re-annotating the complex regulation. Secondly, we introduce the Sub-expression intersection mechanism to solve the problem of the generation of duplicate sub-logic expressions. Last but not least, we demonstrate how PRESEG mechanism utilized the power of the pre-trained language model BERT and the Transformer-based NMT model to generate the basic part in the logic representations. Our semantic parser will be beneficial in tasks such as mapping other legal rules to logic representations.

DOI: https://doi.org/10.1007/978-981-19-8234-7_32

Keywords: pattern mining, inter-transaction pattern mining

dMITP-Miner: An Efficient Method for Mining Maximal Inter-transaction Patterns

Thanh-Ngo Nguyen [0000-0003-3137-8308]

thanh-ngo.nguyen@pwr.edu.pl

Simplified Title

An Efficient Method for Mining Maximal Inter-transaction Patterns

Abstract

In this paper, we propose an efficient algorithm, namely dMITP-Miner, for mining frequent maximal inter-transaction patterns (FMITPs). The proposed method uses diffset to store the information of patterns for efficiently min-ing FMITPs. In addition, we also proposed effective pruning strategies to help in reducing the search space to speed up the runtime and to cut down the memory usage. Experiments have been conducted to compare the effectiveness between the dMITP-Miner and the tMITP-Miner method in terms of runtime and memory usage.

I Introduction

Our proposed algorithm has the following main contributions. We apply our proposed strategies presented in [1] and [2], which help fast pruning infrequent 1-patterns. Then, dMITP-Miner algorithm applies those strategies to reduce the search space in order to quickly find all frequent inter-transaction patterns (FITPs) at the 1-pattern level with their tidsets. Next, the proposed algorithm uses DFS (Depth First Search) traversing to generate all FMITPs with their diffsets. The tidsets of FMPs are no longer used from the 2-pattern level onwards, diffsets are used to store the information of the patterns instead. Finally, experiments are conducted to prove the effectiveness between dMITP-Miner and tMITP-miner [3] algorithms in terms of runtime and memory usage.

II State of the Art

So far, many algorithms have been proposed for mining frequent maximal patterns of items occurring within transactions, but there is no method for mining frequent maximal patterns (FMPs) across transactions in the database. In this study, we propose an efficient algorithm, namely dMITP-Miner, for mining frequent maximal inter-transaction patterns (FMITPs).

III Original Contribution

One of the main types of condensed representation is frequent maximal patterns (FMPs). Although a set of FMPs is a subset of FPs, it still has the necessary properties for generating essential association rules and helps significantly reduce the search space, computation time, and memory usage. So far, many methods have been proposed for mining FMPs, consisting of MaxMiner, DepthProject, GenMax, dGenMax, MAFIA, TDM-MFI, and INLA-MFP.

IV Methodology

We propose an efficient algorithm, namely dMITP-Miner, for mining frequent maximal inter-transaction patterns (FMITPs). The proposed method uses diffset to store the information of patterns for efficiently mining FMITPs. In addition, we also proposed effective pruning strategies to help in reducing the search space to speed up the runtime and to cut down the memory usage. Experiments have been conducted to compare the effectiveness between the dMITP-Miner and the tMITP-Miner in terms of runtime and memory usage.

V Results

Many applications of inter-transaction pattern mining have been developed in recent years, such as the inter-transaction association rules used to predict stock market movements; the use of inter-transaction association rules for multi-dimensional contexts for prediction and their application to studying meteorological data; extended inter-transaction association rules to a more general form of association rules, called generalized multidimensional inter-transaction association rules, that can predict the rules like "*after McDonald and Burger King open branches, KFC will open a branch within two months, and between one and three miles away*"; and using inter-transaction association rules on financial market databases to develop an efficient application of profit rule mining based on the inter-day trading model.

VI EVALUATION

The algorithms applied in the experiments were written using Visual C# 2019 and tested on a computer with the following specifications: CPU Intel(R) Core(TM) i7-8565U processor @ 1.80 GHz, 20GB RAM, and running Windows 10. Experimental databases were downloaded from the Frequent Itemset Mining Dataset Repository (http://fimi.ua.ac.be/data).

We compare the proposed algorithm, dMITP-Miner, with the tMITP-Miner algorithm, regarding the mining time and memory usage. Through the experimental evaluations, we varied the parameters such as *minSup*, and *maxSpan*, in order to accurately assess the effectiveness of the algorithms used in the tests.describe main assumptions of a chosen method for evaluating your research (e.g. experiment, case study, etc.). Indicate, which criteria have been adopted in selecting this method. Describe which conclusions can be drawn from obtained results. The performance of the dMITP-Miner algorithm is better than that of the tMITP-Miner algorithm in terms of runtime and memory usage in the most cases.

VII CONCLUSIONS

In this paper, we proposed an efficient algorithm, dMITP-Miner, for mining frequent maximal inter-transaction patterns. The dMITP-Miner algorithm mines frequent maximal inter-transaction patterns based on the diffset to store mined patterns' information, for reducing computational cost and speeding up mining time.

REFERENCES

[1] T.-N. Nguyen, N. T. T. Loan, B. Vo, and N.-T. Nguyen, "An efficient algorithm for mining frequent closed inter-transaction patterns," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2019, vol. 2019-Octob. doi: 10.1109/SMC.2019.8914208.

[2] T. N. Nguyen, L. T. T. Nguyen, B. Vo, N. T. Nguyen, and T. D. D. Nguyen, "An N-List-Based Approach for Mining Frequent Inter-Transaction Patterns," *IEEE Access*, vol. 8, pp. 116840–116855, 2020, doi: 10.1109/ACCESS.2020.3004530.

[3] T.-N. Nguyen, N. T. T. Loan, B. Vo, and A. Kozierkiewicz, "Efficient Method for Mining Maximal Inter-transaction Patterns," in *Computational Collective Intelligence - 12th International Conference, {ICCCI} 2020, Da Nang, Vietnam, November 30 - December 3, 2020, Proceedings*, 2020, vol. 12496, pp. 316–327. doi: 10.1007/978-3-030-63007-2_25.

DOI: https://doi.org/10.1007/978-981-19-8234-7_33
Link to repository: https://bit.ly/it-career-bot

Keywords: Career Counselling, Personalized Chatbot, Context-Aware Knowledge Base

ITCareerBot: A Personalized Career Counselling Chatbot

Cuong Nguyen Duy, Hanh Dung Dinh Nguyen, Cuong Pham-Nguyen [0000-0002-7057-753X], Thang Le Dinh [0000-0002-5324-2746], Le Nguyen Hoai Nam

1753034@student.hcmus.edu.vn, 1753021@student.hcmus.edu.vn, pncuong@fit.hcmus.edu.vn, Thang.Ledinh@uqtr.ca, lnhnam@fit.hcmus.edu.vn

ABSTRACT

Nowadays, IT students and professionals often need additional knowledge and skills to fulfill market requirements and to target their professional goals to increase their opportunities for growth. This study aims at leveraging a specific type of chatbot to offer intelligent and personalized advising learning services to job seekers as learners by providing information and recommendations of a learning path according to the market trends and learners' profiles . Firstly, a chatbot framework is developed based on a context-aware knowledge model and a recommendation method. Thus, the context-aware knowledge base and its instance were built by analyzing different data sources collected from professional social networks and online education platforms. Furthermore, the recommendation method defines how the chatbot responds to user messages based on the matching between the current job seekers' skills and their career interests. Concerning research evaluation, the effectiveness of the proposed framework is validated by measuring the algorithms. Moreover, some efficient and satisfactory criteria were analyzed and evaluated based on the user feedback from a survey.

I INTRODUCTION

Career counselling is very important for everyone's career path, especially in the higher education sector. Since it helps students understand what they need in each occupation, hence students can decide to follow the job that they feel the most suitable for their personality, interests, abilities, and situation. However, those services still have faced some key issues such as the followings: i) the final year students do not have the right career orientation after graduation, struggled with choosing the right career path, not sure whether their chosen major is suitable for their desired future job; ii) lack of adaptation of the study program to each student's profile leads to a high dropout rate in universities because of losing motivation to learn. To explore this problem, our analysis was conducted in 2021 with 2,860 questions raised by users in the IT field, on the three social websites (Quora, Stack overflow, Stack exchange) shows that the need for asking a learning path of a specific occupation is absolutely dominant (83.4%), following by requesting skills required for an occupation (5.1%). In the context of such increasing demand, there is a need for AI-based solutions such as chatbots for offering career advice to IT professionals. Our study introduces an integrated model, which organizes the domain and context-aware knowledge for context-aware smart service systems [1]. It refines the architecture for chatbots and extends the knowledge representation models for a personalized career-counselling service, which is based on learner skills. Based on this architecture, a specific type of chatbots is proposed, called ITCareerBot for offering career-counselling services [2]. More specifically, ITCareerBot is designed for learners where they can ask for different information such as for career-related courses to achieve a job, for skills gained after completing a course, for time to study a course, and for recommended resources of a specific course. The study includes several parts such as analyzing the data to identify requirements, designing the data model and system architecture, implementing chatbot functionalities, and validating the approach by analyzing user surveys and performance testing.

II STATE OF THE ART

Several studies are conducted to develop chatbot applications that capture and handle contextualization and personalized content in different fields such as education, healthcare, and business transactions. They firstly focused on extracting context information, then presented them in two ways: using a knowledge model or a neural network model. Different methods are used to generate the responses such as machine learning models or pattern matching. Concerning career-counselling chatbots [3], one intelligent career-counselling bot was built that supports the users' decisions on their careers by advising on their future career. However, it has two limitations: i) the responses were built without personalizing the answers based on user profiles, and ii) no evaluation was carried out to show whether the chatbot was effective in the education environment. For this reason, this study aims at filling the gap by proposing an approach for building personalized chatbots for career-counselling services that considers the learner profile and context to provide individualized responses.

III ORIGINAL CONTRIBUTION

The main contributions of the study are as follows: i) An integrated context-aware knowledge base comprises an occupation and skills model, a course model, and a context model and its dataset. The models allow sharing and utilizing context-aware knowledge in a large range of applications thanks to semantic web technology. The dataset is collected and analyzed from social IT professional networks and online education platforms, therefore it represents the job requirements of the IT industry; and ii) A recommendation method, which is integrated into the response generation process, considers the learners' skills, occupation requirements, and course model to provide the relevant advising learning path to the learners.

IV METHODOLOGY

The main research question of this study is "*How to design and implement a personalized career-counselling chatbot based on a context-aware knowledge model?*". The exploratory research methodology is used to conduct the study to investigate the research question, which has not previously been studied in depth. First, the study analyses the data collected from social networks combined with an experimental method on the basis of expanding the knowledge model in our previous studies. Thus, it proposes a solution for building a counselling chatbot in the IT domain that considers a personalization of responses when answering the user's queries. This solution predicts the rating of a user for each targeted learning path concerning an occupation by using three combined factors: i) Weight of user skills covered in both the occupation and the learning path, ii) Weight of skills covered in both the occupation and the learning path, but have not yet been acquired by the user, and iii) Weigh of additional user skills not covered in the occupation.

V RESULTS

The key finding is a chatbot that is integrated a recommendation method to personalize learning paths for the IT field via a dialog fashion. It can serve as a framework to be potentially used in other domains by refining the knowledge base. In this case, the occupation model will be extended and the features of the context model and user model can be redefined to meet the new applications.

VI EVALUATION

Experiments are conducted to evaluate the chatbot based on the three criteria: Effectiveness, Efficiency and Satisfaction. This effectiveness is related to the performance of the algorithms for user intent identification and entity extraction on user messages. To perform the test, the test set is randomly taken 20% and the remaining 80% is the training set. Precision, Re-call, and F-score are used to evaluate the performance of the models. The average precision, recall and F1-score results is always greater than 85%. The efficiency is judged through the relevance of the answer. A survey was conducted on 27 users who gave feedbacks after using the chatbot: the majority of users didn't or rarely found wrong answers from the chatbot (85.2%), and the answers match their expectation. The satisfaction is related to the ease of use, the ease of reading and understanding. The user feedback that was easily understood and the conversations were easy to follow.

VII CONCLUSIONS

The proposed approach integrates a context-aware knowledge base and a recommendation method to generate the customized responses. The dataset was built based on the analysis of data collected from social and professional networks. The application of the approach can be integrated into e-learning or MOOC platforms which can be used in industries or educational institutes. It can be extended for other domains than the IT. Moreover, one of our future research projects aims at enhancing the approach for more complicated and elaborated counselling services, such as refining the recommendation method to take into account other context dimensions (e.g. time to learn a course, online/offline, course fee, etc.) to improve the responses. It is also foreseen to integrate the context-aware knowledge model with the current artificial intelligence techniques, such as deep learning and reinforcement learning, to create new contextual and more efficient career-counselling chatbots based on the current knowledge base.

REFERENCES

[1] LE DINH T., PHAM THI, T. T., PHAM-NGUYEN, C., AND NAM, L. *A knowledge-based model for context-aware smart service systems*. Journal of Information and Telecommunication, Aug. 2021, pp. 141-162.

[2] THI, P., DIEP, H., NGUYEN, T., PHAM-NGUYEN, C., LE DINH, T., AND NAM, L. *Towards An Ontology-Based Knowledge Base for Job Postings*. 7th NAFOSTED Conference on Information and Computer Science (NICS). Nov. 2020, pp. 267–272.

[3] LEE, T. et al. *Intelligent Career Advisers in Your Pocket? A Need Assessment Study of Chatbots for Student Career Advising* . In 25th Americas Conference on Information Systems, AMCIS 2019, Cancún, August 15-17, 2019. Association for Information Systems.

DOI: https://doi.org/10.1007/978-3-031-21743-2_49
Link to repository: https://github.com/toan01-uet/v3MFND

Keywords: Vietnamese fake news detection, multimodal, multi-domain

v3MFND: A Deep Multi-domain Multimodal Fake News Detection Model for Vietnamese

Cam-Van Nguyen Thi, Thanh-Toan Vuong, Duc-Trong Le, Quang-Thuy Ha

(vanntc,19021279,trongld,thuyhq)@vnu.edu.vn

Simplified Title

Vietnamese Multi-domain Multimodal Fake News Detection Model

Abstract

Fake news become a critical problem on the Internet, especially social media. During the worldwide COVID-19 epidemic, social networking sites (SNSs) are primary sources to spread false news, which are incredibly difficult to detect and regulate them since they rapidly grow everyday. With multimedia technology advances, the content of social media news now is manifested via various modalities, such as text, photos, and videos. Approaches that learn the multimodal representation for detecting fake news have evolved in recent years. Additionally, there exist diverse content domains in news platforms. Exploiting data from these domains potentially solve the data sparsity problem as well as simultaneously boosting overall performance. In this paper, we propose an effective Deep Multi-domain Multimodal Fake News Detection model for Vietnamese, **v3MFND** for short. Extensive experiments on a real-life dataset reveal that **v3MFND** improves the performance of multi-domain multimodal fake news detection for Vietnamese considerably. An ablation study is also carried out to evaluate the role of each individual modality in the multimodal model.

I Introduction

Fake news detection becomes a significant and urgent issue that requires attention since the dissemination of fake news over the Internet has become a modern scourge, especially through social networks in the news media. Social media news now includes information in several modalities, such as text, images, and videos. The question of how to learn a joint representation that includes multimodal information has sparked a lot of interest in the scientific community. Furthermore, news platforms provide a variety of news in many domains in real-world scenarios. Levering data from these domains may help alleviate the data sparsity problem while also improving the performance of all domains. These mentioned points motivate us to exploit the fake news detection problem in the direction of multi-domain multimodal methods for Vietnamese.

II State of the Art

II.1 Unimodal to multimodal fake new detection

Methods for detecting fake news have steadily progressed from unimodal to multimodal techniques in recent years. SpotFake (Shivangi et al., 2020) is a multimodal framework for fake news detection using BERT. Song et al.(2021) leverage an attention method to fuse a number of word embeddings and one image embedding to create fused features, and then extract essential features as a joint representation. Generally, the modeling multiple modalities is efficient to improve the fake news detection performance.

II.2 Multi-domain fake news detection

MDFEND (Quiong Nan et al.,2021) model uses domain gate to combine different representations retrieved by mixture-of-experts in order to cope with multi-domain transfer and isolation. The limitation of MDFEND is just use the text feature while ignoring the connected images in each news. For Vietnamese, Tuan, N.M.D (2021) proposed a scaled dot-product attention mechanism to capture the relationship between text features extracted by a pre-trained BERT model and visual features extracted by a pre-trained VGG-19 model.

III Original Contribution

We build a multi-domain, multimodal fake news dataset for Vietnamese named **M2-ReINTEL** up on the ReINTEL dataset, whereby assigns domain names to news items. Subsequently, we propose a multi- domain, multi-method fake news detection framework for Vietnamese named v3MFND using advanced deep learning models to extract features for multi-modal data (text, images) and multi-domain data problem solving.

IV METHODOLOGY

The v3MFND mainly consists of four components: pre-processing phase, modalities representation learning , fusion layer, and a classification layer. The experiments is implemented on dataset M2-ReINTEL including 4825 news, which only 3583 news have visual features. All of news are manually labeled with ten domains specifically *science, health, politics, education, economics, disaster, military, sports, entertainment, society*. The difference in fake/real labels, the quantity of articles between domains, and the heterogeneity in the number of photos all contribute to this imbalance.

IV.1 Proposed model

Firstly, The input dataset is fed through a separate pre-processing phase for each modality. Text data will be cleaned, normalized, then segmented using VNCoreNLP and tokenized using PhoBERT Tokenizer.

Secondly, instead of a vector representation of the text, we leveraged PhoBERT to extract word embeddings contain textual information. In order to extract the news' representations for multiple domains, we use multiple experts networks (TextCNN) in our model. Domain gate with the domain embedding as well as sentence embedding also is employed as input to guide the selection process.

Later, the VGG-19 model is used to learn different visual features. We extract the output of the VGG-19 convolutional network's second last layer, which is passed via a fully connected layer to obtain final visual representation.

In the fusion layer, the two feature vectors received from separate modalities are fused together using a simple concatenation approach. It is not only combines different visual features, but also reflects the dependencies between textual features, visual features and meta-data in the same news. Additionally, we also conducts experiments with some other combination methods besides concat, which are addition and average.

Lastly, the final feature vector of the news is fed into the classifier, which is a Multi-layer Perception network with a SoftMax function to make prediction. The fake news detector's purpose is to determine whether or not the news is fake, the loss function is set to Binary Cross-Entropy.

We conducted experiments using M2-ReINTEL on all 10 domains and the 5 domains with the largest amount of labels due to the imbalance of data regarding domains and labels. After tuning the proposed model on various settings to find the optimal value, the final settings are found separately for each of the above experiments. We report AUC-ROC as the performance measure.

IV.2 Baseline

We seek to compare against two baseline models namely MDFEND (multi-domain model) and SpotFake (multimodal model). Additionally, we conduct more experiments to evaluate the role of metadata, which is also mined by any competing teams at VLSP Shared Task 2020 in the model. We also examine different fusion operations, such as average, concat and sum in our model.

V RESULTS

In the experiment for 10 data domains, MDFEND and v3MFND give better overall results than the SpotFake model. In addition, with the use of more metadata in the v3MFND model, the model not only ineffective, but also reduced the performance of the model. Besides, in the v3MFND model, in most domains, the 'concat' fusion gives higher results than other fusion methods. The MDFEND model scored 0.9753 and 0.9548 in the Disaster and Social domains, respectively, whereas v3MFN scored 0.9294, which was lower than MDFEND's 0.0254. (2.6%). The SpotFake model received the highest score of 1.0 in the Ecocomics domain. The rationale is similar to that of Education and Health.

In the experiment for 5 most popular domians, our proposed model v3MFND as concat fusion gives the best results on most domains and on the entire evaluation set. Specifically, on the Disaster, Social, Economics and Health domain, the v3MFND model achieved 0.9754, 0.8858, 1.0 and 1.0, respectively, with the AUC–ROC measure. As for the entire evaluation set, the v3MFND model reached 0.9375.

VI EVALUATION

From the above results table, we can also see that MDFEND and v3MFND models with multi-domain problem handling give better results in most domains and overall better results than SpotFake model. In addition, with the use of more metadata on the v3MFND model, the overall performance of the model was reduced, but on some domains such as Politics, our model gave the highest result with a precision of 0.9583

VII CONCLUSIONS

Multi-domain multimodal fake news detection model can be applied to social networking or digital news sites, helping to warn users. Early automatic fake news detection helps prevent the spread of misinformation, which leads to unintended consequences, especially in specialized data domains

Link to arXiv: `https://arXiv.org/abs/2104.12473`
DOI: `https://doi.org/10.1007/978-981-19-8234-7_34`

Keywords: Multi-tiered Persistent Storage, Scheduling, Parallel Data Processing, Performance Tuning, Database Management Systems

Scheduling Parallel Data Transfers in Multi-tiered Persistent Storage

Nan Noon Noon[0000-0003-3985-5455], Dr. Janusz R. Getta[0000-0001-6492-5641], Dr. Tianbing Xia[0000-0002-4520-5021]

nnn326@uowmail.edu.au, jrg@uow.edu.au, txia@uow.edu.au

Simplified Title

Scheduling Parallel Data Transfers

Abstract

Multi-tiered persistent storage provides a logical view where all available storage is distributed over a number of levels of different speeds and capacities. Efficient scheduling of parallel data transfers in multi-tiered persistent storage is a significant problem for pipelined data processing. This work considers a class of database applications implemented as sequences of operations that transfer data between persistent storage tiers. We show how to partition the sets of data transfers to reduce the number of conflicts when data transfers are performed in parallel. The paper proposes the new rule-based algorithms for allocating parallel data transfer to the processors to minimize total processing time. The new algorithms evenly distribute the workload among the processors and reduce their idle times. We describe a number of experiments that validate the efficiency of parallel data transfer plans generated by the algorithms presented in the paper.

I Introduction

In a multi-tiered view, data are distributed over many different levels of persistent storage with various capacities and performance characteristics. Data processing in multi-tiered persistent storage is performed while data are transferred between the levels. These are simultaneously running data processing applications that compete for access to the highest levels, where the performance of input/output operations is the best. Efficient resource allocation and scheduling in multi-tiered persistent storage is a significant challenge for efficient storage utilisation [1]. In this work, we look at the problem of efficient scheduling of parallel data transfers [2, 3] between the levels of multi-tiered persistent storage.

II State of the Art

Our earlier research contributed to the invention of automated performance tuning plans with materialisation and indexing of a single layer of multi-tiered persistent storage. In the following research, we proposed a new resource allocation algorithm over multiple layers of multi-tiered persistent storage and presented a new method to discover the query processing plans for predicted workload using a new cost model.

This paper presents new algorithms for efficient resource scheduling of parallel data transfers between the levels of multi-tiered persistent storage. We consider a pipelined data processing model wherein streams of data simultaneously pass through processors located in the nodes of an acyclic-directed graph. The processors read multiple data streams, perform operations on data, and output the data to the following processors in the pipelines. We assume that data can be simultaneously read and written from/to the different multi-tiered persistent storages. Thus, whenever it is possible, the processing of data is performed in parallel.

III Original Contribution

The research contributions of the paper are the following:

1. Database applications are implemented as sequences of operations on data. We show how to convert the sequences of operations into sets of data transfers between the levels of multi-tiered storage.

2. We show how to partition the sets of data transfers to reduce the total number of conflicts between the data transfers.

3. We propose new rule-based algorithms for the allocation of data transfer to the processors to minimise total processing time, to evenly distribute the workload among the processors, and to reduce the idle time of the processors.

4. We show how the pipelined data processing model can be efficiently implemented in a multi-tiered persistent storage model.

IV Methodology

In this paper, we have created a new rule-based algorithm for the efficient scheduling of parallel data transfers through the application of a methodology called *Combination of Scheduling Rules (CSR)*. The scheduling rules are systematically added to the algorithm one by one. Each time a new rule is added to the algorithm, we conduct performance evaluations to estimate the improvements and to identify potential weaknesses. The weaknesses of the algorithm identified at one step contributed to the creation of a new scheduling rule in the next step. CSR is a mixture of theoretical and experimental approaches to the scheduling of parallel data transfers. This methodology allows for the addition of more scheduling rules for handling special cases whenever necessary.

V Results

The application of *CSR* approach with the *Shortest Processing Time (SPT)*, *First Come, First Served (FCFS)*, *Random scheduling (R4)*, and *Optimal Resource Allocation Plan ORAP* scheduling rules required 256 time units to process the given datasets. On the other hand, the application of *SPT*, *FCFS* and *R4* scheduling required 287, 286, and 311 time units, respectively. The Optimal Allocation Plan requires 250 time units. In summary, the application of our scheduling rules yields results close to the optimal one. A combination of scheduling rules yields better results in most experiments than the other existing scheduling methods. However, the algorithm presented in the paper does not always generate the optimal solution. Unfortunately, the generation of optimal plans takes too long to be used in practice. For example, in an extreme case, the time spent on finding the optimal solution can be longer than the processing time of all tasks. Thus, the algorithm compromises the quality of the parallel data transfer plans and the time spent on generating such plans.

VI Evaluation

For the experiments, we created 12 testing datasets. To create a dataset, we found an optimal solution first, then reverse-engineered an optimal solution to generate a set of queries and a set of sequences of transfers. The optimal solution, *Optimal Resource Allocation Plan (ORAP)*, was created by arranging several transfers on available processors such that the allocation was well-balanced on those processors and there were no idle times.

We used several scheduling methods for each dataset and compared the resulting processing time. The existing scheduling methods are *Shortest Processing Time (SPT)*, *First Come, First Served (FCFS)*, and *Random* scheduling *(R4)*. The proposed method, *CSR*, balances the current workload among many processors and minimises both the idle time of the processors and the overall transfer time.

VII Conclusions

The various persistent storage devices available on-site or in the clouds constitute the global *multi-tiered view* of persistent storage. Therefore, the outcomes of this research can be used for managing the cloud and distribution of resources across the world.

References

[1] Blazewicz, J., Ecker, K., Pesh, E., Schmidt, G., Sterna, M., and Weglarz, J. *Handbook on Scheduling From theory to Practice. 2nd edn.* Springer, 2019.

[2] Frachtenberg, E., Feitelson, G., Petrini, F., and Fernandez, J. Adaptive parallel job scheduling with flexible coscheduling. *IEEE Transactions on Parallel and Distributed Systems 16*, 11 (2005), 1066–1077.

[3] Yanyong, Z., Franke, H., Moreira, J., and Sivasubramaniam, A. An integrated approach to parallel scheduling using gang-scheduling, backfilling, and migration. *IEEE Transactions on Parallel and Distributed Systems 14*, 3 (2003), 236–247.

DOI: –

Keywords: neural networks, sensitive information, preserving privacy

Detecting Sensitive Data with GANs and Fully Convolutional Networks[1]

Marcin Korytkowski[0000-0002-6002-2733], Jakub Nowak[0000-0002-1572-3426], Rafał Scherer[0000-0001-9592-262X]

marcin.korytkowski@pcz.pl, jakub.nowak@pcz.pl, rafal.scherer@pcz.pl

Simplified Title

Detecting Sensitive Data with Neural Networks

Abstract

The article presents a method of document anonymization using generative adversarial neural networks. Unlike other anonymization methods, in the presented work, the anonymization concerns sensitive data in the form of images placed in text documents. Specifically, it is based on the CycleGAN idea and uses the U-Net model as a generator. To train the model we built a dataset with text documents with embedded real-life images, and medical images. The method is characterized by a very high efficiency, which enables the detection of 99.8% of areas where the sensitive image is located.

I Introduction

Many entities and private persons want to protect their data against leakage. It can be information about both health and company secrets, e.g. research works. The subject of the processing of sensitive data is also extremely important in the context of EU regulations, e.g. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data and the criminal and financial liability of persons creating and processing such collections of information. It is also worth noting that the theft of sensitive data may be used to assess the health condition of politicians or other decision-makers. We present a system for detecting documents containing sensitive data, also in the cases where they are intentionally hidden there.

For obvious reasons, the classification of data into one of two classes: with and without sensitive data must be automatic. In a situation where nowadays even small entities process gigabytes of information daily, a human is not able to manually verify the content of processed files. In this article, we propose a solution that fully automates this process based on machine learning techniques. The task facing the system is to detect and remove sensitive data from the documents being processed. It is based on the CycleGAN idea and uses the U-Net fully convolutional neural model as a generator. To train the model we built a dataset with text documents with embedded real-life images, and medical images.

II State of the Art

Currently, anonymization with the use of neural networks is used primarily to detect specific phrases in the text. Unlike the text, we will not analyze the context of the text, but the actual information contained in a visual form of images. The detection of sensitive data such as the human face has been very well developed, among others, thanks to the vast amount of data available on social network channels. In the studies cited, very good results were achieved with the use of convolutional neural nets. Our work is aimed at detecting the places of occurrence of sensitive data, such as: results of magnetic resonance imaging, X-rays, etc. However, in the case of this type of data, we usually encounter the problem of small amounts of data available for training. One of the ways to improve the operation of classification algorithms is by generating synthetic data on the basis of the available set [6]. In our research, we tackle the presented problem differently. We want a generative adversarial neural network (GAN) [3] to be able to distinguish between sensitive data itself. To put it simply, the GAN is supposed to generate images without sensitive data.

The presented solution is based on the CycleGAN network model [9]. In its original application, the network was designed to convert graphic images. Among other things, they trained the model to convert horse images to zebra images, and city landscapes at night to city landscapes by day. The great advantage of CycleGAN is that this model can be trained without paired examples, i.e. it does not require sample photos before and after conversion to train the model. For example, it is not necessary to provide the same image of a horse turned into a zebra. The model architecture consists of two generator models: one generator (GeneratorA) for generating images for the first class (ClassA) and a second generator (GeneratorB) for generating images for the second class (ClassB):

GeneratorA → ClassA (documents requiring anonymization),
GeneratorB → ClassB (documents that do not require anonymization).

Generator models perform image translation, which means that the image conversion process depends on the input image, particularly an image from another domain. Generator-A takes an image from ClassB as the input and ClassB takes an image from ClassA as the input:

ClassB → GeneratorA → ClassA,
ClassA → GeneratorB → ClassB.

Each generator has its own dedicated discriminator model. The first discriminator model (DiscriminatorA) takes the true images from ClassA and the generated images from GeneratorA and predicts whether they are true or false. The second discriminator model (DiscriminatorB) takes the true images from ClassB and the generated images from GeneratorB and predicts whether they are true or false. The discriminator and generator models are trained in an adversarial zero-sum process, much like normal GAN models. Our solution uses also the U-Net model [7] as a generator, which was initially used, inter alia, to detect neoplasms in medical images. The architecture stems from the so-called "fully convolutional network" [5].

Generators learn to cheat discriminators better, and discriminators learn better to detect false images. Together, the models find equilibrium during the training process. In addition, generator models are regulated not only to create new images in the target class, but also to create translated versions of the input images from the source class. This is achieved by using the generated images as input to the appropriate generator model and comparing the output image with the original images. The transmission of the image by both generators is called a cycle. Together, each pair of generator models is trained to better recreate the original source image, which is referred to as cycle consistency.

III Original Contribution

Our solution uses the CycleGAN network model [9] and the U-Net model [7] as a generator with data structure described in the next section.

IV Methodology

The training sequence was built on the basis of text documents with embedded medical data. Now we will describe the input data. The training string for the GAN requires two classes of objects — that require anonymisation and do not require anonymisation. The non-anonymised part contained only a grayscale image obtained from a text document (e.g. WORD, PDF) along with random images from the ImageNET dataset [2] inserted in random places. The part requiring anonymisation was additionally provided with data in the form of X-ray images and computed tomography. The image was saved as a JPEG file with a size 1024 × 1024 pixels. The portion of documents that did not required anonymisation did not contain images from the ImageNET database. For any document, the size of the image could not be greater than 80% of the width and height, and less than 20%. The inserted images were additionally scaled by a random factor.

The output of the GAN is also a grayscale image with a size of 1024 × 1024 pixels. In the GAN model, the discriminator as the evaluating part of the generator is also important.

In the conducted research, synthetic data requiring anonymisation were inserted in random places. The network in the learning process retrieved document scans without inserted sensitive images as ClassA. As ClassB, the network received documents with sensitive (e.g. medical) images.

On the basis of the conducted research, the network, apart from detecting the document with the newly generated sensitive image, was able to remove the sensitive image from text documents without losing such elements as a stamp or a barcode.

To prepare input data with sensitive medical objects we used our own X-ray images and images taken from the datasets described in [4], [1] and [8]. Eventually, the composition of the training dataset was as follows:
the number of files without images: 20,100,
the number of files with images from the ImageNet: 15,000,
the number of files with sensitive medical images: 15,000.

V Results

Through the operation of the GAN structure and the proprietary solution of inserting medical photos into the content of various text files, the first class of documents (containing sensitive data) was defined, which consisted of 30,000 graphic files. The second class of documents was created by artificially generating graphic files on the basis of the public ImageNET database (a total of 30,000 randomly selected files) of files.

The first discriminator model (DiscriminatorA) takes the true images from ClassA and the generated images from GeneratorA and predicts whether they are true or false. The second discriminator model (DiscriminatorB) takes the true images from ClassB and the generated images from GeneratorB and predicts whether they are true or false. The discriminator and generator models are trained in an adversarial zero-sum process, much like the standard GAN models. An extremely interesting feature of the trained GAN network, as described above, is that it can be used to remove sensitive medical data (a kind of graphical anonymisation). The percentage of correctly selected sensitive data was 99.8%. The percentage of incorrectly selected insensitive data was 29.3%. That was calculated based on the number of similar pixels.

VI Conclusions

We proposed a method to anonymise documents using generative adversarial neural networks and fully convolutional networks. The anonymisation concerns sensitive data in the form of images placed in text documents. It is based on the CycleGAN idea and uses the U-Net model as a generator. To train the model we built our own dataset with real-life MS Word and PDF text documents with embedded real-life images, and medical images. The method is characterized by a very high efficiency, which enables the detection of 99.8% of areas where the sensitive image is located.

Acknowledgements

The work was supported by The National Centre for Research and Development (NCBR), the project no POIR.01.01.01-00-1431/19.

References

[1] Cohen, J. P., Morrison, P., and Dao, L. Covid-19 image data collection. *arXiv 2003.11597* (2020).

[2] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (2009), Ieee, pp. 248–255.

[3] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems* (2014), Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, Curran Associates, Inc.

[4] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell 172*, 5 (2018), 1122–1131.

[5] Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3431–3440.

[6] Röglin, J., Ziegeler, K., Kube, J., König, F., Hermann, K.-G., and Ortmann, S. Improving classification results on a small medical dataset using a gan; an outlook for dealing with rare disease datasets. *Frontiers in Computer Science* (2022), 102.

[7] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (2015), Springer, pp. 234–241.

[8] Tschandl, P., Rosendahl, C., and Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data 5*, 1 (2018), 1–9.

[9] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2223–2232.

DOI: https://doi.org/10.1007/978-3-031-21967-2_8

Keywords: clustering, outlier detection, clustering quality indices.

The quality of clustering data containing outliers

Agnieszka Nowak - Brzezińska[0000-0001-7238-1170], Igor Gaibei[0000-0002-4708-9036]

agnieszka.nowak-brzezinska@us.edu.pl, igor.gaibei@us.edu.pl

Simplified Title

The quality of clustering data containing outliers.

Abstract

This article evaluates the efficiency and performance of both clustering algorithms: an agglomerative hierarchical clustering *AHC* (with various linkage options and distance measures) and the $K-Means$ algorithm. We assess the quality of clustering using Davies-Bouldin and Dunn cluster validity indices. Our goal is to compare and analyze outlier detection algorithms depending on the applied clustering algorithm. We also wanted to verify whether the quality of clusters without outliers is higher than of those with outliers. In our research, we compare the *LOF* (Local Outlier Factor) and *COF* (Connectivity-based Outlier Factor) algorithms for detecting outliers (selecting 1%, 5%, and 10% of the most outlier instances in a given dataset). Next, we analyze how clustering quality has improved after excluding such outliers. In the experiments, three real datasets were used with a different number of instances. We wanted to investigate whether it is essential what clustering algorithm and outlier detection method we will use? Our goal was to check whether the clustering parameters impact the obtained clustering results. To the best of our knowledge, no research would combine these issues in one study.

I Introduction

Data clustering is one of the most effective tools for dealing with large amounts of data. When there is a lot of data, we cannot manage it or extract valuable knowledge. By creating clusters of similar data, we naturally divide large datasets into homogeneous groups, which allows us to quickly search for groups of objects best suited to what we are currently looking for. Clusters have got representatives reflecting the cluster's content. As we search only within the representatives of these clusters, we may omit the most relevant data even if they exist in this dataset. That is why it is essential to verify the quality of the clusters. There are many possible quality indices for measuring the quality of created clusters. We decided to use the two most popular: *Dunn* and *Davies-Bouldin* indices. Intuitively, if there are natural data clusters (well-separated data) in the dataset, the clustering quality should be high and easy to achieve. When the dataset contains outliers, the clustering process is more complicated and time-consuming and does not guarantee that the partition made is optimal. It affects the clusters undoubtedly. The clustering quality of the dataset should be higher after removing the outliers. We decided to test two clustering algorithms with different parameters because, as we know, they always have a significant impact on clustering quality. Our goal is to check which clustering parameters create clusters with better quality, which method of detecting outliers causes outliers to be detected and whether the input data affects the effectiveness of detecting outliers and improves the quality of clusters after removing outliers. Does the character of the input data influence the effectiveness of the clustering process or outlier detection process? Finally, we will confirm that the more outliers we detect, the more the quality of the clusters improves.

II State of the Art

In the literature, we can find many papers on either the comparison of the $k-Means$ and the *AHC* algorithm, different distance measures or methods of combining clusters, methods of detecting outliers, or finally, methods of analyzing the quality of clustering. None of them answered the questions introduced in this paper.

III Original Contribution

The original contribution of the authors of this work is the implementation of clustering algorithms with outlier detection algorithms for datasets containing potential outliers and cluster quality analysis. This study assesses the quality of the clusters (using dedicated indices for assessing the quality of the clusters) before detecting potential outliers in the data and after their extraction from the knowledge base. The study also aimed to assess the impact of clustering parameters and deviation detection on the quality of clusters. The research confirmed that the quality of the clusters improves after removing the deviations.

IV Methodology

The experiments aimed to check the impact of clustering algorithms, clustering methods, and selected distance measures on the effectiveness of outlier detection, measured by the cluster quality assessment. Our methodology is based on the following steps. First, we cluster all objects in a given dataset and then discover potential outliers. Such outliers are excluded from the dataset, and the remaining objects are clustered. We compare the cluster quality assessment for the whole dataset and the dataset without discovering outliers. We used two indexes for cluster quality assessment: Dunn and Davies-Bouldin indices. We performed experiments on three different real datasets. We changed the number of detected outliers three times, using 1%, 5%, and 10% of the entire dataset as the number of outliers. Six hundred eighty-six experiments were performed in total. In the case of the K-Means algorithm, the K parameter was changed (various numbers of clusters, depending on the size of the set). In the case of the AHC algorithm, the method of combining clusters (singles, complete, average) and a measure of distance (Euclidean, Chebyshev) was changed. This should be multiplied by experiments for the LOF and COF methods and three different values of the outlier numbers, i.e., 1%, 5%, and 10%.

V Results

The thesis that if the dataset contains outliers, it affects the quality of the clusters was confirmed. By removing the outliers we are able to form good-quality clusters from the data. We will also reduce clustering time (as there is no longer any difficulty in a cluster formation). Consequently, the better quality of the created clusters means better-quality of the discovered knowledge. The conducted research concludes that the COF algorithm contributes more often to the improvement of the quality of clusters than LOF after removing the deviations indicated by these algorithms. Considering the clustering algorithms, it turns out that the $K - Means$ algorithm responds much more often, eliminating outliers by improving the quality of clusters, which is probably because it is much less resistant to the appearance of outliers in the set than the AHC algorithm.

VI Evaluation

The research confirmed the original assumption that the more outliers we remove from the set, the better the quality of the clusters would be. However, an essential conclusion seems to be that the type of input data significantly affects the results achieved: the quality of the clusters created for data containing potential outliers.

VII Conclusions

Potential applications of clustering and outlier detection algorithms are extensive. We still need to discover new knowledge in such essential areas of life as medicine or the economy (business). This would not be possible without unsupervised learning algorithms, of which the best examples are clustering algorithms. We have a lot of data in a specific field, and we want to extract valuable knowledge from them. We need data analysis tools that will be able to create good cluster quality (consistent) and discover unusual data as potentially interesting information worth further analysis.

References

[1] Kishan, G.M.; Chilukuri, K.M.; HuaMing H.: Anomaly detection principles and algorithms. Springer, 23–38, (2017)

[2] Ranga Suri N.N.R.: Narasimha, Murty M.; Athithan, G.: Outlier Detection: Techniques and Applications. Springer AG, 3–9, (2019)

[3] Legany, C.; Juhasz, S.; Babos, A.: Cluster validity measurement techniques, Knowledge Engineering and Data Bases, WSEAS, USA, 388–393, (2006)

DOI: `https://doi.org/10.1007/978-3-031-21967-2_36`

Keywords: NDVI, clustering, management zones, precision agriculture, geospatial data visualization

Clustering analysis applied to NDVI maps to delimit management zones for grain crops

Aliya Nugumanova[0000-0001-5522-4421], Almasbek Maulit[0000-0002-0519-3222],
Maxim Sutula[0000-0002-3153-6356]

yalisha@yandex.kz,maulit.almas@yandex.ru,max.sutula@gmail.com

SIMPLIFIED TITLE

NDVI clustering for precision agriculture.

ABSTRACT

This research studies the possibility of applying data mining methods to determine homogeneous management zones in fields sown with cereals. For the study, satellite images of two fields in the East Kazakhstan region were used, obtained by the Sentinel-2 satellite in different periods of time (images of the first field were obtained from May to September 2020, images of the second field - from May to August 2021). Based on these images, a dataset of seasonal NDVI values was formed for each field. Four different clustering algorithms were applied to each of the datasets, the clustering results were visualized and rasterized as color maps, which were then offered for comparison and verification by an expert agronomist. Based on the expert review, recommendations were formulated for determining zones of homogeneous management.

I INTRODUCTION

In this research, we apply cluster analysis to the normalized difference vegetation index (NDVI) data obtained because of processing a series of satellite images of fields sown with grain crops in the East Kazakhstan region in the seasons of 2020-2021. Our motivation is to provide farmers with an accessible yet accurate and sustainable method for identifying homogeneous management areas in the field based on this approach. To this end, we study four clustering algorithms: K-means, K-medians, Hclust, Dbscan, and verify the results obtained.

II STATE OF THE ART

One of the works relevant to our study delimits management zones using clustering methods based on NDVI data, soil characteristics, and long-term crop yields [1]. In our study, we use four clustering algorithms to delimit management zones. The k-means algorithm is one of the popular iterative data clustering methods; it is fast and efficient to use. The literature describes many examples of applying this algorithm to NDVI indicators [5, 6, 4]. The k-medians algorithm is a modification of the k-means algorithm. To calculate the centroids of clusters, not the arithmetic mean but the median is used, due to which the algorithm is considered more resistant to outliers [8]. The Hclust algorithm implements hierarchical clustering in the R environment [3]. The DBSCAN algorithm is a relatively young method compared to the clustering algorithms listed above. It was first published in 1996 [2], while, for example, k-means was developed in the 1950s [7]. The idea of the DBSCAN algorithm is to search for high-density zones, i.e., defining clusters as zones with a close arrangement of points.

III ORIGINAL CONTRIBUTION

Our contribution is in a proposed clustering model and an R code for homogenous fertility zone identification.

IV METHODOLOGY

In our study, we use pixels (field points) as clustering objects and NDVI values by season dates – as features. For clustering, we use four well-established algorithms: k-means, k-medians, Hclust, and Dbscan. We analyze 4 clusters, of which 3 clusters define zones of high, medium, and low fertility, and one cluster is technical for points outside the field. Our study is experimental and can be used as a simple and cheap technology from the point of view of reproduction, available to small agricultural farms.

V RESULTS

Our results apply directly to the determination of uniformity zones in an agricultural field for which satellite imagery of vegetation is available.

VI Evaluation

Our methods have been manually evaluated by an agronomist, and as historical data accumulates, the evaluation can be automated.

VII Conclusions

In this article, we assessed the possibility of using clustering methods to identify fertility zones in the cultivation of grain crops using wheat as an example. In our experiments, the most stable behavior was demonstrated by the k-means and hclust algorithms, and of these two algorithms, the expert agronomist preferred fertility maps based on the hclust algorithm. The k-medians algorithm showed very high instability due to such parameters as the descent rate and the descent step. The dbscan algorithm requires further, more profound research regarding the choice of the parameter, which is obviously closely related to the statistics of the original data. In addition, the proposed methodology requires numerical verification, and in our future work, we plan to evaluate the correlation between fertility assessment and yield in each of the field zones.

References

[1] ALI, A., RONDELLI, V., MARTELLI, R., FALSONE, G., LUPIA, F., AND BARBANTI, L. Management zones delineation through clustering techniques based on soils traits, ndvi data, and multiple year crop yields. Agriculture 12, 2 (2022), 231.

[2] ESTER, M., KRIEGEL, H.-P., SANDER, J., XU, X., ET AL. A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (1996), vol. 96, pp. 226–231.

[3] GIORDANI, P., FERRARO, M. B., AND MARTELLA, F. Hierarchical clustering. In An Introduction to Clustering with R. Springer, 2020, pp. 9–73.

[4] MARINO, S., AND ALVINO, A. Detection of homogeneous wheat areas using multi-temporal uas images and ground truth data analyzed by cluster analysis. European Journal of Remote Sensing 51, 1 (2018), 266–275.

[5] NASER, M. A., KHOSLA, R., LONGCHAMPS, L., AND DAHAL, S. Using ndvi to differentiate wheat genotypes productivity under dryland and irrigated conditions. Remote Sensing 12, 5 (2020), 824.

[6] ROMANI, L., GONC¸ALVES, R., AMARAL, B., CHINO, D., ZULLO, J., TRAINA, C., SOUSA, E. P. M. D., AND TRAINA, A. J. M. Clustering analysis applied to ndvi/noaa multitemporal images to improve the monitoring process of sugarcane crops. In 2011 6th International Workshop on the Analysis of Multi-temporal Remote Sensing Images (Multi-Temp) (2011), IEEE, pp. 33–36.

[7] STEINHAUS, H., ET AL. Sur la division des corps materiels en parties.´ Bull. Acad. Polon. Sci 1, 804 (1956), 801.

[8] WHELAN, C., HARRELL, G., AND WANG, J. Understanding the k-medians problem. In Proceedings of the International Conference on Scientific Computing (CSC) (2015), The Steering Committee of The World Congress in Computer Science, Computer ..., p. 219.

DOI: https://doi.org/10.1007/978-981-19-8234-7_35
Link to repository: http://clip.ipipan.waw.pl/POSMAC

Keywords: OCR post-correction, POSMAC corpus, Polish

Fine-Tuning OCR Error Detection and Correction in a Polish Corpus of Scientific Abstracts

Maciej Ogrodniczuk [[0000-0002-3467-9424]]

maciej.ogrodniczuk@ipipan.waw.pl

SIMPLIFIED TITLE

Correcting OCR Errors in a Polish Corpus of Scientific Abstracts

ABSTRACT

The paper explores the idea of detecting and correcting post-OCR errors in a corpus of Polish scientific abstracts by first evaluating several available spellchecking approaches and then reusing one of the rule-based solutions to eliminate frequent errors most likely resulting from technical problems of the OCR process. The fine-tuning consisted in removing word breaks, rejecting corrections which change the case of the output, removing unnecessary spaces between word segments and restoring Polish letters replaced with spaces whenever the correction resulted in a valid Polish word. The obtained system proved competitive with language model-based solutions.

I INTRODUCTION

The process of OCR (optical character recognition) may result in errors due to many factors such as poor paper quality, wear of printing press or damage caused during use. In spite of applying various compensation mechanisms already during image processing and text recognition, many errors may still remain.

This is also the case with The Polish Open Science Metadata Corpus (POSMAC), a new source of scientific articles (abstract and full texts) acquired from the Polish Library of Science and included in the set of CURLICAT corpora [2]. The corpus contains over 142 thousand files with over 55 million words dated between 1934 and 2020, coming from over 900 Polish scientific journals and books, in most cases scanned and OCR-ed. Since the texts have been recognized in various periods, by various teams and methods, they can still contain many errors.

II STATE OF THE ART

We reviewed several recent spellcheckers for Polish:

- SPELLER — a spellchecker integrating LANGUAGETOOL (the most popular error correction tool for Polish, a multilingual grammar, style, and spell checker), SPACY (a robust Python natural language processing library) and AUTOCORRECT (another multilingual spelling corrector in Python)
- SYMSPELL — a spellchecking service which also provides multi-word correction and word segmentation of noisy text
- ED 3 PL [3] — a tool developed for the most recent spellchecking task for Polish at PolEval 2021 evaluation task [1]
- grammar and spellchecking tools integrated with Microsoft Word and Google Docs (although not intended for interactive use).

III ORIGINAL CONTRIBUTION

We proposed a solution which builds on the results of an existing spellchecker adjusted to scientific texts by:

- removing word breaks
- rejecting corrections which change the case of the output
- removing unnecessary spaces between word segments
- restoring Polish letters which were replaced with spaces due to technical problems of OCR systems.

IV Methodology

We defined two priorities for the process of correcting errors. Firstly, we would like to eliminate errors while not introducing the new ones. This means concentration on precision rather than recall (following the rule: “correct as much as you can but only when you are certain that the change will not result in an error”). Secondly, corrections need to be applied automatically. These requirements define our setup: we need to reuse or construct a spellchecker for Polish which is oriented on precision and is not interactive.

V Results

We started with comparing existing solutions. Their accuracy was sufficiently high (ranging from 85.79 to 96.50) but precision of even the best solution (70.54) was not satisfactory.

The qualitative error analysis showed that low results were partially caused by specifics of scientific texts. Systems frequently (and incorrectly) corrected technical terms, person names and punctuation in citations or foreign fragments.

Then we created a new system based on Speller by making several rule-based adjustments to its corrections to eliminate false positives such as keeping all uppercase words unchanged (since scientific texts contained many named entities). Some corrections required a dictionary lookup. For this purpose we used the list of over 5 million unique Polish inflected word forms after joining two largest electronic dictionaries for Polish. The dictionary was used for example in glueing word segments which were incorrectly separated or adding missing Polish letters which were lost in the process of OCR.

VI Evaluation

Several metrics are used to evaluate spellcheckers, such as word error rate. But still, the most common metric is classification accuracy which we decided to use as our main ranking criterion. We tested the system on a subset of POSMAC corpus by counting true and false positives and negatives, We compared word-aligned original samples, manually corrected samples and system output. Then values of precision, recall and accuracy were calculated.

We created the development set, used to define categories of errors, from 1000 randomly selected sentences from the POSMAC corpus. The categories were: missing diacritics, missing or extra chracters, unnecessary spaces, typos, words glued together, uppercase instead of lowercase (or the other way round), excessive punctuation, replacement of two close letters etc. Then we created the evaluation set by selecting another set of sentences from the corpus and performed their manual correction until we reached the number of 500 errors.

VII Conclusions

Spellcheckers can be used to correct errors made by users but also by computer systems processing or generating texts. Our solution, fine-tuned for correcting OCR errors in scientific texts, may be further adjusted to correct specific categories of errors in different journals or periods.

References

[1] Kobyliński, Ł., Kieraś, W., and Rynkun, S. PolEval 2021 Task 3: Post-correction of OCR Results. In *Proceedings of the PolEval 2021 Workshop* (Warsaw, Poland, 2021), M. Ogrodniczuk and Ł. Kobyliński, Eds., Institute of Computer Science, Polish Academy of Sciences, pp. 85–91.

[2] Váradi, T., Nyéki, B., Koeva, S., Tadić, M., Štefanec, V., Ogrodniczuk, M., Nitoń, B., Pęzik, P., Mititelu, V. B., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., and Repar, A. Introducing the CURLICAT Corpora: Seven-language Domain Specific Annotated Corpora from Curated Sources. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC 2022)* (Marseille, France, 2022), N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds., European Language Resources Association (ELRA), pp. 100–108.

[3] Wróbel, K. OCR Correction with Encoder-Decoder Transformer. In *Proceedings of the PolEval 2021 Workshop* (Warsaw, Poland, 2021), M. Ogrodniczuk and Ł. Kobyliński, Eds., Institute of Computer Science, Polish Academy of Sciences, pp. 97–102.

Link to arXiv: https://arXiv.org/abs/2104.12473
DOI: https://doi.org/10.1007/978-3-031-21743-2_41

Keywords: Automatic Speech Recognition, End-to-end, Connectionist Temporal Classification, Conditional Random Fields, ResNet

Development of CRF and CTC based end-to-end Kazakh speech recognition system

Dina Oralbekova[0000-0003-4975-6493], Orken Mamyrbayev[0000-0001-8318-3794], Mohamed Othman[0000-0002-5124-5759], Keylan Alimhan[0000-0003-0766-2229], Bagashar Zhumazhanov[0000-0002-5035-9076], Bulbul Nuranbayeva[0000-0003-3426-1914]

dinaoral@mail.ru

Simplified Title

Development of an end-to-end Kazakh speech recognition system

Abstract

Architecture end-to-ends are commonly used methods in many areas of machine learning, namely speech recognition. The end-to-end structure represents the system as one whole element, in contrast to the traditional one, which has several in-dependent elements. The end-to-end system provides a direct mapping of acoustic signals in a sequence of labels without intermediate states, without the need for post-processing at the output, making it easy to implement. Combining several end-to-end method types perform better results than applying them separately. Inspired by this issue, in this work we have realized a method for using CRF and CTC together to recognize a low-resource language like the Kazakh language. In this work, architectures of a recurrent neural network and a ResNet network were applied to build a model using language models. The results of experimental studies showed that the proposed approach based on the ResNet architecture with the RNN language model achieved the best CER result with a value of 9.86% compared to other network architectures for the Kazakh language.

I Introduction

Today the end-to-end model became widespread, which trains the components of the traditional model simultaneously without isolating individual elements, representing the system as a single neural network. Different ANN architectures can be used at all recognition stages, making it effective in terms of performance compared to other popular methods.

The E2E system realizes direct reflection of acoustic indicators in a sequence of marks without intermediate states, which does not require further processing at the output. These processes make the system easy to implement. There are several basic types of E2E models, such as connectionist temporal classification (CTC) [1], encoder-decoder with attention models, and Conditional Random Fields (CRF). E2E models require a large amount of speech data for training, which is problematic for languages with limited training data. And one of these languages is the Kazakh language. The Kazakh language has an agglutinative character, in which the dominant type of inflection is agglutination, opposite to the inflectional one. Some research works have shown that the combined use of E2E models like CTC and attention can be trained from start to finish, while this combination gave a very good result, which almost came close to the accuracy of the human level [2]. Based on these studies, we study the end-to-end system of the joint CTC and CRF models.

In this research, we have built a hybrid model based on two end-to-end methods, CRF and CTC, for Kazakh speech recognition.

II State of the Art

Conditional Random Fields (CRF) is a model that allows you to combine local information to predict the global probabilistic model from sequences [3]. This model is considered to be a kind of Markov random field. It proposed an algorithm for estimating parameters for conditional random fields and showed that CRF has a greater advantage over HMM and MEMM (maximum entropy Markov models) for natural language data. In many types of research the CRF model was applied to assess the measurement of accuracy in the problem of phonetic recognition, as well as the accuracy of detecting boundaries between them. The results show that when using transition functions in the CRF-based recognition structure, recognition performance is significantly improved by reducing the number of phoneme deletions.

Currently, the most common in speech recognition are linear and segmental CRF (linear chain and segmental CRF) models. This model is often used to solve the problems of marking and segmenting sequences. Other works demonstrated an implementation of CTC-CRF E2E models. For the experiment, Chinese and English tests, such as Switchboard and Aishell, were applied, thus obtaining the most modern results among the existing end-to-end models with fewer parameters and competitihttps://www.overleaf.com/project/636201b3ef5b98da3ed3e08dve in comparison with the hybrid models DNN-HMM. The reviewed works show us that the joint use of E2E models developed the productivity of the ASR system than using them separately.

III Methodology

The CTC function is used to train a neural network in sequence recognition. CTC eliminates the need for data alignment and allows for quite a few layers, a simple network structure to implement a model that maps audio to the sequence of utterances.

Conditional Random Field (CRF) is a discriminative undirected probabilistic graphical model. In contrast to the Markov model of maximum entropy, this method does not have a label bias problem. CRF and its various modifications have applications in natural language processing, computer vision, and speech recognition.

The CTC model was modified; it uses monotonic alignment between speech and tag sequences and trains the network quickly. Besides, the proposed model will be effective in recognizing speech in long sequences if training takes place in short training data. In addition, CTC helps speed up the process of assessing the desired alignment without the help of rough estimates of this process, which is labor-intensive and time-consuming.

IV Results

For experiments, corpus of speech for the Kazakh language was taken with a total volume of speech data 300 hours, with 90% of the audio data used for training and 10% for model validation. End-to-end models with different variations of neural networks were implemented, as well as the implementation of models separately and jointly.

Audio recordings with transcription from news sites in the Kazakh language, audiobook sites, which is 1600 separate phrases of different speakers and none of the speakers was used simultaneously in both parts. The end-to-end model, when using the CTC function without a language model, reached a CER of 17.45% and a WER of 29.01% (see Fig.2). Integration of the external language model into the CTC end-to-end system improved the CER and WER indicators by 13% and 18%, respectively. Our joint model of CTC and CRF with ResNet architecture showed good results without the use of LM, and CER reached 11.57% and WER - 18.32

V Evaluation

And after adding the language model, the model slightly improved its quality, by almost 1.5%. From obtained results we can see that the models using the ResNet network showed the best result in terms of the coefficients of correctly recognized words and characters.

VI Conclusions

The work considered the joint end-to-end models CTC and CRF for recognizing Kazakh speech. To implement this model, RNN variations were applied, such as LSTM and BiLSTM, as well as the ResNet. Convolutional neural networks were used for feature extraction. The practice works were conducted using the Kazakh language corpus with a volume of 300 speech hours, and the result demonstrated that the system could achieve high results using the ResNet and the use of RNN-based language model. Decoding based on these models does not increase the computational cost, and due to this, the decoding speed does not slow down. Thus, the best CER indicator reached 9.86%, which is a competitive result today. The proposed method is flexible enough and does not require conditional independence of variables. In addition, we can realize that the proposed model can be used to recognize other languages with limited training data, which are part of the Turkic languages.

Now, we target to study insertion-based models for recognizing agglutinative languages.

References

[1] Graves, A., Fernandes, S., Gomez, F., and Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *ICML 2006* (2006).

[2] Hori, T., Watanabe, S., Zhang, Y., and Chan, W. Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm. *Interspeech* (2017).

[3] Lafferty, J., McCallum, A., and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the International Conference on Machine Learning, Williamstown, MA, USA* (2001).

DOI: https://doi.org/10.1007/978-3-031-21743-2_55

Keywords: interface application, multidisciplinary space

Multimedia application for analyzing interdisciplinary scientific collaboration

Veslava Osińska[0000−0002−1306−7832], Konrad Poręba, Grzegorz Osiński[0000-0002-2939-4176], Brett Buttliere[0000-0001-5025-0460]

{wieo,bbutliere}@umk.pl,konrad.poreba@student.uj.edu.pl, grzegorz.osinski@wsksim.edu.pl

Simplified Title

Web analytical dashboard for scholars

Abstract

Information about scientific research can be represented on interactive science maps, which can yield insight into scientific activity at different levels of organization: micro, meso and macro. There are many examples at the meso or macro levels, i.e., research fields. But the micro level – individual analysis - is rather rarely considered in academic writing or app design. The authors designed and tested a web application, Scientific Visualizer, to illustrate an individual's scientific activity, which became a basis for a wide-ranging analysis. The visualizations generated by software will aid researchers in management of their own career, as well as planning effective communication and collaboration. The authors also have discussed their own approach to constructing a disciplinary space of publications based on the current classification of science. They added to scholarship on software and its application in academic community.

I Introduction

The developers of data visualization software prioritize macro analyses dedicated to data on big groups of users or global economic, politic processes. Individual researchers however need to analyze the literature patterns of own scientific activity. There is no opensource tool which would allow for the complete analysis of a single scholar's output. The micro level – individual analysis is rather rarely considered in academic writing or app design To redress this lack, the authors designed a Web application – Scientific Visualizer, dedicated to micro, i.e. individual analysis of scholars' output and for scholars. The goal was not only to make the analyses available, but to also develop ways of visualizing data such that the scholars would understand the data in new ways, impossible in a traditional formats.

II State of the Art

Currently we may distinguish five types of data analysis in relation to academic publishing: 1) quantitative, 2) geospatial, 3) temporary, 4) network, i.e. collaboration between people, and 5) contextual that is relies on different approaches towards topic extraction. Bibliographic metadata derived from scientific literature provide these analytical contexts. The results are visualized on science maps, which can deliver an essential patterns of social, intellectual and technological structure of science and of its evolution. Taking the multidisciplinary approach and various topologies of output space, it is possible to specify research interests of particular researchers on science map.

III Original Contribution

A predefined list of academic journals were the scholars publish their results was mapped onto 2D representation using the novel and effective in the case of current database algorithm. Using given visualization, the authors proposed to calculate the "coverage area" for each scholar analyzed. The appropriate measure Multidisciplinary Area of Scientific Interests (MASI) was defined by disciplinary space of researchers'. One should be calculated the distances from the density center of each research paper published in a journal assigned to a given discipline(s). Moreover the authors' application provides many overlapping contexts of analyses.

IV METHODOLOGY

As the source of data a university bibliographic database (called Expertus) was used. Apart from standard bibliographic metadata describing particular entities: authors, publications, journals and publishers, the record notes how many points the Ministry of Science and Higher Education has awarded to a journal. As a referential dataset the authors used a list of 32,681 journals (both international and national), ascribed to selected disciplines and knowledge domains. According Polish Science Classification, there are 46 disciplines and 8 domain areas. Thus, the data matrix 32,681 × 46 was constructed to represent disciplinary proximity of journals. Next for dimension reduction and visualization *t*-SNE (*t*-distributed stochastic neighbour embedding) was applied as an appropriate algorithm for science mapping.

Obtained science map provided visualization of scholar's articles in relation to overall disciplinary space. The extrinsic nodes of articles form multidisciplinary area of scholars interests. MASI coverage area was calculated within the polygon due to given visualization by using Monte-Carlo algorithm.

The above method is applied in the Web application dedicated to particular scholars. The application is based on the D3.js library - a set of very popular open JS scripts. Additionally, the series of another measures originated from visualizations are proposed and implemented in the Scientific Visualizer.

V RESULTS

This paper presents a web application – Scientific Visualizer that visualizes and analyses an individual's scientific output as an alternative to commercial tools: Scival or InCites. The scholars can study several contexts. The timeline charts allow the researchers to analyse the dynamics of their publishing. The visualizations of collaboration patterns yield information regarding the researcher's network, as well as the networks of their co-authors. The graph of geospatial patterns – an innovation in science mapping – gives the researcher a general overview of the dissemination of their work across the world. The tag cloud, as always, provides an immediate insight unavailable by other means.

Aside from the basic statistical measures (the mean, the yearly count), a series of new nonlinear parameters including quantification of muldisciplinary scope was proposed.

VI EVALUATION

Evaluation of MASI parameters was carried out using a multi-case study. Several examples of particular scientists' having various specialization and interests were chosen. As we can see, MASI can be a basic component of analytical framework. Due to this approach we can identify the shape of the researcher's interests space, and to quantify it.

VII CONCLUSIONS

Web application is available at local university website and any scholar can use it for visual analyses as well as planning future both national and international collaboration. Besides the scholars, management staff of particular department is interested in using proposed measures and calculations for evaluating purposes.

Our web application Scientific Visualizer offers a visualization framework which can resemble Google Analytics dashboard but provides many more possibilities than any commercial platform.

REFERENCES

[1] BÖRNER, K. *The Atlas of Knowledge*. MIT Press, 2016.

[2] OSINSKA, V. A qualitative–quantitative study of science mapping by different algorithms: The Polish journals landscape. *Journal of Information Science* 47(3), 2021, https://doi.org/10.1177/0165551520902738.

[3] VAN DER MAATEN, L.J.P., HINTON, G.E. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* 11, 2008, https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[4] ZHOU, C., CUI, H. Monte Carlo method of polygon intersection test. *Journal of Computational Information Systems* 9(12), 2013, 4707-4713.

DOI: https://doi.org/10.1007/978-3-031-21743-2_54
Link to repository: https://github.com/thanhtrunghuynh93/LBSN2Vec

Keywords: location-based social network, hypergraph embedding, recommendation system.

Social multi-role discovering with hypergraph embedding for Location-based Social Networks

Pham Minh Tam, Hoang Thanh Dat, Nguyen Minh Hieu, Vu Viet Hung,
Huynh Thanh Trung, Huynh Quyet Thang

{tam.pm202708m@sis,dat.ht202714m@sis,hieu.nm2052511m@sis,hung.vv162050@sis,thanghq@soict}.hust.edu.vn,thanh.huynh@epfl.ch

SIMPLIFIED TITLE

Location-based Social Networks analysis with social multi-role by hypergraph embedding.

ABSTRACT

Location-based social networks (LBSNs) have become more and more popular in the recent years. The typical LBSN platforms such as Foursquare, Facebook Local or Yelp allow the user to share their daily digital footprints in the form of check-ins with other people in different communities. The dynamic between users' social context and their mobility plays an important role in LBSN, e.g. users potentially participate with their friends in the same activity. The social interaction also demonstrate in the form of multi-role context, as each user may experience different activities with each particular community. Existing representation learning for LBSNs analysis often fails to fully capture such complex social pattern. In this paper, we propose a representation learning model in which the multi-role social interaction can be captured simultaneously with the mobility information. More specifically, the model first applies a ”persona” decomposition process, where each user node is splitted into several pseudo nodes presenting for his social roles. The process then learns multiple presentations for each persona that reflect the corresponding role by maximizing the collocation of nodes sampling on input user-user edges (friendships) and user-time-POI-semantic hyperedges (check-ins). We conduct experiments on 5 real-world datasets with 7 state-of-the-art baselines to demonstrate the robustness of our model on downstream tasks such as friendship suggestion and location prediction.

I INTRODUCTION

Location-based social networks (LBSNs) such as Facebook Local, Foursquare, Yelp, Brightkite and Gowalla have emerged in the recent years [1]. LBSNs allow the users to share their daily experiences to other people by the check-ins, each includes a location (a.k.a point of interest (POI) such as university, plaza, supermarket), a specific timestamp, a semantic category (e.g. attending class, working out or shopping). LBSNs data contain rich socio-spatial properties of user activities. Existing techniques have not fully exploited the multi-context nature of LBSNs. In LBSNs, the user nodes associate with other key modalities including spatial pattern (POI nodes), temporal pattern (time nodes), and semantic pattern (activity nodes) in the check-ins, results in a high-order modal. In this work, we address the problem of analyzing LBSNs through learning representation for network nodes that capture at the same time multi-role social context and mobility dynamic of the users.

II STATE OF THE ART

Given the great benefit of LBSNs analysis, a rich body of researches has been proposed to model the social interaction to user mobility. Traditional techniques design hand-crafted features extracted from either user mobility data (e.g. co-location rates) or user friendships (e.g. Katz index to investigate the impact from one on the other. Such approaches often require significant human effort and domain knowledge as well as lack of generalizability to different applications. Recent techniques leverage the advances in graph representation learning to embed the nodes into low-dimensional embedding spaces that automatically capture the users mobility and social context, based on the original graph topology and nodes' attribute. However, these approaches simplify the complexity of LBSNs by first breaking the hyperedges into smaller classical edges, then apply the existing representation learning technique for classical graph. Such transformation process might involve information loss and performance degradation.

III Original Contribution

In this work, we propose a **M**ulti-**R**ole Social Interaction aware embedding framework for **LBSN** (MR-LBSN), where we leverage hypergraph embedding to capture the at the same time multi-role social context and mobility dynamic of the users. We propose a multi-context hypergraph embedding method for LBSNs that captures simultaneously both friendship and check-ins information. We specially design a persona decomposition algorithm that considers at the same time user context scope and check-in assignment for persona nodes. We integrate user multi-role context and user mobility under the same modal by combining persona transition and check-in exploration simultaneously in the random walk process. We develop a social multi-role aware embedding that leverage the sampled random walks to integrate simultaneously the check-in and friendship information.

IV Methodology

We evaluate the performance of the techniques on two important downstream tasks: friend suggestion and POI recommendation. For the former task, we leverage the embedding similarity between the users to determine whether they are potential to make friend in the future, then evaluate the result using five metrics: precision, recall, F1-score and nDCG to the top-K predicted friendships [3]. For POI recommendation, for a set of (user, time, semantic) as the query, we choose the POI that having the closet embedding as the result, then apply Hit@K to evaluate the result [2].

V Results

Our technique is able to capture at the same time multi-role social context and mobility dynamic of the users, as well as their correlation. For friend recommendation task, which relates to users' social context, our technique can recognize the ranked candidate list to recommend for the users, which is an essential and popular use case in real-world application. Also, our representations can be used to retrieve better POI recommendation, given a specific user, time and category.

VI Evaluation

There are some detailed evaluation of our technique and other baselines:
+ Friendship suggestion task:

- *MR-LBSN* achieved the gain of nearly 50% in Precision and F1-score compared to the best baseline.
- *MR-LBSN* achieved slightly less significant improvement as for precision and F1-score, but still reached an average enhancement of 21.35%.
- For the nDCG@k metric, our techniques outperforms other baselines by 15-30% in all five datasets.

+ POI recommendation task:

- *MR-LBSN* achieved an average improvements of 30.23% in Hit@5 and 28.46% in Hit@10 to the best baseline, *LSBN2Vec*.
- The inappropriate use of both social relationship and mobility information was less effective that using only the mobility data.

VII Conclusions

In this paper, we propose a LBSN hypergraph embedding technique that captures the holistic interactions of multi-role social context and user mobility. The technique can be used to better analyze LBSN and enhance the downstream tasks such as friend recommendation and POI suggestion. In the future work, we would want to explore mobility dynamic patterns such as sequential effects, cyclic effects, which are crucial for LBSN analysis [1].

References

[1] Kefalas, P., Symeonidis, P., and Manolopoulos, Y. A graph-based taxonomy of recommendation algorithms and systems in lbsns. *TKDE 28*, 3 (2015), 604–622.

[2] Liben-Nowell, D., and Kleinberg, J. The link-prediction problem for social networks. *JASIST 58*, 7 (2007), 1019–1031.

[3] Scellato, S., Noulas, A., and Mascolo, C. Exploiting place features in link prediction on location-based social networks. In *KDD* (2011), pp. 1046–1054.

DOI: https://doi.org/10.1007/978-3-031-21967-2_31

Keywords: Student Monitoring System, Face Recognition, MTCNN, FaceNet

Tracking Student Attendance in Virtual Classes Based on MTCNN and FaceNet

Trong-Nghia Pham[0000-0002-2273-5089], Nam-Phong Nguyen[0000-0002-9153-6286], Nguyen-Minh-Quan Dinh[0000-0002-1838-8627], Thanh Le[0000-0002-2180-4222]

{ptnghia,lnthanh}@fit.hcmus.edu.vn,{giophuongnam.phong,quan2312016vn}@gmail.com

SIMPLIFIED TITLE

Tracking Student Attendance in Virtual Classes Based on MTCNN and FaceNet.

ABSTRACT

All classes are held online in order to ensure safety during the COVID pandemic. Unlike onsite classes, it is difficult for us to determine the full participation of students in the class, as well as to detect strangers entering the classroom. Therefore, We propose a student monitoring system based on facial recognition approaches. Classical models in face recognition are reviewed and tested to select the appropriate model. Specifically, we design the system with models such as MTCNN[2], FaceNet[1], and propose measures to identify people in the database. The results show that the system takes an average of 30 seconds for learning and 2 seconds for identifying a new face, respectively. Experiments also indicate that the ability to recognize faces achieves high results in normal lighting conditions. Unrecognized cases mostly fall into too dark light conditions. The important point is that the system was less likely to misrecognize objects in most of our tests.

I INTRODUCTION

The Covid pandemic is spreading globally and significantly affecting all areas of life, including education. Many universities have switched from onsite instruction to online not to disrupt learning. However, it is difficult for teachers to observe the status of students in class as well as detect strangers appearance. From that problem, we propose a system based on facial recognition to identify and track students in the classroom. The system can be practically implemented in onsite classes and works on several online platforms. This system also facilitates the monitoring of students at the university, restricts strangers, and assesses student participation in courses. External factors such as illumination, posture, expression, glasses, and hairstyle are problems we solve by evaluating state-of-the-art methods to build a powerful and stable system.

II STATE OF THE ART

A face recognition system includes three main components, which are face detection, feature extraction, and face recognition.

II.1 Face detection

Popular face detection methods include Viola-Jones, the histogram of the directional gradient (HOG), and Single-Shot Detector(SSD). Although these methods give fast execution speed, the accuracy is not really good in many situations

II.2 Feature Extraction

Linear discriminant analysis (LDA), principal component analysis (PCA), Scale-invariant feature transform (SIFT), and local binary sampling method (LBP) are feature extraction methods that many people know, but they are not powerful enough for the feature extraction task for face classification.

II.3 Face Recognition

After extracting features from the face, methods like Correlation filters (CF), convolutional neural networks (CNN), and k-nearest neighbors (K-NN) are usually applied in order to solve the face classification problem based on those features.

Reproduced with permission from Springer Nature

III Original Contribution

We survey and evaluate the state-of-the-art approaches, then choose the most suitable models for the proposed system to achieve high accuracy with an acceptable execution time that makes the system work in real environments. We also make adjustments to match the reality of the context. Training should be quick and straightforward to implement for new faces.

IV Methodology

We evaluate the advantages and disadvantages as well as conduct experiments of existing models and methods in face detection and face recognition to be able to meet the requirements we have set for our system in terms of accuracy and execution speed. After selecting, we combine these models into a complete system and evaluate the entire system. Moreover, we also assess the system's capability in a real online classroom held on the Zoom platform.

V Results

Through evaluations, we choose MTCNN[2] for face detection and FaceNet[1] for feature extraction to build our system, which has the experiment results that are also quite good. A comparison experiment between MTCNN and SSD on FDDB (Face Detection Dataset and Benchmark) shows that although SSD is faster, MTCNN still gives higher accuracy with acceptable speed. For the experiment on the famous people dataset, the system gives pretty good facial recognition results of people in the data set and people not in the data set with high accuracy. Moreover, when testing the system in an online classroom held on the Zoom platform, most students were recognized, except for a few whose webcams were too dim.

VI Evaluation

Experiments are performed on datasets of faces based on two main criteria: accuracy and execution speed. Models suitable for the system must have high accuracy while ensuring the execution speed needs to be achieved in real-time to be able to apply in practice.

VII Conclusions

Face recognition is one of the highly applicable problems. We focus on learning methods to solve this problem and apply it to the student monitoring system in the classroom. The implementation of the system helps teachers track students as well as detect intruders, especially in online classes. We have selected algorithms to integrate into the system through surveys and analysis of methods. In particular, two methods are considered the core for the system to work well. They are the MTCNN algorithm to identify face area and FaceNet to recognize a face. We have also adjusted and added measures such as cosine similarity to increase the ability to identify faces the database. The system can be integrated into online classroom platforms directly for use.

References

[1] Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 815–823.

[2] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters 23*, 10 (2016), 1499–1503.

DOI: https://doi.org/10.1007/978-981-19-8234-7_37
Link to repository: https://github.com/dungpham98/url_gan

Keywords: URL Classfier, Text Generation, Text Diversity

En-SeqGAN: An Efficient Sequence Generation Model for Deceiving URL Classifiers

Pham Tuan Dung [0000-0002-2183-4640], Pham Thi Thanh Thuy [0000-0003-3985-3599], Ta Viet Cuong [0000-0001-8058-5915]

dungpt98@vnu.edu.vn, thanh-thuy.pham@mica.edu.vn, cuongtv@vnu.edu.vn

Simplified Title

En-SeqGAN: An Efficient Sequence Generation Model tailored towards Attacking URL Classifiers.

Abstract

Generative Adversarial Networks (GANs) are recently used to generate URL patterns to fool the phishing URL classifiers. Some of these works use Wasserstein GAN (WGAN) to generate domain samples for deceiving phishing URL detectors. However, WGAN-based models are designed to work mainly on continuous data and cannot capture the diverse set of patterns in a URL sequence. In order to overcome this issue, we propose En-SeqGAN which works on discrete data to generate full URL sequences. The proposed model is based on the standard SeqGAN with the addition of entropy regularization to encourage the model to produce diverse URL samples. Several intensive experiments are done to prove that the URL samples generated by the proposed model can evade the gray-box phishing detectors of LSTM and Random Forest. The efficiency of gray-box attack by En-SeqGAN on these URL classifiers outperforms both methods of SeqGAN and WGAN. Moreover, En-SeqGAN can generate well-structured URL samples with various URL sequence lengths.

I Introduction

The growth of the World Wide Web (WWW) has attracted the attention of cybercriminals who use the web to spread malware to compromise the networks of individuals and organizations. URLs are effective means of phishing attacks, with the two most common types being email spoofing, and fake websites.

One of the recent popular methods of attackers is to use GAN networks to generate URL samples sophisticated enough to fool classifiers. Recent works try to generate some part of a URL string, such as its domain or its feature. The current problem of generating full URL sequences in recent works is that the generated URL strings might not have a standard URL structure. One of the main reasons is the difficulty in balancing between the diversity of the generated patterns and the relationship between each character in a URL string. In order to tackle this issue, in this work, an efficient discrete GAN model, named En-SeqGAN, is proposed. It is the extension of SeqGAN with additional entropy regularization in the objective function of the policy gradient training.

II State of the Art

GAN-based models have been proposed in several works to generate URL samples for deceiving the classifiers. Basically, GANs are used to generate (1) domain names in URL and (2) full URL strings. The models of Domain-GAN , WGAN, DeepDGA, etc., are for domain name generation. Other works try to generate a text string as a URL string, but the generated results do not present the standard structure of a URL string. Moreover, in these works, there is a lack of detailed evaluations of the quality of the generated URLs.

III Original Contribution

The contributions are as follows:

- Introduce En-SeqGAN model, a Sequence Generation Model which utilizes Entropy Regularization in its objective function.

- Propose a new pipeline that contains two scenarios to evaluate the efficiency of the proposed En-SeqGAN model.

IV Methodology

In this work, we apply SeqGAN [1] model and make a certain improvement to it. SeqGAN in origin uses Policy Gradient with REINFORCE [2] to train the generator while the discriminator is a classifier to discriminate real and generated text and provides reward signals for the generator updates. The samples generated by SeqGAN rely on an appropriate number of times for Monte Carlo search to update the generator and maximize the expected end reward. But when SeqGAN focused on maximizing the reward, the framework might ignore the possibilities to explore more diverse samples. In order to apply SeqGAN to generate complete and well-structured URL strings, entropy regularization is added to the objective function of SeqGAN. This improved SeqGAN helps to fool URL classifiers in an effective way while maintaining the diversity of the character patterns in the URL string.

V Results

Table 1: Number of samples in training set and classifier accuracy.

Model	**β value**	**Set A (%)**		**Set B (%)**		**Set C (%)**	
		LSTM	RF	LSTM	RF	LSTM	RF
Baseline	N/A	97.00	71.92	95.67	77.72	96.71	78.03
WGAN		85.64	99.99	87.53	99.99	95.74	99.99
SeqGAN		80.79	45.20	80.72	39.20	92.35	30.75
En-SeqGAN	7e-3	68.00	28.30	78.85	27.15	88.51	45.34
	1e-2	71.36	38.50	69.19	33.40	**84.96**	27.86
	2e-2	**60.63**	**16.83**	**83.12**	**20.52**	85.53	**25.76**

As shown in Table 1, our proposed method outperforms WGAN, SeqGAN and Baseline evaluations. For set A, with the LSTM classifier, the detection accuracy of our proposed method is lower than other methods, from approximately 36% (compared to the Baseline method) to 25% (WGAN) or 20,16% (SeqGAN). The experimental results with RF are better than LSTM for all cases. The detection accuracy of our En-SeqGAN is only 16.83%, which is much lower than other methods. At the settings of B and C, En-SeqGAN gains better results in comparison with other ones at both LSTM and RF classifiers, and the distinguishability of RF is much lower than LSTM classifier. Overall, the detection results gained from the proposed method tend to increase with the number of testing samples.

VI Evaluation

The defensibility of the classifiers is evaluated on phishing URLs generated by GANs. It is calculated by the number of correctly classified URLs over the total number of URLs, or Accuracy.

In order to evaluate the distribution of the URL samples generated by GANs, we perform an Ngram Analysis on different GAN model by calculating and analyzing the bi-gram distributions of the first 30 bi-grams for each GAN model. On order to evaluate the structure of URL strings generated by GANs, we consider two criteria that penalize the structural properties of an URL, which is the Top-Level Domain score and Parsing Score. The results show that WGAN bi-gram distribution is much different from the training set compared to the ones of SeqGAN and EnSeqGAN. En-SeqGAN has a more uniform bi-gram distribution than SeqGAN due to the entropy regularization.

This encourages En-SeqGAN to generate more diverse URL strings. On the small and medium training size (set A and B), the adding entropy regularization has negative effects in modeling the URL structures, which leads to our En-SeqGan model having a lower Parsing score than the standard SeqGan model. On the large training (set C), our proposed model starts to learn the URL structures and has a better score than the standard one.

VII Conclusions

From our results, ones could benefit their defend mechanism against other gray-box attack frameworks. Since the generated samples from our proposed method are full-length URL strings, these URLs could be used, and analyzed to enhance various URL defend frameworks.

References

[1] Yu, Lantao, Weinan Zhang, Jun Wang, and Yong Yu. ”Seqgan: Sequence generative adversarial nets with policy gradient.” In Proceedings of the AAAI conference on artificial intelligence, vol. 31, no. 1. 2017.

[2] Williams, Ronald J. ”Simple statistical gradient-following algorithms for connectionist reinforcement learning.” Machine learning 8, no. 3 (1992): 229-256.

DOI: https://doi.org/10.1007/978-981-19-8234-7_38

Keywords: Traffic monitoring, Distributed Machine Learning

Towards Communication-efficient Distributed Background Subtraction

Hung Ngoc Phan, Synh Viet-Uyen Ha, Phuong Hoai Ha [0000-0001-8366-5590]

hung.n.phan@uit.no, hvusynh@hcmiu.edu.vn, phuong.hoai.ha@uit.no

SIMPLIFIED TITLE

A communication-efficient distributed approach to road traffic monitoring

ABSTRACT

Road traffic monitoring is one of the essential components in data analysis for urban air pollution prevention. In road traffic monitoring, background subtraction is a critical approach where moving objects are extracted via facilitating motion information of interest to static surroundings, known as backgrounds. To work with various contextual dynamics of nature scenes, supervised models of background subtraction aim to solve a gradient-based optimization problem on multi-modal sequences of videos by training a convolutional neural network. As video datasets are scaling up, distributing the model learning on multiple processing elements is a pivotal technique to leverage the computational power among various devices. However, one of major challenges in distributed machine learning is communication overhead. This paper introduces a new communication-efficient distributed framework for background subtraction (CEDFrame), alleviating the communication overhead in distributed training with video data. The new framework utilizes event-triggered communication on a ring topology among workers and the Partitioned Globally Address Space (PGAS) paradigm for asynchronous computation. Through the new framework, we investigate how training a background subtraction tolerates the trade-offs between communication avoidance and accuracy in model learning. The experimental results on NVIDIA DGX-2 using the CDnet-2014 dataset show that the new framework can reduce the communication overhead by at least 94.71% while having a negligible decrement in testing accuracy (at most 2.68%).

I INTRODUCTION

Road traffic is one of the primary sources contributing to urban air pollution. Therefore, road traffic monitoring is an appealing application in computer vision in which background subtraction or change detection is a significant approach to motion analysis in video processing. The technique aims to segregate desired objects, called foregrounds, from the background scenes. The backgrounds are devoid of moving elements which are not of interest to the system (e.g., streets, houses, trees).

II STATE OF THE ART

The typical attention of background subtraction is to speculate underlying scenes' properties by presenting learning-based models that are basically grounded on learning towards collected data [1]. One of the most popular approaches to achieving the generalization of multi-modal data is to formulate background subtraction as probabilistic estimation for prediction in which an optimal model is obtained by performing the gradient-based method with a large-scale of training samples. Besides concerns on accuracy due to the incompleteness of data, performing a model learning on a great deal of sampling data requires an appropriate scheme of resource usage that takes advantage of computational parallelism among leading-edge processing units to gain highly efficient model learning. As the scales of training datasets and model sizes increase dramatically, one of the popularly-used approaches is to decompose the mini-batch learning into multiple concurrent optimization pipelines. In this context, data exchange between processing units potentially imposes a significant overhead, degrading the effectiveness of computing tasks.

Distributed machine learning has become a pivotal research field, in which communication-efficient algorithms are proposed to diminish communication overhead among parallel processing units and maintain a high ratio of the computation to communication. There are two typical strategies towards communication-efficient systems: *Asynchrony* and *Communication-efficiency*. In both theoretical analysis and applicable implementation, accommodating these two approaches simultaneously in distributed training of deep neural networks imposes a critical trade-off between model accuracy and communication-avoidance. Background subtraction is a specific problem where we need to cope with a large scale of contextual dynamics encountered by moving objects on various camera scenes

Reproduced with permission from Springer Nature

or video sequences. Communication efficiency in distributed training towards this research area is indispensable to achieving a scalable learning model.

III Original Contribution

To gain insights into the trade-off between model accuracy and communication-avoidance, we have proposed a new asynchronous, server-free and communication-efficient framework for training a background subtraction model. Our framework leverages event-triggered control to reduce communication among processing elements (PEs) in a ring topology, and exploits asynchrony to eliminate synchronization overhead in model training. With this framework, we have enabled investigating the balance between accuracy and communication-avoidance for distributed background subtraction.

IV Methodology

First, we have developed a new framework named CEDFrame capable of transforming traditional *centralized* background subtraction models to communication efficient *distributed* schemes for investigating and optimizing the trade-off between accuracy and communication avoidance towards decentralized training. The framework adopts an event-triggered scheme [2] to regulate the amount of message transmission across various control thresholds. We have leveraged the Partitioned Globally Address Space (PGAS) paradigm for asynchronous computation while utilizing computational capability of GPUs to accelerate computation in model training. Delegating communication and computation to CPUs and GPUs respectively, we have tailored the design of the frameworks to utilize the computational capability and the memory of large training systems.

Second, based on the new framework, we have introduced a new communication-efficient decentralized algorithm for background subtraction named Distributed Motion Segmentation (D-MoSeg). We have employed the framework to train a background subtraction model to examine the generalization of the model across multi-contextual dynamics with respect to the effect of communication avoidance.

V Results

We have conducted experimental evaluation of the new algorithm on a state-of-the-art AI server NVIDIA DGX-2. Experimented on CDnet-2014, a large-scale database of video sequences for background subtraction with various scenarios, the proposed framework can reduce the communication overhead by at least 94.71% while having a negligible decrement in testing accuracy (at most 2.68%).

VI Evaluation

We have employed our framework to train a compact neural network for multi-modal background subtraction in a distributed fashion across various visual dynamics. We have performed the evaluation on a NVIDIA DGX-2 server with two Intel Xeon Platinum 8168 CPUs (24-core, 2.7 GHz), 1.5TB of memory and 16 NVIDIA Tesla V100 GPUs. The framework has been implemented with UPC++, a PGAS library, and PyTorch C++ API.

VII Conclusions

Road traffic significantly contributes to urban air pollution, and therefore road traffic monitoring is a crucial component in air pollution prevention. This paper has introduced a communication-efficient distributed approach to road traffic monitoring based on background subtraction. Background subtraction is a specific problem where data-driven models cope with varying scenes for multi-contextual generalization. However, as training datasets scale up, communication overhead emerges as a bottleneck in distributed models. This research has introduced a new framework CEDFrame to transform traditional centralized background subtraction models into communication-efficient distributed models. The framework utilizes an event-trigger method and asynchronous communication to reduce the communication overhead. Using the new framework, we have developed a new communication-efficient decentralized algorithm for background subtraction D-MoSeg. The experimental evaluation on a large-scale video dataset has showed that our new framework could significantly reduce communication while having a negligible decrement in training and testing accuracy. For future work, examining the effect of stragglers and the staleness of model parameters during training is potential research.

References

[1] Bouwmans, T., Javed, S., Sultana, M., and Jung, S. K. Deep neural network concepts for background subtraction:a systematic review and comparative evaluation. *Neural Networks 117* (2019), 8–66.

[2] Ghosh, S., and Gupta, V. EventGraD: Event-Triggered Communication in Parallel Stochastic Gradient Descent. In *2020 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC)* (2020), pp. 1–8.

DOI: https://doi.org/10.1007/978-3-031-21743-2_47

Keywords: inverse text normalization, automatic speech recognition, neural network models, rule-based systems

Neural Inverse Text Normalization with Numerical Recognition for Low Resource Scenarios

Tuan Anh Phan, Ngoc Dung Nguyen[0000-0002-1141-0975], Huong Le Thanh, Khac-Hoai Nam Bui[0000-0002-3427-8460]

(anhpt161,dungnn7,nambkh)@viettel.com.vn, huonglt@soict.hust.edu.vn

Simplified Title

Neural Inverse Text Normalization for Low Resource Scenarios

Abstract

Neural inverse text normalization (ITN) has recently become an emerging approach for automatic speech recognition in terms of post-processing for readability. In particular, leveraging ITN using neural network models has achieved remarkable results instead of relying on the accuracy of manual rules. However, ITN is a highly language-dependent task that is especially tricky in ambiguous languages. In this study, we focus on improving the performance of ITN tasks by adopting the combination of neural network models and rule-based systems. Specifically, we first use a seq2seq model to detect input sentences' numerical segments (e.g., cardinals, ordinals, and date). Then, detected segments are converted into written form using rule-based systems. Technically, a major difference in our method is that we only use neural network models to detect numerical segments, which are able to deal with the low resource and ambiguous scenarios of target languages. Regarding the experiment, we evaluate different languages to indicate the advantages of the proposed method.

I Introduction

ITN is one of the Natural language processing (NLP) tasks of transforming the written form into spoken form. The conventional approach for addressing ITN is rule-based systems, for instance, finite state transducer (FST) based models. The major problem with this approach is the scalability problem, which requires complex accurate transformation rules [3]. Recently, NN-based models, typically seq2seq, have achieved high performances and become state-of-the-art models for the ITN problem [2]. In this regard, the data-hungry problem (i.e., low resource scenarios) is an open issue that needs to take into account for improving performance. Furthermore, due to the significant difference between written and spoken forms, handling numbers correctly is a central problem in this research field. In particular, to be able to read the numeric values, the models should be worked on both consecutive tasks such as recognizing the parts that belong to numeric values and combining those parts to precise numbers. In this study, we take an investigation to improve the performance of Neural ITN in terms of low resources and ambiguous scenarios.

II State of the Art

Pusateri et. al. [1] presents a data-driven approach for ITN problems with a set of simple rules and a few hand-crafted grammar to cast ITN as a labeling problem. Authors in [2] propose a combination of transformer-based seq2seq models and FST-based text normalization for data preparation. Although can be able to overcome common recoverable errors, all of them do not consider ITN problems under data-hungry and complex data scenarios.

III Original Contribution

The main contributions of our method are two folds: i) We propose a novel hybrid approach by combining a neural network with a rule-based system, which is able to deal with ITN problems in terms of low resources and ambiguous scenarios; and ii) the proposed method is evaluated with different language such as English and Vietnamese with promising results.

IV Methodology

IV.1 General Framework

We propose a novel hybrid model using seq2seq neural network and rule base systems. In the first stage, each sentence is put into a transformer-based seq2seq model for detecting numerical segments by using tag ¡n¿ and ¡/n¿, in which *n* represents numerical classes. Then, a set of rules is employed to convert tokens, which be wrapped by tag to number, into the written form. Otherwise, all parts of a sentence, which are not in the tag are conserved.

Essentially, NN is only utilized for distinguishing which is in the number and which is not. After that, when the model has candidates for numbers, they are transformed into the correct form by the set of rules.

IV.2 Training Model

We implement two training models which are RNN-based and transformer-based seq2seq models. For the RNN model, we employ a bi-LSTM as encoder and an LSTM as decoder, respectively.

IV.3 Rule-based Systems

The output of NN models with detected segments is transformed into the written form using a set of rules. With each numerical token, we used the word2number1 python package[1] for converting spoken numbers into written numbers. We also extended the tool in order to handle negative cardinals and larger numbers and construct the extra modules for reading the number, which belongs to other classes such as MEASURE, DATE, PHONE, and TIME.

V Evaluation

Tab. 1 shows the comparison results of our experiments on the test set, which bold parts are the best results. Based on the reported results, there are several assumptions can be concluded as follows:

Dataset	**English**				**Vietnamese**			
	100k	200k	500k	1m	100k	200k	500k	1m
RNN	0.7405	0.7909	0.7706	0.7959	0.6422	0.6794	0.6921	0.6777
Transformers	0.7048	0.7919	**0.8558**	**0.9138**	0.6755	0.7201	0.7447	0.7594
RNN (our)	**0.8334**	**0.8353**	0.8377	0.848	**0.7019**	0.7144	0.718	0.711
Transformers (our)	0.741	0.8188	0.8394	0.8933	0.6774	**0.7286**	**0.7667**	**0.7885**

Table 1: Comparison of models on test set with BLEU scores. Bold texts indicate the best results.

- Our method outperforms baseline models in the case of low resource scenarios (i.e., 100k and 200k) and is able to achieve competitive results in the case of higher resources (i.e., 500k and 1000k) with the English language.
- For the Vietnamese language, our method is able to achieve the best results in all cases.
- Recurrent-based seq2seq models with attention achieve better performance compared with Transformer in the case of low resource scenarios. Meanwhile, Transformer-based models are able to achieve the best results by increasing the number of training samples.

VI Conclusions

In this study, we introduce a new method for the neural ITN approach. Specifically, the difference from previous works, we divide the neural ITN problem into two stages. Particularly, in the first stage, neural models are used to detect numerical segments. Sequentially, the written form is extracted based on a set of rules in the second stage. In this regard, our method is able to deal with low-resource scenarios, where there is not much available data for training. Furthermore, we showed that our method can be easily extended to other languages without linguistic knowledge requirements.

References

[1] Pusateri, E., Ambati, B. R., Brooks, E., Plátek, O., McAllaster, D., and Nagesha, V. A mostly data-driven approach to inverse text normalization. In *Proceeding of the 18th Annual Conference of the International Speech Communication Association (Interspeech)* (2017), ISCA, pp. 2784–2788.

[2] Sunkara, M., Shivade, C., Bodapati, S., and Kirchhoff, K. Neural inverse text normalization. In *Proceeding of the 46th International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2021), IEEE, pp. 7573–7577.

[3] Zhang, H., Sproat, R., Ng, A. H., Stahlberg, F., Peng, X., Gorman, K., and Roark, B. Neural models of text normalization for speech applications. *Comput. Linguistics 45*, 2 (2019), 293–337.

[1] https://pypi.org/project/word2number/

DOI: https://doi.org/10.1007/978-3-031-21967-2_30

Keywords: Segmentation, Medical Image processing, Generative adversarial network, Brain tumor

A Research for Segmentation of Brain Tumors Based on GAN Model

Linh Khanh Phung, Sinh Van Nguyen[0000-0003-0424-5542], Tan Duy Le, Marcin Maleszka

nvsinh@hcmiu.edu.vn

SIMPLIFIED TITLE

Segmentation of Brain Tumors on the medical images based on generative adversarial network

ABSTRACT

Analysis of medical image is a useful method that can support doctors in medical diagnosis. The development of deep learning models is essential and widely applied in image processing and computer vision. Application of machine learning and artificial intelligent in brain tumor diagnosis brings an accuracy and efficiency in medical treatment field. In this research paper, we present a method for determining and segmenting the brain tumor region in the medical image dataset based on 3D Generative Adversarial Network (3D-GAN) model. We first explore the state-of-the-art methods and recent approaches in such field. Our proposed 3D-GAN model consist of three steps: (i) pre-processing data, (ii) building an architecture of multi-scaled GAN model, and (iii) modifying loss function. The last our contribution is creating an application to visualize 3D models that representation of medical resonance brain images with the incorporation of the chosen models to determine exactly the region containing brain tumors. Comparing to the existing methods, our proposed model obtained better performance and accuracy.

I INTRODUCTION

The methods for images analyzing and processing from the medical dataset like MRI, CT scanner are proved obtaining advantages in both diagnosis and treatment. It helps deceasing the time of treatment and risk; increasing health recuperation after operating and saving treatment fee. Brain tumor is a very dangerous disease for anyone. The techniques in image processing can be applied to analyze, recognize and classify the tumor regions. In medical image segmentation, a majority of GAN-based segmentation model is implemented using two components: generator and discriminator. This research paper presents a proposed method for segmentation of Brain Tumors based on GAN Model.

II RELATED WORK

The first method is based on the graphical techniques, geometric modeling and image processing that has been proposed in [1]. Other methods for medical image processing are proposed to reconstruct the data objects from a DICOM dataset. The advent of CNN has contributed to the high quality outcome of automatic medical image segmentation. However, the current network models like UNet, VNet still meet limitations in variations of object scales in medical images and demanding inept weight ensemble Therefore, improving the current state of the 3D neuro-imaging network is crucial and can be put into practice under the refinements of some frameworks such as GAN [2].

III ORIGINAL CONTRIBUTION

We proposed a method for segmentation of Brain Tumors on medical images based on GAN Model. We used the BRATS 2021 dataset [3] to train and validate. Comparing to the several methods, our proposed method obtained better performance and accuracy.

IV METHODOLOGY

The main idea of our proposed method is presented as follows. We use a GAN model with improved steps in architecture design and loss function. The method includes four modules: Input data; Generator architecture; Bridge and Discriminator. The constrains on discriminator should be made to enhance the optimization of losses. Therefore, we developed a simple decision algorithm to help tweak the epochs dynamically based on the adversarial loss trends. In general, the idea is based on linear regression to construct a line $Y = \alpha * X + \beta$ from the set of iterations X and its corresponding adversarial loss Y. The increase of the loss values when the slope α is positive, or the large y-intercept β when the loss is stabilizing (i.e. $\alpha \approx 0$) indicate the discriminator's excellence in detecting the genuineness of the masks. Therefore, training discriminator in these circumstances is crucial and usually contributes to the segmentation loss.

Model	Label	Dice (%)	Jaccard (%)	Haudorff(mm)	ASSD (mm)
Vox2Vox	ET	79.745	74.823	18.892	6.855
	WT	83.238	75.199	30.996	12.370
	TC	78.210	72.440	26.013	9.6722
Vnet	ET	78.332	73.581	14.918	5.527
	WT	86.411	79.153	18.109	6.351
	TC	75.478	68.846	18.491	7.066
Our ResUnet (without GAN)	ET	77.964	72.190	**6.382**	**2.349**
	WT	88.103	80.930	16.505	6.0193
	TC	80.366	75.569	***8.431***	***3.523***
Our multi-scale GAN	ET	**82.318**	**78.487**	9.616	3.367
	WT	**88.599**	**81.466**	**13.323**	**5.123**
	TC	***83.108***	***79.705***	13.180	5.416

Table 1: Comparison of models' accuracy. The best results of ET, WT, and TC are highlighted in underlined-bold, bold, and italics-bold, respectively.

V Results

The obtained results of our proposed method are presented in Table 1. The bold numbers are shown advantages comparing to existing methods.

VI Evaluation

Our model achieved competitive results. Our model can be argued that models that are insensitive may result in losing information in some cases. Furthermore, our model showed competency in detecting abnormal tumor with variety of shapes. Finally, for cases that are considered difficult for all models. Regarding the quantitative results shown in Table 1, our models were compared against 2 original CNN-based, i.e. U-Net and V-Net, and 2 GAN-based models, i.e. VoxelGAN and Vox2Vox. Although our residual U-Net alone can achieve adequate results, its scores were relatively smaller than those of V-Net and Vox2Vox. By using the proposed methodology, our multi-scale GAN gained highest scores and smallest distances in whole tumor assessment and increased the baseline model's scores of enhancing tumor and tumor core classes significantly (roughly 3 to 6%)..

VII Conclusions

In this research, we proposed a segmentation model using GAN models. The novelty of our method focused on the construction of multi-scale GAN, the modification of training process and loss functions. Furthermore, we performed data pre-processing, developed a visualization application for 3D brain MRI and tumors, evaluated our model quantitatively and qualitatively against other 4 models. Our model achieved the most prominent results of all graded models.

References

[1] Sinh Van Nguyen, Ha Manh Tran and Marcin Maleszka. Geometric modeling: Background for processing the 3D objects. Journal of Applied Intelligence Vol. 51, pp. 6182–6201, 2021.

[2] Zhaoa, ZengShun, at. al. Semantic Segmentation by Improved Generative Adversarial Networks. Computer Vision and Pattern Recognition (cs.CV), FOS: Computer and information sciences, FOS: Computer and information sciences. doi 10.48550/ARXIV.2104.09917, 2021.

[3] Baid, Ujjwal et. al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv, doi. 10.48550/ARXIV.2107.02314, 2021.

DOI: https://doi.org/10.1007/978-3-031-21967-2_28
Link to repository: https://bitbucket.org/indigoYoshimaru/facemask-web

Keywords: Face Mask Detection, Covid-19, Single-stage, Real-time, Face Mask Dataset, Deep Learning, YOLO, Web Application

Single-stage real-time face mask detection

Linh Phung-Khanh, Bogdan Trawiński, Vi Le-Thi-Tuong, Anh Pham-Hoang-Nam, Nga Ly-Tu

phungkhanhlinh.iu@gmail.com, bogdan.trawinski@pwr.edu.pl, lttvi1822@gmail.com, phna0220@gmail.com, ltnga@hcmiu.edu.vn

Simplified Title

Single-stage real-time face mask detection

Abstract

With the battle against COVID-19 entering a more intense stage against the new Omicron variant, the study of face mask detection technologies has become highly regarded in the research community. We still noticed three research gaps that our contributions could possibly suffice. Firstly, despite the introduction of various mask detectors over the last two years, most of them were constructed following the two-stage approach and are inappropriate for usage in real-time applications The second gap is how the currently available datasets could not support the detectors in identifying correct, incorrect and no mask-wearing efficiently without the need for data pre-processing. The third and final gap concerns the costly expenses required as the other detector models were embedded into microcomputers such as Arduino and Raspberry Pi. In this paper, we will first propose a modified YOLO-based model that was explicitly designed to resolve the real-time face mask detection problem; during the process, we have updated the collected datasets and thus will also make them publicly available so that other similar experiments could benefit from; lastly, the proposed model is then implemented onto our custom web application for real-time face mask detection. Our resulted model was shown to exceed its baseline on the revised dataset, and its performance when applied to the application was satisfactory with insignificant inference time.

I Introduction

The paper amplifies the need to have a fast and precise face mask detector under certain circumstances such as the COVID-19 pandemic. Even when the pandemic is gone, in hazardous environments, i.e. hospitals, laboratories, and chemical exposure environments, workers should also be under surveillance to ensure workplace safety. Furthermore, we addressed our concerns in terms of real-time prediction and the reasons behind choosing a single-staged detector over two-staged models. Last but not least, the datasets' imbalance and lack of data sources, and the setup of the face mask detection model in real-life applications are among the motivations for this paper. The assumption includes the crowded areas with a variety of types of face masks to be our system surveillance environment. Our system is also assumed to run over the network so it is fundamental that the areas have stable networks and sufficient servers.

II State of the Art

The number of currently obtainable single-staged face mask detectors are still surprisingly limited such as RetinaFaceMask [1], whose results was in pure precision and recall for each class (face and mask), and SE-YOLOv3 [2]. However, most of current work are two-staged models that could either slow or expensive when a large number of stations or set-up places is required [3, 4].

III Original Contribution

The proposed method includes the modification of YOLOv5 architecture, datasets update and creation, and web-based implementation. The modified model was trained on both original and new datasets and tested against other versions of YOLO family. The proposed method showed 1.3% increase in AP@50 score comparing to baseline and promises an industrial approach rather than stationary hardware implementation.

IV Methodology

The method starts with using Kmeans to cluster the 12 anchor boxes that represents the 4 scaled output of the model. Then, we use the Ghost Bottleneck and Ghost Convolution as the main blocks of the network. Ghost Convolution is applied using linear, cheap operations to create a more enriched result with less computation complexity. The output of the Ghost Convolution can be summarized as the concatenation of 2 convolution block by filters:

$$Y = (y_1, y_2) = (conv(X), \Phi(y_1))$$

We then modified the dataset PWMFD to minimize the dataset imbalance by (1) crawled 240 images of incorrectly-worn masked-faces from multiple sources, (2) performed data augmentation on them to give a total of 1200 images, (3) auto labelling using RetinaFace model to detect all available faces, (4) manually evaluate and adjust the annotation.

Finally, we integrated our model to a simple web application using Flask, SocketIO and IP camera connection.

V Results

While performing real-time detection on the application, our model managed to achieve an average of 50 FPS on the testing laptop with inference time as little as 0.0027 seconds. Our web application also include notifications for frames that have violations of the condition defined int the paper. Final results are shown in Fig. 5 and Fig. 6 of the paper.

VI Evaluation

As shown in Table 1, our model achieved relative high result of 97.5% for AP_{50} and 86.9% for AP_{75} after training 100 epochs. After having been trained for 200 epochs, our model attained the highest result of 97.6% AP_{50} and 88.4% AP_{75} while still maintaining reasonable inference time. Additionally, our model is relatively light-weighted comparing to YOLOv3 and YOLOv4.

Model	**Size (MB)**	**Avg. inference time per image (seconds)**	AP_{50}	AP_{75}	AP
YOLOv3	234.9	0.0946	0.860	0.405	-
YOLOv4	244.2	0.1040	0.967	0.732	-
YOLOv5s - 100 epochs	13.7	0.0281	0.962	0.853	0.717
YOLOv5s - 200 epochs	13.7	0.0273	0.969	0.875	0.739
Our model - 100 epochs	13.9	0.0318	0.975	0.869	0.735
Our model - 200 epochs	13.9	0.0321	**0.976**	**0.884**	**0.749**

Table 1: Face mask detectors comparison on PWMFD

VII Conclusions

We first proposed a modified YOLO-based face mask detection model using CSP and Ghost module, which was able to detect the wearing of masks among many people and determine whether the masks' statuses are valid with respect to the regulations from WHO. Along with the development of this model, we also presented an updated version of the PWMFD dataset and experimented with a synthetic dataset called FMITA. A web application was also built to allow the proposed model to demonstrate its mentioned abilities in real-time with very low inference time per image.

References

[1] Jiang, M., Fan, X., and Yan, H. Retinamask: A face mask detector. *arXiv preprint arXiv:2005.03950* (2020).

[2] Jiang, X., Gao, T., Zhu, Z., and Zhao, Y. Real-time face mask detection method based on yolov3. *Electronics 10*, 7 (2021), 837.

[3] Susanto, S., Putra, F. A., Analia, R., and Suciningtyas, I. K. L. N. The face mask detection for preventing the spread of covid-19 at politeknik negeri batam. In *2020 3rd International Conference on Applied Engineering (ICAE)* (2020), pp. 1–5.

[4] Yadav, S. Deep learning based safe social distancing and face mask detection in public areas for covid-19 safety guidelines adherence. *International Journal for Research in Applied Science and Engineering Technology 8*, 7 (2020), 1368–1375.

DOI: https://doi.org/10.1007/978-3-031-21743-2_13

Keywords: driver maneuvers, driver profiling

A stable method for detecting driver maneuvers using a rule classifier

Piotr Porwik[0000-0001-8989-9478], Tomasz Orczyk[0000-0002-4664-8369], Rafal Doroz[0000-0001-6103-1175]

piotr.porwik@us.edu.pl,tomasz.orczyk@us.edu.pl,rafal.doroz@us.edu.pl

Simplified Title

A stable method for detecting driver maneuvers using a rule classifier

Abstract

Traffic accidents and vehicle mishandling are significant problems in road transportation, affecting human lives. Various studies suggest that driver behavior is a key factor in the most road accidents and contributes significantly to fuel consumption and emissions. Improvements in driver behavior can be achieved by providing feedback to drivers on their driving behavior. The identification of risky and wasteful maneuvers allows the evaluation of driver behavior. This allows the elimination of irresponsible drivers who pose a danger in traffic, and at the same time, it allows the reduction of maintenance and repair costs of the vehicle fleet. This paper presents the first stage of a driver profiling method based on the analysis of signals coming from the vehicle CAN bus and auxiliary device containing a GPS receiver and an IMU unit. No additional equipment is needed, what is an advantage of the proposed method.

I Introduction

Around the world, road traffic is growing at a tremendous rate, year after year. This is due to economic needs, supply chains to hard-to-reach places, or the convenience of the individual car, including car long- or short-term leasing and rental services. Unfortunately, the increase in traffic is also associated with an increase in the number of accidents involving drivers and other road users. Every year, thousands of people lose their lives or are seriously injured in road accidents. Road accidents generate costs related to the treatment and rehabilitation of people, repair of cars, or repair of road infrastructure. Any erroneous decision by a driver can lead to a dangerous traffic incident that must be effectively addressed. This problem becomes even more complex when the traffic is heterogeneous and involves different types of vehicles.

The issue of vehicle damage and insurance is very important from the business point of view of car rental and insurance companies because it always causes losses for them. After an accident, the car must be repaired, and tangible and intangible damages must be covered, which is often the cause of legal disputes. This is also important for car rental companies, where customers do not take care of the rental cars and often damage them through irresponsible driving or mishandling. Some of these behaviors are difficult to detect and only manifest themselves after some time in poor condition of the vehicle. This user behavior is increasingly prompting companies to install devices in vehicles that monitor the driver's driving [1]. This makes it possible, in particularly extreme cases, to refuse to rent the vehicle again.

II State of the Art

Most studies devoted to analyzing driver behavior using machine learning methods focus on analyzing the age and gender of the driver, looking at the impact of using additional devices while driving, checking for the presence of vehicles on side streets, the intensity of pedestrian, bicycle or vehicle traffic in front of, from behind or from the opposite direction. These solutions can mainly be used in laboratory simulators, where the prediction of driver behavior can be measured, and then some recommendations can be formed. As a result, this leads to increased road safety by imparting this knowledge during driver training. Some driving simulator studies have analyzed driver behavior under various road conditions. In addition, the study focused on driver behavior at intersections with traffic lights. Longitudinal acceleration and longitudinal speed were measured, as well as throttle pressure, deceleration during braking, and brake pedal force. However, these measurements were carried out on simulators and have not been conducted under actual road conditions.

III Original Contribution

In this paper, a telemetry-based method for detecting maneuvers is proposed. The goal of this method was to construct a relatively simple method of identifying maneuvers, using a composition of telemetry data streams (CAN, IMU, GPS), which could be implemented in a so-called On-Board Unit module. This could be considered as a pre-processing stage of a cloud-based driver assessment system.

IV Methodology

The research had an experimental nature. Data for identifying (and scoring) various maneuvers were collected on a closed test track with two types of surface (concrete slabs, and skid plate), by several drivers assisted by a pilot using a proprietary, multichannel data logging device. Drivers were following the route prepared by road safety specialists. Additionally, during the experiment, the driver assistant was marking the beginning and end of each maneuver using a digital tablet with specialized software. During the drive, data from the CAN bus, IMU and GPS were recorded. Based on these signals, it is possible to detect the type of maneuver the car driver is performing at any given time. Based on these measurements, a rule classifier model was built.

V Results

Automatic detection of maneuvers has been done using experimentally established rules, which are based on observing the telemetry data. It has been shown that maneuver recognition based on the rule classifier is more accurate than GPS detection and assistant-pilot indications. Obtained overall maneuver detection accuracy was above 95

VI Evaluation

The correctness of the maneuver detection rules was verified on 7 loops of a 3 km long public road under normal urban traffic conditions.

VII Conclusions

Presented algorithm - maneuvers detection is only the first stage for assessing the safety of maneuvers and evaluating the style of driving. It allows marking fragments of the telemetry data stream for secondary feature extraction and further evaluation. In future research, we plan to find an optimal set of parameters for detecting maneuvers and determining driver driving style. The driving style will be evaluated based on additional parameters such as driving speed, steering wheel jerking, or acceleration.

References

[1] Karri, S. L., De Silva, L., Lai, D. T. C., and Yong, S. Y. Identification and classification of driving behaviour at signalized intersections using support vector machine. *International Journal of Automation and Computing (18)* (2021), 480–491.

DOI: https://doi.org/10.1007/978-3-031-21967-2_2

Keywords: natural language processing, book classification, clustering, text analysis, machine learning

A comparative study of classification and clustering methods from text of books

Barbara Probierz[0000-0002-5122-2645], Jan Kozak[0000-0002-2128-6998], Anita Hrabia[0000-0002-1282-1995],

barbara.probierz@ue.katowice.pl, jan.kozak@ue.katowice.pl, anita.hrabia@ue.katowice.pl

Simplified Title

A comparative study of classification and clustering methods from text of books

Abstract

We propose to create an appropriate model system, the use of which will allow for automatic assignment of books to appropriate categories by analyzing the text from the content of the books. Our research was tested on a database consisting of 552 documents. Well-known techniques of natural language processing (NLP) were used for the proper preprocessing of the book content and for data analysis. Then, two different machine learning approaches were used: classification and clustering in order to properly assign books to selected categories. Measures of accuracy, precision and recall were used to evaluate the quality of classification. In our research, good classification results were obtained, even above 90% accuracy. Also, the use of clustering algorithms allowed for effective assignment of books to categories.

I Introduction

Book collections in libraries are an important means of information, but without proper assignment of books into appropriate categories, searching for books on similar topics is very troublesome for both librarians and readers. This is a complex problem due to the analysis of large sets of real text data, such as the content of books. The attempt to make electronic texts available in the public domain is Project Gutenberg, initiated by Michael Hart in 1971. It is an online library of free e-books. As a volunteer effort, Project Gutenberg aims to digitize, archive and distribute literary works. The mission of the project is to encourage all interested people to create e-books and help in their dissemination. Therefore, we propose to develop a method of automatic classification of books based on their content, to facilitate the search for books on similar topics.

II State of the Art

A very important problem in NLP is to understand the text you read and based on it to classify the documents [3]. For this purpose, researchers are trying to develop a model for automatic document categorization using NLP and machine learning methods such as classification or clustering [2]. The combination of classification and clustering methods was used by the authors [1] to develop a method for classifying scientific articles based on the subject of clusters. This solution is aimed at grouping scientific articles into categories in order to overcome the difficulties faced by users looking for scientific papers.

III Original Contribution

Our hypothesis is that, based on the text of the book, it can be efficiently and automatically categorized into appropriate categories. In this article, we define categorization as assigning a book to one of five categories, i.e., art, biology, mathematics, philosophy and technology, regardless of the machine learning method used. We transformed text from the content of a book into a decision table containing information about the occurrence of words in the text. For this purpose, by using one of the measures (TF-IDF, TF, Binary) to analyze the content of books, we obtain word weight vectors, which we transform into a decision table. The columns in the table are words selected by the word weighting measures. There are subsequent books in the lines, and the intersection of the line and column is the word score for the book (depending on the word weights selected). The last column is the decision, on which the book's appropriate category is written.

IV Methodology

The proposed approach is to automatically assign books to the appropriate categories based on the performed text analysis from the content of the books by applying NLP techniques and machine learning algorithms. In the first step of our approach, the content of all the books should be read along with the actual categories and then the data preprocessing process. First, natural language processing (NLP) techniques were used to transform text from the content of a book into a decision table containing information about the occurrence of words in the text. Then, the prepared decision table was used for supervised learning (classification) and for unsupervised learning (clustering). The algorithms used for the classification were CART, Bagging, Random Forest, AdaBoost and SVM, while the K-means algorithm was used for clustering. Accuracy, precision and recall measures were used to evaluate the quality of the classification. An additional aim is to see how the quality of the approach is affected by the size of the training set and the size of the test set.

V Results

For the classifiers, we used a training set containing 90% of all books and a test set containing the remaining 10% of books. This is a model case, which allows us to present differences between specific word weighting measures. The results were divided due to the measure of the quality of classification. Thus, in terms of accuracy, TF-IDF is the best measure. Only for the Bagging and AdaBoost algorithms slightly better results are obtained for the other measures, but the average value is for: TF: 0.794; TF-IDF: 0.848; Binary: 0.798. An analogous situation is in the case of recall, where the mean values are for TF, TF-IDF and Binary, respectively: 0.732, 0.798 and 0.724. On the other hand, in the case of precision, the best results are obtained for Binary. The average values are TF: 0.804; TF-IDF: 0.814; Binary: 0.840. The results for the K-means algorithm are slightly different. For the TF measure, the books were assigned to two clusters. For the Binary measure, only math books were assigned to one cluster, and the most books were assigned to the other cluster. For the TF-IDF measure, documents can be clearly grouped by using unsupervised learning methods, and the obtained results are close to reality. Additionally, as in the case of the Binary measure, many books were assigned to the last cluster.

VI Evaluation

The aim of our experiments was to test whether, on the real data set (prepared by us on the basis of the books from Project Gutenberg), the proposed approach achieves good performance as measured by various measures of classification quality, such as accuracy, precision and recall. In addition, we have shown that the use of clustering algorithms allows for the effective assignment of books to categories. Thus, it can be seen that TF-IDF is a measure whose application allows for potentially good results. In the case of Bagging, Random Forest and SVM algorithms, it allows prediction with an accuracy of around 90%. However, in the case of Random Forest and SVM also recall, and precision is close to 90%. For unsupervised learning methods, documents can be clearly grouped by using the K-means algorithm, and the obtained results are comparable to reality.

VII Conclusions

On the basis of the experiments carried out, it was confirmed that the application of NLP for the prediction of a book category, based on the analysis of its content, allows to achieve accuracy, precision, and recall of 90%. This is achieved when using the TF-IDF measure and the Random Forest algorithm. However, it should be noted that higher precision can be achieved with the Binary measure and the Bagging algorithm. Additionally, it should be noted that in the case of the TF-IDF measure, it is possible to obtain similar results using both the classification and clustering algorithms. In the case of book clustering, good results were obtained by combining the K-means algorithm (for the number of clusters corresponding to the actual number of decision classes) with the TF-IDF measure. The resulting clusters largely correspond to the actual categories of books.

References

[1] Jalal, A. A., and Ali, B. H. Text documents clustering using data mining techniques. *International Journal of Electrical & Computer Engineering (2088-8708) 11*, 1 (2021).

[2] Selivanova, I., Kosyakov, D., Dubovitskii, D., and Guskov, A. Expert, journal, and automatic classification of full texts and annotations of scientific articles. *Automatic Documentation and Mathematical Linguistics 55*, 4 (2021), 178–189.

[3] Yao, L., Mao, C., and Luo, Y. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence* (2019), vol. 33, pp. 7370–7377.

DOI: https://doi.org/10.1007/978-3-031-21743-2_11

Keywords: Neurodegeneration, rough set theory, intelligent predictions

Rough set rules (RSR) predominantly based on cognitive tests can predict Alzheimer-related dementia

Andrzej W. Przybyszewski[0000-0002-0156-7856], Kamila Bojakowska, Jerzy P. Nowacki[0000-0001-7912-4716], Aldona Drabik, BIOCARD Study Team

przy@pjwstk.edu.pl, kamila.bojakowska@cskmswia.pl, nowacki@pjwstk.edu.pl, adrabik@pjwstk.edu.pl

Simplified Title

AI can predict dementia from cognitive tests

Abstract

Technology progress helped us to live better and longer, but aging is the major factor related to ND (neurodegenerative diseases) such as Alzheimer's or Parkinson's disease. Alzheimer's disease (AD) correlated neurodegenerative processes begin over 30 years, whereas cognitive changes begin about 15-11 years, before the first AD symptoms. The purpose of our study was to predict if 'healthy' subjects might get AD dementia soon. We have analyzed BIOCARD data from the project that started with 349 normal subjects and followed over 20 years with over 150 different attributes. Subjects were evaluated every year by neurologists with the global score CDR (Clinical Dementia Rating) parameters to determine if a particular individual is normal or has Mid Cognitive Impairment (MCI) or dementia. We have used classification based on CDRSUM (sum of boxes) as a more precise and quantitative general index than the global score to provide more information on patients with mild dementia. CDRSUM values for prodromal patients are 0.0: cognitive normal; (0.5-4.0): questionable cognitive impairment; (0.5-2.5): questionable impairment; (3.0-4.0): very mild dementia; (4.5-9.0): mild dementia. We have obtained rough set rules (RSR) from Model1: 149 patients classified as AD, MCI, and normal; and Model2: 40 patients with AD. By using Model1 classified by neurologists as 21 cognitive normal (CDR=0) subjects, with our classification based on RSR, we have obtained 8 subjects with CDRSUM $>$ 0: all 8 subjects were above 0.75, one subjects between 0.75 and 1.25, and 5 subjects between 0.75 and 2.25, and two subjects were above 2.25. These subjects might have questionable cognitive impairment. Using Model2 we found with RSR that two subjects had CDRSUM between 4.5 and 6.5, which means they might have mild dementia (4.5-9.0). RSR consist of algorithms that might predict future cognitive AD-related impairments in individual, normal, healthy subjects.

I Introduction

Cognitive changes are dominating the most common neurodegenerative disease (ND) - Alzheimer?s disease (AD). In most cases of AD, neurodegeneration starts from the hippocampus and frontal cortex, and it is related to memory and orientation problems. With the disease progression, other brain regions become also affected. As each patient has dissimilar neurodegeneration development and compensation in consequence symptoms might be various and finding optimal treatment is an art for an experienced neurologist. We have estimated disease progression with sets of psychophysical attributes found as the most meaningful in patients from the BIOCARD study [1] and combined them with the results of the APOE. The risk of AD increases and the age-at-onset decreases with the number of APOE4 alleles. This study is the continuation of the rough set theory application to follow predominantly the cognitive changes in neurodegenerative diseases (ND) such as Parkinson's [3] and now Alzheimer's diseases..

II State of the Art

We are living longer and in consequence, the prevalence of Alzheimer's disease (AD) has increased. Chances to get AD-related dementia are increasing with the age and there are about 50% for ages above 90 years. There is no cure for AD as the neurodegeneration brain processes begin several decades before the first symptoms. When a patient notice problems with orientation and/or episodic memory, a large portion of her/his brain is already dead, and we do not know how to revive dead cells. Therefore, a possible solution is to find early biomarkers when the neurodegeneration processes are in the not-advanced phases. Actually, there are several such early biomarkers like changes in the brain structures (MRI) or in the cerebrospinal fluid (CSF). The main problem is that they are not practical, as it is difficult to get brain scans or test CSF in all people after 60 years of age. There is another biomarker related to cognitive changes, but the question is how to differentiate the subtle cognitive problems related to aging from that related to the beginning of AD.

III Original Contribution

Our method is based on the brain's visual system where we easily recognize objects because we know them. Therefore, we have compared normal patients with two Model groups: 1) Model1 consists of mixed of normal, in early AD (MCI - mild cognitive impairment), and AD; 2) Model2 consists of only AD patients. By using the Granular Computing (CG) method we have found rules that connect the results of cognitive tests with the CDRSUM (cognitive dementia rating sum) that quantifies the actual stage of the patient. By using these rules in the cognitive test results of normal patients we can estimate if their results indicate early symptoms of AD. The rules from Model1 indicate very early AD symptoms and the rules from Model2 more advanced early symptoms.

IV Methodology

Granular Computing is well established AI method that divides data sets into abstract granules that might be seen as elementary sets (atoms) with distinctive functional properties gathering together many different features. Our study is using GC to estimate the meaning of different cognitive granules.

V Results

We have evaluated 11 cognitive test results and APOE genotype (as 0, 1). As the cognitive data are noisy and often contradictory we have used a Rough Set implementation [2] of the GC. Rough Set Rules (RSR) can be certain when all data are consistent, or uncertain when some of the data are contradictory (lower and upper approximation). Generally, our rules are indicative (uncertain) as they are from different patient populations, but we have increased their confidence by choosing different attribute subsets and obtaining consistent results. We have tested 21 normal subjects and found with help of Model1 that 8 might get questionable impairment and two of them very mild dementia. On the basis of Model2, we have found that two normal patients might actually have mild dementia.

VI Evaluation

Researchers in [1] found that on the basis of CSF results, MRI scans, and cognitive tests they can predict at a 5-years period the likelihood of progression from cognitively normal to MCI. We have used only cognitive test results and on our model basis, we have demonstrated that some cognitive normal (clinical criteria) individuals might have already shown the beginning of the cognitive impairments to a different extent.

VII Conclusions

Granular Computing by finding very early and early cognitive impairments in individual cognitive-normal subjects from predominantly cognitive online tests give a chance to find a cure for Alzheimer's disease.

References

[1] Albert, M., Zhu, Y., Moghekar, A., Mori, S., Miller, M., Soldan, A., Pettigrew, C., Selnes, O., Li, S., and Wang, M.-C. Predicting progression from normal cognition to mild cognitive impairment for individuals at 5 years. *Brain : a journal of neurology 141* (01 2018).

[2] Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning about Data.* 1991.

[3] Przybyszewski, A., Kon, M., Szlufik, S., Szymański, A., Habela, P., and Koziorowski, D. Multimodal learning and intelligent prediction of symptom development in individual parkinson's patients. *Sensors 16* (09 2016), 1498.

DOI: https://doi.org/10.1007/978-981-19-8234-7_40

Keywords: Facial expression, human-robot interaction, edge device, deep learning

An Efficient Multi-view Facial Expression Classifier Implementing on Edge Device

Muhamad Dwisnanto Putro[0000-0002-1785-1018], Duy-Linh Nguyen[0000-0001-6184-4133], Adri Priadana, Kang-Hyun Jo[0000-0002-4937-7082]

dputro@mail.ulsan.ac.kr, ndlinh301@mail.ulsan.ac.kr, priadana3202@mail.ulsan.ac.kr, acejo@ulsan.ac.kr

Simplified Title

Multi-pose facial expression recognition on low-cost device

Abstract

The robotic technology demands human-robot interaction to implement a real-time facial emotion detector. This system has a role in recognizing the expressions of the user. Therefore, this application is recommended to work quickly to support the robot's capabilities. It helps the robot to analyze the customer's face effectively. However, the previous methods weakly recognize non-frontal faces. It is caused by the facial pose variations only to show partial facial features. This paper proposes a multi-view real-time facial emotion detector based on a lightweight convolutional neural network. It offers a four-stage backbone as an efficient feature extractor that discriminates specific facial components. The convolution with Cross Stage Partial (CSP) approach was employed to reduce computations from convolution operations. The attention module is inserted into the CSP block. These modules also support the detector to work speedily on edge devices. The classification system learns the information about facial features from the KDEF dataset. As a result, facial emotion recognition achieves comparative performance to other methods with an accuracy of 97.10% on the KDEF, 73.95 on the FER-2013, and 84.91% on the RAFDB dataset. The integrated system using a face detector shows that the system obtains a data processing speed of 30 frames per second on the Jetson Nano.

I Introduction

Vision and perception are essential fields and have been applied to improve the Human-robot interaction method. Especially for a social robot that requires facial expression recognition to understand human emotion. The robot can get the feeling information based on this gesture. Face provides a basic component to recognize emotion. In addition, there are six human facial expressions, such as anger, fear, disgust, sadness, surprise, and happiness [4]. Facial movements affect each facial expression differently. An expression has a set of distinctive features that produce a distinctive form. Thus, this method is adopted by a vision-based facial expression system to recognize every facial emotion. Feature extraction is the core work of a recognition system. It separates important information from useless features [15, 14, 16]. Besides, valuable features help the classification system obtain high accuracy. The convolutional neural network has proven to be a powerful feature extractor for capturing the distinctive features of a face. A weighted filter is applied to separate facial features that affect the recognition system. Many architectures have been present, which have been employed as the backbone of the model [18, 13, 5]. However, they need a graph accelerator to operate quickly in the inference stage. Therefore, they tend to produce a large number of parameters and computations. On the other hand, a recognition system is required to be implemented in an edge device such as a Jetson Nano [11]. This device is compatible with sensors and actuators commonly used in robots. This device is friendly with sensors and actuators, and many robots use it.

II Original Contribution

Real-time facial expression recognition is proposed to identify multiple poses on human facial emotions. The main contributions of this work are as follows:

1. A novel facial expression recognizer efficiently identifies the human facial feeling in different face views.
2. A light architecture using CNN architecture upgrades Cross Stage Partial (CSP) to reduce the computational operations and parameters. It also proposes a novel excitation module to enhance the backbone performance.
3. The classification model achieves high performance and can operate in real-time processing speed of 30 FPS on a Jetson Nano.

III Methodology

A real-time system consists of a face detection LWFCPU [12] and facial expression classification. A CNN-based classification system employs the backbone as an essential module to produce specific features in the classification model. A four-stage light backbone was offered using a sparse convolution operation shown in Fig. 2.

A superficial model is weak in distinguishing facial elements at a high level. Thus, the attention module is used to capture specific features related to each facial expression [17]. Therefore, the proposed model develops a depthwise excitation module inserted at each skip connection of the CSP method.

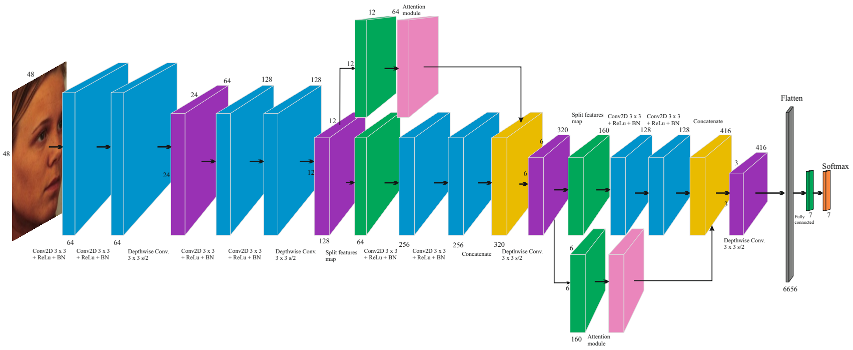


Figure 1: The proposed facial expression classification model.

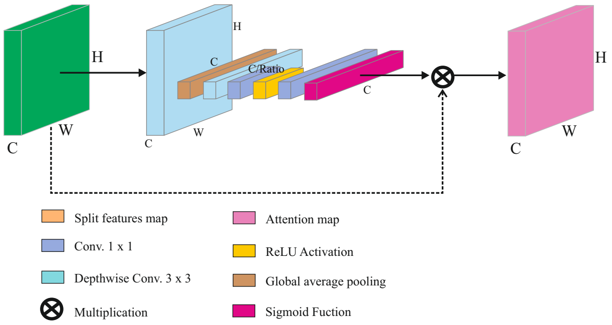


Figure 2: A depthwise enhancement module.

This model was trained and evaluated on several datasets, including the (KDEF) dataset [2], FER-2013 dataset [6], and RAFDB [8] dataset. To enrich the variety of data, it uses random contrast, brightness, saturation, hue, rotation, and flip. In addition, we use Categorical cross-entropy to calculate the loss model, Adam (Adaptive Moment Estimation) as optimizer. It uses K-fold cross-validation to split and evaluate the model on the KDEF dataset on 50 epochs. The learning rate in the training process is 10^{-4} on all datasets. We set 500 epochs in the FER dataset and 200 epochs for the RAFDB dataset.

IV Results

Table 1 shows that the proposed classification model is trained and evaluated on the KDEF dataset, achieving an accuracy of 97.10%. This result outperforms other facial expression methods. In addition, we also train and examine the model for difficult challenges using the FER-2013 dataset, obtaining an accuracy of 73.95%. it is more lower than Ensemble MLCNN [10] and AM-NET [7]. However, our model outperforms SSNs and single MLCNN [10]. Furthermore, we also examine the model accuracy on the RAFDB dataset, which is superior to other architectures, as shown in Table 1.

A real-time face emotion detector was tested on the Jetson Nano with input from a webcam. The proposed detector's speed is compared with competitors, integrated with a face detection [12]. Face detection helps to generate facial regions to avoid disturbance from background elements. Multi-model fusion method generates a small number of parameters than our model. However, the proposed model is more accurate and can operate fast with a data processing speed of 29.58 FPS.

V Conclusions

This paper proposes a facial emotion recognizer to predict seven classes of multi-profile facial expressions in real time. Additionally, it focuses on implementing an edge device. The proposed model contains a four-stage

Table 1: Evaluation of proposed architecture compared to other methods on KDEF, FER-2013 and RAFDB dataset

KDEF dataset	
Method	**Accuracy**
O-FER [3]	91.42
CCFN [19]	91.60
Multi-Xception	94.29
Resnet-19	94.49
Multi-C-Xception	94.63
DFSD-LDA-AdaBoost	95.06
Akhand et al [1]	96.51
Proposed	97.10
FER-2013 dataset	
Multi-scale CNN	72.82
SNNs	73.00
Single MLCNN [10]	73.03
Ensemble MLCNNs [10]	74.09
AM-Net [7]	75.82
Proposed	73.95
RAFDB dataset	
MobileNetV1	81.62
PG-CNN	82.27
DLP-CNN	84.22
A-MobileNet [9]	84.49
Proposed	84.91

Table 2: Runtime efficiency compared to competitors on Jetson Nano

Method	**Paremeter**	**Accuracy(%)**			**Speed of Integrated (FPS)**
		KDEF	**FER-2013**	**RAFDB**	
Multi-model fusion [13]	1,206,279	93.42	-	-	26.38
AM-NET [7]	24,904,204	-	75.82	-	5.71
MLCNN [10]	20,787,783	-	73.03	-	10.89
Ensemble MLCNN [10]	92,825,543	-	74.09	-	Out of memory
Akhand et al [1]	28,907,943	96.51	-	-	Out of memory
A-MobileNet [9]	3,321,513	-	**-**	84.49	20.35
Proposed detector	**2,006,759**	**97.10**	**73.95**	**84.91**	29.58

backbone to efficiently extract the essential elements and a depthwise excitation module to enhance the valuable features. As a result, the proposed model achieves high accuracy with competitive performance compared to the previous methods. The speed of the integrated system achieves 29.57 FPS when working on a Jetson Nano.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the government(MSIT).(No.2020R1A2C200897212)

References

[1] Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., and Shimamura, T. Facial emotion recognition using transfer learning in the deep cnn. *Electronics 10*, 9 (2021).

[2] Calvo, M., and Lundqvist, D. Facial expressions of emotion (kdef): Identification under different display-duration conditions. In *Behavior Research Methods* (1998), vol. 40, p. 109–115.

[3] Dong, J., Zhang, L., Chen, Y., and Jiang, W. Occlusion expression recognition based on non-convex low-rank double dictionaries and occlusion error model. *Signal Processing: Image Communication 76* (2019), 81–88.

[4] Ekman, P. Facial expressions of emotion: New findings, new questions. *Psychological Science 3*, 1 (1992), 34–38.

[5] Fareed, K., Sultan, F., Khan, K., and Mahmood, Z. A robust face recognition method for expression and pose variant images. In *2020 14th International Conference on Open Source Systems and Technologies (ICOSST)* (2020), pp. 1–6.

[6] Goodfellow, I. J., Erhan, D., Luc Carrier, P., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., and Bengio, Y. Challenges in representation learning: A report on three machine learning contests. *Neural Networks 64* (2015), 59 – 63. Special Issue on "Deep Learning of Representations".

[7] Li, J., Jin, K., Zhou, D., Kubota, N., and Ju, Z. Attention mechanism-based cnn for facial expression recognition. *Neurocomputing 411* (2020), 340 – 350.

[8] Li, S., and Deng, W. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing 28*, 1 (2019), 356–370.

[9] Nan, Y., Ju, J., Hua, Q., Zhang, H., and Wang, B. A-mobilenet: An approach of facial expression recognition. *Alexandria Engineering Journal 61*, 6 (2022), 4435–4444.

[10] Nguyen, H.-D., Kim, S.-H., Lee, G.-S., Yang, H.-J., Na, I.-S., and Kim, S.-H. Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks. *IEEE Transactions on Affective Computing 13*, 1 (2022), 226–237.

[11] Pathak, R., and Singh, Y. Real time baby facial expression recognition using deep learning and iot edge computing. In *2020 5th International Conference on Computing, Communication and Security (ICCCS)* (2020), pp. 1–6.

[12] Putro, M. D., Nguyen, D., and Jo, K. Lightweight convolutional neural network for real-time face detector on cpu supporting interaction of service robot. In *2020 13th International Conference on Human System Interaction (HSI)* (2020), pp. 94–99.

[13] Qi, A., Wei, J., and Bai, B. Research on deep learning expression recognition algorithm based on multi-model fusion. In *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)* (2019), pp. 288–291.

[14] Rao, Q., Qu, X., Mao, Q., and Zhan, Y. Multi-pose facial expression recognition based on surf boosting. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (2015), pp. 630–635.

[15] Rujirakul, K., and So-In, C. Histogram equalized deep pca with elm classification for expressive face recognition. In *2018 International Workshop on Advanced Image Technology (IWAIT)* (2018), pp. 1–4.

[16] Santra, B., and Mukherjee, D. P. Local saliency-inspired binary patterns for automatic recognition of multi-view facial expression. In *2016 IEEE International Conference on Image Processing (ICIP)* (2016), pp. 624–628.

[17] Sun, W., Zhao, H., and Jin, Z. A visual attention based roi detection method for facial expression recognition. *Neurocomputing 296* (2018), 12 – 22.

[18] Webb, N., Ruiz-Garcia, A., Elshaw, M., and Palade, V. Emotion recognition from face images in an unconstrained environment for usage on social robots. In *2020 International Joint Conference on Neural Networks (IJCNN)* (2020), pp. 1–8.

[19] Ye, Y., Zhang, X., Lin, Y., and Wang, H. Facial expression recognition via region-based convolutional fusion network. *Journal of Visual Communication and Image Representation 62* (2019), 1–11.

DOI: https://doi.org/10.1007/978-3-031-21967-2_5

Keywords: Hierarchical clustering, Bayesian Rose Tree, Hierarchy constraining

Covariance Controlled Bayesian Rose Trees

Damian Pęszor[0000-0002-3754-2212], Eryka Probierz[0000-0002-6588-1975]

damian.peszor@polsl.pl,eryka.probierz@polsl.pl

Simplified Title

Hierarchical clustering using covariance-based constraints

Abstract

This paper aims to present a modified version of Bayesian Rose Trees (BRT). The classical BRT approach performs data clustering without restricting the resulting hierarchy to the binary tree. The proposed method allows for constraining the resulting hierarchies on the basis of additional knowledge. Thanks to this modification, it is possible to analyse not only the raw structure of the data, but also the nature of a cluster. This allows an automatic interpretation of the resulting hierarchies while differentiating between clusters of different magnitudes, or those that extend significantly through one pair of dimensions while being coherent in a different one. On the basis of the resulting modifications, it is possible to analyse the depth level as a function of likelihood. The developed method allows maximising customisation possibilities and comparative analysis between the nature of clusters. It can be applied to the clustering of different types of content, e.g. visual, textual, or in a modern approach to the construction of container databases.

I Introduction

The purpose of the presented research was to allow the user to create hierarchical structures from data points, that are constrained, so that different nodes at the same depth of hierarchy represent the same concepts, such as day, month or city in spatiotemporal data. The important factor in conducted research was to allow the constraints to be provided in a way that can be interpreted by domain experts rather than controlled strictly by parameters that require tuning to obtain desired results.

II State of the Art

To discover the hierarchical nature of a given set of data, one has to employ an algorithm that will not only cluster such data but also do it hierarchically. Such hierarchies are a subject of assumptions about the way the data is organised. One such method, Bayesian Hierarchical Clustering (BHC) [3] is a standard tool used to find a hierarchy in a given data set. BHC is representative of a common approach to the problem, wherein clusters at one level of the hierarchy are joined together as a higher-level cluster based on some measure of similarity. Since the information about the number of clusters is not present, BHC is a bottom-up approach, so it finds out larger clusters on its way to unify all data as a single, all-encompassing cluster. Such an approach results in a hierarchy represented as a binary tree of clusters. The important issue arrives with such a representation. Since there is a limited number of children at any given node, a very coherent cluster of very similar data points must be represented as either a balanced tree, a cascade, or something in between. Similarly, a data set containing a few clusters that are easily distinguishable from each other, while each includes a closely related set of data points, has to be represented in the same manner. It is, therefore, impossible to tell based on the resulting hierarchy, which of the tree branches correspond to coherent clusters and which correspond to loosely related data, which has to be organised in the same manner due to the restrictions of using the binary tree.

Bayesian Rose Trees (BRT) [2, 1] is a method that solves the issue arising from the usage of binary trees by allowing multiple children for every cluster in a rose tree structure. This results in a multilevel hierarchy that models similar data points as children of a node representing the cluster, while relations between clusters are represented by the entirety of the constructed tree. The resulting hierarchical representation is one of the many possible options, therefore, the BRT algorithm uses the likelihood ratio in order to determine which pair of clusters to merge to build the tree using the bottom-up approach.

While the hierarchy obtained from BRT represents the data well, it does depend on hyperparameters. One can use them to shape the resulting tree, but they require careful tuning and are not intuitively understood by users. Moreover, different levels of the resulting tree correspond to different sizes of clusters and therefore to different concepts. The resulting tree is consequently hard to analyse, especially with automated solutions.

III Original Contribution

The contribution of the presented research is the modification of the Bayesian Rose Trees algorithm which allows to differentiate between depth levels of the resulting hierarchy by constraining the clusters represented at each level by using the covariance matrices to model the possible clusters at each depth. This allows to create trees that clearly represent different hierarchical concepts that can be found in the data.

IV Methodology

The proposed method represents user requirements in the form of L different cluster characteristics over N-dimensional data as a hypermatrix K of order 3. K is, therefore, a concatenation along the third dimension of covariance matrices such that $K = \left[k_{i,j,l}\right] = \left[\Sigma_1|\Sigma_2|\cdots|\Sigma_L\right]$, wherein $i \in [1 \dots N]$, $j \in [1 \dots N]$, $l \in [1 \dots L]$.

The third dimension of K defines a sequence of minimal requirements for a cluster characteristic to correspond to a given level of overarching clustering hierarchy, wherein consecutive covariance matrices correspond to levels of hierarchy. Therefore, all clusters of level l_x incorporate clusters of level $l_y : l_y > l_x$ and are incorporated by clusters of level $l_z : l_z < l_x$. Such a hierarchical approach results in constraints on the covariance matrices so that the modules of covariances never increase as the depth level l increases.

Let the cluster of level l be defined as a rose tree T of which leaves $\mathscr{D} = \text{leaves}(T)$ are such that any element of the covariance matrix of $\mathscr{D}$ is greater than the corresponding element of K, with l being the highest of those which fulfil that criteria.

Then the four possible merge operations defined for BRT; join, two symmetrical absorbs, and collapse, which are compared by BRT using the likelihood ratio between the possible merged cluster and existing ones can be controlled by an additional factor depending on cluster depth levels.

- Join, wherein $T_m = \{T_i, T_j\}$ is always available.
- Absorb, wherein $T_m = ch(T_i) \cup \{T_j\}$ requires that $l(T_m) = l(T_i)$.
- Absorb, wherein $T_m = \{T_i\} \cup ch(T_j)$ requires that $l(T_m) = l(T_j)$.
- Collapse, wherein $T_m = ch(T_i) \cup ch(T_j)$ requires that $l(T_m) = l(T_i) = l(T_j)$.

Note that for the classical BRT $k_{i,j,k} = 0$ or alternatively $L = 0$. The proposed algorithm is, therefore, a generalisation of BRT.

V Conclusions

This article presented a modified constrained hierarchical clustering algorithm based on BRT. The classical BRT approach was enhanced with the constraining method based on the range of covariances. Such an application is essential when analysing real-world datasets, which allows the data to be created and explained by simple models. It is an unquestionable advantage of BRT that it allows for better data representation; however, simply modelling the data may not be sufficient in many cases, where additional processing is done automatically. This is especially true when there is a need for a comparative analysis of the clusters obtained. This problem can be solved using the proposed modification, which, based on the range of covariance, allows for a deeper and more meaningful contextual analysis of the data. It is also extremely important that the possibility of constraining the resulting hierarchies is formulated in a way that allows experts to use contextual domain knowledge. The proposed solution allows to define the range of the resulting hierarchies by means of understandable constraints described by covariance matrices, which makes the solution useful in a wide range of applications. BRT solutions offer great opportunities on the topics of autonomic taxonomy construction, text-based causality modelling, and hierarchy extraction. The hierarchical nature of the clustering allows for use in computer graphics, wherein a multiresolution representation of the shape is needed to provide the best quality without the overhead. Similarly, the hierarchical structure is useful for recognising planes and obstacles under noisy circumstances in the case of computer vision algorithms or when controlling the hierarchical shape by animation using segmented surfaces. However, these need some constraints to be useful, which can be applied using the proposed solution.

References

[1] Blundell, C., Teh, Y. W., and Heller, K. A. Discovering non-binary hierarchical structures with bayesian rose trees. *Mixtures: Estimation and Applications* (2011), 161–185.

[2] Blundell, C., Teh, Y. W., and Heller, K. A. Bayesian rose trees. *arXiv preprint arXiv:1203.3468* (2012).

[3] Heller, K. A., and Ghahramani, Z. Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning* (2005), pp. 297–304.

DOI: https://doi.org/10.1007/978-3-031-21967-2_9

Keywords: Multi-class classification, Multiclass classification, Multinomial classification, Performance, Sensitivity, Specificity, Accuracy.

Aggregated performance measures for multi-class classification

Damian Pęszor[0000-0002-3754-2212], Konrad Wojciechowski[0000-0003-4679-2667]

damian.peszor@polsl.pl, kwojciechowski@pjwstk.edu.pl

SIMPLIFIED TITLE

Performance estimation in multinomial classification

ABSTRACT

This paper aims to present an approach to the generalisation of performance measures commonly used in binary classification to the field of multinomial classification to use them in hyperparameter estimation for various machine learning methods and similar techniques. The classical approach is to use a binary classification in which each representative of any incorrect class is considered a representative of an umbrella class that is a union of all incorrect classes. Such an approach leads to the removal of important information from the classification process and therefore to the lower value of each experiment for determining the gradient when trying to optimise hyperparameters. We propose aggregated performance measures that can be thought of as an analogue of classical ones. The proposed measures better represent the multinomial nature of such algorithms and obtain more valuable information that allows selecting the correct direction while analysing the gradient of the resulting measures.

I INTRODUCTION

Classification is one of the most prevalent problems in the applying machine learning or statistics-based techniques. It is widely used in many different fields, from the classification of obstacles of flying UAVs, to the classification of patients potentially suffering from neurodegenerative diseases to facial recognition. Many classification algorithms depend on hyperparameters that have to be adjusted in order for the algorithm to work properly in a given domain. Such adjustment is rarely based on domain knowledge, often done automatically by searching the hyperparameter space for the optimum performance of the classification obtained.

II STATE OF THE ART

The most common tools for evaluating algorithm parameterisations are the measures derived from binary classification, such as True Positive Rate (TPR) and True Negative Rate (TNR). These are also used in multinomial cases, where all points except those where the correct class is the most probable are labelled incorrect, with no distinction between the second-best and the worst fit. Some multivalue measures aggregate the values between multiple classes [2]. However, this can hardly be used to indicate whether a given hyperparameter value improved or worsened classification. Measures can be averaged to obtain a single value, which is called macro-averaging [1]. Alternatively, micro-averaging can be used, where the summation is performed separately for the numerator and denominator [3]. In micro-averaging, more numerous classes for which there is a stronger belief that the value of the measure is representative, have a stronger influence over the aggregated measure. However, the reason for classification might be based on the assumption that the less-represented classes, e.g. uncommon diseases, are interesting ones, so such an approach might suppress the important objectives of the classification. Since averaged measures are still based on the underlying binary classification, the change of position of class in ranking is irrelevant as long as the position is not the first one. Therefore, the values of measures are often the same, despite the classification results being different, and can be highly influenced by a single case that is opposite to the entirety of classification results in the case of multinomial classification.

III ORIGINAL CONTRIBUTION

In the study presented, the aggregated performance measures were proposed, aiming to capture the performance of multi-class classification instead of a binary one. The proposed approach leverages the order of the list of classes sorted by the probability of being the correct classes for each data point.

This allows for a better selection of the parameterisation of a classification algorithm due to a higher sensitivity to the overall behaviour of the method rather than only to the binarized result of classification being either correct or incorrect in terms of the given sample. This approach is especially useful when a human user is to determine the final result of classification based on the options selected by the algorithm, such as in the case of facial re-identification or counselling systems.

IV Methodology

The presented study is primarily theoretical, as it presents a new methodology that can be used to evaluate multinomial classification more effectively. The proposed method assumes that for each data point in the data set $\forall i : o_i \in O$, a point belongs also to a single class O_k, which belongs to the set of all classes $\forall k : k \in K$. In the binary case, the first position of the correct class for a given sample, $p_k(o_i) = 1$, will indicate that a sample is a True Positive for a given class k. In the multinomial case, the proposed method uses a threshold t, so that $|TP_{k,t}|$ is the number of all cases in which $p_k(o_i) \leq t$. The proposed method can be used for various measures, such as TPR. For any such measure, the average over all classes for a given threshold is divided by this threshold and summed as in Eq. 1

$$TPR = \sum_{t=1}^{|K|-1} \frac{\overline{TPR_t}}{t} = \frac{1}{|K|} \sum_{t=1}^{|K|-1} \sum_{k} \frac{|TP_{k,t}|}{|O_k|t} = \frac{1}{|K|} \sum_{k} \left(\frac{1}{|O_k|} \sum_{o_i \in O_k} \sum_{t=p_k(o_i)}^{|K|-1} \frac{1}{t} \right) \tag{1}$$

This can be rewritten using harmonic numbers as in Eq. 2, wherein vacuous summation defines H_0.

$$TPR = H_{|K|-1} - \frac{1}{|K|} \sum_{k} \left(\frac{1}{|O_k|} \sum_{o_i \in O_k} H_{p_k(o_i)-1} \right) \tag{2}$$

The values of the harmonic number H can be pre-calculated in an array, which allows easy access. Therefore, the entirety of the value is quite efficient to calculate, despite the added complexity behind the aggregation of multiple thresholds. While in the case of parameter tuning, this is not needed, one can normalise the range of values as in Eq. 3.

$$T\hat{P}R = 1 - \frac{1}{|K|H_{|K|-1}} \sum_{k} \left(\frac{1}{|O_k|} \sum_{o_i \in O_k} H_{p_k(o_i)-1} \right) \tag{3}$$

V Conclusions

The presented approach is important in many real-life scenarios, in which humans make the final decision and the machine learning approach is used to narrow the range of possibilities to consider. Among others, such scenarios include facial recognition, where the human operator is informed by the system, but his decision might be influenced by another factor, and neurodegenerative disease diagnosis, where extra tests can be performed that will differentiate between a few potential diseases or between the best fitting "healthy" class and the slightly less probable disease that would remain ignored using binary classification. An approach that considers not only the most probable result, but also further ones leads to an algorithm that positions the correct result as high as possible, even if for some reason it might not be the best fit, while otherwise, the correct result might be in a position that does not allow for recognition by a human operator.

The second effect is crucial for the machine learning process. In the one-vs.-all approaches to classification, adjustments to hyperparameters that result in the change of the position of the correct label in the resulting ordering are ignored unless the first position is involved. In the presented approach, the value of the measure is always affected, while traditionally it might not respond to changes in parameters or be driven by fluctuations of a few edge cases.

References

[1] Hossin, M., and Sulaiman, M. N. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process 5*, 2 (2015), 1.

[2] Lachiche, N., and Flach, P. A. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using roc curves. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (2003), pp. 416–423.

[3] Sokolova, M., and Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information processing & management 45*, 4 (2009), 427–437.

Link to arXiv: `https://arxiv.org/abs/2209.14008`
DOI: `https://doi.org/10.1007/978-981-19-8234-7_41`
Link to repository: `http://clip.ipipan.waw.pl/POSMAC`

Keywords: keyword extraction, T5 language model, POSMAC, Polish

Keyword Extraction from Short Texts with a Text-To-Text Transfer Transformer

Piotr Pęzik[0000-0003-0019-5840], Agnieszka Mikołajczyk[0000-0002-8003-6243],
Adam Wawrzyński[0000-0002-1698-2390], Bartłomiej Nitoń[0000-0003-3306-7650],
Maciej Ogrodniczuk[0000-0002-3467-9424]

piotr.pezik@uni.lodz.pl, agnieszka.mikolajczyk@voicelab.ai, adam.wawrzynski@voicelab.ai, filip.zarnecki@voicelab.ai, bartek.niton@gmail.com, maciej.ogrodniczuk@ipipan.waw.pl

Simplified Title

Generation of Keywords based on the Text of a Scientific Article

Abstract

The paper explores the relevance of the Text-To-Text Transfer Transformer language model (T5) for Polish (plT5) to the task of intrinsic and extrinsic keyword extraction from short text passages. The evaluation is carried out on the new Polish Open Science Metadata Corpus (POSMAC), which is released with this paper: a collection of 216,214 abstracts of scientific publications compiled in the CURLICAT project. We compare the results obtained by four different methods, i.e. plT5kw, extremeText, TermoPL, KeyBERT and conclude that the plT5kw model yields particularly promising results for both frequent and sparsely represented keywords. Furthermore, a plT5kw keyword generation model trained on the POSMAC also seems to produce highly useful results in cross-domain text labelling scenarios. We discuss the performance of the model on news stories and phone-based dialog transcripts which represent text genres and domains extrinsic to the dataset of scientific abstracts. Finally, we also attempt to characterize the challenges of evaluating a text-to-text model on both intrinsic and extrinsic keyword extraction.

I Introduction

The main problem discussed in this paper is extraction or generation of keywords from short text passages. That is, when you have a text such as an abstract of a research paper, the task is to generate a small set of words or phrases which describe its content. Approaches to this problem can be *extractive* or partly *abstractive*. In the former case, keywords are extracted and possibly normalized more or less directly from the text of the sample. Abstractive methods can assign labels that may not have occurred in the original text.

Apart from generating keywords using various methods we evaluated their results on the new Polish Open Science Metadata Corpus (POSMAC) which contains over 200 thousand abstracts of scientific publications. In the last step, we showed that our solution created for scientific articles can be successfully applied to other domains: news stories and phone dialogues.

II State of the Art

Several approaches have been proposed so far to tackle the problem of keyword generation:

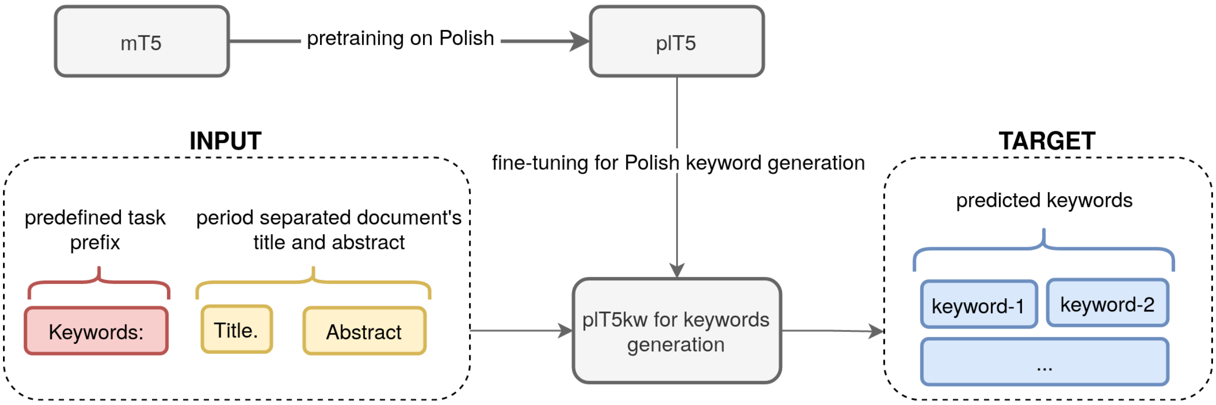
- T5 and plT5 [1] — language models using text-to-text operations,
- FastText — a popular text classification library which uses vector representations of words as input to a relatively simple neural network
- extremeText — FastText extension optimized for large taxonomies
- TermoPL [2] — a statistical terminology extraction tool
- KeyBERT — a keyword extraction library utilizing BERT language model representations.

III Original Contribution

The novelty of the proposed solution consists in the fact that we tested the relevance of transformer models to the task of keyword generation, and we evaluated and released a dataset showing how our system can be transferred to other domains and languages.

IV Methodology

After comparing the results of the state-of-the-art tools we used the plT5-based model trained on six reference corpora of Polish, calling it PLT5KW.



For calculating the upper bound of purely extractive keyword identification we included TermoPL in our experimental study.

V Results

The tuned PLT5KW solution outperforms the other approaches when tested on the original dataset of scientific abstracts. Furthermore, a preliminary analysis of keywords assigned to text from very different domains (news stories and speech transcripts) showed that the proposed solution is capable of generating relevant, properly formatted and well-abstracted keywords.

We envisage further challenges which need to be addressed in future research on this problem. For example, it seems reasonable to assume that open-set keyword extraction could benefit from distributional vector-based techniques of normalizing semantically equivalent keywords. Also, there are potential benefits of zero- or few-shot fine-tuning of text-to-text keyword extraction models to the target domain, which need to be considered more systematically. Finally, the results obtained in this study may vary for different languages, which requires further evaluation, possibly on multilingual variants of the T5 model used.

In addition to evaluating the main approaches, we also assessed several baseline keyword extraction approaches, including FirstPhrases, TopicRank, PositionRank, MultipartiteRank, TextRank, KPMiner and TfIdf with some adjustments aimed at boosting their performance (such as lemmatizing input text). However, the results obtained for all of those methods were very poor.

VI Evaluation

We evaluated the above-mentioned set of complementary keyword generation approaches on the original POSMAC dataset. Abstracts annotated with keywords were split into a training and test set with a ratio of 70/30%.

The relevance and coverage of keyword assignments are evaluated in terms of micro- and macro- precision and recall values, as well as their harmonic means (F_1-scores) averaged over the documents in the test set.

These scores are measured at several ranks (k=1, 3, 5 and more) for each approach in two different scenarios: a) using the full set of keywords assigned in the training and test set and b) training and/or evaluating only on keywords which occur at least 10 times in the dataset.

VII Conclusions

The proposed solution can be successfully used in all applications requiring high-level specification of the content of a document such as index generation, query refinement, text summarization, author assistance, etc.

References

[1] Chrabrowa, A., Dragan, Ł., Grzegorczyk, K., Kajtoch, D., Koszowski, M., Mroczkowski, R., and Rybak, P. Evaluation of Transfer Learning for Polish with a Text-to-Text Model. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)* (Marseille, France, 2022), European Language Resources Association, pp. 4374–4394.

[2] Marciniak, M., Mykowiecka, A., and Rychlik, P. TermoPL — a Flexible Tool for Terminology Extraction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (2016), N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds., European Language Resources Association, pp. 2278–2284.

DOI: https://doi.org/10.1007/978-981-19-8234-7_42
Link to repository: https://github.com/NLP-KazNU

Keywords: machine translation, morphology model, complete set of endings, Turkic languages

Development Kazakh-Turkish Machine Translation on the Base of Complete Set of Endings Model

Aitan Qamet[0000-0003-0392-2732], Kamila Zhakypbayeva[0000-0002-4385-6656],
Aliya Turganbayeva[0000-0001-9660-6928], Ualsher Tukeyev[0000-0001-9878-981X]

qametaitan@gmail.com, zhakypbaeva18@gmail.com, turganbayeva16@gmail.com,
ualsher.tukeyev@gmail.com

Simplified Title

Kazakh-Turkish Machine Translation on Complete Set of Endings.

Abstract

This article discusses Kazakh-Turkish machine translation based on the morphology model of a complete set of endings (CSE). To collect the necessary materials and implement the algorithm, a lot of research and careful work was carried out during the study. The main ones include: creation of a complete set of endings in the Kazakh and Turkish languages; morphological description of endings and creation of a correspondence table of morphological description of endings for two languages; creation of a correspondence table of Kazakh and Turkish stem dictionaries; creation a correspondence table of Kazakh and Turkish stop-word dictionaries; develop Kazakh-Turkish machine translation algorithm based on the CSE model; create software based on this algorithm. The main part of the study consists of fragments of the collected material, as well as the main algorithms and results. Scientific contribution of the article: machine translation technology for Kazakh-Turkish translation based on the CSE model is developed. Experiment results shoes possibility of proposed technology of machine translation for Turkic languages.

I Introduction

The purpose of this research is to create a machine translation technology based on the CSE (Complete Set of Endings) morphology model and demonstrate its results. Many Turkic languages are among the languages with limited resources, and this factor may lead to a limitation of the scope of application of Turkic languages in the period of modern digitalization. This problem encourages researchers to consider various methods and algorithms and increase the number of resources needed. Nowadays, when information exchange is going on at a very fast pace, it is significant to develop Kazakh-Turkish machine translation, as an example for the Turkic languages, and get high-quality results from it. The assumptions of the conducted research are the syntax of the sentences of Turkic languages is similar; the main difference between the sentences of the languages lies in the morphology of the words. Therefore, the hypothesis of this study is that machine translation of Turkic languages sentences can be solved using the CSE-model of morphology.

State of the Art

The problem of machine translation of the Turkic languages has been little studied because most of the languages of the Turkic group belong to low-resource languages. The most studied languages in terms of NLP are the state languages: Azerbaijani, Kazakh, Kirghiz, Turkish, Uzbek and Tatar. These languages are supported by well-known translation system Google Translate. However, these languages also have a different level of the development NLP models and methods. Due to the development of neural technologies, neural machine translation has become a state-of-the-art method in the field of machine translation of natural languages. However, the lack of sufficient electronic linguistic resources (parallel corpora) for many languages is a very urgent issue of machine translation of natural languages, including Turkic languages.

II Original Contribution

Scientific contribution of the article is a new machine translation technology based on the CSE morphology model and relational decision tables. The proposed machine translation technology for agglutinative languages is not required big volume of parallel corpus for learning as neural machine translation. Proposed machine translation technology is data-driven technology. For using this technology to new languages it is needed to prepare CSE-model and relational decision tables for endings, stems and stop words of languages pair and to use universal programs for machine translation.

III METHODOLOGY

In this work a new solution for Kazakh-Turkish machine translation is presented. Although the development of machine translation is currently at a high level, the results of translation among Turkic languages and low-resource languages are not very high. Kazakh and Turkish have many common features, as they are Turkic languages. Kazakh and Turkish belong to the group of morphology rich languages, which provides with excellent conditions for translation using the CSE morphology model.

The steps for Kazakh-Turkish machine translation based on the CSE model are as follows:

1. Development of a complete set of endings for the Kazakh and Turkish languages using the morphological model CSE.
2. Creating of Morphological analysis of Kazakh and Turkish endings based on the complete set of endings.
3. Creating of correspondence table of Kazakh and Turkish endings and its morphology descriptions.
4. Creation a corresponding table of Kazakh and Turkish word stems.
5. Creation a corresponding table of Kazakh and Turkish stop words.
6. Development of an algorithm for Kazakh-Turkish machine translation based on the CSE model.
7. Creating software based on the base of the machine translation algorithm.

The morphological model CSE based on inferring of the complete set of endings [1]. Universal Programs for Stemming, Segmentation and Morphological Analysis of Turkic Words are described [2]. Therefore, proposed MT technology is data-driven technology. This technology needs creating complete set of endings, table of morphology description of endings, corresponding tables of stems and stop words for language pair.

IV RESULTS

The main advantage of the CSE model is the method of creating linguistic tables, which is useful for linguists and suggests using universal programs for different languages for processing. The paper shows how to create a complete set of language endings, a table of correspondence of endings and their morphological descriptions of a language pair, a table of correspondence of stems and stop words of a language pair, the use of universal machine translation programs for the Kazakh-Turkish language pair. The results of this work show the possibility of using this machine translation technology for other low-resource Turkic languages.

V EVALUATION

Due to the time constraints of scientific and technical work, a small translation model with 500 stems was created. In testing the operation of the machine translation algorithm based on the model of a complete set of Kazakh and Turkish endings, a case consisting of 45 parallel sentences in the Kazakh and Turkish languages was used. Experiments with this machine translation technology showed a BLEU metric score of 20% (state-of-the-art metric in machine translation), and human estimation showed a level sufficient to understand the text. Obviously, the higher the vocabulary, the higher the translation result. Therefore, future experiments assume an increasing of a vocabulary of stems.

VI CONCLUSIONS

The results of this research can use in machine translation of other Turkic languages and maybe for other agglutinative languages. Developed universal programs can used for other agglutinative languages, if for these languages will prepared CSE-model and correspondence tables of endings morphology descriptions, stems, stop words.

REFERENCES

[1] TUKEYEV, U., KARIBAYEVA, A. *Inferring the Complete Set of Kazakh Endings as a Language Resource.* In: Hernes M., Wojtkiewicz K., Szczerbicki E. (eds) Advances in Computational Collective Intelligence. ICCCI 2020. Communications in Computer and Information Science, vol 1287, pp.741-751. Springer, Cham. https://doi.org/10.1007/978-3-030-63119-2_60

[2] TUKEYEV, U., KARIBAYEVA, A., TURGANBAYEVA, A., AMIROVA, D. *Universal Programs for Stemming, Segmentation, Morphological Analysis of Turkic Words* // In: Nguyen N.T., Iliadis L., Maglogiannis I., Trawiński B. (eds) Computational Collective Intelligence. ICCCI 2021. Lecture Notes in Computer Science, – Springer, Cham, 2021. – vol 12876. – P. 643–654. https://doi.org/10.1007/978-3-030-88081-1_48

DOI: https://doi.org/10.1007/978-3-031-21967-2_12
Link to repository: https://github.com/udayRage/pykit_old/tree/master/traditional/3peclat

Keywords: pattern mining, periodic patterns, columnar databases

Towards Efficient Discovery of Partial Periodic Patterns in Columnar Temporal Databases

Penugonda Ravikumar[0000-0001-9124-9781], Venus vikranth raj[0000-0003-4915-5946], P. Likhitha[0000-0003-3032-9061], R. Uday Kiran[0000-0002-5417-0289], Yutaka Watanobe[0000-0002-0030-3859], Sadanori Ito[0000-0002-8266-8463], Koji Zettsu[0000-0003-4062-2376], Masashi Toyoda[0000-0001-9473-5531]

{raviua138,venusvikranthraj12,likhithapalla7,uday.rage,y.watanobe}@gmail.com,{ito, zettsu}@nict.go.jp,toyoda@tkl.iis.u-tokyo.ac.jp

Simplified Title

Partial periodic patterns identification in columnar temporal databases

Abstract

Finding partial periodic patterns in temporal databases is a challenging problem of great importance in many real-world applications. Most previous studies focused on finding these patterns in row temporal databases. To the best of our knowledge, there exists no study that aims to find partial periodic patterns in columnar temporal databases. One cannot ignore the importance of the knowledge that exists in very large columnar temporal databases. It is because real-world big data is widely stored in columnar temporal databases. With this motivation, this paper proposes an efficient algorithm, Partial Periodic Pattern-Equivalence CLass Transformation (3P-ECLAT), to find desired patterns in a columnar temporal database. Experimental results on synthetic and real-world databases demonstrate that 3P-ECLAT is not only memory and runtime efficient but also highly scalable. Finally, we present the usefulness of 3P-ECLAT with a case study on air pollution analytics.

I Introduction

Partial periodic pattern mining (PPPM) [1] is a vital knowledge discovery technique in data mining. The primary purpose of PPPM is to identify some of the frequently co-occurring events (or patterns) that will happen only during a particular time, such as only on weekends, in a specific time of the day, and on a special day of a month. However, these events will happen regularly. The scope of this research is not limited to only a few applications. It plays a crucial role in many real-world applications, such as market-basket analysis, air pollution analysis, and traffic congestion analysis. For example, if we look at the database of a supermarket, we might see that most customers buy meat often and other standard items on the weekends. In the real world, data will be stored either in a row or columnar format, and both have their importance. Uday et al. [1] have already discussed the partial periodic patterns (3Ps) in row databases. However, they can only extract 3Ps from columnar databases by transforming them into a row database. Unfortunately, this transformation will take substantial computational resources. With this motivation, this paper aims to develop an efficient algorithm to find 3Ps in a columnar database.

In this paper, we have used two user-specified constraints named *minimum periodic-support* (*minPS*) and *periodicity* (*per*) to discover 3Ps effectively in columnar temporal databases. The *minPS* controls the minimum number of periodic occurrences of a pattern in a database. The *per* controls the maximum inter-arrival time of a pattern in the database. Therefore, all the 3Ps derived from a database are defined as follows.

I.1 Problem Definition

Given a temporal database (TDB), a set of items, *period* (*per*), and *minimum period-support* (*minPS*), our aim is to discover all 3Ps in TDB that have *period-support* no less than *minPS*.

II State of the Art

In the literature, a 3P-growth algorithm is already presented to discover 3Ps. It can be used to calculate 3Ps in columnar temporal databases as well. However, the limitation is that we need to translate the columnar database into a row database, which can only take row databases as input. This naïve transformation process is costly in terms of memory and runtime if we use extensive columnar databases. Therefore, developing novel algorithms that produce 3Ps in both row and columnar databases is highly recommended.

III Original Contribution

The contributions of this paper are as follows: (*i*) This paper proposes a novel algorithm to find 3Ps in a columnar temporal database. We call our algorithm Partial Periodic Pattern-Equivalence CLass Transformation (3P-ECLAT). (*ii*) To the best of our knowledge, this is the first algorithm that aims to find 3Ps in a columnar temporal database. A key advantage of this algorithm over the state-of-the-art algorithm (3P-growth) is that it can also be employed to find 3Ps in a horizontal database. (*iii*) Experimental results on synthetic and real-world databases demonstrate that our algorithm is both memory and runtime efficient and highly scalable. We will also show that 3P-ECLAT outperforms the state-of-the-art algorithm while finding 3Ps in a row database. (*iv*) Lastly, we use data about air pollution to show how valuable our algorithm is.

IV Methodology

The 3P-growth algorithm can take the row database as input. In this regard, we have also used the row database as input to 3P-ECLAT. Therefore, first, we have identified the one-length 3Ps while transforming the row database to a columnar database and discarded the uninteresting patterns whose *period-support* is less than the *minSup* value. In this step, we will scan the entire database and build a 3P-list, and all the items will be sorted according to their *period-support.* Next, we use one-length 3Ps, which are present in the 3P-list, to generate their super sets as 3Ps if they satisfy the user-specified *minPS* constraint. In this step, we have recursively followed the depth-first search strategy to get the whole set of 3Ps from the database. We compared the 3P-ECLAT algorithm to the state-of-the-art 3P-growth algorithm on several benchmark databases.

V Results

The number of 3Ps generated by both algorithms is the same in every database. The following are some of the noteworthy findings that can be observed from these databases: (*i*) In the case of sparse databases, both algorithms were able to generate the 3Ps, whereas, in dense databases, 3P-ECLAT can produce interesting patterns, but 3P-growth is running out of memory due to the huge search space, especially at low *minPS*. (*ii*) An increase in the *minPS* value shows a decrease in every database's total number of patterns. As a result, both algorithms' runtime and memory consumption decrease while increasing the *minPS* value. More importantly, the 3P-ECLAT algorithm is much faster and consumes less memory than the 3P-growth algorithm at any *minPS* value. (*iii*) An increase in the *Per* value shows the increase in the total number of patterns. This makes both algorithms take longer to run and use more memory while the *Per* value increases. (*iv*) We have used an extensive columnar temporal database to discover the efficacy and productivity of the proposed algorithm. If we keep increasing the database size, both algorithms' runtimes and memory requirements will increase almost linearly. But the 3P-ECLAT algorithm needs less time to run and less memory than the 3P-Growth algorithm, no matter how big the database is.

VI Evaluation

We have evaluated the performance of the proposed algorithm with the help of a case study on air pollution analytics: The Ministry of Environment, Japan, has set up a sensor network system called SORAMAME to monitor air pollution throughout Japan. The raw data produced by these sensors is transformed into a binary columnar database if the raw data value is $\geq$15. Then, the transformed data is provided to the 3P-ECLAT algorithm to identify all sets of sensor identifiers in which pollution levels are high. We have observed that most of the sensors were situated in the southeast of Japan. Thus, it can be inferred that people working or living in the southeast parts of Japan were periodically exposed to high levels of PM2.5. Such information may be useful to ecologists in devising policies to control pollution and improve public health. Please note that more in-depth studies, such as finding highly polluted areas on weekends or during particular time intervals of a day, can also be efficiently carried out with our algorithm.

VII Conclusions

The proposed algorithm applies to several real-world problems, such as air pollution analytics, traffic congestion analytics, and market basket analysis. For example, suppose we extract interesting patterns from the congestion database by applying the proposed algorithm. In that case, people living near the sensor device locations will frequently and regularly experience traffic congestion. Such knowledge of the 3Ps (or heavily congested roads) can benefit the traffic control room by diverting traffic, suggesting police patrols, and alerting pedestrians on the road. The scope of this research is not limited to a particular application.

References

[1] Kiran, R. U., Shang, H., Toyoda, M., and Kitsuregawa, M. Discovering partial periodic itemsets in temporal databases. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (2017), SSDBM '17.

DOI: https://doi.org/10.1007/978-3-031-21967-2_7

Keywords: IDS-Intrusion Detection System, DoS- Denial of Service, Cyber space, NetFlow

A Survey of Network Features for Machine Learning Algorithms to Detect Network Attacks

Joveria Rubab, Hammad Afzal and Waleed Bin Shahid

jrubab@live.com, hammad.afzal@mcs.edu.pk, Waleed.shahid@mcs.edu.pk

Simplified Title

Defending Cyber-space using Artificial Intelligence

Abstract

The immeasurable amount of data in network traffic has increased its vulnerability. Therefore, monitoring and analyzing traffic for threat hunting is inevitable. Analyzing and capturing real-time network traffic is challenging due to privacy and space concerns. However,many simulated datasets are available. Machine-learning based intrusion detection systems are trained on these datasets for attacK detection. Selection of correct features has significant importance in determining the efficiency of various Ml-based algorithms. Hence, this paper provides a literature survey of the various machine learning based IDS. Features, attacks, machine learning algorithms and their corresponding datasets are identified in the survey. The survey may help researchers in identifying benchmark features correlated to network attacks.At the time of writing this paper there is no such IDS that associates network features to attacks.

I Introduction

Information technology is integral part of our life, from connecting to collaborating we require Internet. This dependence on internet has increased the threat of cyber attacks.It puts privacy,health, life, finances, governments, property and security at risk. Crime is always one step ahead of the law. Numerous researchers are working to find ways for defending cyberspace.Artificially intelligent solutions are being presented to design robust system. Machine learning models can detect malicious patterns that may threaten security. Each intrusion leaves a unique set of patterns that assists in its classification. For supervised learning, a labeled dataset is required, where network data flows are labeled as benign or attack traffic. Real-network traffic is hard to obtain therefore, researchers have designed network testbeds to generate synthetic datasets. Research community has applied many models; however, deployment of such models practically is still scarce due to the lack of a common set of network features across all datasets. .

II State of the Art

Machine learning classification algorithms work on features. The choice of accurate features affects the results of classification. Various intrusion detection systems have used different number of feature.Howeever, two NIDs are need special mention,as they would be used in further research.Sarhan converted four widely known datasets into NetFlow [2]. NetFlow is an industry-standard protocol for network traffic collection [1]. UNSW-NB15,ToN-IoT, BOT-IOT and CSE-CIC-IDS2018 were converted into netflows using different software.08 common features, were selected from all the datasets. A binary and multiclass classifier was used for detection of attacks. Binary classifiers performed better on all four datasets. However, the multi-class classifier did not perform well on some attacks i.e. fuzzers, analysis, exploits.The authors therefore decided to increase the number of features, for better performance.

As a result, they increased the number of features from 08 to 43. [3]. Increased number of features improved classification results. The authors claim, these 43 features may prove to be a “benchmark feature set” lacking for NIDS, previously. They believe these common feature sets may help in the practical deployment of the intrusion detection systems which currently is scarce. However, no justification for the selection of these features is provided in the paper.

III Original Contribution

Network attack occur at different layers of the network. For identification of each attack unique network features are required. Our survey shows that similar features are being used for multiple attacks by most NIDS. The results of classification, therefore, can either be biased or uncertain. Also, little or no justification is provided by many of the authors for selection of feature .

IV Methodology

The study is a literature review of various surveys.

V Results

As, depicted from the survey most NIDs mentioned in the paper use same number of features for same attack.As [2] has selected 8 features, for 11 network attacks and also [3] has selected 43 features, for all network attacks.Whereas, network attacks have different nature, therefore applying equal and similar features for all attacks may not be a good idea. Also, the attack dimensions are evolving rapidly, identification of common features set for multiple attacks is necessary. A common feature set with proper justification, associated to individual attacks, will enhance deployment of machine learning based NIDs.

VI Evaluation

The study initially is a survey, we are also, working on practical application of this survey and prove via experimentation the results of our survey. It was mandatory to find out working of various Network Intrusion Detection system, before formally performing the experiment to prove our hypothesis.

VII Conclusions

The current research uses same mirror to for all attacks. Different attacks are targeted at different layers therefore similar features cannot be used for all attacks. Researchers need to identify unique features correlated to each attack for better performance of machine learning algorithms. Domain knowledge must be included for the recognition of features. A unique set of features correlated to each attack will help in the practical and physical deployment of NIDs. It will increase confidence of users on the system as network attacks are a serious threat to everyone.

References

[1] `https://www.eginnovations.com/blog/what-is-netflow/`.

[2] Sarhan, M., Layeghy, S., Moustafa, N., and Portmann, M. Netflow datasets for machine learning-based network intrusion detection systems. *arXiv preprint arXiv:2011.09144* (2020).

[3] Sarhan, M., Layeghy, S., Moustafa, N., and Portmann, M. Towards a standard feature set of nids datasets. *arXiv preprint arXiv:2101.11315* (2021).

DOI: https://doi.org/10.1007/978-3-031-21967-2_39

Keywords: U-net, CT Image, Medical Image Processing, Artificial Intelligence, Deep Learning

Pre-processing of CT images of the lungs

Talshyn Sarsembayeva[0000-0001-7668-2640], Madina Mansurova[0000-0002-9680-2758], Adai Shomanov[0000-0001-8253-7474], Magzhan Sarsembayev[0000-0003-2139-2456], Gassyrbek Rakhimzhanov, Symbat Sagyzbayeva

sarsembayeva.talshyn@gmail.com, M_Mansurova@kaznu.kz, adai.shomanov@gmail.com,magjan@kaznu.kz, r.gasyrbek01@gmail.com, sagyzbaevasymbat@gmail.com

Simplified Title

Pre-processing of CT images of the lungs.

Abstract

Respiratory diseases are one of the primary causes of death in today's population, and early detection of lung disorders has always been and continues to be critical. In this sense, it is critical to evaluate the condition of the lungs on a regular basis in order to avoid disease or detect it before it does substantial harm to human health. As the most popular and readily available research tool in diagnosis, radiography is critical. Despite all of the benefits of this technology, diagnosing sickness symptoms from photos is a challenging task that necessitates the involvement of highly experienced specialists as well as significant time investment. The difficulty arises from the incompleteness and inaccuracy of the initial data, particularly the presence of numerous image distortions such as excessive exposure, the presence of foreign objects, and so on. The U-net technique was used to do early processing of CT images of the lungs using a neural network during the research. The current status of study in the field of X-ray and CT image identification employing in-depth training methodologies demonstrated that pathological process recognition is one of the most significant tasks of processing today.

I Introduction

Today, respiratory diseases are one of the leading causes of death in the population, and early diagnosis of lung diseases has been and remains very important. In this regard, it is necessary to regularly monitor the condition of the lungs to prevent the disease or to detect it before causing significant harm to human health. Despite all of the benefits of this technology, diagnosing sickness symptoms from photos is a challenging task that necessitates the involvement of highly experienced specialists. The work [3] proposes a review of the literature on CT scans of the lungs and states, preprocessing ideas, segmentation of a variety of pulmonary arrangements, and Feature Extraction to identify and categorize chest anomalies. In addition, research developments and disagreements are discussed, as well as instructions for future investigations. This work examines the segmentation of various pulmonary structures, registration of chest images, and applications for the detection, classification, and quantification of chest anomalies in CT scans.

II State of the Art

Using in-depth machine learning methods to analyze medical images, such as computed tomography (CT), magnetic resonance imaging (MRI), and chest radiography, allows for a more tailored and effective approach to diagnosing and treating lung ailments. A significant decrease in time costs and an improved burden for healthcare personnel are advantages of adopting machine learning technologies for picture recognition. To date, a large amount of biomedical data has been collected, which contributes to radically changing the therapeutic experience by using machine learning algorithms for processing.

III Original Contribution

The key problem is that they only assess the number of true positives, false positives, and false negatives, ignoring the alleged location. As a result, the contour's average distance and the surface's average distance are more appropriate. The evaluation was based on test data that had not been used during the training phase. Jacquard had a score of 0.926, and Dice received a score of 0.961. The reduction path's high-resolution capabilities are paired with the output chosen to forecast a more accurate result based on this information, which is the architecture's core idea.

IV Methodology

There are two types of exceptions: local and integrative. Local features have the advantage of being adaptable and not changing in response to variations in brightness and illumination, but they are not unique. Changes in the object’s structure and difficult lighting conditions have no effect on the integral features that characterize the image of the thing as a whole. When the item sought is modeled on a number of regions, each of which is characterized by its own set of features - local texture descriptor - there is a combined method - the use of local features as pieces of an integrated description. The object as a whole is described by a set of such descriptors. By studying the vector of symbols obtained in the preceding stage and splitting the relevant space into subdomains representing the appropriate class, classification is defined as identifying whether an object belongs to a certain class. There are numerous classification methods: neural networks, statistical, key trees and forests, metric, and kernel. The membership of two classes is evaluated for the purpose of identifying the object in the image: the class of images that contain the object and the class of images that do not contain the object. The set of selected characteristics, their capacity to discriminate between photographs of objects of different classes, is a crucial aspect influencing the quality and stability of the classification. The simpler the space of possibilities is, and the classifier can have a simpler form, the more separable the functions are. Many ”simple” qualities can be combined to approximate a few ”complex” characteristics.

V Results

Using these concepts, a modern and effective method for combining all stages of image analysis has been developed: pre-processing, simultaneous acquisition of a set of ”simple” signs, and classification based on their optimization using a multi-layer deep learning training base of convolutional neural images. The capacity allocation mechanism takes place on the ground floor (CNN), which is part of the classifier; the structure of features is produced automatically during the learning process and is governed by the network’s model and architecture. CNN can be seen as a broad approach to modeling functional space, but it requires an image exercise model that represents all of the relevant computational resources and objects. This is because, according to the rules of network design, the specific model of the object in CNN is built solely on the basis of information included in the training database. The model of distinction is developed by the researcher using apriori information on the nature of the pictures of the object in the classical approach. As a result, the typical sub-issue of constructing an efficient collection of functions becomes the challenge of developing an optimal CNN architecture, which in turn generates the required features from the images in the classification problem [1], [2]. This work was funded by the Committee of Science of the Republic of Kazakhstan AP09260767 Development of an intellectual information and analytical system for assessing the health status of students in Kazakhstan (2021-2023).

VI Evaluation

The model’s performance is estimated using Jaccard and Dice indicators, which are well-known for this type of computer vision problem. The intersection across the union is known as jacquard, and it is the same size as bone F1.

VII Conclusions

The following tasks were completed during the work: 1) research into the principles of developing machine learning algorithms; 2) creation of a database for training and testing an automated diagnostic system (ACD) for CT images of the lungs; and 3) evaluation of indicators of the developed system’s information content.

References

[1] LAKHANI, P., AND BASKARAN, S. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 284 2, 2017.

[2] MANSUROVA, M., SARSENOVA, L., KADYRBEK, N., SARSEMBAYEVA, T., TYULEPBERDINOVA, G., AND SAILAU, B. Design and Development of Student Digital Health Profile. 10.1109/AICT52784.2021.9620459, 2021.

[3] VIJAYARAJ, J., AND LOGANATHAN, D. Various Segmentation Techniques for Lung Cancer Detection using CT Images: A Review. 2021.

DOI: https://doi.org/10.1007/978-3-031-21967-2_40

Keywords: Asymmetric information, Blockchain, Used car, Hyperledger, Smart contract.

Application of Hyperledger blockchain to reduce information asymmetries in the used car market

Chien-wen Shen[0000-0002-3792-0818], Agnieszka Maria Koziel[0000-0003-0913-813X], Chieh Wen

aga.koziel@g.ncu.edu.tw

Simplified Title

Blockchain application to reduce used car market information asymmetries.

Abstract

The used car market has long been an example of a market rife with information asymmetries because most consumers have insufficient experience in buying cars and can only use the car history information provided by car dealers to make prepurchase judgments. Such a problem can be alleviated by blockchain technology, which is a decentralized distributed database using nodes of a computer network to record the historical information of a car, where the chain of data cannot be falsified. Accordingly, we propose a Hyperledger-based approach to illustrate the development of a used car blockchain application to reduce information asymmetries. In Hyperledger Fabric, all of the transactions are managed by smart contracts written in chaincode that are invoked when the external application needs to interact with the blockchain ledger. All business network transactions are recorded on the smart contracts, allowing the records to coexist among the participants, which include dealer, maintenance plant, motor vehicle office, police office, and customers. A simulation about the time reduction caused by our proposed Hyperledger blockchain application was also discussed. Our findings suggest that blockchain technology shortened time obtaining historical information and increase trust among buyers in the used car market.

I Introduction

Information asymmetries and quality uncertainties are problems faced by used car market buyers at any time before making a purchase decision. Buyers suffer from problems such as the car being in worse condition than initially indicated, accident damage that is not disclosed, fraud, vaulted products, etc. To cope with those problems, buyers spend a lot of time seeking information. Asymmetry of information leads to increasing used car costs and buyer dissatisfaction that causes damage to customers' rights and interests, resulting in disputes between buyers and dealers. For instance, in Taiwan in 2019, the number of used cars doubled the new car market size, but there is still a problem of quality uncertainties and a lack of trust in the second-hand car market. Buyers, to ensure the quality of the car planning to purchase, must put much effort into seeking information that might be falsified, incorrect, or missing. Issues that buyers face during the information-seeking process might be alleviated by using blockchain-based solutions. Blockchain technology can provide a solution to mitigate information asymmetries and more efficiently reduce quality uncertainty, therefore changing relationships between buyers and sellers to be more reliable. This technology is making progress in the automotive industry with a significant impact on development in keeping all car records transparent and secure during the whole car life cycle. Therefore, it might be highly beneficial for the used car market to provide a complete vehicle history.

II State of the Art

In the used car market, consumers can only rely on their own car buying experience or the vehicle information provided by the car dealers to make a judgment on the vehicle condition, whereas, because the used car market has a serious information asymmetry problem, it results in many disputes. Several methods are addressing the information asymmetry problem. One intuitive solution for consumers and competitors is to act as a monitor for each other. Consumer Reports, Underwriters Laboratory, public notaries, and online review services help bridge gaps in information. Another solution is to have manufacturers provide warranties, guarantees, and refunds. In addition to seller-granted warranties, third-party companies can offer their own in the form of insurance that comes at some cost to the consumer. The government can also regulate the quality of goods sold on the market. Many institutions provide car-related documents, but none of them are connected to each other. It results in unequal amounts of information required to make an informed transaction decision. Therefore, the problem of asymmetric information still exists and needs to be solved by different methods.

II.1 Blockchain technology in the used car market

Recently, the application of blockchain technology in the used car market has received increasing attention as it can effectively reduce information asymmetry by storing the past usage status of vehicles. There are already several studies proposing a blockchain technology framework to reduce the information asymmetry issue based on vehicle history. However, they focus on resolving only a specific problem, such as odometer fraud, falsifying car mileage data, insurance information, or vehicle accident documentation which cannot ultimately reduce the information asymmetry. Blockchain technology promises to automatize the tracking of cars through their historical and current life and provide reliable information at any point in time it is needed.

III Original Contribution

To tackle the problem of information asymmetry in the used car market, we apply the Hyperledger approach to build a blockchain framework and illustrate the development of Hyperledger Fabric-based applications. In Hyperledger Fabric, all business network transactions are recorded on smart contracts, allowing the records to coexist among the participants. Smart contracts can reduce administration costs and risk, increase trust and improve the efficiency of used car market business processes. The smart contract records a new node according to the transaction updates, and while the buyer makes a purchase smart contract updates the node's database, all of that happens in real-time, Therefore, we suggest using smart contracts to solve the information asymmetry problem and provide a reliable transaction environment for the used car market.

IV Methodology

In our experimental study, we build the Hyperledger Fabric environment system sample network to demonstrate the situation of car transactions on the blockchain. The used car network can be run through the transaction rules, and finally, the transaction query function can be implemented directly into the Hyperledger API. Fabric is used to run the second-hand Car chain network (car trade), where the internal transaction rules are defined by smart contracts. Smart contracts are in use when the dealer registers a used car to sell on the blockchain, when other players such as maintenance plant, motor vehicle office, or police office update data about the car, or when the buyer makes a decision about buying a vehicle.

V Results

The simulation results show that the blockchain network can reduce the acquisition time of the vehicle's historical data and decrease information-seeking effort. If consumers want to know the historical information of the vehicle, the dealer can obtain the information through the blockchain query, reducing the time for personnel exchanges, waiting and written application, etc. Moreover, experiment results show that if the vehicle industry wants to introduce blockchain applications in the future, they should be implemented by the car dealer first. When the vehicle is sold, the information is recorded on the blockchain, and every time the vehicle enters the original maintenance plant, the records can also be uploaded to the blockchain, and establish cooperation with government agencies.

VI Evaluation

This experimental study records the historical information of the car by building a business network through Hyperledger Fabric Framework to define business network transaction rules. Through the blockchain network, all transaction records of vehicles in the past are kept in smart contracts and are enforced automatically when transaction conditions are satisfied. Smart contracts guarantee appropriate access control where all records are secure and reliable, and the history is tracked through the past ledger making data more transparent. On the used car market, when customers want to buy a car, they can make a pre-purchase evaluation according to the car condition information existing on the blockchain. Blockchain networks can increase the transparency and fairness of transactions and protected the rights and interests of consumers.

VII Conclusions

The results of this research and the proposed framework can be useful for car dealers, maintenance plants, government agencies, and other organizations playing a crucial role in the used-car market industry. Based on the results and for the sake of information transparency, it is recommended that the blockchain application should be implemented by car dealers as the first institution in cooperation with government agencies. However, other institutions that might be possibly involved in car sales and maintenance can make use of this research and should be encouraged to join the used-car market blockchain ecosystem.

Link to arXiv: `https://arXiv.org/abs/2104.12473`
DOI: `https://doi.org/10.1007/978-981-19-8234-7_46`

Keywords: boxer detection, video processing, image processing, dynamic sport analysis.

The problem of detecting boxers in the boxing ring

Piotr Stefański[0000-0003-1229-327X], Jan Kozak[0000-0002-2128-6998], Tomasz Jach[0000-0002-9463-5562]

{piotr.stefanski,jan.kozak,tomasz.jach}@ue.katowice.pl

Simplified Title

The problem of detecting boxers in the boxing ring

Abstract

Modern technology is strongly associated with sports. A perfect example of machine learning in sports is a support of detection of specific events or situations. Such a problem is present in boxing, where boxers' moves need to be precisely detected. However video analysis is labor intensive but may provide valuable information. The paper presents the problem of processing recordings of boxing boxers, in which the dynamics is at an extremely high level and some events last for fractions of seconds. Additionally, the competition is often watched by spectators blocking the view. The goal of this paper is to present accurate, precise and quick method of detecting the presence of pugilists in the ring. This will allow to evaluate and score the boxing fight later. To validate the experiment, relevant material had to be collected – the authors recorded real boxing bouts and prepared the complete training set. The proposed solution will be used to automatically filter-out uninteresting parts of footage, where boxers are not engaged in close-combats situation.

I Introduction

Nowadays, cameras recording videos can be found at every step, both outside and inside buildings belonging to the public and private infrastructure. Current science and technology allows to search for valuable information in recorded footage. For instance, machine learning solutions provide automatic car counting on highways, measurements of distance between people or verification of presence of facial masks.

Camera recordings are also used to analyse sports football games, tennis and many other sports [1]. A similar problem is pursued in this work, as it is devoted to the use of camera images to detect boxers in a boxing ring and the preparation of a suitable test set.

The purpose of this paper is to prepare a solution to correctly and quickly detect the boxers while ignoring other people outside the ring. This will allow for further work on boxing fight evaluation. For the purpose of this paper, the authors set up a recording environment and then recorded real boxing fights. A special type of equipment had to be considered, as boxing is a highly dynamic game. The detection will be based on micro-movements during close-combat, therefore high resolution and high frame rate is required. Better the quality, bigger the video files are which leads to increasing the computational complexity of the task. Only data collected in this way (in the form of recordings) could be used for further analysis, and finally for detecting boxers in the ring. As the bouts are quick (some of them are under 40 milliseconds), single static images provide no value for this task.

II Methodology

Due to the lack of publicly available stable and suitable for vision computing recordings of boxing fights, the solution to the problem of detecting boxers in the boxing ring had to start with the collection of such material. As noted before, the proper equipment is required for acquiring high quality footage. The authors had analysed multiple solutions and choose four GoPro Hero8 sport cameras along with tripods that were placed behind each corner of the boxing ring. The setup allowed boxing fights to be recorded from four different points. In order to prepare a diverse train-set, the authors recorded fights with boxers of different age.

It is worth noting, that with this spacing of equipment, one of the cameras always has a good view of the boxers at any given time. If the boxers are covering each other for two of the cameras, the other two are placed in a good position for the boxers' profiles. Having four cameras the blind-spots are minimised. The combined footage allows to observe the fight with more details than the three human referees can do on their own.

After collecting the necessary material, it is possible to solve the problem of detecting boxers in the ring. Due to the complexity of the problem, the Authors propose to divide it into several stages. The various steps are reflected in the Algorithm 1, which receives an image as input and returns a list of boxers and their positions as output. The algorithm starts from detecting all people in the picture, limiting the detections in the next step to people inside the ring. Finally, the referee is also excluded.

Algorithm 1: Proposed approach for detecting boxers in an image

```
Input: image – one frame from video
Output: boxers – list of coordinates of detected boxers
persons := find_persons_in_image(image);
persons_on_ring := get_only_persons_from_ring(persons);
boxers := get_only_boxers_from_ring(persons_on_ring);
result boxers;
```

III Results

The results of the obtained experiments allow to evaluate the described approaches. Through this research, it is possible to confirm that it is possible to quickly identify the boxers in the ring while filtering out the other people. This is to be used for work related to the calculation of boxing fight statistics.

IV Conclusions

The experiments confirm that the introduced solution allows for an improvement in the accuracy of detecting boxers in the ring. Starting from a simple approach that detected all people in a video frame, we were able to achieve an effect where only boxers in the ring were detected. In order to train any classifying system, the most time and resource consuming part, is the labelling of data. We have achieved the goal – that is to decrease the time spent by humans labelling each frame or video sequence only to the promising or relevant parts. Coverage values need further improvements, as not all boxers are always detected.

References

[1] Thomas, G., Gade, R., Moeslund, T. B., Carr, P., and Hilton, A. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding 159* (2017), 3–18.

DOI: https://doi.org/10.1007/978-3-031-21743-2_22

Keywords: Computer Vision, Background Subtraction, Convolutional Neural Network, Product Recognition

The Combination of Background Subtraction and Convolutional Neural Network for Product Recognition

Tin-Trung Thai[1], Synh Viet-Uyen Ha[2-†], Thong Duy-Minh Nguyen, Huy Dinh-Anh Le, Nhat Minh Chung, Quang Qui-Vinh Nguyen, Vuong Ai-Nguyen,

tttin@hcmiu.edu.vn,hvusynh@hcmiu.edu.vn

Simplified Title

Background Subtraction and Deep Learning for Product Recognition.

Abstract

Multi-class retail product recognition is an important Computer Vision application for the retail industry. Track 4 of the AICITY challenge is introduced for the retail industry. This track focuses on the accuracy and efficiency of the automatic checkout process. However, due to the lack of training data for retail items in the real world, a synthetic data set is usually generated based on the 3d scanned items to produce training data for an automated checkout system. To overcome the difference informative representative appearance between training data and the real-world scenario in the test set provided by the AICITY organizer, our research focuses on analyzing and recognizing retail items by combining the traditional method and state-of-the-art Convolutional Neural Network (CNN) approach. This paper presents our proposed system for product counting and recognition for automated retail checkout. Our proposed method is ranked top 8 in the experimental evaluation in the 2022 AI City challenge Track-4 with an F1-score 0.4082.

I Introduction

Product recognition problem is vital in a lot of Computer Vision applications nowadays. Barcode is now widely applied in industries. However, this method takes time to locate the barcode position printed on each item. With the development of electronic devices, product recognition has become a vital issue in product recognition applications.

II State of the Art

II.1 Object Detection

Object detection is an essential key in object behavior analysis, where the detection module can support many tasks: object tracking, object counting, object recognition, etc. Convolutional Neural Network (CNN) has been presented to generate the best result for many image classification datasets. However, most of the CNN approaches trade-off time execution for accuracy and require super hardware devices for training and testing.

II.2 Multi-Object Tracking

Multi-Object Tracking (MOT) is one of the most widely investigated tasks as it is important in many computer vision applications. The typical motif of MOT problems is Tracking–by–detection (TBD) which combines a object detector and a re–ID model. TBD methods have two separate models for detection and tracking, therefore they are usually computationally expensive.

II.3 Object Counting

Object counting is now more important because of its wide range of real-world applications. Two presentative approaches for object counting are the detection base method and regression-based methods. However, most of these methods rely on a large training data set to train a counting model.

III Original Contribution

In our work, we research and analyze a framework that can provide both efficient performance and accuracy. Our contributions to this paper are described as follows:

Firstly, detect the proposal region using the Background Subtraction model and Skin removal.

Secondly, classify each proposal region by the Inception–ResNet–v2 model.

Finally, track and count objects by simple algorithms.

Rank	Team Name	Score
1	BUPT-MCPRL2	1.0000
2	SKKU Automation Lab	0.4783
3	The Nabeelians	0.4545
...	...	...
8	**HCMIU-CVIP**	**0.4082**
9	CyberCore-Track4	0.4000
10	UTE-AI	0.4000
11	KiteMetric	0.3929
12	AICLUB@UIT	0.3922

Table 1: The final ranking results of the track 4 on the test set A

IV Methodology

IV.1 Background Subtraction and Skin Removal

We adopted the K-nearest neighbors (KNN) background subtraction for moving object detection [2]. After extracting the moving object mask, the skin removal module helps to erase the hand region when the customer is holding the item.

IV.2 Product Classification

After extracting the proposal region in the previous step, the problem is now Object Classification instead of Object Detection. Very deep convolutional neural networks have achieved sucessful results in the field of computer vision, especially in image recognition performance in recent years. Our solution takes advantage of one of the state-of-the-art object classification models, the Inception-ResNet–v2 [1] for product classification

IV.3 Product Tracking and Counting

The tracking and counting algorithm is one of our major contributions to the proposed system. In our system, each tracklet consists of a bounding box of objects that appears from frame t to $t+n$. Each bounding box contains 4 values $[x, y, w, h]$ – the position of the top left corner (x, y) and the width (w) and high (h). The value of classification–class, and probability of each object is also assigned to each box of each tracklet from frame t1 to tn. In each tracklet, we select the most occurrence class and assign the recognition value to each tracklet. Inspired by the runner-up of track 1 in the 2021 AICITY challenge, we adopt a simple tracking algorithm for product tracking: distance between the center point of bounding boxes and IoU matching. After the process of classifying and tracking, we analyze the tracklet and conduct the final result in each video in the test set. At this stage, we merge the tracklets together. Given two tracklets A and B with class a and b respectively, if a and b are the same class and the two tracklets are less than *thresh* frame apart, then join two tracklets into one tracklet.

V Results

Our proposed system achieve the rank 8th on the final scoreboard, however, our solution is promising and does not require a super-powerful hardware device for execution. Our solution does not require extreme computation and a hand-labeled dataset for training a model.

VI Evaluation

Our team achieved a final score of F1 = 0.4082 and ranked eight on the final Leaderboard. The final ranking results of challenge track 4 are shown in table 1.

VII Conclusions

We propose a system for Multi-Class Product Counting and Recognition for Automated Retail Checkout. The system also has a competitively fast execution speed, so that it can apply in the real-world application.

References

[1] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence* (2017).

[2] Zivkovic, Z., and Van Der Heijden, F. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters 27*, 7 (2006), 773–780.

DOI: https://doi.org/10.1007/978-3-031-21967-2_48

Keywords: Mobile Banking System, Fraudulent Transaction, Anti-phishing Model, Verification

AntiPhiMBS-TRN: A New Anti-Phishing Model to Mitigate Phishing Attacks in Mobile Banking System at Transaction Level

Tej Narayan Thakur[1][0000-0003-0762-3895] and Noriaki Yoshiura[2][0000-0001-8831-1871]
tejnarayanthakur@gmail.com, yoshiura@fmx.ics.saitama-u.ac.jp

SIMPLIFIED TITLE

Anti-Phishing Model in Mobile Banking System at Transaction Level after Success of Authentication.

ABSTRACT

With the continuous improvement and growth at a rapid pace in the utility of mobile banking payment technologies, fraudulent mobile banking transactions are being multiplied using bleeding-edge technologies sharply and a significant economic loss is made every year around the world. Phishers seek new vulner-abilities with every advance in fraud prevention and have become an even more pressing issue of security challenges for banks and financial institutions. How-ever, researchers have focused mainly on the prevention of fraudulent transac-tions on the online banking system. This paper proposes a new anti-phishing model for mobile banking systems at the transaction level (AntiPhiMBS-TRN) that mitigates fraudulent transactions in the mobile banking payment system. This model applies a unique id for the transactions and an application id for the bank application known to the bank, bank application, users, and the mobile banking system. In addition, AntiPhiMBS-TRN also utilizes the international mobile equipment identity (IMEI) number of the registered mobile device to prevent fraudulent transactions. Phishers cannot execute fraudulent transactions without knowing the unique id for the transaction, application id, and IMEI number of the mobile device. This paper employs a process meta language (PROMELA) to specify system descriptions and security properties and builds a verification model of AntiPhiMBS-TRN. Finally, AntiPhiMBS-TRN is suc-cessfully verified using a simple PROMELA interpreter (SPIN). The SPIN verification results prove that the proposed AntiPhiMBS-TRN is error-free, and banks can implement the verified model for mitigating fraudulent transactions in the mobile banking system globally.

I INTRODUCTION

The “Mobile Banking System” in this paper refers to the concept of a mobile banking system in general. With the advancement of mobile technologies, most modern commerce payments depend on the mobile banking system that is always open 24/7, 365 days a year for financial transactions. However, the unfortunate truth is that fraudulent transactions are also around-the-clock operations. With the continuous improvement and growth at a rapid pace in the utility of mobile banking payment technologies, fraudulent mobile banking transactions are being multiplied using bleeding-edge tech-nologies sharply and a significant economic loss is made every year around the world. 2019 Iovation financial services fraud and consumer trust report [1] show that 61% of financial transactions originate from mobile and 50% of suspected fraudulent trans-actions seen by Iovation are from mobile devices. Fraudsters seek new vulnerabilities with every advance in fraud prevention and have become an even more pressing issue of security challenges for banks and financial institutions. Fraud-the Facts 2021 [2] revealed that mobile banking fraud losses increased by 41% in 2020 in the UK.

Mobile banking users download phishing apps unknowingly, install them on their mobile devices, and input login credentials (Username and password) in the phishing app unintentionally. They follow the links in the phishing emails/SMS and are redirected to the phishing login interface, and they input the login credentials in the phish-ing login interface. Thus, phishers steal login credentials using phishing apps or phishing login interfaces from mobile banking users and employ the stolen login credentials to login into the mobile banking system. As the login credentials are valid, phishers login into the mobile banking system. Phishers request a transaction, and MBS sends a one-time password (OTP) for the security of the transaction. However, phishers reply to the OTP, and they execute the fraudulent transactions. This paper presents a new anti-phishing model for mobile banking systems at the transaction level (AntiPhiMBS-TRN), and the objective of this research is to mitigate fraudulent transactions executed using stolen login credentials and stolen/lost mobile devices. Banks and financial institutions can implement AntiPhiMBS-TRN to mitigate fraudulent transactions in the mobile banking industry.

II State of the Art

Researchers have worked on the prevention of fraudulent transactions in digital banking. Vishwakarma, Tripathy, and Vemuru proposed a layered approach for near field communication (NFC) enabled mobile payment system to prevent fraudulent transactions. Delecourt and Guo utilized potential reactions of fraudsters into con-sideration to build a robust mobile fraud detection system using adversarial examples. Zhou, Chai, and Qiuintroduced several traditional machine learning algorithms for fraud detection in the mobile payment system. Pracidelli, and Lopes proposed the artifacts capable of minimizing electronic payment fraud problems using unsupervised and supervised algorithms. Kargari and Eshghi proposed a semi-supervised combined model based on clustering algorithms and association rule mining for detecting frauds and suspicious behaviors in banking transactions. Sarma, Alam, Saha, Alam, Alam, and Hossain proposed a system to detect bank fraud using a community detection algorithm that identifies the patterns that can lead to fraud occurrences. Gyamfi and Abdulai used supervised learning methods to support vector machines with spark (SVMS) to build models representing normal and abnormal customer behavior for detecting fraud in new transactions. These mentioned works do not mitigate fraudulent transactions using stolen login credentials and phishing apps in the mobile banking system. Our paper proposes a new anti-phishing model to mitigate fraudulent transactions in the mobile banking system in the world of mobile payment transactions. Banks and financial institutions can implement this model to mitigate phishing attacks in the mobile banking system.

III Original Contribution

The original contribution is to propose a new anti-phishing model for mobile banking systems at the transaction level (AntiPhiMBS-TRN) and to mitigate fraudulent transactions executed using stolen login credentials and stolen/lost mobile devices. Banks and financial institutions can implement AntiPhiMBS-TRN to mitigate fraudulent transactions in the mobile banking industry.

IV Methodology

This paper proposes an anti-phishing model for mobile banking systems at the transaction level (AntiPhiMBS-TRN). This paper also uses model checker SPIN to verify that the model AntiPhiMBS-TRN satisfies basic properties for the banking protocol. Some security properties which are described in linear time temporal logic are also verified in the proposed model.

V Results

Model checker SPIN shows that the proposed model satisfies basic properties and some security properties. Since the verification requires much memory and time, the verification in this paper check up to 50 users.

VI Evaluation

This paper discusses the possibility of transaction level attacks against the proposed model. It follows that the proposed model prevents the transaction level attacks. The verification results of model checker SPIN shows that the proposed model is available for banking system.

VII Conclusions

This paper proposed a new anti-phishing for mobile banking system at the transaction level (AntiPhiMBS-TRN) to mitigate fraudulent transactions globally. Phishers exploit stolen login credentials for fraudulent transac-tions using a new mobile device. However, AntiPhiMBS-TRN detects the new mobile device and queries the IMEI number of the old mobile device to execute the transac-tion. The phishers cannot deliver an IMEI number, a unique id for the transactions and cannot succeed in the fraudulent transactions in the mobile banking system. The phishing apps cannot provide a valid application id to the mobile banking system, and phishers fail to execute the fraudulent transactions in the mobile banking system using phishing apps. Phishers get stolen/lost mobile devices and request transactions using mobile banking systems installed in those devices. However, AntiPhiMBS-TRN employs a unique id for the transaction system, and phishers cannot provide a unique id for transactions. Hence, phishers fail to execute the fraudulent transactions using stolen/lost mobile devices. In future research, we will propose a new secured model to detect the change of locations and mitigate other probable attacks such as man in the middle (MITM) attack, SQL injection attack, man in the browser (MITB) attack, replay attack in the mobile banking system.

References

[1] 2019 Iovation financial services fraud and consumer trust report, https://content.iovation.com/resources/2019-iovation-Financial-Services-Fraud-and-Consumer-Trust-Report.pdf Accessed 2021/12/14

[2] Fraud-The Facts 2021: the definitive overview of payment industry fraud report, https://www.ukfinance.org.uk/policy-and-guidance/reports-publications/fraud-facts-2021. Accessed 2022/4/14

DOI: https://doi.org/10.1007/978-3-031-21743-2_36
Keywords: KD-Tree, Image Retrieval

Semantic Relationship-based Image Retrieval using KD-Tree Structure

Nguyen Thi Dinh[0000-0003-3428-3101], Thanh The Van[0000-0001-8408-2004], Manh Thanh Le[0000-0002-0949-222X]
dinhnt@hufi.edu.vn, thanhvt@hcmue.edu.vn, lmthanh@hueuni.edu.vn

SIMPLIFIED TITLE

Image Retrieval using KD-Tree Structure

ABSTRACT

The semantic relationship of visual objects plays an important role in determining the context and semantics of an image. In this paper, a method of classifying semantic relationships between objects on the image is proposed and applied to a semantic-based image retrieval system. Firstly, the visual objects on an input image are extracted and classified using the R-CNN network model. Secondly, a semantic description of the image is determined by the KD-Tree structure. From that, a model of classifying semantic relationships and extracting semantic descriptions for an input image is proposed to retrieve a set of similar images by semantics. To prove the correctness of the proposed theoretical basis, an experiment was built on the COCO and Flickr image data sets with an average image retrieval performance of 0.6972, and 0.7794, respectively. Experimental results are compared with other works on the same data set to demonstrate the effectiveness of our proposed method and can be applied to multi-object image data sets.

I INTRODUCTION

Determining the semantic relationships of objects on an image and retrieving images based on the semantic relationship is a challenge for multi-object image retrieval problems. Based on the KD-Tree structure for image classification [3], the KD-Tree structure is built for classifying semantic relationships between objects performed in this work. While several published works determine the relationship between objects on the image using a Scene Graph or Knowledge Graph. Identifying the semantic relationship between objects on the image is performed on multi-object image sets by building a triple describing the relationship between two objects, building a query SPARQL, and retrieving on Ontology.

II STATE OF THE ART

To implement this problem, some the tasks need to be performed and proposed, including:

II.1 Object detection using R-CNN network model

In the R-CNN model network, each object is defined by the region proposal, and the proposed regions containing the object bounded are extracted by bounding boxes. Based on the feature, extracted bounding boxes and object classification by assigning labels to each extracted image region using the Mask R-CNN network model. The experimental result for object classification by Mask R-CNN is high.

II.2 Building a KD-Tree structure for semantic relationship classification

The process of building a KD-Tree structure for semantic relationship is presented in steps:

(1) Each region of the object is extracted by an m-dimensional vector feature $f_{ik} = (f_{ik1}, f_{ik2}, \dots, f_{ikm})$;

(2) The node is different from the leaf and stores a $(2 * m) - dimensional$ weight vector randomly;

(3) Each vector $f_{ik} = (f_{ik1}, f_{ik2}, \dots, f_{ikm})$ is inserted into the KD-Tree from the root to find the storage location at the $leak_k$ on the KD-Tree by using the $Sigmoid$ transfer function at the node;

(4) $Leak_k$stores a word to describe semantic relationship;

(5) Based on the label at $leak_k$, extracting a semantic relationship between the objects of the input image.

II.3 Training weight vector process on KD-Tree

A process of adjusting a set of classifier vectors at the nodes on the KD-Tree structure is performed to improve the classification efficiency on KD-Tree as follows:

(1) Calculating the correct classification performance for the set of image regions with the initial weight vectors.

(2) Finding $node_i$ is the wrong location.

(3) Adjusting the vector at $node_i$ so that vector f_i is at the correct leaf.

(4) Comparing the classifier performance after adjusting the weight vector and the classifier performance when using the set of random weight vectors.
(5) This process repeats until a given number of iterations is reached or the target performance is reached.

II.4 Extraction semantic relationship of an image using KD-Tree

Each object image region is extracted as a feature vector and combined into a feature vector for the input image to store at the leaf. On this basis, the semantic relationship is extracted based on the KD-Tree structure in steps: (1) Calculate the performance of semantic relationship classification with w_{kt}; (2) Determine the wrong path of the $f_k$vector on the KD-Tree; (3) Determine the right path of the $f_k$vector on the KD-Tree; (4) Find the position where Nodei misplaces the f_k vector; (5) Adjust the vector stored at $node_i$to f_k the right path; (6) Recreate the KD-Tree according to the adjusted vector set; (7) Calculate the classifier performance after vector tuning at $node_i$.

III ORIGINAL CONTRIBUTION

The contributions of this article are (1) Extracting and classifying visual objects of an image using R-CNN; (2) Building a KD-Tree structure to determine a semantic relationship between objects of an input image; (3) Proposing a model to classify objects and extracting semantic relationships between objects; (4) Building experimental and proving the feasibility and correctness of the proposed method using the COCO [3], Flickr [4] image data sets. The contribution of this work is to build a triplet describing the relationship of the objects using the KD-Tree. This is a possible application method for multi-object image sets.

IV METHODOLOGY

The work has carried out theoretical research and proposed a model combined with experiments to prove the correctness and effectiveness of the proposed theory. The proposed theoretical content includes: (1) Segmentation and classification of object images based on the R-CNN network; (2) proposing a balanced multi-branch KD-Tree structure built by supervised learning method to classify image semantic relationships at leaf nodes; (3) Performing semantic relationship classification between visual objects; (4) formation of triplets describing the image; (5) construct a SPARQL query from the triple; (6) adding data of experimental sets of COCO, Flickr to the ontology; (7) search semantically similar image set to input image based on built ontology.

V RESULTS

With the proposed technique and method, some results were obtained, including (1) building a KD-Tree structure for classifying semantic relationships between objects on images; (2) training KD-Tree to improve the performance of semantic relationship classification; (3) experimental image retrieval using the defined semantic relationship by searching on the ontology; (4) experimental results are higher than some published works [1, 2]; (5) SR-KDT image query system can be applied to multi-object image sets such as Visual Genome.

VI EVALUATION

The results of system semantic relationship-based image retrieval using the KD-Tree structure is higher than other works because: (1) The SR-KDT system uses the R-CNN network for image segmentation and object classification, so the classification performance is high; (2) The training process of weight vector training to classify semantic relationships on KD-Tree should be highly effective. Besides, the approach to classifying semantic relationships between objects by KD-Tree has proven the correctness of the proposed theoretical.

VII CONCLUSIONS

In this paper, a model of semantic relationships-based image retrieval has been implemented using KD-Tree, R-CNN, and Ontology. Based on the semantic relationship of an input image, the SPARQL query is built to retrieve a set of similar images by semantics on ontology. The average image retrieval accuracy results for each COCO and Flickr image data sets were **0.6972** and **0.7794**, respectively. Experimental results have proved that a method of combining the R-CNN, KD-Tree structure, and ontology for semantic-based image retrieval is feasible and effective. In the next development direction, we will combine semantic relationships for KD-Tree Random Forest with ontology to improve performance retrieval.

REFERENCES

[1] Wang Z., Liu X., Li H., Sheng L., Yan J., Wang X., Shao J.: *Camp Cross-modal adaptive message passing for text-image retrieval*. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 5764-5773, (2019).

[2] Yoon, S., Kang, W. Y., Jeon, S., Lee, S., Han, C., Park, J., Kim, E. S.: *Image-toimage retrieval by learning similarity between scene graphs*, arXiv preprint arXiv, 4322, 2012.14700, 2020.

[3] Dinh, N. T., and Le, T. M.: *An Improvement Method of Kd-Tree Using k-Means and k-NN for Semantic-Based Image Retrieval System*. In World Conference on Information Systems and Technologies, Springer, Cham

DOI: https://doi.org/10.1007/978-3-031-21743-2_38

Semantic-based Image Retrieval using R^S-Tree and Knowledge Graph

Le Thi Vinh Thanh[0000-0003-4270-4412], Thanh The Van[0000-0001-8408-2004], and Thanh Manh Le[0000-0002-0949-222X]

thanhltv@bvu.edu.vn, thanhvt@hcmue.edu.vn, lmthanh@hueuni.edu.vn

SIMPLIFIED TITLE

Image Retrieval using R^S-Tree and Knowledge Graph

ABSTRACT

High-level semantic image retrieval is the problem which has applied in many different fields. In this paper, we approached a method of similarity image retrieval and image semantic extraction based on the combination of R^S-Tree and knowledge graph. The main tasks include: (1) building a knowledge graph to store objects, attributes, and relationships of objects on images; (2) searching a set of similar images based on the low-level features stored in R^S-Tree; (3) extracting a scene graph of image using Visual Genome dataset; (4) generating SPARQL query based on the scene graph to extract high-level semantic of the image from the built knowledge graph. In order to evaluate the efficiency as well as compare the accuracy of R^S-Tree, the COREL and Wang image datasets are used. On the basis of the proposed method, a knowledge graph is built based on the Visual Genome dataset. The experimental results are compared with related works to demonstrate the effectiveness of the proposed method. Therefore, our proposed method is feasible in semantic-based image retrieval systems.

I INTRODUCTION

In recent years, many works have approached the knowledge graph and the scene graph to describe the semantics of images [1-3]. In this paper, we propose a model of semantic-based image retrieval using R^S-Tree and a knowledge graph. In this model, there are two main tasks including (1) retrieving a set of similar images based on low-level visual features using R^S-Tree; (2) extracting high-level semantics and relationships of objects on images using a knowledge graph. Firstly, the low-level features of images in the dataset are extracted, such as color, position, shape. Then, the set of similar images is retrieved using R^S-Tree. Secondly, a knowledge graph is built based on the Visual Genome dataset and WordNet. To perform this, we extract the objects, attributes, and relationships on the image from the Visual Genome dataset. With a query image, the system extracts a set of similar images and the scene graphs based on VG dataset. After that, the SPARQL query is generated from the scene graphs. Then, the high-level semantic concepts and the semantic descriptions for the image are extracted from the knowledge graph.

II STATE OF THE ART

The scene graph contains objects, attributes, and relationships to represent the image. Scene graph generation consists of two main tasks (1) extracting visual objects in images; (2) creating relationships between them. The recent works approached scene graphs for semantic-based image retrieval to enhance the effectiveness of the image retrieval process. Johnson, J. et al. proposed a semantic-based image retrieval framework using the concept of the scene graph. The authors used scene graphs to retrieve semantically related images. Wang, S. et al. introduced an image retrieval model using scene graphs, including visual scene graphs (VSG) and text scene graphs (TSG). Schroeder, B. et al. presented a method that uses scene graph embedding as the basis for an approach to image retrieval. The visual relationships are directed subgraphs of the scene graph with a subject and object as nodes connected by a predicate relationship. Yoon, S. et al. proposed an approach for image-to-image retrieval using scene graphs by graph neural networks. In this work, graph neural networks were trained to predict the image relevance measure computed from human-annotated captions using a pre-trained sentence similarity model. Qi, M. et al. proposed a new framework for online cross-modal scene retrieval based on binary representations and semantic graphs. Ramnath et al. introduced a neural-symbolic approach for a one-shot retrieval of images from a large-scale catalog, given the caption description. Quinn, M. H. et al. described a novel architecture for retrieving instances of a query visual situation in a collection of images. The results of those works show that applying scene graphs for the problem of semantic-based image retrieval is feasible. However, these works have not built knowledge graphs to describe the semantics of images. On the other hand, the works have not performed queries based on SPARQL query language to query scene graphs on knowledge graphs.

III ORIGINAL CONTRIBUTION

The article's contributions include (1) building a semantic-based image retrieval model; (2) building a knowledge graph to store objects, attributes, and relationships of objects on images using Visual Genome dataset; (3) extracting a scene graph of image using Visual Genome dataset; (4) generating SPARQL query based on the scene graph to extract high-level semantic of the image from the built knowledge graph.

IV Methodology

In this paper, the $\mathrm{R^S}$-Tree data structure is used to cluster similar image data to enhance performance and improve retrieval time. Besides, a knowledge graph is built to describe and store the level height semantics of the image data set. *Firstly*, the process of image retrieval by content was performed on $\mathrm{R^S}$-Tree. *Secondly*, a knowledge graph is generated from the Visual Genome dataset's components using RDF/XML triple language. *Thirdly*, for every input image, we extract the scene graphs of the image in the VG dataset. *Finally*, we created a SPARQL query and executed the query on the knowledge graph using RDF triple language to retrieve similar images and extract the semantics of the input image.

$\mathrm{R^S}$-Tree is a data partition structure including a root, a set of nodes, and a set of leaves. Internal node denoted S_{node} is of the form $\langle MBS, p\rangle$. Where MBS is a sphere that has center denoted $\vec{c}_{node}$, and radius r_{node}, p is the link pointer to the child nodes. This sphere covers the spheres of nodes in each sub-branch of the tree. Each S_{node} has a minimum element of 2 and a maximum of N. The leaf node denoted S_{leaf} is of the form $\langle MBS, element\rangle$. Where MBS is a sphere that has center denoted $\vec{c}_{leaf}$ and radius r_{leaf} contains a set of elements. Each element $spED$ is of the form $\langle MBS, oid\rangle$. Where MBS is a sphere that has center denoted $\vec{c}_{sp}$, and radius r_{sp}, contains the object space, oid is an identifier $\vec{f} = (v_1, v_2, v_3, \dots, v_d)$. Each leaf node S_{leaf} has the maximum number of elements called M and the minimum $m(1 < m < M/2)$.

A knowledge graph is a graph $G = O, A, R >$, where $O = \{o_1, o_2, \dots, o_n\}$ is a set of objects that are vertices on the graph, o_i is an object class label, concept, or instance; $A = \{a_1, a_2, \dots, a_n\}$ is the set of attributes, a_i is the attribute of the concept or instance; $R = \{r_1, r_2, \dots, r_n\}$ is the set of relationships that are the edges of the graph, r_i is a relationship between concepts and instances, or the relationship between concepts and attributes, or relationship between instances and attributes.

Components in knowledge graph structure include: (1) Node Types: Classes (class), inClass (individual), Objects (individual), Images (individual); (2) Relationships: Object Properties (opOBJinv, opIMGinv, opIMGobj), Relations between Objects (ON, IN, OF, WEAR, RIDE, …); (3) Data Properties: Object (dpOBJWordNet, dpOBJObjectID, dpOBJAtribute, dpOBJWidth, dpOBJHeight, dpOBJSynsetID, dpOBJName) Image (dpIMGImageID, dpIMGURL, dpIMGLocation, dpIMGDataset, dpIMGWidth, dpIMGHeight); (4) Annotations: Relationship (anoRELSynsetID, anoRELPredicate, anoRELRelationID, anoRELWordNet, anoRELDescription).

V Results

The precision, recall, and F-measure is used to evaluate the effectiveness of the proposed method. The experimental result on COREL is precision at 80.29%, recall 65.75%, F-measure 72.29%; the image set of Wang is accuracy at 77.31%, recall 60.36%, F-measure 67.79%; the image set of Visual Genome is precision at 69.66%, recall 59.00%, F-measure 63.89%. We compare the results with previous works on the COREL, Wang dataset to demonstrate the effectiveness of the proposed method. The comparison results showed the precision and effectiveness of the proposed model and algorithm.

VI Evaluation

The experimental result shows that the performance of the proposed method is quite high, because of the following reasons: (1) the $\mathrm{R^S}$-Tree is built based on feature vectors in the form of spheres to optimize storage space and enhance the effectiveness of clustering similar images; (2) the knowledge graph is built to store the semantics relations of objects in an image to improve retrieval efficiency.

VII Conclusions

From the above evaluations, they show that our proposed method is effective in solving the problems of semantic-based image retrieval and can be applied in many fields.

References

[1] Johnson, J., Krishna, R., Stark, M., Li, L. J., Shamma, D., Bernstein, M., & Fei-Fei, L. (2015). Image retrieval using scene graphs. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3668-3678).

[2] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123(1), 32-73.

[3] Zareian, A., Karaman, S., & Chang, S. F. (2020, August). Bridging knowledge graphs to generate scene graphs. In European Conference on Computer Vision (pp. 606-623). Springer, Cham.

DOI: https://doi.org/10.1007/978-3-031-21743-2_51
Link to repository: https://github.com/trantin84/WiGCN

Keywords: Recommender System,Collaborative signal,Graph Neural Network ,Embedding Propagation.

Improvement Graph Convolution Collaborative Filtering with Weighted addition input

Tin T. Tran[0000-0003-4252-6898], V. Snasel[0000-0002-9600-8319]

trung.tin.tran.st@vsb.cz,vaclav.snasel@vsb.cz

Simplified Title

Building a relationship matrix between users to improve item recommendations.

Abstract

Graph Neural Networks have been extensively applied in the field of machine learning to find features of graphs, and recommendation systems are no exception. The ratings of users on considered items can be represented by graphs which are input for many efficient models to find out the characteristics of the users and the items. From these insights, relevant items are recommended to users. However, user's decisions on the items have varying degrees of effects on different users, and this information should be learned so as not to be lost in the process of information mining.
In this publication, we propose to build an additional graph showing the recommended weight of an item to a target user to improve the accuracy of GNN models. Although the users' friendships were not recorded, their correlation was still evident through the commonalities in consumption behavior. We build a model WiGCN (Weighted input GCN) to describe and experiment on well-known datasets. Conclusions will be stated after comparing our results with state-of-the-art such as GCMC, NGCF and LightGCN.

I Introduction

Recommendation systems are an important research area of Information Systems. Based on previous interactions, such as purchases or reviews, of multiple users on items, the system will look for features and point out similar users and similar items to offer a set of recommendations. Combine products and recommendations to target customers. Collaborative filtering is a method of assessing the similarity between users based on a series of past behavior.

The user-item relationship matrix is usually a sparse matrix, because the number of items a user has purchased or is interested in accounts for only a very small portion of all items. Reducing the number of matrix dimensions helps to create embedded matrices of much smaller size without losing the features of users and items. Social recommendations are systems that consider the social relation between users. Users can also be friends with each other in real life or on social networks; a friend's advice is always taken with higher trust.

On the other hand, the user and item relations can also be naturally represented by graphs, and can be exploited by Graph neural network. The process of capturing collaborative signals can go through two stages, decomposer and combiner. In this publication, we propose a model that uses Graph neural network to receive signals from two separate matrices: the implicit matrix that records the user's attention on the item and the weight matrix. The number of interactions between users.

II State of the Art

Due to the history of interaction between users and items, we can model a bi-parties graph $G = \{V, E\}$, where set of vertices $V = \{U, I\}$ is union from set of users and set of items; the E contains edge $e_{i,j} = (u_i, i_j)$ if user u had interaction on item i. From the bi-parties graph G, we can define a matrix $R \subseteq U \times I$ and it's elements has a binary value.

The weighted user references matrix $W_U = R \times R^T$ shows how many common items that user u_i and u_j have interacted by value of $W_{U_{i,j}}$. Because the number of interactive items of each user is different, matrix W should be normalized by the least absolute deviations. Similarly, the weighted item references matrix $W_I = R^T \times R$ indicates how many users the same two items were referenced by.

The Laplacian matrices for user-item relation should be formed in the propagation rule because all embedding vectors could be updated simultaneously. The additional input matrix Δ collects Weighted references values of both users and items.

$$\Delta = D^{-\frac{1}{2}} B D^{-\frac{1}{2}} \quad \text{and} \quad B = \begin{bmatrix} W_U & 0 \\ 0 & W_I \end{bmatrix} \tag{1}$$

III Original Contribution

In the proposed model, a matrix representing the relationships between users has enriched information for the machine learning process. Due to the fact that the users are not friends, their relationship is considered from the set of items of mutual interest.

IV Methodology

We define users embeddings and items embeddings as $e_u \in \mathbb{R}^d$ and $e_i \in \mathbb{R}^d$ with d is the size of the embeddings vector. In the first loop of propagation, embedding layer $E^{(0)}$ need an initial state of He normal weight initialization method which has good performance with Rectified Linear Unit activation function. The embedding should be trained though each round of propagation of a LeakyReLU function.

The collaborative filtering method will capture signals inside the graphs' structure and training the embeddings of both users and items. As a message-passing model of GNN, we extract the signals and then make an aggregation for the embedding at the output.

With messages received from the neighborhood of a user u, we aggregate all of them to refine the representation of u.

$$e_u^{(1)} = LeakyReLU\left(m_{u\leftarrow u} + \sum_{i \in \aleph_u} m_{u\leftarrow i}\right) \tag{2}$$

After a number of propagation iterations, the embedding vector E^* will be acquired, and the predicting score between user u_i on item i_j can be calculated by $\widehat{y}_{ui} = e_{u_i}^{*\top} e_{i_j}^{*}$. We build the loss function with Bayesian Personalized Ranking because it is the best suitable method for implicit feedback datasets.

V Results

To demonstrate the result, we compare our proposed model with the following state-of-the-art methods: **LightGCN**[1], **NGCF**[2] and **GCMC**[3] in both of precision and recall values. The overall performance comparison shows in below table. Our proposed model gives better scores in precisions, recall and NDCG.

Table 1: Overall Performance Comparisons

Dataset	**Gowalla**			**Amazon-book**			**Yelp2018**		
	precision	recall	ndcg@20	precision	recall	ndcg@20	precision	recall	ndcg@20
GCMC	0.03690	0.11799	0.09042	0.01122	0.02539	0.02033	0.02320	0.05114	0.04141
NGCF	0.04080	0.13123	0.11149	0.01713	0.41160	0.03045	0.02192	0.04865	0.03917
LightGCN	0.03961	0.12637	0.11032	0.01181	0.02657	0.02114	0.01975	0.04262	0.03462
Our model	0.04167	0.13503	0.11648	0.01772	0.04335	0.031619	0.02225	0.04894	0.03954

VI Conclusions

In this work, we have proposed a method of weighting the influence among users, which can be considered equivalent to the Social Relation of Trust in the item recommendation problem. The input complement matrix has been calculated to enrich the interactions between the user and the item, and it serves as a moderator of the signals during the extraction and synthesis of messages. By contributing a machine learning model, we believe we have created an inspiration for future studies on the Recommender systems.

References

[1] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. Lightgcn: Simplifying and powering graph convolution network for recommendation, 2020.

[2] Wang, X., He, X., Wang, M., Feng, F., and Chua, T.-S. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (jul 2019), ACM.

[3] Wang, X., Wang, R., Shi, C., Song, G., and Li, Q. Multi-component graph convolutional collaborative filtering. *Proceedings of the AAAI Conference on Artificial Intelligence 34*, 04 (Apr. 2020), 6267–6274.

Link to arXiv: https://arxiv.org/abs/2210.14607
DOI: https://doi.org/10.1007/978-3-031-21743-2_46

Keywords: skill extraction, named entity recognition, text embedding, text ranking

A practical method for occupational skills detection in Vietnamese job listings

Viet-Trung Tran, Hai-Nam Cao, Tuan-Dung Cao

{trungtv,namch,dungct}@soict.hust.edu.vn

SIMPLIFIED TITLE

A practical method to recognize occupational skills in Vietnamese job listings

ABSTRACT

Vietnamese labor market has been under an imbalanced development. The number of university graduates is growing, but so is the unemployment rate. This situation is often caused by the lack of accurate and timely labor market information, which leads to skill miss-matches between worker supply and the actual market demands. To build a data monitoring and analytic platform for the labor market, one of the main challenges is to be able to automatically detect occupational skills from labor-related data, such as resumes and job listings. Traditional approaches rely on existing taxonomy and/or large annotated data to build Named Entity Recognition (NER) models. They are expensive and require huge manual efforts. In this paper, we propose a practical methodology for skill detection in Vietnamese job listings. Rather than viewing the task as a NER task, we consider the task as a ranking problem. We propose a pipeline in which phrases are first extracted and ranked in semantic similarity with the phrases' contexts. Then we employ a final classification to detect skill phrases. We collected three datasets and conducted extensive experiments. The results demonstrated that our methodology achieved better performance than a NER model in scarce datasets.

I INTRODUCTION

Vietnamese job portals have been considered an important bridge between recruitment managers and job seekers. Over the years, these portals have accumulated a growing amount of digital labor-related market data such as job listings and applicants' resumes. To enable advanced analysis, it is imperative to have a model that can automatically detect occupational skills from labor market-related data.

II STATE OF THE ART

In [3], the authors consider this skill detection task as a Named Entity Recognition (NER) task in natural language processing. It has a common drawback: many labeled sentences are needed to train the NER models in a supervised setting. [1] detects skills from a given document by performing a direct match between n-gram sequences and terms in the target taxonomy. This approach, however, does not work for the Vietnamese language as there is no such a taxonomy yet.

III ORIGINAL CONTRIBUTION

In this paper, we present a practical approach for skill detection in Vietnamese job listings. We model the task as a ranking problem. Our approach exploits the structural property of a job description: any skill mentioned found in a requirement section will have a high semantic similarity score with the section itself.

IV METHODOLOGY

In comparison to the traditional NER approach, our methodology is more practical and less expensive in terms of manual efforts. It is a pipeline composed of 4 layers:

1. **Phrase mining** Occupational skill mentions can be multi-word or single-word phrases (e.g., "Java", "data mining"). Thus, a crucial step in our pipeline is phrase mining, which aims at extracting high-quality phrases in a given document. The output of this layer is considered as skill mention candidates for the next ranking layer. We leverage a semi-supervised, weak supervision approach for this layer to reduce manual effort.

2. **Text embedding** This layer is responsible to output the corresponding embedding vectors, given a word, a phrase, a sentence, and a paragraph. The output of the embedding layer will be used to compute ranking similarity scores. For this layer, we can leverage powerful embedding methods such as SIF, and BERT. Thus, this layer requires no manual labeling effort.

3. **Phrase ranking** This layer ranks the importance of a phrase w.r.t. its outer context, the parental requirement section in the job description. Apparently, it can be observed from practical experiments that skill-related phrases often achieve high rankings, while low-ranked phrases are undoubtedly not occupational skills. Therefore, ranking important phrases are capable of discarding clusters of words that are irrelevant to the occupational topic. This layer does not require labeling effort.

4. **Occupational skill classification** Generally, there are cases where many extracted phrases in the previous steps are not occupational skill terms. Therefore, this layer is necessary as a binary classification model to identify truly occupational skill terms. This layer requires a labeled dataset. Beginning with a subset of the output phrases in the ranking layer, our labeling workers are required to prepare two subsets: skill and non-skill ones. In contrast, NER labeling workers must carefully select the phrase spans and assign the corresponding labels. Thus, the dataset construction effort is generally cheaper and faster in our methodology than in the NER approach.

V Results

We implement our methodology with these chosen components: (1) AutoPhrase [2] method for quality phrase detection; (2) various embedding methods; (3) ranking phrases by cosine similarity scores; (4) a small multi-layer neural network for classification.

VI Evaluation

We prepare multiple datasets in which **Vn_requirements_NER** consists of 625 requirement paragraphs of the job descriptions, with manual labeling of skill terms. We measure precision, recall, and F1 scores in complete match and partial match assessments.

VII Conclusions

Our methodology is practical so that it does not require expensive manual labeling datasets. Skill mentions are first detected through an automated phrase detection component that relies on limited positive-only terms. Then a ranking component based on text embeddings is used to filter out non-related skill terms. The remaining terms are fed to a classification model to finalize the skill detection pipeline. Our methodology can achieve comparable performance to a popular industrial-ready NER model in the Vn_resumes_NER dataset while being superior in the smaller Vn_requirements_NER dataset.

References

[1] Javed, F., Hoang, P., Mahoney, T., and McNair, M. Large-scale occupational skills normalization for online recruitment. In *Twenty-ninth IAAI conference* (2017).

[2] Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C. R., and Han, J. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering 30*, 10 (2018), 1825–1837.

[3] Vasudevan, S., Singh, M., Mondal, J., Peran, M., Zweig, B., Johnston, B., and Rosenfeld, R. Estimating Fungibility Between Skills by Combining Skill Similarities Obtained from Multiple Data Sources. *Data Science and Engineering 3*, 3 (2018), 248–262.

DOI: https://doi.org/10.1007/978-981-19-8234-7_49
Link to repository: https://github.com/ttcgiang/ACIIDS_Paper_262

Keywords: Robotics, Coverage Path Planning, Energy Constraints , Approximation, Boustrophedon Decomposition

Hybrid Boustrophedon and Partition Tree Group Algorithm for Coverage Path Planning Problem with Energy Constraints

Tran Thi Cam Giang[1,2], Huynh Thi Thanh Binh[2],

camgiang2010@gmail.com,binhht@sis.hust.edu.vn

Simplified Title

B-WZone Algorithm for Coverage Path Planning Problem with Energy Constraints

Abstract

Coverage path planning (CPP) finds the shortest feasible paths to cover completely an environment while avoiding inner obstacles. This study focuses on the CPP problem for an energy-limited mobile robot that satisfies two main optimization objectives: the total distance and the number of repeated cells. We propose a hybrid algorithm between the Boustrophedon and the partition tree group algorithm, namely B-WZone, for the CPP with energy constraints. The experimental results showed that the B-WZone algorithm helps the robot reduce energy consumption and traveling time in all the tested environments compared with the existing methods.

I Introduction

Coverage path planning (CPP) is a fundamental problem in robotics with abundant real-world applications and many other potential applications, such as vacuum cleaning, painting, demining, and window cleaning robots [2, 1]. Its aim is to find paths that allow the robot to cover entirely the working environment. This problem is divided into two main versions: OfflineCPP, where robot(s) have prior knowledge of the environment, and OnlineCPP, where the robot does not know the environment's details and can accumulate knowledge of the environment during the coverage. This paper has some significant contributions to achieving the maximum coverage and minimizing the coverage path in current OfflineCPP applications using a battery-charged robot.

We propose a new coverage algorithm called B-WZone, which is the combination of Boustrophedon Algorithm and the Partition Tree Group Algorithm with a new approach to divide the environment by split lines to obtain better results.

II State of the Art

Many previous algorithms have been presented to solve the CPP problem with simple assumptions such as the environment is a polygon, the robot is a unit point, and it can only move in four directions with an unlimited energy budget and visit all the accessible locations in an environment including obstacles by a unique single path. These assumptions obviously are unrealistic when applied in real-life situations.

III Original Contribution

Our system is composed of a battery-powered robot that has pre-known about the working space and a recharging station where the robot starts. The goal is to fully pass the whole given environment while avoiding obstacles or being out of energy in the middle of the work. By using our method, we can achieve better optimizations in terms of the number of times the robot returns to the charging station as well as the total traveling distance compared with some existing algorithms.

IV Methodology

To solve our problem, this study introduces a new algorithm that is based on the ideas of the Boustrophedon Algorithm and Partition Tree Group Algorithm, which we named the B-WZone Coverage Algorithm. The B-WZone algorithm includes three key steps: (1) Environment Decomposition by the Boustrophedon method, (2) Grouping nodes of the partition tree into a working zone and (3) Coverage by Boustrophedon Motions.

The first step is to divide the environment into small subareas by using Boustrophedon Decomposition proposed by Choset et al. in 1997. This paper benefits from the unique ability of the Boustrophedon to deal with cell over-generation since we only use vertices where a vertical line can go to both left and right of the vertex. These vertices are connected to create a ”*spit* − *line*”. The more number of subareas is reduced, the shorter the absolute coverage path is.

The next step is building the partition tree T to group subareas into zones for easier coverage. Each node in the tree represents a divided subarea in the previous step. The root node N_1 is the one containing the charging station. If a subarea represented by the node N_i has a split-line that collides with the split-line of the other m subareas in which nothing is the parent node of N_i, the m nodes that denote those subareas will be the child nodes of N_i. If the split-lines of m different subareas (represented by m other nodes) collide together and collide with the split-line of a subarea N_j (not the parent node of these m subareas), then node N_j representing that subarea will be arbitrarily set as a child node of 1 of m nodes mentioned above.

The last step is covered partitioned zones in turn by Boustrophedon Motion. In the beginning, the robot moves to the closest corner of the zone and starts Boustrophedon moving before returning to the charging station. This process is repeated until the robot visits all divided zones.

V Results

The scenario has ten different environments, shown in Fig. 1. The first environment is the same shape with two inner obstacles in the experiment of Wei [2]. The remaining environments are randomly generated. The size of the workspace, the density, and the position of obstacles are different to compare the coverage capacity of B-WZone and the Log-algorithm.

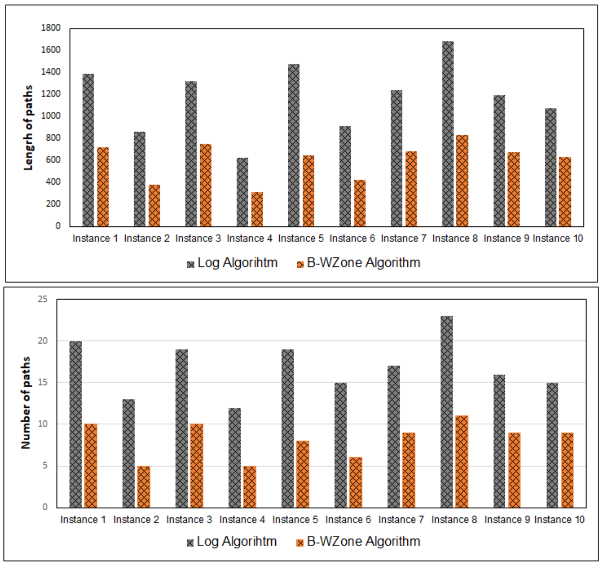


Figure 1: Comparison between the Log-Algorithm and our B-WZone Algorithm

VI Evaluation

We find that in all test cases, our proposed method outperforms the algorithm of Wei and Isler [2] in both optimization criteria by approximately 50%.

VII Conclusions

This paper proposed the B-WZone coverage algorithm that combines the Boustrophedon environmental decomposition and the tree grouping algorithm to solve the OfflineCPP with energy constraints. This algorithm presented its actual performance: reducing the total number of paths, the length of the paths, and the number of repeating cells during the covering process. Our proposed B-WZone algorithm returns better results than the method introduced by Wei and Isler.

References

[1] Galceran, E., and Carreras, M. A survey on coverage path planning for robotics. *Robotics and Autonomous systems 61*, 12 (2013), 1258–1276.

[2] Wei, M., and Isler, V. A log-approximation for coverage path planning with the energy constraint. In *Twenty-Eighth International Conference on Automated Planning and Scheduling* (2018).

DOI: https://doi.org/10.1007/978-981-19-8234-7_50

Keywords: Disease prediction, Data mining, Network analysis, Graph database, Machine learning

Predictive modelling of diseases based on a network and machine learning approach

Quang-Tuan Truong, Nghia Le, Bac Le[0000-0002-4306-6945]

truongquangtuanit@gmail.com, nghialh@uit.edu.vn, lhbac@fit.hcmus.edu.vn

SIMPLIFIED TITLE

Disease prediction modelling

ABSTRACT

Chronic diseases have become the first prioritized concern of the health industry, so understanding the disease progression is necessary for predicting, planning and preparing resources to prevent and cure the diseases most effectively. Basing on patients' medical history, this research analyzes and builds disease network to exploit hidden information showing the disease relations and progressions, applys machine learning models to assess the risks of morbidity and predicts the risk of contracting cardiovascular diseases (CVD) in patients with type-2 diabetes (T2D). The research data includes 249,809 medical histories of 65,337 patients in Ho Chi Minh City, Vietnam. The accuracies of the four prediction models (SVM, DT, RF and KNN) range from 78% to 80%. The predicted data can be used promisingly as a reference for medical specialists to provide effective healthcare guidance to patients as well as for healthcare service providers to use their data effectively and enhance their service quality.

I INTRODUCTION

World Health Organization (WHO) defines chronic diseases as non communicable diseases (NCDs), which are not transmissible directly from person to person and are associated to long duration and slow progression. They have become the leading cause of fatality in most countries. According to WHO, 41 million people die from an NCD annually, which accounts for 77% of all deaths worldwide. Moreover, almost 50% out of these deaths are premature deaths (under 70 years old) and this percentage is growing. Because most NCD patients are still in their working age, chronic diseases reduce labor productivity and increase pressure on the medical system, resulting in considerable financial burdens to the economy of countries, especially low- and middle-income ones. Among four main types of NCDs, Type 2 diabetes is the most dangerous because it has the highest morbidity rate in the world. In fact, it is the top cause of fatality and it increases the risk of contracting other chronic diseases such as cardiovascular diseases [2].

Understanding the progression of chronic diseases and making predictions relating to them can provide important information useful for the prevention and management of the diseases. Thus, many models have been proposed to forecast the morbidity risk of chronic diseases, which can be divided into two main approaches: (i) Traditional statistical approaches and (ii) Machine learning and data mining approaches. However, both methods require a very big and full administrative data set with a great number of medical records, which are extremely difficult to collect because patients' medical data are usually privately secured, discrete and unshareable. Besides, there exist some limitations in the exploitation of the relations among diseases and the progression of them. Also, the above approaches are very different in the method of implementation.

Therefore, in the research scope, this study proposes a new method by widening and combining the above two approaches in order to build a risk prediction model for chronic diseases. Specifically, this study will predict the morbidity risk of cardiovascular diseases in type 2 diabetes patients.

II STATE OF THE ART

The rise of chronic diseases increases the health and financial burden. Therefore, there have been many studies on the improvement of patients' health through prevention, early identification, and proper treatment. However, most approaches have not shown the progression and association between diseases. They also have many limitations in accessing data because the patient's health data is often confidential, discrete, and not shared. For example, some regression models are not suitable because medical data has too many objects such as patients, diseases, doctors, drugs, etc. Therefore, model building is often complicated because there are many independent variables. Although machine learning methods are rich in techniques and tools, they do not show the necessary time and relationship factors in disease prediction. Models based on network analysis are often developed to find hidden patterns and relationships between objects. However, in the medical field, clinical diagnoses are often based on disease factors such as medical history, demographic information, vital signs, etc. These factors can also have a major influence on patient outcomes. In most cases, this information has not been exploited. Therefore, this study proposes a new approach by extending and combining existing methods, with the overall goal of focusing on understanding the progression of chronic disease through disease networks and identifying which patients are at risk of other chronic diseases in the future by applying machine learning algorithms to develop disease prediction models [1].

III Original Contribution

- Instead of approaching the traditional statistical methods to build model, the study applies graph theory and network analysis to develop a disease network, and then applies machine learning methods to build a disease prediction model basing on the data extracted from the disease network.

- The disease network created in this study presents the relations and the progressions of chronic disease comorbidities, as well as providing a simple way to visualize the health trajectory of chronic disease patients.

IV Methodology

The study is divided into two main parts, (i) understanding disease progression through network analysis approach and (ii) developing disease prediction models through machine learning algorithms. To build the disease network, two cohorts (the group of patients with T2D only and the group of patients with both T2D and CVD) were aggregated from their medical history to create two corresponding base disease networks. The final disease network is generated by aggregating the above two base networks. This disease network provides detailed information on the progression and relation of CVD in patients with T2D. Based on these insights, the development of a disease prediction model was also the second goal of this study. To develop the model, the study has extracted five features, three of which are from the disease network (node, edge and cluster) and the other two are from the data set. (gender and age). These features are then used to develop predictive models of CVD risk in patients with T2D.

V Results

The built disease network includes 19 chronic diseases and the edges represent disease progression. Four machine learning algorithms are used: SVM, DT, RF and KNN. Overall, the accuracies of the four models are almost similar. However, the study found that the predictive model based on SVM had the best accuracy of 80.02%, followed by the DT with an accuracy of 79.12%.

VI Evaluation

This study uses 249,809 medical histories of 65,337 patients in a hospital in Ho Chi Minh City. 65% of patients in each cohort were randomly selected to build a disease network. The remaining 35% was divided into two datasets: the training and the test datasets. With the training dataset, 10-fold cross-validation was used to train and test the model and the test dataset was used to evaluate the performance of the models. The study used a confusion matrix to calculate the common measures such as accuracy, precision, recall, F1 score and ROC curve to evaluate the efficiency. Despite using only the disease classification code combined with some other basic features (age and gender), the approach gives quite good accuracy. Therefore, this is expected to be a new development direction in the current context of very little and sparse medical data.

VII Conclusions

Disease prediction is an important research in the health industry and especially in preventive medicine. In the context of the rapid increase of chronic diseases, especially in low- or middle-income countries, and the increasing burden of disease due to population aging in many parts of the world, the results of disease prediction models are important inputs for planning and preparing resources for effective disease prevention. Moreover, this research could be useful to individuals and organizations in the healthcare field as it can help healthcare providers use their data more efficiently and improve their services.

References

[1] Arif Khana, Shahadat Uddina, U. S. *Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression.* 2018.

[2] VietNam, M. O. H. O. *National Strategy for prevention and control of non communicable diseases, period 2015-2025.* 2022.

DOI: https://doi.org/10.1007/978-3-031-21743-2_37

Keywords: Video coding, Video quality, MPEG, codecs, VMAF

Preliminary Study on Video Codec Optimization Using VMAF

Syed Uddin, Mikolaj Leszczuk, Michal Grega
syed.uddin, mikolaj.leszczuk, michal.grega@agh.edu.pl

SIMPLIFIED TITLE

Video codec optimization and Quality of Experience

ABSTRACT

The growth in video streaming has been an exponential one for the last decade or so. High-resolution videos require high bandwidth to transport the videos over the network. There has been a growing demand for compression technologies to compress videos while simultaneously maintaining quality. Video codecs are used to encode and decode video streams. These codecs have been developed by MPEG, Google, Microsoft, and Apple Inc. There are many encoding parameters that affect bitrate and video quality. These performance parameters must be exploited, evaluated, and modeled to find the best possible solutions. This paper demonstrates some preliminary results for video coding sets with selected bitrates. The objective video multimethod assessment fusion (VMAF) metric is calculated for the encoded video versions. In this study, the quality of the encoded videos was evaluated and estimated using VMAF. The results confirm a strong relationship between bitrate and VMAF estimates. This study shows the impact of coding parameters on the VMAF values and provides the foundation for building robust models in the field of video quality analysis.

I INTRODUCTION

In this research, video streams are encoded. The quality of the encoded videos was evaluated and estimated using video multimethod assessment fusion (VMAF) [1]. The video sequences are encoded at different bit rates. These data rates are constant bit rate (CBR) and variable bit rate (VBR) [2]. The application of these data rates has an impact on file size, encoding time, and video quality. In addition to data rates, there are also quantization parameters that also affect video quality. The main goal of this work is to encode the video stream with different video codecs. The VMAF between the original video and the distorted version was calculated and a performance comparison is made between video streams.

II STATE OF THE ART

A number of video codec technologies are utilized to reduce bit rate and enhance video quality. The Versatile video coding (VVC) is used for compression to reduce bit rates from 30 to 40 % compared to High-efficiency video coding (HEVC). The MPEG-4/AVC codec is another promising codec for compression. The alliance for open Media (AOMedia) develops open-source video codecs. Google developed the VP9 standard in order to compete with MPEG. The performance of video coding algorithms is evaluated by comparing their rate-distortion (RD) or rate quality (RQ) performance on a variety of test sequences. Objective and subjective evaluation methods are used to evaluate video quality. Describe briefly existing solutions to the undertaken research problem. Point out boundaries of existing knowledge. Do not do a literature review, and do not make any references or citations.

III ORIGINAL CONTRIBUTION

The VMAF objective model is utilized in order to calculate video quality. The results demonstrate a strong relationship between bitrate and VMAF estimates. In this study, quantization parameters were explored and tested. This study provides the foundation for building robust models in the field of video quality analysis.

IV METHODOLOGY

An experimental study is conducted where Video sequences in YUV format (raw format) are acquired from Big Buck Bunny (BBB). The resolution of the video sequence corresponds to Full high definition (FHD). The FFMPEG software tool is utilized for building and running scripts. The video sequence was encoded using different input bitstreams. The VMAF objective model was used for calculating VMAF values.

V RESULTS

In this study, video sequences were encoded using different codecs. The results demonstrate that the MPEG and VP9 codecs are potential codecs for better compression under specific conditions. The HEVC/H.265 codec is a better codec as evidenced by the encoding settings and corresponding VMAF quality scores. A preliminary model is developed and tested based on quantization parameters and VMAF estimates. This model is used to measure the quality of video sequences based on an objective model.

VI EVALUATION

The experimental study was carried out. The study is based on testing conditions utilizing an objective-based model. A number of codecs were acquired from MPEG and Google. The codecs are configured in FFMPEG tool. The video quantization parameters (QP) utilized and programmed in FFMPEG. The parameters were tested under different conditions. Those parameters are exploited, evaluated, and modeled to find the best optimal solutions.

VII CONCLUSIONS

This study provides the best codec which can be opted for in video optimization. This also provides the foundation framework which can be extended and optimal solutions can be generated. This study provides the implementation of VMAF model used for quality estimation. The study tests quantization parameters under different conditions which is crucial in the field of video quality optimization. The results are utilized for building robust models in the field of video quality analysis.

REFERENCES

[1] Zhi Li, Christos Bampis. VMAF: The Journey Continues. Netflix Technology Blog | Netflix TechBlog. https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12

[2] Katsavounidis, Ioannis & Guo, Liwei. (2018). Video codec comparison using the dynamic optimizer framework.

Link to arXiv: https://arxiv.org/abs/2207.14636
DOI: https://doi.org/10.1007/978-3-031-21743-2_48
Link to repository: https://github.com/sonlam1102/vispamdetection

Keywords: spam reviews, text classification, deep neural models, transformer models, dataset, annotation guidelines

Detecting Spam Reviews on Vietnamese E-commerce Websites

Co Van Dinh, Son T. Luu, Anh Gia-Tuan Nguyen

19521293@gm.uit.edu.vn, sonlt@uit.edu.vn, anhngt@uit.edu.vn

Abstract

The reviews of customers play an essential role in online shopping. People often refer to reviews or comments of previous customers to decide whether to buy a new product. Catching up with this behavior, some people create untruths and illegitimate reviews to hoax customers about the fake quality of products. These are called spam reviews, confusing consumers on online shopping platforms and negatively affecting online shopping behaviors. We propose the dataset called ViSpamReviews, which has a strict annotation procedure for detecting spam reviews on e-commerce platforms. Our dataset consists of two tasks: the binary classification task for detecting whether a review is spam or not and the multi-class classification task for identifying the type of spam. The PhoBERT obtained the highest results on both tasks, 86.89%, and 72.17%, respectively, by macro average F1 score.

I Introduction

In recent years, e-commerce in Vietnam has had strong growth. Besides the advantages of online shopping, these e-commerce platforms face several challenging problems that cannot be fully resolved, such as fake products and fraud qualifications. Customers deciding whether to buy a product on an e-commerce site depends a lot on the reviews of previous users. Understanding this, some users or stores have created fake or spam reviews that affect the user's experience when shopping on these e-commercial platforms. Therefore, detecting these spam reviews will protect both sellers and customers from the risk of low-quality products and preserve the reputation of the sellers. In this paper, we propose a method to detect spam reviews about products on online shopping platforms. First, we constructed a corpus for spam detection from users' reviews by texts. Then, use machine learning approaches to build the classification models for detecting spam comments and evaluate classification models' performances on the constructed dataset.

II State of the Art

Transformers is an architecture that has been proposed in recent years and is currently in widespread use. The appearance of BERT [1] helps many downstream tasks in NLP attain high-performance results while training on a small dataset. BERT and its variances become the baseline approaches in many NLP tasks, which is called BERTology. In the Vietnamese language, there are two kinds of BERTology approaches: multilingual and monolingual models. As a result, the monolingual obtained better results than the multilingual models for the text classification and sequence-to-sequence tasks. In this paper, we use these models as a more effective solution for classifying spam reviews than conventional methods, and the results have also been proven in our experiments.

III Original Contribution

Currently, in Vietnamese, the dataset for detecting spam reviews on e-commerce sites is vital. Therefore, we have introduced a Vietnamese by constructing a corpus for spam detection from users' reviews by texts. The dataset contains more than 19,000 reviews of users from Vietnamese E-commerce websites and is manually annotated by humans. Besides, we also use machine learning approaches to build classification models for detecting spam comments and evaluate classification models' performances on this dataset. This experiment evaluates a part of our problem. Moreover, the error analysis and proposing methods to improve the results of the models are also helpful in future studies.

IV Methodology

Our problem comprises two tasks, Task 1 and Task 2. It is defined as follows: Task 1 is the binary classification task for classifying whether a review is a spam or not, and Task 2 is the multiclass classification task for identifying the type of spam. We solve those two tasks as text classification problems. Word embedding is a vector space used to represent text data that can describe the relationship, semantic similarity, and data context. In text classification, the fastText pre-trained embedding obtained robust results. Therefore, we chose this for our empirical results.

Deep neural network models: We use Text-CNN [2] to extract text features for classification. At the same time, we also use the Long Short Term Memory (LSTM) [3] and Gated Recurrent Unit (GRU) [4] models. These are two variants of the Recurrent Neural Network (RNN). It helps the model remember the previous information for a long time, a restriction faced by the RNNs. In particular, GRU is an attempt to reduce the complexity of the LSTM model, which reduces the computational cost and training time of the model.

Transformers: Besides using deep neural network models for the training model, we also applied two monolingual BERT models, including PhoBERT [5] and BERT4News [6] for our problem of detecting spam reviews.

V RESULTS

We have trained the Text CNN, LSTM, GRU models, and transformers with PhoBERT and BERT4News for our tasks. Then, we use the Accuracy and macro-averaged F1-score metrics to evaluate the performance of baseline models. The experimental results of the models are shown in Table 1. PhoBERT is the model that gives the best evaluation results on both tasks in our experiments.

Table 1. The empirical results of classification models on the dataset

Model	**Accuracy (%)**		**F1 score (%)**	
	Task 1	**Task 2**	**Task 1**	**Task 2**
Text-CNN	84.18	83.42	77.89	64.74
LSTM	82.97	83.35	77.24	66.58
GRU	83.50	82.84	77.67	66.51
PhoBERT	**90.01**	**88.93**	**86.89**	**72.17**
BERT4News	86.39	86.20	86.16	62.62

VI EVALUATION

In the prediction results of the best model - the PhoBERT, the error predictions are significant on the second task – the spam types detection. Most of the SPAM-2 reviews are predicted as NO SPAM, and the number of the wrong prediction is higher than the accurate prediction. The proportion of the incorrect prediction of SPAM-1 reviews is also very high, in which most SPAM-1 reviews are predicted as no spam. However, this error prediction is not too much in the whole test set. The reviews with type SPAM-3 are the same as type SPAM-1. In general, most wrong predictions are caused by the doubt between NO SPAM label and other labels. Thus, the challenge of classification models on our dataset for this task is not only to determine whether the reviews are spam or not but also to identify the type of spam reviews.

VII CONCLUSIONS

This paper provided the ViSpamReviews - a dataset for spam review detection on Vietnamese online shopping websites with more than 19,000 reviews annotated by humans with a strict annotating process. We applied robust classification models to the dataset. The PhoBERT model obtained the highest result with 86.89% by F1-score for the spam classification task and 72.17% by F1-score for the spam types detection task. From the error analysis, we found that it is necessary to integrate extra metadata about the product as well as the previous reviews to boost up the classification models. Our subsequent study is to extend the dataset for detecting spans of spam in the reviews and identify the opinion of users on the specific characteristic of products and their relevant services. Finally, based on the current results, the dataset can be used for developing an application to help shop owners filter spam reviews from users.

REFERENCES

[1] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: *BERT: Pre-training of deep bidirectional transformers for language understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

[2] Kim, Y., *Convolutional neural networks for sentence classification*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[3] Hochreiter, S., Schmidhuber, J.: *Long short-term memory*. Neural computation 9, 1997.

[4] Cho, K., van Merri¨enboer, B., Bahdanau, D., Bengio, Y. *On the properties of neural machine translation: Encoder–decoder approaches*. Proceedings of SSST 8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014.

[5] Nguyen, D.Q., Tuan Nguyen, A.: *PhoBERT: Pre-trained language models for Vietnamese*. Association for Computational Linguistics, 2020.

[6] Nguyen, T.C., Nguyen, V.N. *Compose transformer pretrained models for reliable intelligence identification on social network*, Vietnamese Language and Speech Processing, 2020.

DOI: https://doi.org/10.1007/978-3-031-21967-2_15

Keywords: MFCCs, Mel-spectrogram, Tempogram, Speech emotion

An Empirical Experiment on Feature Extractions Based for Speech Emotion Recognition

Binh Van Duong, Chien Nhu Ha, Trung T. Nguyen, Phuc Nguyen, Trong-Hop Do

{18520505, 18520527}@gm.uit.edu.vn,
trung.nguyendx@hcmut.edu.vn,Phn2012@nyu.edu,hopdt@uit.edu.vn

SIMPLIFIED TITLE

An Empirical Experiment on Feature Extractions Based for Speech Emotion Recognition.

ABSTRACT

In recent years, the virtual assistant has become an essential part of many applications on smart devices. In these applications, users talk to virtual assistants in order to give commands. This makes speech emotion recognition to be a serious problem in improving the service and the quality of virtual assistants. However, speech emotion recognition is not a straightforward task as emotion can be expressed through various features. Having a deep understanding of these features is crucial to achieving a good result in speech emotion recognition. To this end, this paper conducts empirical experiments on three kinds of speech features: Mel-spectrogram, Mel-frequency cepstral coefficients, Tempogram, and their variants for the task of speech emotion recognition. Convolutional Neural Networks, Long Short-Term Memory, Multi-layer Perceptron Classifier, and Light Gradient Boosting Machine are used to build classification models used for the emotion classification task based on the three speech features. Two popular datasets: The Ryerson Audio-Visual Database of Emotional Speech and Song, and The Crowd-Sourced Emotional Multimodal Actors Dataset are used to train these models.

I INTRODUCTION

Lately, many researchers have studied Machine Learning and Deep Learning methods to recognize humans' emotions through writing. They achieved considerable achievements. Besides, speech resources are also noticed to be employed. Especially, speech emotion recognition is a promising field for scientific research and practical applications [1]. This paper presents the experimental results for the purpose of comparing the effect of each extracted feature. It also aims to yield a resource for choosing features for the speech emotion recognition problem. The extracted features Mel-spectrogram, Mel-frequency cepstral coefficients, and Tempogram are used for finding out the most suitable feature for each dataset. Deep Learning and Machine Learning based models have been experimented on the extracted features for comparison. The research focuses on the comparison of the effectiveness of different kinds of speech features for the speech emotion recognition task.

II STATE OF THE ART

Researchers are researching the comparison on different speech features. However, those authors only used one model for classification and the datasets used in those research were quite limited, which led to the lack of experimental results. This research learns from the previous ones and extends to using various types of models to run experiments. Besides, the three different types of speech features and their variants would provide much more experimental results. This paper contributes more empirical experiment results on the effectiveness of features in the speech emotion recognition problem which could be a valuable resource for researchers working on the task of speech emotion.

III ORIGINAL CONTRIBUTION

The research provided experiments on three kinds of speech features (i.e., MFCCs, Mel-spectrogram, Tempogram, Combined feature) and two datasets (i.e., RAVDESS, CREMA-D). The authors run experiments on different Machine Learning/ Deep Learning algorithms and a stacking model so as to give empirical results for the research.

IV METHODOLOGY

The experiment results come from six different models (i.e., MLP, Light-GBM, sDL, Conv-1d + LSTM, Conv-2d, stacking model). In order to train the models, there are four emotions (anger, fear, happiness, and sadness) chosen for classification. Because those emotions all appear in both datasets and the difference between them is pretty clear. This could help to evaluate the models better than using other emotions. The models as sSL, Light-GBM , and MLP would accept the one-dimensional input. Other models use two-dimensional input with a fixed length of 3 seconds per audio. So that audio signal having a duration longer than that fixed-length would be pruned or padded before moving to the features extraction stage.

V RESULTS

In addition to evaluating three types of features (Mel-spectrogram, MFCCs, and Tempogram) on designed models, some options in the dimension of MFCCs experimented on several models result in more comparative evaluation numbers. This is a plus point in making comparisons and discussing the strengths and weaknesses of different sets of parameters as well as implemented features. The results of implementations on three sorts of features along with RAVDESS using Light-GBM and MLP. It is noteworthy that the Combined feature (synthesized from MFCCs, Mel-spectrogram, and Tempogram) brings the best scores for both accuracy and F1-score.

VI EVALUATION

Regarding the single-feature approach, the authors highly recommend using MFCCs or Mel-spectrogram in this context. Besides, Tempogram is a valuable feature, although this feature requires more fine-tuning to compete with other mentioned ways of extractions. The Tempogram is not a preferable choice because its implicit extracted information is fiercely related to acoustic repetition often seen in musical beat research. In terms of the MFCCs feature, it is evident that the extracted dimensions, the number of MFCCs coefficients, should be modified to gain the best result.

The authors have also conducted a binary approach by using only two kinds of emotion with the desire to achieve better results. However, this strategy brought back even worse evaluated numbers; therefore, the research concentrates on four types of emotion as an ideal quantity to carry out further research.

VII CONCLUSIONS

The experimental results obtained from this research provide a better understanding of the features in the speech emotion recognition problem. The study furnishes the details of each feature extraction and the basic knowledge of the research field. It can be seen that utilizing different features at once (like the Combined feature) could nourish the models so much.

In the future, this paper can be updated in the feature selection and models. Many other kinds of speech features might be employed as well as the way to combine them, and the models' architectures can be developed better. Furthermore, a Vietnamese emotional speech dataset could be created to supply a data resource for the speech emotion recognition problem.

REFERENCES

[1] Cen, Ling and Wu, Fei and Yu, Zhu Liang and Hu, Fengye, A real-time speech emotion recognition system and its application in online learning, 2016.

DOI: `https://doi.org/10.1007/978-3-031-21967-2_21`

Keywords: Software Enhancement Effort Estimation, Function Point Analysis (FPA), Gaussian Mixture Model, k-means, clustering algorithm, machine learning

Analyzing the effectiveness of the Gaussian Mixture Model clustering algorithm in Software Enhancement Effort Estimation

Vo Van Hai, Ho Le Thi Kim Nhung, Zdenka Prokopova, Radek Silhavy, Petr Silhavy

`{vo_van, lho, prokopova, rsilhavy, psilhavy}@utb.cz`

Simplified Title

Analyzing the effectiveness of a clustering algorithm in software enhancement effort estimation.

Abstract

Background: The influence of data clustering on the effort estimating process has been studied extensively. Studies focus on partitioning and density-based clustering, and some use hierarchical clustering, but most focus on software development effort estimation. Aim: We focus on the Gaussian Mixture Model algorithm's effectiveness in the software enhancement effort estimation. Meth-od: We used the Gaussian Mixture Model clustering algorithm to cluster the dataset into clusters and then applied the IFPUG FPA method for effort estimation on these clusters. The ISBSG dataset was used in this study. The number of clusters is determined using the Elbow method with the Distortion score. Besides, the k-means algorithm was also used as the comparative algorithm. The baseline model was determined by using the FPA method on the entire dataset without clustering.

Result: With the number of clusters selected as 4, on six evaluation criteria, MAE, MAPE, RMSE, MBRE, and MIBRE, the experimental results show the estimated accuracy using the FPA method on clustered data significantly better when compared with no clustering. Conclusion: the software enhancement effort estimation can be significantly improved when using the Gaussian Mixture Model clustering algorithm.

I Introduction

Essentially, in the early phase of a software project, the information available about the project's features is often incomplete, leading to inaccurate estimates. In addition, some software project specifications, such as varying customer requirements, the flexible relationship between project features, the high diversity of development techniques, and the rapid development of hardware platforms and other related problems, make the estimation process difficult. These difficulties are the challenge that also motivates researchers to propose different software estimation methods, especially in software development efforts.

This study focused on the effectiveness of clustering on software enhancement effort estimation using the Gaussian Mixture Model (GMM) clustering algorithm. Besides, we also use the k-means algorithm as the comparative algorithm. This study aims to answer three research questions:

1. RQ1. How does the data clustering using clustering algorithms affect the FPA method estimation accuracy?
2. RQ2. Does the FPA method on clustered datasets outperform this on the non-clustered dataset?
3. RQ3. Which clustering algorithm between the GMM and k-means gives the higher estimation accuracy?

In addition, a statistical pairwise t-test comparison was performed to validate the proposed method's accuracy. The hypothesis relates to confirming a significant difference in estimation capability between using the FPA method on non-clustering and clustering datasets using the GMM algorithm. In other words, the estimation accuracy of the FPA method on the clustered dataset is significantly different from that on the non-clustered dataset.

II State of the Art

Some previous studies have shown that data segmentation has certain effects on the accuracy of software effort estimation. Several solutions have been proposed by applying specific segmentation algorithms and have achieved certain results. GMM is a clustering algorithm that has been successfully applied in the field of image recognition.

However, there has not been a study using this algorithm in the software effort estimation aspect. Therefore, this study applies this algorithm and evaluates its influence on the field of software effort estimation.

III ORIGINAL CONTRIBUTION

The essential contributions of this study are as follows: 1) The effect of clustering has been demonstrated, allowing the IFPUG FPA to be applied to clustered data with the benefit of increasing the accuracy of effort estimation; 2) By applying the GMM and k-means clustering algorithms, the software enhancement effort estimation's accuracy is significantly improved; and 3) A comparison between applying GMM and k-means clustering algorithms on the tested dataset

IV METHODOLOGY

Describe methods that have been utilized to achieve the presented results. Moreover, indicate whether your study is theoretical, experimental, a case study, or a literature review.

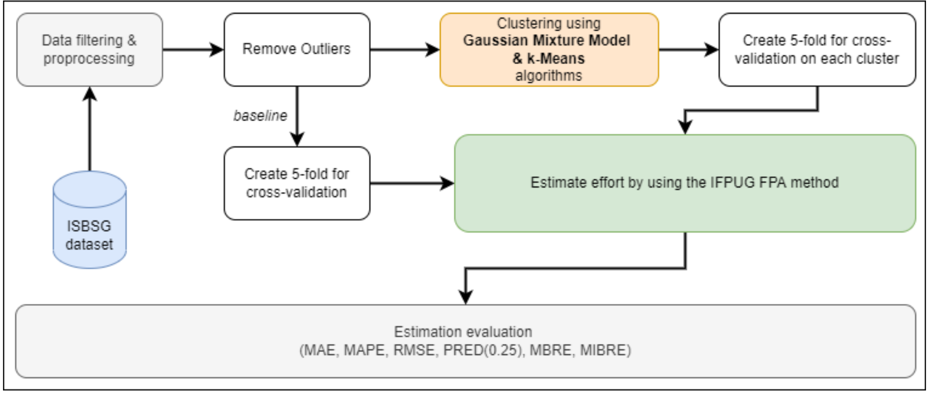

Data filtering & proprocessing
Remove Outliers
Clustering using Gaussian Mixture Model & k-Means algorithms
Create 5-fold for cross-validation on each cluster
baseline
Create 5-fold for cross-validation
Estimate effort by using the IFPUG FPA method
ISBSG dataset
Estimation evaluation (MAE, MAPE, RMSE, PRED(0.25), MBRE, MIBRE)

Fig. 1. Experiments setup

There are three main comparisons in this study: 1) applying FPA on clusters using the k-means algorithm versus baseline; 2) applying FPA on clusters using the k-means algorithm versus baseline; 3) applying FPA on clusters using the k-means algorithm versus using the GMM algorithm.

V RESULTS

The three research questions proposed in this study were proven with the proposed experiment. Using the FPA method on data segments using two k-means and GMM segmentation algorithms achieves more accurate estimation than applying the FPA method in the entire non-clustered dataset. In addition, the GMM algorithm achieves statistically better results than the k-means algorithm between the two applied clustering algorithms.

VI EVALUATION

With six evaluation criteria applied (MAE, MAPE, RMSE, PRED (0.25), MBRE, and MIBRE), the comparison results between the experimental arms are shown in Table 1. With this result, it is easy to see that using data segmentation in the software estimation process provides higher accuracy. Especially with the GMM clustering algorithm, the estimation accuracy gained from 43.17% to 71.08%, depending on the evaluation criterion, compared to not applying data segmentation.

Table 1. Comparison between methods

	MAE	MAPE	RMSE	PRED	MBRE	MIBRE
baseline	521.792	20.702	1,143.450	0.636	0.209	0.139
Means of the k-means results	304.918	12.322	569.142	0.794	0.125	0.099
Means of the GMM results	177.644	9.444	330.724	0.852	0.097	0.079

VII CONCLUSIONS

In this study, we have focused on evaluating the effect of clustering on software effort estimation using the FPA method. Two clustering algorithms, k-means and GMM, have been assessed, and the results of this procedure are compared with the baseline model. The experimental results show that the estimation accuracy will significantly improve when applying the clustering algorithm. Of the two algorithms used, the GMM algorithm achieves higher accuracy in estimation than the k-means algorithm.

DOI: `https://doi.org/10.1007/978-981-19-8234-7_51`
Link to repository: `https://drive.google.com/drive/folders/1vE5gAOu2wTrH7V143s9ehsCl_e_juFCt`

Keywords: Neural network, augmentation, DeepDream, ensemble, algorithm

DeepDream algorithm for data augmentation in a neural network ensemble applied to multiclass image classification

Dmitrii Viaktin[0000-0002-3617-3101], Begonya Garcia-Zapirain [0000-0002-9356-1186], Amaia Mendez Zorrilla[0000-0002-0539-4753]

dmitrii.viatkin@opendeusto.es, mbgarciazapil@deusto.es
amaia.mendez@deusto.es

SIMPLIFIED TITLE

How to train a neural network to find unusual features in images and use it in the model ensemble.

ABSTRACT

This paper presents the application of Deep-Dream algorithm for data augmentation applied to images. This algorithm analyzes the image by a trained neural network and hides main features of the image. Importance and place of the features is estimated based on neural network layers output values. The new neural network trained on the processed data are forced to search and train a new set of features in the data, since the known features have been hidden. Trained on original and processed data neural networks are used in an ensemble. Experiments were conducted with a balanced images dataset with 5000 images in 10 classes. Experiment was conducted with a neural network based on InceptionV3 architecture in two variations: with non-pretrained weights and with pretrained weights. The neural network received a 256x256 pixel image as input. Training was conducted using categorical cross entropy loss function, accuracy metric, Adam optimizer with a learning rate of 0.0001. The improvements of the algorithm are almost insignificant when classifying an ensemble of neural network models for a small number of classes, but the impact of the algorithm increases as the number of classes increases. For binary classification there may be no improvement in ensemble accuracy, but when the number of classes increases and becomes more than 5, the influence of the algorithm on the accuracy of the final ensemble increases. The improvement in ensemble accuracy can be 0.5-4%, depending on the initial training conditions without the use of other types of data processing and augmentation.

I INTRODUCTION

Neural networks are trained to find and analyze sets of features in the data during training. Data augmentation allows making minor changes to the data and enable neural networks to find the most stable feature sets in the data and improve the values of the loss function and target metrics of the neural network [1]. Depending on the staring conditions, hyperparameters, and training data structure, each neural network finds a unique set of features for analyzed data. These feature sets for trained neural networks are similar for the same data and for the same task, but not absolutely repeated. This difference makes it possible to combine neural networks into ensembles to improve the values of target metrics and loss function [2].

The explored and developed Deep-Dream-based algorithm can process images and hide the image's main features based on the output layer values of trained neural networks. As it becomes difficult for the neural network to detect the main features on processed images, the neural network is trained to analyze combinations of secondary features that are ignored when training on the source unprocessed data. As a result, a neural network trained on the processed data usually has worse performance than the model trained on the original source data. However, combining a model trained on processed images and a model trained on source images into an ensemble improves the performance of the final ensemble more than just combining the two neural network models because neural networks are trained to detect different feature sets.

II STATE OF THE ART

If the feature sets detected by the neural networks are very similar, there will not be many benefits from combining them into an ensemble. Otherwise, if neural networks are trained for similar tasks, but detect different sets of features, their combining into an ensemble will significantly improve the final neural network model.

The improvement of loss functions and target metrics values for the ensemble of neural networks increase when the difference between the detected feature sets of pre-trained neural networks in this ensemble increases.

Classical image augmentation techniques such as geometric transformations, noise, and the use of generative neural networks expand the number of main detectable features in the data. But hidden combinations of secondary features can be missed in the training process because the main features have a much greater influence on the loss function and target metrics than the secondary features. The proposed Deep-Dream based algorithm can increase the diversity of detected features.

III ORIGINAL CONTRIBUTION

This paper describes an image augmentation algorithm based on the Deep-Dream algorithm and the possibility of using it to train neural networks to find unusual features in data. The possibility of combining neural networks trained using this algorithm into ensembles was also investigated.

IV METHODOLOGY

The research developed an algorithm and experimentally tested it on open data. In this subsection, the details about the developed algorithm and the neural networks used in this research are described. The order of application of the methods used is shown below (see Fig. 1).

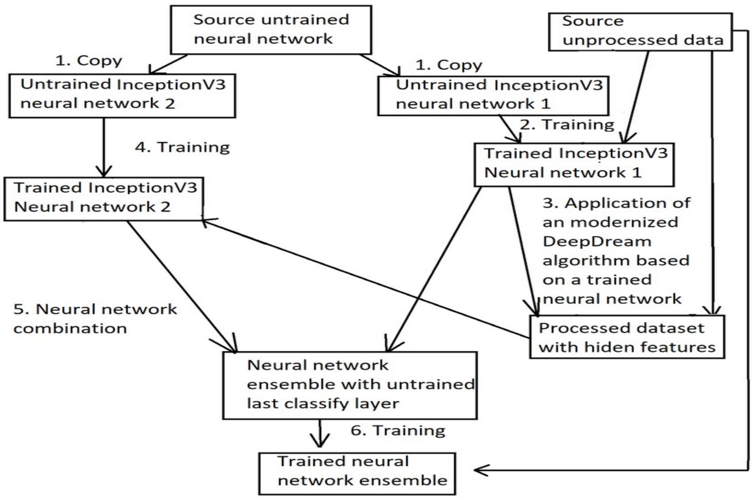

Source untrained neural network
Source unprocessed data
1. Copy
1. Copy
Untrained InceptionV3 neural network 2
Untrained InceptionV3 neural network 1
2. Training
4. Training
Trained InceptionV3 Neural network 1
Trained InceptionV3 Neural network 2
3. Application of an modernized DeepDream algorithm based on a trained neural network
5. Neural network combination
Processed dataset with hiden features
Neural network ensemble with untrained last classify layer
6. Training
Trained neural network ensemble

Fig. 1. Order of application of the methods.

V RESULTS

As a result of applying this algorithm in the training of the neural networks and combining these neural networks into an ensemble, the ensemble accuracy increases up to 0.5-1% for ensembles based on pre-trained neural networks and up to 5% for neural networks with random initialization of initial weights

VI EVALUATION

Experiments were conducted with a balanced image dataset with 5000 images in 10 classes. Experiment was conducted with a neural network based on InceptionV3 architecture in two variations. Training was conducted using categorical cross-entropy loss function, accuracy metric, and Adam optimizer with a learning rate of 0.0001. The improvements of the algorithm are almost insignificant when classifying an ensemble of neural network models for a small number of classes, but the impact of the algorithm increases as the number of classes increases. For binary classification, there may be no improvement in ensemble accuracy, but when the number of classes increases and becomes more than 5, the influence of the algorithm on the accuracy of the final ensemble increases.

VII CONCLUSIONS

The developed algorithm and methods of its application can be used to hide features in the data. Neural networks trained on such processed and original data, when combined into an ensemble, can increase the target metrics for the original data.

REFERENCES

[1] LUIS P., JASON W. *The Effectiveness of Data Augmentation in Image Classification using Deep Learning*. arXiv, 2017

[2] HUI L., XUESONG W., SHIFEI D. *Research and development of neural network ensembles: a survey*. Artificial Intelligence Review, 49(4), 455–479, 2017

DOI: https://doi.org/10.1007/978-3-031-21743-2_33
Keywords: keyphrase extraction, thematic vector

A Semantic-based Approach for Keyphrase Extraction from Vietnamese Documents using thematic vector

Linh Viet Le, Tho Thi Ngoc Le

levietlinh@gmail.com,ltn.tho@hutech.edu.vn

Simplified Title

Vietnamese keyphrase extracting with thematic vector.

Abstract

Keyphrase extraction plays an important role in many applications of Natural Language Processing. There are many effective proposals for English, but those approaches are not completely applicable for low resources languages such as Vietnamese. In this paper, we propose a Semantic-based Approach for Keyphrase Extraction (SAKE), which improved the TextRank algorithm [1]. In SAKE, we apply semantic to the phrases and incorporates the semantic to the ranking process. Technically, a document is represented as a graph, in which ver tices are words and edges are relations among words. In each document, we get a representative thematic vector by computing the average of word embedding vectors. Each vertex has a similarity score to the thematic vector and this score will be involved to the scoring in the ranking process. The important vertices are highly weighted not only by their relationships to other vertices but also by the similarity to the document theme. We experimented our proposed method on Vietnamese news articles. The result shows that our SAKE improved TextRank for Vietnamese text by achieving 1.8% higher of F1-score.

I Introduction

Adapt the unsupervised method to extract the keyphrase from Vietnamese news articles. Due to the nature of the language, existing approaches for English are not fully applicable to Vietnamese. There are some previous works for Vietnamese keyphrase extraction. However, most of them are supervised approaches or require ontology resource. Therefore, we motivate to adapt the general unsupervised approaches to extract keyphrases from Vietnamese text.

II State of the Art

Various approaches to keyphrase extraction for Vietnamese document have been explorered.

II.1 Supervised approach

Treat keyphrase extraction as classification problem. The features for training are words and theirs corresponding POS tags. The corpus for training is built by labelling each word as one of three classes I (In) – O(Out) – B(Begin) of a phrase. Then, classifying model such as SVM is applied to learn a classifier.

II.2 Base on ontology

Utilize ontology to extract keyphrases. Use the Vietnamese Wikipedia as an ontology and some specific characteristics of the Vietnamese language for the keyphrase selection step.

II.3 Base on deep learning technique

Use hybrid the superior of CNN and LSTM models to learn for keyphrases.

III Original Contribution

We propose a Semantic-based Approach for Keyphrase Extraction (SAKE) to extract phrases that ultilizes the thematic vector. This work is motivated from work by Key2Vec [2], proposed by Mahata et al. in 2018. Key2Vec is a method for training phrase embeddings which are used to represent for candidate keyphrases. However, this method is not immediately applicable to Vietnamese dataset due to the lack of Vietnamese phrase dataset to train model. Therefore, we are not able to implement this study for Vietnamese documents. Key2Vec motivates us on using semantic concept to extract phrases in Vietnamese news documents.

Our proposed SAKE algorithm includes four steps:

- Pre-process the text for cleaning, only get words that are potential candidates
- Choose words from first sentences to create average vector that represent the semantic context of the text.
- Build a graph from words in text and from top words ranking, combining the single words to phrases that appears in text.

  By using combination of single words to generate phrase, we can collect the potential phrases.
- Build a graph with vertices are single word and potential phrases that we generated, apply weighted graph-based algorithm to rank the top keys.

IV METHODOLOGY

Our SAKE solution includes four steps:

- Step 1 is preprocessing to tokenize words, get POS.
- Step 2 is creating an average vector from a collection of words in the first sentences.
- Step 3 is using graph-based method to select candidates that includes keyword and keyphrase.
- Step 4 is basing on word-embedding, we score each candidate vector with a group of keyword's average vector. The score is used as cosine similarity. Using graph-based algorithm's ranking for getting top T keyphrases, we ranked the candidate.

V RESULTS

V.1 Dataset

We used a dataset that was crawled from three online news websites (Thanh Nien, Phu Nu, Lao Dong), which contains 13,149 articles which we collected from May, 2019 to August, 2021.

V.2 Comparing to TextRank

For comparing with our algorithm, we reimplemented TextRank. The result shows ours result is 2.3% higher on the Recall and 1.8% higher on the F1-score.

VI EVALUATION

SAKE is semantic-based ranking algorithm that is inspired by the popular TextRank algorithm. This paper considers the domain of Vietnamese news articles. The relation of words/phrases to the theme of document is important, so that we use this relation to extract more valued keyphrases.

VII CONCLUSIONS

Our proposed approach, SAKE, has been evaluated on Vietnamese news articles. The result shows that, the keyphrases from SAKE are matched more words than the baseline solutions. It indicates that semantic vector can enrich the graph representation of document for ranking the keyphrases.

REFERENCES

1. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411. Association for Computational Linguistics (2004)
2. Mahata, D., Kuriakose, J., Shah, R.R., Zimmermann, R.: Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 634–639. Association for Computational Linguistics (2018). doi: 10.18653/v1/N18-2100

DOI: https://doi.org/10.1007/978-3-031-21743-2_45

Keywords: Question generation, BERT, Transformer, UIT-ViQuAD, UIT-ViNewsQA

Enhancing Vietnamese Question Generation with Reinforcement Learning

Nguyen Vu, Kiet Van Nguyen[0000-0002-8456-2742]

18520323@gm.uit.edu.vn,kietnv@uit.edu.vn

Simplified Title

Retriever-Reader Question Answering System for Vietnamese

Abstract

In this paper, we evaluate different powerful question generation systems in two benchmark Vietnamese datasets: UIT-ViNewsQA and UIT-ViQuAD. First, we conduct experiments on deep neural networks and sequence-to-sequence approaches based on a context and an answer to generate a question. In addition, in order to investigate several powerful approaches, we utilize two strong language models (LM): the monolingual language model PhoBERT and a massively multilingual pre-trained language model mT5. To obtain higher performance, we enhance LM-based methods with reinforcement learning during the decoding process. Our experiments show that the best model (PhoBERT with reinforcement learning) achieves the BLEU 4 scores of 19.77 on UIT-ViNewsQA and 20.43 on UIT-ViQuAD.

I Introduction

Based on Question Answering (QA) - predicting the answer by text and question has attracted the natural language processing community in the world, many datasets, and many approaches to the task of reading comprehension. However, Question Generation (QG) depends on the text, and the answer has not resulted in many methods and datasets. Especially with a language with few data sources like Vietnamese, there have been no published implementation methods and approaches to the QG until now. We approach and build the system for the task as the predecessors, creating questions based on the text. However, this paper explicitly constructs a baseline for the Vietnamese corpus, a language that is difficult to handle because of its semantic complexity. In this paper, we build models for Vietnamese question generation.

II Original Contribution

This paper evaluates robust question-generation systems in two benchmark Vietnamese datasets: UIT-ViNewsQA [3], and UIT-ViQuAD [2]. First, we conduct experiments on deep neural networks and sequence-to-sequence approaches based on a context and an answer to generate a question. In addition, to investigate several powerful approaches, we utilize two strong language models (LM): the monolingual language model PhoBERT and a massively multilingual pre-trained language model mT5. To obtain higher performance, we enhance LM-based methods with reinforcement learning during the decoding process. Our experiments show that the best model achieves the BLEU 4 scores of 19.77 on UIT-ViNewsQA and 20.43 on UIT-ViQuAD.

III Methodology

In this paper, question generation systems combining seq2seq, mT5, and PhoBERT with reinforcement learning are implemented for Vietnamese question generation models.

IV Results

On both the UIT-ViNewsQA dataset and the UIT-ViQuAD dataset, PhoBERT enhanced with Reinforcement Learning (RL) has the best performance for the Question Generation task using the sentence-level context. Besides, our systems have performances approximately each other. The RNN encoder-decoder architecture with the seq2seq model has the lowest performance. PhoBERT [1] with reinforcement learning achieves the best performance on BLEU 4 metric.

V Conclusions

In this study, we evaluate different powerful models in natural language processing that generate any question automatically from the given context (which is a sentence or a text) and the answer. In particular, three different models: seq2seq based on RNN encoder-decoder architecture, PhoBERT, and mT5 based on a pre-trained transformer model, were investigated. We especially proposed a transformer-based question generation approach enhanced with reinforcement learning for predicted questions closely to humans. Currently, our approach using the sentence-level context has a better performance than text-level context across the two benchmark datasets: UIT-ViQuAD and UIT-ViNewsQA, in terms of the BLEU 4 metric.

We would like to improve the performance of our proposed approach on both sentence level and text level. Moreover, with the appearance and growth of the transformer model, we want to use more and more pre-trained models to generate questions. We are also interested in further that the standard QA models can answer the automatically generated questions and the multi-task machine reading comprehension as automatically extracted answers, then generate questions on the context and answers, and then the machine can answer questions.

References

[1] Nguyen, D. Q., and Nguyen, A. T. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744* (2020).

[2] Nguyen, K., Nguyen, V., Nguyen, A., and Nguyen, N. A vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics* (2020), pp. 2595–2605.

[3] Van Nguyen, K., Van Huynh, T., Nguyen, D.-V., Nguyen, A. G.-T., and Nguyen, N. L.-T. New vietnamese corpus for machine reading comprehension of health news articles. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* (feb 2022). Just Accepted.

DOI: https://doi.org/10.1007/978-3-031-21967-2_53

Keywords: Digital signal processing, Parallel computing, Echo cancellation

Multi-pass, non-recursive acoustic echo cancellation sequential and parallel algorithms

Maciej Walczyński[0000-0002-5126-7851]

maciej.walczynski@pwr.edu.pl

Simplified Title

Parallel multi-pass acoustic echo cancellation algorithms.

Abstract

With few exceptions, the echo is a common phenomenon that occurs more or less during our talks. We then hear our own voice as a sound wave that bounced off the floor, walls, or other objects and returned to our ears. If the reflected sound comes back after a very short period of time, then it is perceived not as an echo, but as a kind of spectral distortion or reverberation, consisting of extending the sound. Depending on the situation, the occurrence of reverberation may or may not be desirable. When the time of arrival of reflections exceeds a certain value, it occurs as a separate sound that interferes with the reception of sound information. This paper will present methods of acoustic echo cancellation in telecommunication networks, in which parallel and distributed computing environments will be applied using multi-pass parallel algorithms based on non-recursive adaptive filtering.

I Introduction

The purpose of this work was to create a design, implement in a GPGPU environment, and test the performance and quality of multi-pass parallel adaptive filtering algorithms in the acoustic echo cancellation problem.

II State of the Art

For many decades, the main way to increase the computing power of computers was to increase the CPU clock frequency. However, for the past decade or so, we have seen a change in this approach in favor of using multicore processors. This approach necessitates the definition of a new class of algorithms that can take advantage of such multithreading. There are few items in the literature on parallel optimal filtering, especially its implementation on GPGPU platforms [1].

III Original Contribution

In the work presented here, the main focus is on two aspects. The first was to improve the performance of sequential algorithms of the LMS family by extracting those parts of the algorithms that are parallelizable and parallelizing them. The second aspect was to improve the quality of adaptive filtering by repeatedly filtering the original signal.

IV Methodology

TODO The paper presents the author's theoretical considerations on the possibility of parallelizing selected acoustic echo cancellation algorithms in their single-pass and multi-pass forms. A methodology for the parallelization of adaptive filtering algorithms is presented, with a particular emphasis on demonstrating those parts of the algorithms whose parallelization will bring the greatest benefits understood as a reduction in the execution time of the algorithm on a p-processors machine.

V Results

Initial performance tests of the proposed algorithms have shown their correct operation and increasing effectiveness in eliminating acoustic echoes in subsequent runs. One can see an increase in the value of the ERLE parameter for subsequent runs of the algorithm. Further studies are planned to verify the subjective perception of the effects of the described algorithms on a control group of listeners. Example results are presented in Figure 1. One can see an increase in the value of the ERLE parameter for subsequent runs of the algorithm. Further studies are planned to verify the subjective perception of the effects of the described algorithms on a control group of listeners.

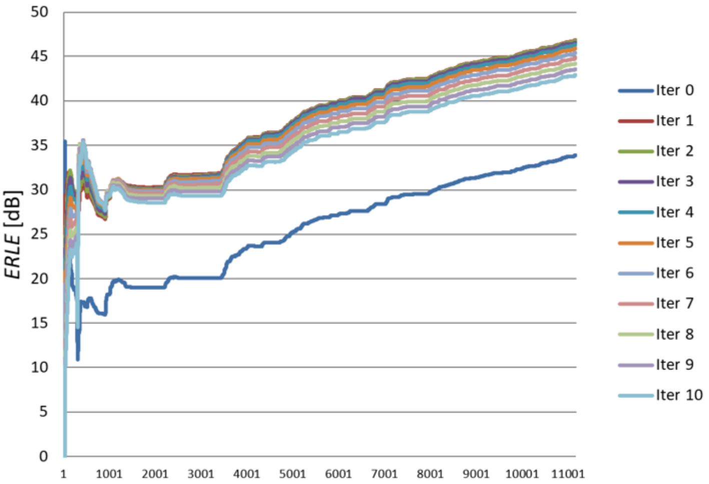


Figure 1: Acoustic echo reduction. Numbers on the abscissa axis indicate sample numbers.

VI Evaluation

For some contemporary architectures of parallel computing environments, such as *GPU*s, it may not be possible to eliminate acoustic echoes using real-time adaptive filters from the *LMS* family. FFor these types of architectures, which have a large number of relatively inefficient computing units, this paper proposes an acoustic echo cancellation model based on the *ParParPipelineLMS* and *ParParPipelineNLMS* algorithms. The *ParParPipelineLMS* calculation algorithm is a development of the *ParSeqPipelineLMS* concept. The characteristic feature of this algorithm is the pipelined execution of the *ParLMS* [2] algorithm on a group of $p_{ParLMS} = O(\frac{N}{\log n})$ processors, where N is the number of *ParLMS* filter coefficients. One iteration of such an algorithm, as shown in the proof of Theorem 1, can be executed in $O(\log N)$ time on a *CREW PRAM* machine using $O(\frac{N}{\log N})$ processors. Thus, for a machine having more than $O(\frac{N}{\log n})$ processors, it becomes possible to perform multiple runs of the *ParParPipelineLMS* algorithm. Assuming one has a $p = O(N)$-processor *PRAM* machine, it is possible to perform $P = \lfloor logN \rfloor$ runs of the algorithm.

VII Conclusions

The aim of this paper was to show that by using a multiprocessor computing system it is possible to obtain measurable, both qualitative and quantitative, gains of using a concurrent environment. This objective has been largely achieved. Methods of acoustic echo cancellation have been proposed that are faster than the currently known methods through implementation in a concurrent computing environment. Also, assuming a fixed computation time and using concurrent environments, it is possible to obtain better quality results than sequential algorithms by applying sequential, sequential-parallel and parallel multi-pass adaptive filtering algorithms. These methods are original procedures proposed by the author. It was also possible to determine the theoretical computational complexity of the key elements of the echo cancellation algorithms in such a way that the whole computational process is cost-optimal, having a fixed number of processors p by means of a theoretical model of a parallel computer.

References

[1] Bożejko, W., Dobrucki, A., and Walczyński, M. Parallelizing of digital signal processing with using gpu. In *Signal Processing Algorithms, Architectures, Arrangements, and Applications SPA 2010* (2010), pp. 29–33.

[2] Dobrucki, A., Walczynski, M., and Bozejko, W. Family of parallel lms-based adaptive algorithms of echo cancellation. *Computational Methods in Science and Technology 21*, 4 (2015), 191–200.

DOI: `https://doi.org/10.1007/978-3-031-21743-2_27`

Keywords: Intrusion Detection, Internet of Things, Distributed Ledger, Cybersecurity, Blockchain.

Collaborative Intrusion Detection System for Internet of Things using Distributed Ledger Technology: A Survey on Challenges and Opportunities

Aulia Arif Wardana[1][0000-0003-2201-0464], Grzegorz Kołaczek[1][0000-0001-7125-0988], Parman Sukarno[2][0000-0002-2565-3580]

aulia.wardana@pwr.edu.pl, auliawardan@telkomuniversity.ac.id

Simplified Title

CIDS for IoT using DLT: A Survey on Challenges and Opportunities.

Abstract

This review presents the current state-of-the-art of the Distributed Ledger Technology (DLT) model used in the Collaborative Intrusion Detection System (CIDS) for anomaly detection in Internet of Things (IoT) network. The distributed IoT ecosystem has many cybersecurity problems related to anomalous activities on the network. CIDS technology is usually applied to detect anomalous activities on the IoT network. CIDS is suitable for IoT network because they have the same distributed characteristic. The use of DLT technology is expected to be able to help the IDS system accelerate detection and increase the accuracy of detection through a collaborative detection mechanism. This review will look more deeply at the placement strategies, detection method, security threat, and validation & testing method from CIDS with DLT-based for IoT network. This review also discusses the open issue and the lesson learned at the end of the review. The result is expected to produce the next research topic and help professionals design effective CIDS based on DLT for the IoT network.

I Introduction

The IoT is made up of a large-scale network with heterogeneous characteristics. The vast amount of data that is sent over the IoT network is vulnerable to forgery and data manipulation, causing the integrity and confidentiality of the data to be disrupted. IDS on a distributed network is not enough to work independently for detecting coordinated attacks but must also be collaborative to accelerate detection in distributed systems. Therefore, CIDS is a suitable system to detect anomalies on the distributed system. There are many types and approaches to CIDS in the network; one of them is using DLT for built CIDS system to detect anomalous activities on the network. This review aims to discuss open issues and future research related to the use of DLT for CIDS in the IoT network.

II State of the Art

Research [1] aims to support the researcher by giving insight into the overview of the integration of blockchain and CIDS to detect anomalies in IoT networks. The review [2] is a survey of IDS in IoT. One of the IDS models discussed in this survey is CIDS which applied multi-IDS agents for IoT networks. The other review from [3] explains about security of IPv6 Routing Protocol for Low Power and Lossy Network (RPL). One strategy to secure the IPv6 RPL-based IoT network is using CIDS. Based on all the previous reviews, this research motivates us to review about CIDS on IoT network. The usage of DLT also enables integration with CIDS for anomaly detection in IoT network.

III Original Contribution

The contribution from this literature survey is summarized as follow:

- This research will review the integration DLT technology with CIDS in IoT networks for storage sharing, trust & audit management, and information exchange.
- This research is also review about the usage of DLT for collaborative learning in IDS to accelerate detection on IoT network.
- This research will review all two points before using the point of view of the placement strategies, detection method, security threat, and validation & testing method.

IV Collaborative Intrusion Detection Systems based on Distributed Ledger Technology in Internet of Things

Fig. 1 illustrates the taxonomy of CIDS in IoT. Based on the taxonomy diagram, this review creates a short explanation in this section to summarize all studies related to CIDS DLT-based IoT network.

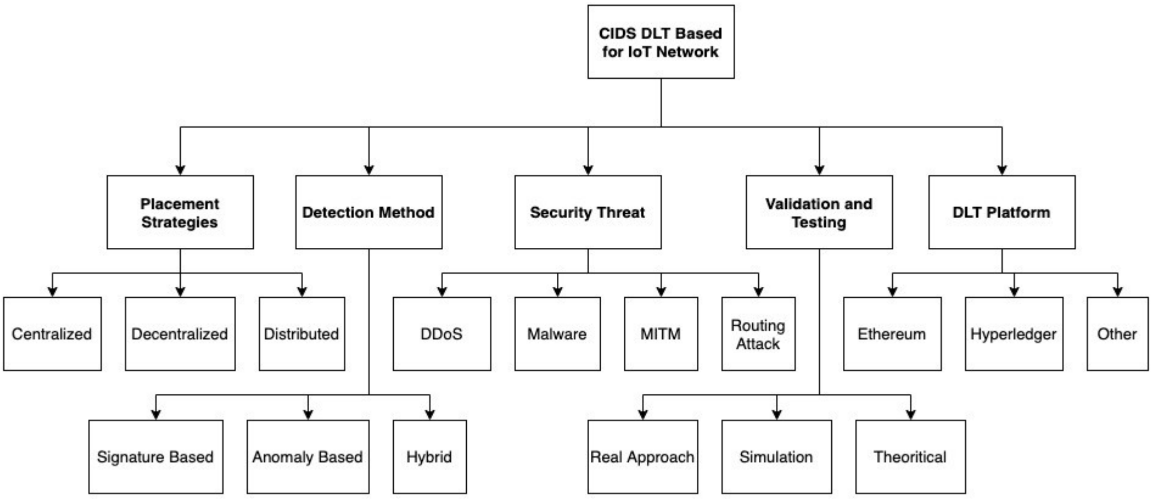


Fig. 1**.** Taxonomy from CIDS DLT-based in IoT.

CIDS has three types of CIDS architecture. It is centralized, decentralized, and distributed. There are three types of detection methods in IDS, signature, anomaly, and hybrid method. CIDS is an excellent IDS system for dealing with coordinated attack types. There are several types of coordinated attacks that often attack, such as DDoS, worm outbreak, and large-scale stealthy scans. The validation method in this review consists of three types: real approach, simulation, and theoretical. Ethereum and Hyperledger are examples of open source DLT platforms that are easy to use.

V FUTURE AND OPPORTUNITIES

V.1 Challenge in Research

- **Resource and Cost**: IoT has limited resource and energy in its devices, but DLT concept needs more computational power and energy. This condition requires the implementation of CIDS DLT-based model in IoT need to consider resource and cost.
- **Insider Attack**: CIDS is used to detect outsider attacks, but insider attack needs to be considered too.
- **Complexity and Performance**: IoT is one of the distributed systems that integrate heterogeneous technology (software and hardware) into their ecosystem. The implementation of CIDS DLT-based in IoT must be simple and effective with good performance.

V.2 Future Research and Opportunities

Decentralized learning is one of the future research agendas for collaboration between AI and DLT. The development of decentralized learning will create three research opportunities. Firstly, the good design of the consensus algorithm in DLT will affect machine learning model training. Secondly, a smart contract is programable, so simple AI computation in smart contract is possible. Lastly, computation in smart contract is a heavy task while doing on-chain.

VI CONCLUSIONS

The combination of DLT with CIDS to detect anomalous activities in the IoT network will improve the CIDS detection process from the CIDS. This paper presented an overview of some research in that field. This literature review paper can help researchers and industries to see the challenge and future direction for the adoption of CIDS with DLT system to detect anomalies in the IoT network. There is some consideration about the adoption of DLT in CIDS, like signature & data sharing, collaboration alert & reporting, adaptive architecture for large system, collaborative real-time monitoring, and deployment simplification.

REFERENCES

[1] H. Benaddi and K. Ibrahimi, “A Review: Collaborative Intrusion Detection for IoT integrating the Blockchain technologies,” *Proc. - 2020 Int. Conf. Wirel. Networks Mob. Commun. WINCOM 2020*, 2020, doi: 10.1109/WINCOM50532.2020.9272464.

[2] B. B. Zarpelão, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, “A survey of intrusion detection in Internet of Things,” *J. Netw. Comput. Appl.*, vol. 84, no. February, pp. 25–37, 2017, doi: 10.1016/j.jnca.2017.02.009.

[3] A. Verma and V. Ranga, “Security of RPL Based 6LoWPAN Networks in the Internet of Things: A Review,” *IEEE Sens. J.*, vol. 20, no. 11, pp. 5666–5690, 2020, doi: 10.1109/JSEN.2020.2973677.

DOI: 10.1007/978-3-031-21967-2_41
Keywords: anomaly detection, unsupervised machine learning, algorithm quality assessment, data dimensionality

Excess-Mass and Mass-Volume quality measures susceptibility to intrusion detection system's data dimensionality

Arkadiusz Warzynski[0000-0002-0452-7518], Łukasz Falas[0000-0001-5547-2070], Patryk Schauer[0000-0003-0912-4010]

arkadiusz.warzynski@pwr.edu.pl,lukasz.falas@pwr.edu.pl,patryk.schauer@pwr.edu.pl

Simplified Title

Verification of unsupervised machine learning methods quality in intrusion detection systems

Abstract

In spite of ever-increasing volume of network traffic, unsupervised intrusion detection methods are one of most widely researched solutions in the field of network security. One of the key challenges related to development of such solutions is the proper assessment of methods utilized in the process of anomaly detection. Real life cases show that in many situations labeled network data is not available, which effectively excludes possibility to utilized standard criteria for evaluation of anomaly detection algorithms like Receiver Operating Characteristic or Precision-Recall curves. In this paper, an alternative criterion based on Excess-Mass and Mass-Volume curves are analyzed, which can enable anomaly detection algorithms quality assessments without need for labeled datasets. This paper focuses on the assessment of effectiveness of Excess-Mass and Mass-Volume curves-based criteria in relation to intrusion detection system's data dimensionality. The article discusses these criteria and presents the intrusion detection algorithms and datasets that will be utilized in the analysis of data dimensionality influence on their effectiveness. This discussion is followed by experimental verification of these criteria on various real-life datasets differing in dimensionality and statistical analysis of the results indicating relation between effectiveness of analyzed criteria and dimensionality of data processed in intrusion detection systems.

I Introduction

Numerous anomaly detection methods proposed in various papers utilize labeled datasets and standard evaluation criteria, e.g., criteria using Receiver Operating Characteristic (ROC) or Precision-Recall (PR) curves. However, real-life cases show that in most distributed network systems, labeled datasets are not available due to a lack of funds or simply lack of capabilities to label even a small part of a large network traffic dataset. This article focuses on verifying the applicability of assessment criteria of anomaly detection methods based on Excess-Mass and Mass-Volume curves in systems where labeled data is not available.

II State of the Art

The authors of [1], motivated by a wide range of applications, identified how to rank observations in the same order as induced by the density function, which they called anomaly scoring. For the discussed problem, they proposed a performance criterion in the form of the MV curve. [2] shows that the Mass Volume curve is a natural criterion to evaluate the accuracy of decision rules regarding anomaly scoring. In response to this work, in the article [4], an alternative approach to the anomaly scoring problem based on Excess-Mass Curves was presented. Another work in which criteria based on the Excess-Mass (EM) and Mass-Volume (MV) curves were proposed in [3]. The method is based on feature sub-sampling and aggregating and compares results to ROC and PR criteria.

III Original Contribution

This paper discussed research aimed at verifying and indicating if alternative criteria for evaluating anomaly detection methods based on Excess-Mass and Mass-Volume curves can be effectively utilized for low-dimension data and if they are susceptible to data dimensionality. Conducted research focused on comparing standard ROC and PR curves with EM and MV based curves to assess if EM and MV based metrics will indicate the same algorithms best and worst algorithms in the field of network traffic anomaly detection.

IV Methodology

To examine the effect of feature selection algorithms on the applicability of EM and MV measures compared to ROC and PR algorithms, we decided to repeat the experiments on the same datasets using identical anomaly detection algorithms to determine whether, with a limited number of dimensions, all validation measures point to the same best anomaly detection algorithm. Tests were carried out for LOF, SVM, Isolation Forrest, ABOD algorithms, and an example implementation of an autoencoder and compared to the original study [5]. Popular benchmark datasets for intrusion detection systems were used for the tests: NLS-KDD, UGR'16 and UNSWNB15. We decided to use three different algorithms to avoid the influence of the choice of a specific feature selection method. With each feature selection method, we would select a subset of 3, 5 and 7 features for each dataset.

V Results

Despite testing several feature selection methods and limiting themselves to a small number of dimensions, the results obtained do not allow us to conclude that for the selected algorithms, it is possible to perform effective validation using methods that do not require labels. Only in one case, all validation methods pointed to the same anomaly detection algorithm as the best one. At the same time, it turned out that in most cases, reducing the number of dimensions did not cause the ROC and PR methods to significantly deviate from the values obtained using all the features available in the datasets. It is, therefore, possible to maintain comparable classification quality even when reducing the number of dimensions, but this does not increase the reliability of using EM and MV curves to assess classification performance.

VI Evaluation

The evaluation of the results consisted of collecting the results from each evaluation method tested using each algorithm and each dataset after applying different feature selection methods. In the matrix of results, we marked which classification algorithm obtained the best and worst result for each dataset according to each result validation method. The experimentation results prove that the number of dimensions, in this case, is not crucial, as it did not positively influence the comparison of results with validation methods using labels. The problem presented here requires additional research to determine whether the data normalization method used in the pre-processing step might not play a key role in the results obtained by the EM and MV algorithms. It may well be that by normalizing the data using the methods presented in the original research, it is impossible to process every dataset related to the anomaly detection problem, which would lead to the need to adjust pre-processing methods and would be a considerable limitation for practical application.

VII Conclusions

The results of the experimentation have shown that the reduction of dataset dimensionality did not result in the expected increase of quality of EM and MV based metrics indication of the best algorithms. This leads to the conclusion that while these metrics can be susceptible to data dimensionality, additional data preprocessing, which was not stated in the original article, is probably required in order to enable the effective utilization of such measures in the development of unsupervised anomaly detection methods for intrusion detection systems.

References

[1] Clemenc´ ¸on, S., and Jakubowicz, J. Scoring anomalies: a m-estimation formulation. In *Artificial Intelligence and Statistics* (2013), PMLR, pp. 659–667.

[2] Clemenc´ ¸on, S., and Thomas, A. Mass volume curves and anomaly ranking. *Electronic Journal of Statistics 12,* 2 (2018), 2806–2872.

[3] Goix, N. How to evaluate the quality of unsupervised anomaly detection algorithms? *arXiv preprint arXiv:1607.01152* (2016).

[4] Goix, N., Sabourin, A., and Clemenc´ ¸on, S. On anomaly ranking and excess-mass curves. In *Artificial Intelligence and Statistics* (2015), PMLR, pp. 287–295.

[5] Warzynski, A., Falas, Ł., and Schauer, P. Excess-mass and mass-volume anomaly detection algorithms applicability in unsupervised intrusion detection systems. In *2021 IEEE 30th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* (2021), IEEE, pp. 131– 136.

DOI: `https://doi.org/10.1007/978-981-19-8234-7_54`

Keywords: Agent based modeling, Teaching model, Group modeling, Multi-agent simulation

Agent Based Model of Elementary School Group Learning

Barbara Wedrychowicz, Marcin Maleszka[0000-0001-6989-2906], Nguyen Van Sinh

barbara.wedrychowicz@pwr.edu.pl,marcin.maleszka@pwr.edu.pl,nvsinh@hcmiu.edu.vn

Simplified Title

Agent Based Model of Elementary School Group Learning

Abstract

The paper describes an agent model built to simulate the behaviour of primary school students during group work. The model is focused on predicting the change of student knowledge due to cooperation, compared to individual learning. We conducted extensive observation studies of real students and prepared a basic model based on that data. We compared this real data model with data gathered during simulation. Some preliminary findings on how to improve learning during groupwork are currently being tested with the same group of real students.

I Introduction

Working in groups is a common tool in education. Due to the lack of research on the topic, groups are often selected at random, or teachers allow students to create groups on their own. This motivated us to develop a model to simulate these groups and optimize their composition. We verify the thesis that on the basis of theoretical modelling and observation of real groups, it is possible to develop a model of students working in a group, complete with emergent behaviours that sometimes occur during their cooperation.

II State of the Art

The approaches to describe elementary student learning are developed using informal or descriptive methods, mostly from the perspective of education sciences. One of the classical approaches states that the more active and the more participating the student is, the more effective is his learning. Other works discuss that interaction between students leads to higher grades, faster memorization, higher satisfaction, and higher chances of continuing education in the next tier. The influence of working in group is considered as more of a sociology related problem. Some researchers point out that the full potential of the group can only be achieved when taking into account positive interactions, personal and group responsibility, a will and an ability to cooperate, and the openness of students. Different levels of these attributes in different students lead to different approaches to groupwork, and selecting students by these parameters could be a basis for group formation.

III Original Contribution

Our research was intended to bridge the gap between the formal models of computer science and the research done in education and social science areas. The developed multiagent model of students learning while working in groups contains multiple formalizations of concepts that were previously described without mathematical apparatus. Additionally, with this approach, the model was further evaluated in simulated environment, as described below. This allows additional step for experiments, when direct application in learning environment may be discruptive.

IV Methodology

We developed a multiagent model of students during groupwork based on the observations and analysis of the real group. Students were observed during their activities in normal ICT class activities. The students were described mostly by subjective attributes: age, knowledge before groupwork, knowledge after groupwork, communicativeness, leader potential, interest in the topic, and ease of learning.

During some classes student attributes were estimated, and during others they worked in groups. While groupwork students were doing task intended for increasing their knowledge or for repetition (better remembering) of said knowledge.

The observed attributes of students were analysed in terms of their relation to initial knowledge. We used multiple regression analysis.It has shown a strong correlation between the ease of learning and interest in the topic with the initial knowledge. It was also checked that the attributes are not related, so they were all used in the later parts of the analysis.

We determined a simple formula for individual learning that takes into account not only individuals attributes, but also parameters of the group. The most important ones, based on literature and our analysis, are communicativeness, leader potential, interest in specific topic, and general ease of learning.

V EVALUATION

The GAMA agent environment was used to implement the multiagent model. It provided not only numerical results, but also a visual representation of the simulations. We compared the agent simulation to real observations of student learning. Specifically, we compared the sets of differences between initial and final knowledge in the observations, and the set of differences between initial and learned knowledge in the agent model, for different sizes of groups. The simulation is run and observed by discrete simulations, each representing one minute of class time.

VI RESULTS

Based on the observations and tests of the obtained results, even working with a small number of student attributes gives a very good representation of the real group. On this basis, it is possible to simulate a group of students, their characteristics, and their groupwork, and then select appropriate methods to improve the learning process.

The base experiment was done in simple two-person groups. Agents are initialized with random parameters and not grouped. Individual parameters are randomized based on distribution observed in real groups. Simulation follows the standard lesson plan of an elementary school, that is students grouping first, then working in pairs. Observations of state are conducted each minute, with more details gathered at the beginning and end of simulation. In observations, we have determined that after theoretical introduction the groupwork starts between 15th and 20th minute of class. Previous to that, the students are creating groups, or discussing cooperation. In simulation, the grouping occurs between 10th and 20th iteration (minute). Following that, the agents learn on their own and from others, according to proposed model. As with real groups, the knowledge increase is not linear but increases over time. Knowledge level at the end of the simulation (end of class) varies, but very rarely there is none, which also fits the observational data.

Similar simulations have been conducted for larger (3 or 4 agents) groups. The groups are formed between 17th and 24th iteration (17th and 24th minute of class) and at the end of simulation the knowledge level varies. The difference is that the learning is much more varied: most agents attain a high level of knowledge, but some learn almost nothing. Agents with low interest and less communicative learn much slower in larger groups, as with real groups of students.

We compared the agent simulation to real observations of student learning. Specifically, we compared the sets of differences between initial and final knowledge in the observations, and the set of differences between initial and learned knowledge after 45 iterations in the agent model, for different sizes of groups. For initial knowledge the standard error is 0.99 and for the final knowledge it is 1.07. The agent model does not increase the difference significantly.

VII CONCLUSIONS

The aim of the overall research is to improve student learning by changing parameters that depend on the teacher. While the student's interest in the topic is independent of the teacher, the group size or method of division could be modified to speed up knowledge acquisition. Based on the simulated model we have determined that similar communicativeness is more conductive to learning that similarities or differences in knowledge level and that two-person groups are most effective.

The results from the simulation are currently being tested in a real classroom environment, while we work to improve the model. Followup results show that the proposed model needs to be extended further, as more student attributes have been observed in real classrooms, as compared to those proposed in literature.

DOI: https://doi.org/10.1007/978-981-19-8234-7_55

Keywords: cataract surgery videos, deep learning, video summarisation, surgeon training

Clinically-relevant summarisation of cataract surgery videos using deep learning

Jesse Whitten, James McKelvie and Michael Mayo

mmayo@waikato.ac.nz

Simplified Title

Clinically-relevant summarisation of cataract surgery videos using deep learning

Abstract

Cataract surgery is one of the most frequently performed medical procedures worldwide, an estimated 20 million such surgeries occurring annually. However, the training required to become a competent cataract surgeon takes years due to its challenging technical nature. This limits the supply of capable surgeons. One aspect of modern cataract surgery is that video recordings are routinely taken using microscope cameras, and these recordings can be used to review errors and improve technique throughout surgical training. However, reviewing raw surgery video footage is tedious and may not lead to actionable insights improving surgeon performance. To tackle this issue, a novel artificial intelligence (AI)-based framework for the extraction of detailed surgery video summary statistics directly from the raw surgery footage is proposed. The input to the system is a video of a cataract surgery procedures and the output is a summary report. The approach uses deep learning models (ResNet-50, ResNet-152 and InceptionV3 were tested) to identify and time surgical instrument activity. Additionally, a unique dataset consisting of 57,422 hand-labelled frames extracted from a new locally-sourced video dataset of 29 retrospective cataract surgery recordings was created. Testing these predictive models with 4-fold cross validation across ten different surgical instruments resulted in a best mean testing prediction area under the ROC curve of 97.6%, and a mean testing sensitivity of 96.6%. Given these high levels of accuracy, the reports generated by our system are high quality and could be used to provide actionable insight into surgical technique during surgical training.

I Introduction

This paper proposes a novel deep learning-based framework for the summarisation of cataract surgery videos, and additionally provides an analysis of a new dataset of cataract surgery videos. Our focus is detecting and summarising the duration and patterns of surgical instrument use, along with periods of idle time in which no surgical tools are in use. In our proposal, recognition and tracking instruments during surgery is performed using deep learning models; surgery summaries are then presented after labelling the video footage in the form of visual instrument sequence charts and detailed tabular summaries of instrument usage statistics. Such a process compresses a typically long and unedited video into a concise, representative clinical analysis report that would take no more than a couple of pages, making it useful for cataract surgeon training and other types of quantitative analysis. The primary time saving benefit for surgeons and trainees is that viewing the entirety of the surgery videos is not needed.

II State of the Art

In general, there are two broad applications in the literature for the analysis of cataract surgical videos: phase identification and surgical tool identification. A phase refers to a particular stage in the surgery. Phase detection from a single frame or even a sequence of consecutive frames is difficult because the same surgical tools could be used in different phases; accurate phase detection therefore requires a holistic analysis of the entire surgery. Surgical tool identification, on the other hand, is used to accurately determine which tool or tools are in use in each localised portion of the video, a problem more related to traditional object detection. Surgical tool detection can either be viewed as a standalone problem by itself, or as an additional input to a phase detection process. Phase detection, however, does not necessarily require surgical tool recognition and has been performed without it. Prior to the deep learning revolution, a handful of research works starting circa 2011 considered traditional computer vision handcrafted features and statistical approaches for surgical video analysis. However, most modern approaches are deep learning-based and a variety of different approaches are reviewed in the paper. Several general critiques can be levelled at the field: firstly, there are no standardised benchmarks, and most authors use a different set of evaluation data and/or metrics, making their results incomparable; and additionally, there is no clear agreement on what the phases and tools to be recognised actually are. Therefore it is not clear that high-accuracy methods from the literature will perform well in practice in a local setting.

III Original Contribution

The main novelties of this paper are twofold: firstly, techniques similar to those proposed in the literature were a applied to a local dataset of cataract video surgeries with some interesting insights and results as a consequence; secondly we extended the pipeline beyond what most other works in the literature do by assembling the predictions across an entire report into a single page summary report that was deemed clinically relevant.

IV Methodology

We first of all assembled a dataset of 29 cataract surgery videos. Frames were sampled at a rate of 6FPS and labelled with surgical tool. Image preprocessing techniques such as Hough circle transform was then used to localise and segment the pupil area, and standardize with respect to level of zoom and image size. The labelled data was then divided on a per-video basis into four mutually exclusive subsets so that four-fold cross validation experiments could be performed for model selection. A total of twelve different deep neural network model configurations were then defined, each configuration being used to train and test a model four times. Results were averaged over the four folds of the cross validation to ensure robustness. Test predictions from the most accurate model during testing was then used to generate summaries of the surgery for subjective evaluation.

V Results

The best deep neural network model was found to be highly accurate at surgical tool recognition, and the reports were evaluated (by an expert cataract surgeon) and found to have a high degree of clinical significance.

VI Conclusions

The long term aim of this research project is to construct an AI-based system for cataract surgery video analysis. The primary purpose is to improve the efficiency of cataract surgeries by providing actionable insight and feedback for both educational purposes and surgeon self-improvement. Summarisation of the video into a single page report can also save significant surgeon time when reviewing previously recorded surgeries.

DOI: https://doi.org/10.1007/978-981-19-8234-7_57

Keywords: Candidate Screening, Asymmetric Error Costs

Cost-Oriented Candidate Screening Using Machine Learning Algorithms

Shachar Wild 1[0000-0000-0000-0000], Mark Last 2[0000-0003-0748-7918]

{wildsha,mlast}@post.bgu.ac.il

Simplified Title

Saving the Costs of a Candidate Screening Process

Abstract

Choosing the right candidates for any kind of position, whether it is for academic studies or for a professional job, is not an easy task, since each candidate has multiple traits, which may impact her or his success probability in a different way. Furthermore, admitting inappropriate candidates and leaving out the right ones may incur significant costs to the screening organization. Therefore, such a candidate selection process requires a lot of time and resources. In this paper, we treat this task as a cost optimization problem and use machine learning methods to predict the most cost-effective number of candidates to admit, given a ranked list of all candidates and a cost function. This is a general problem, which applies to various domains, such as: job candidate screening, student admission, document retrieval, and diagnostic testing. We conduct comprehensive experiments on two real-world case studies that demonstrate the effectiveness of the proposed method in finding the optimal number of admitted candidates.

I Introduction

In various screening processes (e.g., university admission, web search, diagnostic testing), the problem of determining how many of top-ranking candidates to admit remains open, while the cost of rejecting a successful candidate may be much higher than the cost of admitting a failing one. As such, the main purpose of this research is to identify an optimal entry bar, which given a candidates ranked list and a cost function, will retrieve a sub-list of admitted candidates resulting in a minimal overall cost. The proposed solution incorporates both false rejection and false admission costs, machine learning algorithms, and it can be applied to a wide range of candidate screening problems.

II State of the Art

Most previous studies focused on the information retrieval domain where they addressed the problem of truncating a ranked list of web search results while optimizing a given search accuracy measure. In addition to being computationally expensive, none of these studies considered the problem of predicting the optimal truncation point with respect to the costs of missing a relevant document (analogous to rejection of a good candidate) and retrieving an irrelevant document (analogous to admission of a failing candidate).

III Original Contribution

In contrast to the previous studies, we treat the general task of predicting the optimal truncation point in a ranked list of candidates as a cost optimization problem considering both false rejection and false admission costs incurred by the screening decisions. The proposed solution is demonstrated on two case studies: selecting candidates for COVID-19 diagnostic testing and selecting credit card transactions for manual fraud screening.

IV Methodology

This is an experimental work, which includes two case studies. We propose an algorithm which, given a ranked list and the ratio between false rejection and false admission costs, will predict the optimal entry bar. This problem can be seen as a regression task, since the value of the predicted truncation point k is numeric (position number in a ranked list). Hence, in the training step, we fit a regression model which, given the ranking score of each candidate in a given list, predicts the optimal truncation threshold k. The model is trained on candidate groups with known outcomes (success or failure) of each candidate.

V Results

Our prediction models achieve a significant improvement over the baseline fixed-threshold approach for all examined costs ratios and ranking algorithms. Furthermore, having a higher false rejection cost, compared to the false admission cost, results in a higher optimal truncation cutoff point but at a lower optimal cost. In addition to that, across the examined time period, the optimal and real truncation points share a relatively similar trend and the predicted truncation points result in lower costs, compared to the costs incurred by the actual screening policy.

VI Evaluation

Our evaluation experiments are conducted to answer the following research questions: (1) Can we identify group-specific truncation thresholds, which result in a lower total cost than a fixed cut-off point baseline? (2) Will the predicted truncation points result in significantly shorter candidate lists than the initial number of candidates?

We based our experiments on two publicly available datasets: 'COVID-19 Testing Results in Israel' [1] and 'Credit Card Fraud Detection' [2]. For the former, a 'group' can be seen as the number of daily tests and 'candidates' can be seen as people who are susceptible to the disease. The truncation threshold represents the number of people who will be recommended for testing. For the latter, candidates can be seen as credit card transactions. The truncation threshold represents the number of most suspicious transactions, which require manual screening.

VII Conclusions

In this study, we have shown that our generic approach, which treats candidate screening as an error cost minimization problem, can be applied to various categories of candidate screening processes. In our evaluation experiments on real-world data, we have achieved nearly optimal error-cost performance in two case studies and using various prediction models. The performed sensitivity analysis showed a strong and stable performance across a range of cost ratio values and candidate ranking algorithms. Thus, we can deduce that the proposed methodology can significantly improve the cost-efficiency of many real-world candidate screening tasks.

References

[1] Israel COVID-19 Data Tracker - tested individuals. https://data.gov.il/dataset/covid-19/resource/d337959a-020a-4ed3-84f7-fca182292308, 2021. [Online; accessed 21-July-2021].

[2] The world's largest data science community with powerful tools and resources. https://www.kaggle.com, 2021. [Online; accessed 19-July-2021].

DOI: https://doi.org/10.1007/978-3-031-21967-2_6

Keywords: Lung cancer, Survival analysis, Metastasis, Radiomics features

Potential of radiomics features for predicting time to metastasis in NSCLC

Agata Wilk[0000-0001-7554-1803], Damian Borys[0000-0003-0229-2601], Krzysztof Fujarewicz[0000-0002-1837-6466], Andrea d'Amico[0000-0003-4632-2139], Rafał Suwiński[0000-0002-3895-7938], Andrzej Świerniak[0000-0002-5698-5721]

agata.wilk@polsl.pl

SIMPLIFIED TITLE

Predicting time to metastasis in lung cancer based on radiomics

ABSTRACT

Lung cancer is the most deadly malignancy, with distant metastasis being a major negative prognostic factor. Recently, interest is growing in imaging data as a source of predictors due to the low invasiveness of their acquisition. Using a cohort of 131 patients and a total of 356 ROIs we built a Cox regression model which predicts metastasis and time to its occurrence based on radiomic features extracted from PET/CT images. We employed several variable selection methods, including filtering based on correlation, univariate analysis, recursive elimination and LASSO, and obtained a C-index of 0.7 for the best model. This result shows that radiomic features have great potential as predictors of metastatic relapse, knowledge of which could aid clinicians in planning treatment.

I INTRODUCTION

Lung cancer is the most common cause of cancer-related death worldwide. One of the reasons is its high metastatic potential, with secondary tumours appearing primarily in brain, liver and bones. Occurrence of distant metastasis determines treatment options and is a major negative prognostic factor. The aim of this research was to predict time to metastasis basing on PET/CT images, which are relatively non-invasive and routinely used in clinical practice.

II STATE OF THE ART

Medical images provide valuable knowledge about tumor location, size and shape which can then be used for clinical decision-making. However, these high resolution images contain a much greater amount of information, most of it undetectable by human eye. A solution to this problem is radiomics - extraction of numerical features from the images. To date, radiomic features have been successfully used in medical scenarios for cancer diagnostics (benign/malignant tumors or cancer subtyping), staging and prediction of survival or recurrence. The issue of metastasis has also been addressed, mostly for local (lymph node) spread, but also detection of distant metastases [1].

III ORIGINAL CONTRIBUTION

In the present study, we used radiomic features to not only predict the occurrence of metastasis, but also time to its onset.

IV METHODOLOGY

For a cohort of 131 patients with non small cell lung cancer, we extracted 105 radiomic features for between 1 and 11 regions of interest (ROI). We normalized the data, and used unsupervised analysis (correlation, clustering and principal component analysis) to explore the relationships between samples and features. We modelled metastasis free survival using Cox regression, with several feature selection methods applied: pre-filtering based on univariate analysis, pre-filtering based on correlations between variables, recursive elimination, and LASSO. Due to high intra-patient heterogeneity, we repeated the analysis taking into account only the largest ROI for each patient.

V EVALUATION

We assessed performance of the models using C-index and time-dependent ROC curve (for time = 900 days) for resubstitution on the training set and visualized the results on Kaplan-Meier plots.

VI Results

Hierarchical clustering did not result in emergence of distinct groups corresponding to tumour stage or subtype, which suggests that radiomic features carry information not contained in commonly used clinical features. While the best overall performance was achieved by recursive elimination selection with pre-filtering based on univariate analysis, the LASSO selection yielded a model only slightly worse with considerably fewer predictors. The best model achieved a C-index of 0.7 using seven features, three of which showed statistical significance: Maximum2DiameterColumn, Minimum and RunLengthNonUniformityNormalized.

For the dataset containing only one ROI per patient, dimensionality reduction proved more challenging. Models which were best in terms of quality metric had a large number of features, which presents a risk of overfitting. A selection process with multiple pre-filtering steps resulted in a model with similar C-index and only eight features.

VII Conclusions

Our results confirm the potential of radiomics for predicting time to metastasis in lung cancer. Such knowledge can be used by clinicians for adjusting therapy intensity and introducing systemic treatment to prevent the spread of the cancer. Although the obtained C-index is relatively high, and the simulated Kaplan-Meier curve is very close to observed one, these metrics are only cumulative. Therefore in our future work we plan to employ other, more sophisticated methods to predict metastasis for individual patients, as well as to develop a solution for integration of multiple ROIs

References

[1] Ferreira Junior, J. R., Koenigkam-Santos, M., Cipriano, F. E. G., Fabro, A. T., and de Azevedo-Marques, P. M. Radiomics-based features for pattern recognition of lung cancer histopathology and metastases. *Computer Methods and Programs in Biomedicine 159* (2018), 23–30.

DOI: https://doi.org/10.1007/978-3-031-21967-2_37

Keywords: Parkinson's disease, Medical diagnosis, Feature extraction, Hand-drawn spirals

Features of hand-drawn spirals for recognition of Parkinson's disease

Krzysztof Wrobel[0000-0001-7339-1100], Rafal Doroz[0000-0001-6103-1175], Piotr Porwik[0000-0001-8989-9478], Tomasz Orczyk[0000-0002-4664-8369], Agnieszka Betkowska Cavalcante[0000-0003-2236-5265], Monika Grajzer

{krzysztof.wrobel,rafal.doroz,piotr.porwik,tomasz.orczyk}@us.edu.pl,{a.b.cavalcante,m.grajzer}@gidolabs.eu

Simplified Title

Features of hand-drawn spirals for recognition of Parkinson's disease

Abstract

In this paper, a method for diagnosing Parkinson's disease based on features derived from hand-drawn spirals is presented. During drawing of these spirals on a tablet, coordinates of points of the spiral, pressure and angle of the pen at that point, and timestamp were registered. A set of features derived from the registered data, by means of which the classification was performed, has been proposed. For testing purposes, classification has been done by means of several of the most popular machine learning methods, for which the accuracy of Parkinson's disease recognition was determined. The study has proven that the proposed set of features enables the effective diagnosis of Parkinson's disease. The proposed method can be used in screening tests for Parkinson's disease.

The experiments were conducted on a publicly available "Parkinson Disease Spiral Drawings Using Digitized Graphics Tablet Data Set" database from the UCI archives. This database contains drawings of spirals made by people with Parkinson's disease as well as by healthy people.

I Introduction

One of the symptoms of Parkinson's disease is trembling of the whole body, especially noticeable in the hands [1, 3]. Affected people may have difficulty drawing even the simplest figures. Because of that, it is possible to diagnose Parkinson's disease using a freehand drawing analysis. The advantage of this diagnostic method is the fact that it is non-invasive. A person can draw a picture on paper, scan or photograph it, and send it for medical analysis. An alternative method is to draw the image on a tablet, which makes it possible to obtain a digital image directly. Such an image can then be analyzed using various numerical methods [2].

II State of the Art

Techniques developed up to date for diagnosing Parkinson's disease using hand-drawn spirals have used various data analysis methods. These were mostly standard machine learning or deep learning methods. From the raw data describing the spiral, attempts were made to extract sets of features for which classification results were the best. However, neither the method nor the set of features that gave the best results has been extracted. Such an attempt was made in this paper.

III Original Contribution

The paper proposes a new set of features, based on hand-drawn spirals, for diagnosing Parkinson's disease. A total of 14 features were derived from the raw data, describing both the smoothness and shape of the spirals drawn.

The presented procedure can be performed remotely, which means that the patients do not need to leave their homes, enabling wide access to this diagnostic method for older people, or people with motor disabilities. It also makes it suitable for screening tests on a large scale.

IV Methodology

The method assumes that successive reference spiral points are generated, and the following characteristics are calculated at each discrete spiral point: pressure Δp, velocity Δv, acceleration Δa and pen angle Δl. For these values, the standard deviation is zero. Sick people, due to hand tremor, unconsciously increase the pressure, change the pen angle or change the pen speed. Therefore, the values of standard deviations calculated for the data describing the analyzed spirals are greater than zero, and this value increases as the disease progresses. The features determined in this way form a new group of features. A total of 14 features were generated. The new features are analyzed by a set of popular classifiers used in machine learning. The classification quality of 6 classifiers was examined.

V RESULTS

A computerized method of assessing the condition of a patient suffering from Parkinson's disease has been proposed. It is a non-invasive method. It is possible to study the progression of the disease, which is a great convenience for the doctor. The developed method of data acquisition and presentation of results can be implemented in centers dealing with degenerative brain diseases. The method can also be modified to diagnose other neurological diseases, but this requires cooperation with a neurologist.

VI EVALUATION

The article proposes a method to help evaluate the progression of Parkinson's disease. On the basis of selected features, the effectiveness of diagnosing the disease using various single classifiers was tested. The main goal of the experiments conducted was to test whether simple data acquisition methods would recognize the disease using machine learning methods. A simple data acquisition method was proposed, from the point of view of the patient and the doctor. In the experiments conducted, the most useful features were selected for high efficiency of disease recognition.

VII CONCLUSIONS

The article presents a new method for diagnosing Parkinson's disease based on hand-drawn spirals and features generated on their basis. Experiments showed high classification accuracy of the presented approach, which encourages further work. In the next stages, we plan to generate more features and propose a new feature reduction method based on ranking different statistical methods. Ultimately, only those features will be selected for which the effectiveness of classification will be greatest. In order to increase the effectiveness of classification, attempts will be made to create an ensemble of the most suitable classifiers. The competence of the classifiers in analyzing individual features will also be tested. The disadvantage of the test data on which the experiments were conducted is that they are not balanced. They are characterized by a preponderance of data from people with Parkinson's disease. Future work will use methods to eliminate this phenomenon.

REFERENCES

[1] GOLBE, L. I., MARK, M. H., AND SAGE, J. I. *Parkinson's disease handbook.* American Parkinson Disease Association, Inc., 2014.

[2] ISENKUL, M., SAKAR, B., AND KURSUN, O. Improved spiral test using digitized graphics tablet for monitoring parkinson's disease. In *The 2nd International Conference on e-Health and Telemedicine (ICEHTM-2014)* (2014), pp. 171–175.

[3] PARKINSON, J. *An essay on the shaking palsy.* London, 1817.

DOI: https://doi.org/10.1007/978-981-19-8234-7_58
Link to repository: https://github.com/erabrave/remote-health-monitoring.git

Keywords: health monitoring, internet of things, physiological signs monitoring, ambient assisted living, telemedicine.

Design of IoT based Patient Health Monitoring System

Yerlan Zaitin[0000-0002-2819-4527], Madina Mansurova[0000-0002-9680-2758], Octavian Postolache[0000-0001-5055-6347]

yzaitin@gmail.com,Madina.Mansurova@kaznu.edu.kz,Adrian.Postolache@iscte-iul.pt

Simplified Title

Design of IoT based Patient Health Monitoring System

Abstract

Internet of Things allows health care providers to break out of traditional clinical settings. Home monitoring systems allow patients and physicians to monitor a person's health when they are not in the doctor's office to prevent unnecessary and expensive visits to the doctor. Many years of research have shown that remote monitoring of patients is one of the most effective treatments for chronic diseases such as diabetes, heart failure and chronic obstructive pulmonary disease (COPD), but also increases patient participation and reduces emergency admissions. Assisted and individually adapted environments will be possible through the introduction of technologies that provide individual medical care to anyone living in an environment of their choice. In this article, we consider several requirements for the development of such systems, in particular sensors for automatic detection of physiological parameters of the patient's health data which will be send to the cloud where in real time artificial intelligence will process the data and present it to the doctor in the form of infographics that can be easily analyzed, the mobile app where a patient can see own health parameters and contact with a doctor and the website where the doctor, based on real-time data, can examine the patient and, in an emergency, contact with a patient and warn against future diseases.

I Introduction

The Internet of Things provides a continuous platform for facilitating interaction between people and various physical and virtual entities, including individual healthcare domains. The lack of access to medical resources, the growing number of older people with chronic diseases and their need for remote monitoring, rising medical costs and the desire for telemedicine in developing countries make IoT an interesting topic in healthcare [3]. This work aims to consider the components required to create a system that collects physiological data from patients without the involvement of the healthcare system. This study aims to discuss the solutions for remote patient monitoring presented in the literature describing vital sign monitoring systems and identifying the most important physiological parameters that need to be considered to ensure a viable diagnosis of health status, and create a prototype of a long- acting system and the methods for detecting early changes in patients' health parameters, which could help to provide measures for disease prevention. The novel diagnostic system will help to implement a relatively simple and inexpensive health control of a patient over a long period.

II State of the Art

Up to now, several researchers have considered several ways and methods of remote monitoring of patients. With all the prospects for the development of synchronous home monitoring using Mobile Health and the Internet of Things technologies, it is necessary to note the not fully resolved issues, both of a functional nature and at the level of the sensors themselves. The main unresolved problem is the integration of sensors into a single complex, convenient for use by the patient.

III Original Contribution

In this work we designed and developed the mobile APP because of its ability of connecting to the microcontrollers and to the cloud services with Bluetooth and wi-fi where a patient can see own health parameters, send his/her data in real-time to the doctor and contact with a doctor.

IV Methodology

The IoT diagnostic system is described by a combination of a three-tier architecture with reception, network and application applications, as well as a cloud architecture. In the architecture of the IoT diagnostic system, cloud computing seems convenient, as it provides flexibility and scalability for users and developers [1], [2]. Users can access services such as servers, databases, data processing and storage tools. Developers can work and use the necessary data generation, artificial intelligence and visualization tools through the cloud [5] .

V Results

To create this device, we decided to use the ESP8266 NodeMCU v3 microcontroller, since ESP8266 is already installed in this board, which will allow us to send information via the Internet. In addition, one of the advantages of this board is the presence of an I2C bus, which allows us to connect useful and accurate sensors, such as the MAX30102 sensor, which can accurately determine the heartbeat and blood saturation. Another sensor that also uses the I2C bus for connecting to NodeMCU v3 is the MAX30205. It can measure a person's temperature and does it with an accuracy of 0.1C. In order to record data using these sensors, the patient simply needs to fix all the sensors on his arm, and then press the button, which will be located on the body of the utility model. After that, all the necessary data will be collected within a minute and sent to the database. We decided to use FireBase for the database, since its API is very convenient to use for all the tasks that we need, which include sending and reading data from a mobile application, a website and a microcontroller [4].

VI Evaluation

A utility model has been created for this project, with the help of which the patient will collect analyzes, after that all the results will be sent to a database, from where, using a website or a mobile application, the doctor will be able to monitor Patient physiological parameters.

VII Conclusions

In the first stage of our research, we created architecture of the Monitoring system based on the IoT method and includes a three-tier architecture with perception, network, application layers, and cloud-based architecture. This work examines and analyzes the requirements for remote patient monitoring. The devices needed to collect physiological data from patients were identified. At the next stage, in coordination with health care organizations, the system will be tested and experimented with remote monitoring of the health of volunteer patients. Data will be collected and analyzed to see how effective the developed patient monitoring system is. This work was funded by Committee of Science of Republic of Kazakhstan AP09260767 «Development of an intellectual information and analytical system for assessing the health status of students in Kazakhstan» (2021-2023).

References

[1] Jacob Rodrigues, M., Postolache, O., and Cercas, F. *Physiological and behavior monitoring systems for smart healthcare environments: A review.* Sensors, 2020.

[2] Kuņicina, N., Zabašta, A., Bruzgiene, R., Čaiko, J., and Patļins. *The Resilience of Automatic Wireless Meters Reading for Distribution Networks in Smart City Model.* 2018 IEEE 59th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON2018): Conference Proceedings, Latvia, Riga, 2018.

[3] Mostafa, H., Mona, M., Mohammad, N., Parvaneh, A., and Ebrahim, M. *A systematic review of IoT in healthcare: Applications, techniques, and trends.* Journal of Network an Computer Applications, 2021.

[4] Zhang, H., Muhammad, B., Muhammad, U., and Mian, A. *Safe City: Toward safe and secured data management design for IoT enabled smart city planning.* IEEE Access, 2020.

[5] Zholdas, N., Mansurovam, M., Postolache, O., Kalimoldayev, M., and Sarsembayeva, T. *A Personalized mHealth Monitoring System for Children and Adolescents with T1 Diabetes by Utilizing IoT Sensors and Assessing Physical Activities.* International Journal, 2022.

DOI: `https://doi.org/10.1007/978-3-031-21967-2_56`

Keywords: Cognitive load, Sternberg memory test, diagrammatic reasoning test, EEG, ECG, GSR, NASA-TLX questionnaire

Multimodal Approach to Measuring Cognitive Load Using Sternberg Memory and Input Diagrammatic Reasoning Tests

Patient Zihisire Muke [0000-0001-7860-5067], Zbigniew Telec [0000-0003-2539-9452], Bogdan Trawiński [0000-0002-2956-6388]

{patient.zihisire, zbigniew.telec, bogdan.trawinski}@pwr.edu.pl

Simplified Title

Cognitive Load Measurement Using Memory and Reasoning Tests

Abstract

Results of a study of cognitive load using multimodal biometric techniques including electrocardiography, electroencephalography and galvanic skin response are presented in the paper. Thirty student volunteers took part in an experiment conducted on the iMotions integrated biometric platform in a laboratory setting. Two types of tests were employed as research stimuli, namely the Sternberg memory test and input diagrammatic reasoning test. Data were collected using participant activity measures, Single Ease Question (SEQ) and NASA Task Load Index (NASA-TLX) self-report questionnaires, and biometric measurements. In total, 21 metrics were calculated, including two performance, eight subjective, four electrocardiographic, three encephalographic, and four galvanic skin response metrics based on the collected experimental data. The nonparametric Wilcoxon tests were applied to find statistically significant differences between individual metrics for the Sternberg memory tasks and input diagrammatic reasoning tasks for easy and hard difficulty levels. The conducted research allowed to make many interesting observations and showed the usefulness of various measures in the analysis of the cognitive load associated with memory and reasoning tasks.

I Introduction

Cognitive Load Theory (CLT) is a cognitive learning theory designed by the psychologist John Sweller in the 1980s to provide educators with a novel pedagogical framework which addresses the limitations of the human cognitive architecture in education [1]. Nevertheless, education is not the only field of expertise where CLT can help. User interface design can as well benefit significantly from CLT, as the way information is presented when communicating with computer systems can have a huge impact on the quality of the user experience [2].

The main aim of this study was to understand and quantify different levels of cognitive load from a multimodal approach type of measurement based on specified cognitive tasks prepared in a special custom software application. In this research experiment, there were two different task types, each with two difficulty levels: easy and hard. These tasks were presented to participants in a random difficulty order of tasks, i.e., from easy to hard and vice versa. The Sternberg memory task consists of a memory test in which participants are asked to remember an exact number of words and asked to reselect those words from a number of options. The diagram input task consisted of a reasoning test in which participants were presented with a series of questions with words and some rules for finding answers.

The biometric research platform from the iMotions company [3] was used and experiments were conducted on 30 volunteer participants, young people with higher educational backgrounds. Participants' cognitive load was assessed based primarily on metrics extracted from three cognitive load measurement techniques, including subjective, task-performance, and psychophysiological measures, which in turn included ECG, EEG and GSR. Statistical analyses were performed using nonparametric tests to compare different metrics.

State of the Art

Sternberg memory and diagrammatic reasoning-like tasks have been used as the so-called loading tasks to explicitly control the mental workload (or memory load) of subjects with the purpose of studying how it impacts their performance in a such a way that it should be possible to reduce the cognitive load of subjects and conflict resolution, decision making, and collaborative problem solving through external representations. In most of cases, only one sensor was employed (EEG sensor) to keep track of neural activities, particularly in the frontal, parietal and occipital regions of the brain while performing either Sternberg memory or diagrammatic reasoning tasks. To ensure deep involvement in the memory and cognitive processes of subjects in term of performance, a real-time multimodal approaches using deeper computational methods should be considered.

II ORIGINAL CONTRIBUTION

To fully understand and quantify different levels of cognitive load by extracting the most useful metrics from various methods to evaluate cognitive load according to the cognitive load theory, a multimodal measurement approach based on specific cognitive tasks associated with memory and reasoning tasks prepared in a standalone application was employed in this study. The most useful metrics within utilized measure were extracted and evaluated.

III METHODOLOGY

This research is an experimental study where 30 subject' cognitive load was assessed while performing memory and reasoning tasks based primarily on three cognitive load measurement techniques, including subjective (which included the Single Ease Question (SEQ) and NASA Task Load Index (NASA-TLX) self-report questionnaires), task-performance (which included the task completion time and rate), and psychophysiological measures, which in turn included metrics extracted from the electroencephalography (EEG), electrocardiography (ECG), and galvanic skin response (GSR) biometric sensors. Statistical analyses were also performed using nonparametric tests to compare different metrics.

IV RESULTS

Two research questions were formulated, which became the basis for the study presented in this paper: RQ1: What metrics best quantify cognitive load among performance, subjective, and psychophysiological measures when performing the Sternberg memory tasks and Input Diagrammatic Reasoning Tasks? RQ2: Is there any statistical difference between the easy and hard tasks load among individual performance, subjective and psychophysiological metrics while performing the Sternberg memory tasks and Input Diagrammatic Reasoning Tasks?

To respond to RQ1, the following metrics are most appropriate for measuring cognitive load when participants complete the Sternberg memory and Input Diagrammatic Reasoning tasks: *Tct*, *Tct, Seq, Nxo, NXc, NXp, NXm*, *NXf*, *Csd, Eal, Gpc, Gpa* and *Gav* metrics. To answer the RQ2 of the study we performed the nonparametric Wilcoxon tests for two variables comparisons. In the case of the Sternberg memory tasks, the differences between easy and hard tasks were statistically significant for: *Tcr*, *Tct*, *Seq*, *NXo*, *Nxc*, *NXf*, *NXm*, *Gpa*, and *Gav metrics*. Regarding the Input Diagrammatic Reasoning Tasks, the level of differences between easy and hard tasks were statistically significant for: *Tct*, *Seq*, *NXp*, *Gpc*, *Gpa*, and *Gav* metrics.

V EVALUATION

For better insight into the cognitive load processes, the research experiment presented in this paper was conducted in a laboratory setting due to fact that data were collected using both post hoc questionnaires and real-time biometric sensors which had to be placed on participants for more accurate results. The conducted experiment allowed us to make many interesting observations and showed the usefulness of various measures in the analysis of the cognitive load associated with memory and reasoning tasks.

VI CONCLUSIONS

This study demonstrated the usefulness of both Sternberg Memory and Diagrammatic Reasoning tasks as stimuli and the performance, subjective, and psychophysiological metrics we used to measure cognitive load. Therefore, future work will focus on incorporating machine learning techniques to extract significant models from physiological data and link them to various states of cognition, engagement, emotion, and more. In that manner, intelligent interactive computer systems can be designed to continuously monitor and measure the current mental state of users and adjust their interactions accordingly.

REFERENCES

[1] Sweller, J.: Cognitive load during problem solving: effects on learning. Cogn. Sci. 12(1), 257–285 (1988).

[2] Karczewska, B., Kukla, E., Zihisire Muke, P., Telec, Z., Trawiński, B.: Usability Study of Mobile Applications with Cognitive Load Resulting from Environmental Factors. In: Nguyen, N.T. et al. (eds.) ACIIDS 2021. LNCS (LNAI), vol. 12672, pp. 851–864, Springer, Cham (2021).

[3] iMotions Biometric Research Platform (8.1), iMotions A/S, Copenhagen, Denmark (2021)

DOI: https://doi.org/10.1007/978-3-031-21743-2_9

Keywords: machine learning, eye moves, smile, happiness, facial emotions, mimicry, emotion authenticity, noise, chaos, nonlinear dynamical systems, eye move chaosity, eye movement chaosity

Detecting true and declarative facial emotions by changes in nonlinear dynamics of eye movements.

Albert Śledzianowski [0000-0002-3592-6829], Jerzy P. Nowacki [0000-0001-7912-4716], Andrzej W. Przybyszewski [0000-0002-0156-7856], Krzysztof Urbanowicz

albert.sledzianowski@pjwstk.edu.pl,przy@pjwstk.edu.pl

Simplified Title

Detecting facial emotions with correlations of chaotic dynamics of eye movements

Abstract

In our previous work we have showed that we can improve classifications of facial emotions (FE) by extending a dataset with chaotic dynamics parameters of eye movements (EM). This time we wanted to confirm our results using public and independently created video sources and for this purpose we chose the Affectiva-MIT Facial Expression Dataset (AM-FED). Our purpose was to find out whether we can estimate Happiness through non-linear dynamics of EM also in independent video data. We have calculated EM chaotic dynamics in video recordings of the AM-FED database and performed estimations of Happiness calculated with parameters provided by the Open Face library (OF). We also calculated correlation between these parameters and parameters attached to the AM-FED database using our own method based on sliding windows and proposed a method of using its output parameter with a short algorithm. We have observed that true Happiness was connected to a moderate value of negative correlation with EM chaotic dynamics in the case when smile was not present, while for declarative "Smile" parameters we observed a moderate positive value. By using EM chaotic dynamics correlation we have estimated the difference between posed smiles and true Happiness with the XGBoost classifier, with accuracy results of 0.75 (ROC-AUC 0.9) and precision of 0.8 (tested with dataset of 0.33). We are proposing EM chaotic dynamics parameters as an extension for estimations of Happiness based only on facial muscles activity. We think that this approach can confirm the authenticity of Happiness in various cases and also introduce the distinction between real and declarative FE into Computer Vision. It also can bring solution in cases when lower part of the face is hidden, i.e. when is covered by a protective mask.

I Introduction

This is the continuation of our research on the complex dynamical system describing EM presented in our previous article [3], where we have proved a positive, statistically significant correlation between the value of EM chaotic dynamics and the intensity of the Happiness (as FE). In this research we tested our findings on public domain videos with FE classified by Facial Action Coding System (FACS). For this purpose we have chosen the "AM-FED" database created by researchers of the Affectiva Inc, the MIT Media Lab and the Robotics Institute of Carnegie Mellon University [2]. We wanted to compare the differences in EM chaotic dynamical properties between people who show the true (Happiness) or only the declarative FE (Smile). We wanted to propose to use the EM chaotic parameters as an additional data in the method of FE classification. In particular, we were interested if EM chaotic dynamics can improve estimations of authenticity of expressed Happiness and if it can help to solve the common mistakes in the automated FE classifications, like mistaking a worry face of pain with a smile.

II State of the Art

In the context of EM chaotic dynamics context different researches tried to describe EM as nonlinear dynamical system dependent on different aspects but there is lack of any publications related to the EM chaotic dynamics during different FE, especially in the context of their use in automated FE estimations. Also the difference between EM during posed smiles and during experiencing true Happiness are not enough described.

III Original Contribution

We proved ability to estimate the differences in EM chaotic dynamical properties between people who show the true (Happiness) or only the declarative FE (Smile). We are proposing to use the EM chaotic parameters as an additional data in methods of FE classification. Our current results can improve estimations of authenticity of expressed Happiness (as FE) which can also help to solve the common mistakes in the automated FE classifications, like mistaking a worry face of pain with a smile.

IV Methodology

In our experimental study we have used the OpenFace (OF) library for Action Units (AU) estimations and for estimations of gaze vectors used to calculate the chaotic dynamics of the EM [1]. We used an algorithm basing on the FACS for Happiness estimations in the AM-FED videos. For Smile estimations we used parameters provided with the "AM-FED" database [2]. Methods of the EM chaotic dynamical analysis used in this research was described in our previous works where we present our calculation method of chaos, noise and linear and window size [3]. We also calculated Pearson's correlation coefficient for "Smile" and "Happiness" and recreated correlation in the dataset through the dependency between EM chaotic dynamics and chaos means shifted to different positions in the window. We used the XGBoost classifier for estimations of "Smile" and "Happiness" with created dataset and decision-tree visualizations.

V Results

In our results a smile is accompanied by a positive correlation with the chaotic EM, while in the case of Happiness we can see negative correlation. For both Smile and Happiness the opposite direction is visible when analyzing the correlation results with the chaotic EM (while maintaining similar low-average levels). In our opinion, this parameter distinguishes between posed smile and true FE of Happiness. The chaotic EM is visible for all data positively correlated with Happiness, negative correlation is only visible for Happiness not confirmed by results of "Smile" estimations. This observation was earlier confirmed with very similar positive statistically significant correlations levels [3].

VI Evaluation

The main assumptions for developing our method was to find out whether various face emotional states are accompanied by changes in nonlinear dynamics of eye movements. In this article we wanted to extend our findings from previous experiment described in [3] by repeating measurements and analysis on publicly available, verified database. In context of obtained results, we're proposing to introduce the chaotic EM parameters into the FACS-based automated methods of Happiness estimation along with a decision mechanism that would determine, whether the intensity of AU06 and AU12 can be classified as true or not.

VII Conclusions

The method presented in this article uses EM chaotic dynamics correlation together with the FACS estimations. We see possibilities for presented method in helping in FE classification of people with mimicry problems, like Parkinson's disease patients with facial effects of the Bradykinesia ("poker face") or just people having problems with proper Happiness mimicry. This method can also bring solution in Happiness detection for partially covered face i.e. by a protective mask.

References

[1] Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)* (2018), pp. 59–66.

[2] McDuff, D., el Kaliouby, R., Senechal, T., Amr, M., Cohn, J. F., and Picard, R. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected "in-the-wild". In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2013), pp. 881–888.

[3] Sledzianowski, A., Urbanowicz, K., Glac, W., Slota, R., Wojtowicz, M., Nowak, M., and Przybyszewski, A. Face emotional responses correlate with chaotic dynamics of eye movements. *Procedia Computer Science 192* (2021), 2881–2892. Knowledge-Based and Intelligent Information And Engineering Systems: Proceedings of the 25th International Conference KES2021.

DOI: https://doi.org/10.1007/978-3-031-21967-2_19

Keywords: fBIM data, machine learning, civil engineering modelling

Machine Learning Methods for BIM Data

Grażyna Ślusarczyk[0000-0003-1032-1644], Barbara Strug[0000-0002-2204-507X]

grazyna.slusarczyk@uj.edu.pl,barbara.strug@uj.edu.pl

Simplified Title

Machine Learning Methods for Building Information Modeling Data

Abstract

This paper presents a survey of machine learning methods used in applications dedicated to building and construction industry. A BIM model being a database system for civil engineering data is presented. A representative selection of methods and applications is described. The aim of this paper is to facilitate the continuation of research efforts and to encourage bigger participation of researchers in database systems in the filed of civil engineering.

I Introduction

Building Information Modeling (BIM) is nowadays widely used in architecture, engineering and construction industry (AEC). The building and construction industry employs currently about 7 percent of the world's working-age population and is one of the world economy's largest sectors. It is estimated that about $10 trillion is spent on construction-related goods and services every year. In the last decade, the acceptance and actual use of BIM has increased significantly within the building community. It has largely contributed to the process of eliminating faults in designs. BIM allows architects and engineers to create 3D simulations of the desired structures which contain significantly more information on the actual structures, than drawings produced using traditional Computer Aided Drafting CAD systems. As a result, BIM is becoming more and more present in the construction industry.

II State of the Art

The information about a building created in any software can be exported to different formats. Each commercial application has its own file type to store building data, but all of them can also export building information to an IFC file. The file format IFC has become de facto a standard way of interchanging and storing BIM data. It is an interoperable BIM standard for CAD applications, which supports a full range of data exchange among different disciplines and heterogeneous applications. Information retrieved from IFC files can be used by many different applications.

IFC specifies different types of building entities and their basic properties. It defines an EXPRESS based entity-relationship model, which consists of several hundred entities organized into an object-based inheritance hierarchy. All the entities in IFC are divided into rooted and non-rooted ones. While the first ones are derived from IfcRoot and have an identity (a GUID), attributes for name, description, and revision control, the other ones (non-rooted) do not have identities and their instances exist only if they are referenced from a rooted instance directly or indirectly. IfcRoot is subdivided into three concepts: object definitions, relationships, and property sets:

1. IfcObjectDefinition captures tangible object occurrences and types
2. IfcRelationship captures relationships among objects
3. IfcPropertyDefinition captures dynamically extensible properties of objects

III Results

Information important from the point of view of many applications lies implicitly in the interrelation between building elements. Therefore several approaches directed at extracting implicit data from building models have been presented.

III.1 Learning semantic information

A method of extracting features which are then used in the affinity propagation clustering algorithm to get spaces with similar usage functions, is proposed. The method allows for automatic learning of functional knowledge from building space structures. The physical properties of each space and their boundary relationships in BIM model are extracted from the IFC file based on BIM data. Then boundary graphs with space boundary relationships, where properties of each space propagate along the edges, are build. Features of building spaces are extracted from the space boundary graphs. Based on these features and the graph representation of the building structure, the adapted affinity propagation algorithm performs building space clustering analysis, in order to get representative samples of building spaces. The experimental results performed on a real world BIM dataset containing 595 spaces from a 20-storey building show that building spaces with typical usage functions, like senior offices, open offices and circulation spaces, can be discovered by an unsupervised learning algorithm.

III.2 Semantic enrichment of BIM models

The other group of papers considers the use of machine learning algorithms for semantic enrichment of BIM models obtained from point cloud data. In this way a time-consuming process of manually creating 3D models useful for architectural and civil engineering applications can be avoided. Semantic enrichment encompasses classification of building objects, aggregation and grouping of building elements, implementing associations to reflect connections and numbering, unique identification, completion of missing objects, and reconstruction of occluded objects. Then the classification of the model as a whole, or of particular assemblies or objects within the model in respect to code compliance, can be performed. BIM objects derive many of their properties from their classes, making object classification crucial for reuse in different analysis tasks, like spatial validation of a BIM model, quantity take-off and cost estimation.

III.3 Building condition diagnosis

Machine learning methods are also used for defect classification in masonry walls of historic buildings. First, the process of Scan-to-BIM, which automatically segments point clouds of ashlar masonry walls into their constitutive elements, is presented. Then the machine learning based approach to classification of common types of wall defects, that considers both the geometry and colour information of the acquired point clouds, is described. The found defects are recorded in a structured manner within the BIM model, which allows for monitoring the effects of deterioration. A supervised logistic regression algorithm has been employed to classify different types of decay using parameters of roughness of stones and dispersion of colour in stones. Stones labelled as 'defective' by experts are used for training the classifier, which is subsequently employed to label new data. The proposed approach has been tested on data from the main façade of the Royal Chapel in Stirling Castle, Scotland. For the training process samples of three classes of decay (erosion, mechanical damage and discolouration) were used. 15 samples (5 of each class) were included in the test set, obtaining a global accuracy of 93.3% in the classification.

IV CONCLUSIONS

Traditional methods for modeling and optimizing complex structure systems require huge amounts of computing resources. Artificial-intelligence-based methods can often provide valuable alternatives for efficiently solving problems in architectural and engineering design, construction and manufacturing.

Machine learning techniques have considerable potential in the development of BIM. The application of classification algorithms would enable machines to do tasks usually done by hand. Results of application of machine learning methods on architectural datasets provide a relevant alternative view to explicit querying mechanisms and provides useful insights for more informed decisions in the design and management of buildings. Machine learning might facilitate less experienced users to query complex BIM datasets for project specific insights.

It was shown that machine learning algorithms can learn key features of a building belonging to a certain category, and this acquired knowledge could be used in the future. The proposed techniques can be applied for retrieval, reference, and evaluation of design process. The presence of the historical data combined with the acquired knowledge of the building type key features could help in developing methods for automatic design of building structures with required characteristics. The presented methods can be extended to further subdivide the BIM main categories into sub-categories that could represent different areas of interest in these structures.

REFERENCES

Author Index

WUST Publishing House prints
can be obtained via mailorder:
zamawianie.ksiazek@pwr.edu.pl
www.ksiegarnia.pwr.edu.pl
ISBN 978-83-7493-289-9