

Key Information Extraction from Mobile-Captured Vietnamese Receipt Images using Graph Neural Networks Approach

Van Dung Pham
Deep Learning and Application
Ho Chi Minh City, Vietnam
dungpv.ai@gmail.com

Le Quan Nguyen
Sejong University
Seoul, South Korea
quannl71290@sju.ac.kr

Nhat Truong Pham
Institute for Computational Science
Ton Duc Thang University
Ho Chi Minh City, Vietnam
phamnhattruong@tdtu.edu.vn

Bao Hung Nguyen
SK-Global, JSC.
Ho Chi Minh City, Vietnam
baohung.ng@gmail.com

Duc Ngoc Minh Dang
Computing Fundamental Department
FPT University
Ho Chi Minh City, Vietnam
ducnm2@fe.edu.vn

Sy Dzong Nguyen*
Institute for Computational Science
Ton Duc Thang University
Ho Chi Minh City, Vietnam
nguyensydzong@tdtu.edu.vn

Abstract—Information extraction and retrieval are growing fields that have a significant role in document parser and analysis systems. Researches and applications developed in recent years show the numerous difficulties and obstacles in extracting key information from documents. Thanks to the raising of graph theory and deep learning, graph representation and graph learning have been widely applied in information extraction to obtain more exact results. In this paper, we propose a solution upon graph neural networks (GNN) for key information extraction (KIE) that aims to extract the key information from mobile-captured Vietnamese receipt images. Firstly, the images are pre-processed using U²-Net, and then a CRAFT model is used to detect texts from the pre-processed images. Next, the implemented TransformerOCR model is employed for text recognition. Finally, a GNN-based model is designed to extract the key information based on the recognized texts. For validating the effectiveness of the proposed solution, the publicly available dataset released from the Mobile-Captured Receipt Recognition (MC-OCR) Challenge 2021 is used to train and evaluate. The experimental results indicate that our proposed solution achieves a character error rate (CER) score of 0.25 on the private test set, which is more comparable with all reported solutions in the MC-OCR Challenge 2021 as mentioned in the literature. For reproducing and knowledge-sharing purposes, our implementation of the proposed solution is publicly available at https://github.com/ThorPham/Key_information_extraction.

Index Terms—Graph Neural Networks, Key Information Extraction, Optical Character Recognition, Text Detection, Text Recognition.

I. INTRODUCTION

Information extraction (IE) is a process to automatically extract unstructured and semi-structured documents into structured information, then save them to the database for future retrieval. IE aims to draw out related information to pre-specified types such as location/address, person, organization

names, and the relationship between these entities, from any types of data like text, image, audio, and video. In document analysis, IE is a subset of optical character recognition (OCR) that not only detects and recognizes the text from the image but also extracts the key information from them.

IE has a significant role in a wide range of applications like web-based searching, curriculum vitae (CV) parser, clinical reports, invoices, receipts, and business documents. Lohani *et al.* [1] proposed a novel approach using graph convolutional networks (GCN) to extract information in invoice reading and extracting systems. Oral *et al.* [2] investigated biLSTM conditional random fields (CRFs) and graph-based extraction to draw out text-intensive and visually rich scanned bank documents. Bhatia *et al.* [3] deployed an end-to-end CV parsing system to obtain candidates' relevant information and rank them to choose the most suitable one for a given job description using the BERT classification model.

MC-OCR Challenge 2021 [4] was an OCR challenge in conjunction with The 15th IEEE-RIVF International Conference on Computing and Communication Technologies. This challenge aimed to detect and recognize the text from the structured and semi-structured receipts and invoices captured by mobile devices. Two main tasks in this challenge were quality evaluation of receipt images and OCR recognition. The first task was to evaluate the quality of the mobile-captured receipt images in the range (0, 1) where the highest quality is 1 and the lowest one is 0. The second task addressed the extraction of key information from the mobile-captured receipt and invoice images. In this paper, we propose a GNN-based approach to solve the second task of the MC-OCR Challenge 2021 which aimed to extract KIE from mobile-captured Vietnamese receipt images with the following steps:

- Text detection;

* Corresponding author.

- Text recognition;
- Key information extraction.

where KIE is the most important step. It mostly influences the final score calculated by character error rate (CER) which is the evaluation metric for the OCR recognition task in the MC-OCR Challenge 2021. As a result, our proposed solution achieves a very high CER score of 0.25 which is more competitive than other solutions reported in the MC-OCR Challenge 2021. Furthermore, using the GNN-based approach is an extremely potential solution for extracting key information from structured and semi-structured documents and mobile-captured receipt images.

The contribution of our proposed solution is summarized as follows:

- Using U²-Net [5] for salient object detection to pre-process receipt images;
- Using CRAFT [6] and implementing TransformerOCR¹ namely *VietOCR* to detect and recognize texts in the receipt images;
- Designing a GNN-based model by combining residual gated graph convnets (RG-GCN) [7] and graph normalization (GN) [8] to extract key information from recognized texts;
- Using spelling correction and regular expression to post-process the extracted key information.
- Using multiple text detection methods and *pseudo labels* as data augmentation to generate and synthesize more data in terms of adapting the model during training progress and reducing over-fitting.

The rest of the paper is organized as follows. The relevant literature research is discussed in Section II. The proposed methodology is addressed in detail in Section III. In Section IV, the experimental results are presented and discussed. Finally, concluding our solution and future works are summarized in Section V.

II. RELATED WORK

Most traditional approaches have utilized rule-based, regular expression (regex), and template matching methods for KIE. Dhakal *et al.* [9] proposed a novel algorithm namely a one-shot template matching to automatically extract data from other business documents that have the same format as the given annotated document. Although all standard template-based approaches require a predefined template structure in advance, the authors in [10] designed the rule-based IE systems to focus on rapidly processing the information.

In the past decade, neural networks and deep learning have succeeded in a wide range of applications, such as computer vision, natural language processing, signal processing, OCR, and even COVID-19 detection [4], [11]–[14]. As a result, numerous studies have used deep learning for key IE. For instance, Karim *et al.* [15] proposed a named entity recognition

system for the Bangla language by combining densely connected network, bidirectional long short-term memory (biLSTM), and word embedding. Qian *et al.* [16] designed an IE framework based on graph representation namely GraphIE that learns not only the local but also non-local representations to improve the performance of the word-level predictions. Yu *et al.* [17] proposed a framework namely PICK (processing key information extraction) by combining the graph convolution with graph learning for key information extraction from documents.

Recently, the MC-OCR Challenge [4] focused on detecting, recognizing, and extracting key information from the mobile-captured receipt images based on the quality evaluation and OCR recognition tasks. In which, several solutions were been proposed for key information extraction. Bui *et al.* [18] used Faster R-CNN for text detection, and then used Transformer for text recognition. Nguyen *et al.* [19] used YOLO-V5 for text detection and implemented TransformerOCR for text recognition. Le *et al.* [20] used pixel aggregation network (PAN) for text detection and implemented TransformerOCR for text recognition. Nguyen *et al.* [21] used CRAFT [6] for text detection and TransformerOCR for text recognition, then applied several techniques including support vector machine (SVM), PhoBERT, and rule-based for key information extraction. Hieu *et al.* [22] used the same models as the solution in ref. [21], but without using SVM and PhoBERT models. Nguyen *et al.* [23] used PaddleOCR for text detection, MobileNet-V3 for rotating correction, implemented TransformerOCR for text detection, and PICK for key information extraction. However, there was only one solution using a GNN-based approach for key information extraction from mobile-captured Vietnamese receipt images and its accuracy was very high among all solutions reported in the challenge. Moreover, the use of the implemented TransformerOCR is very important in almost all solutions.

III. METHODOLOGY

Motivated by the recent studies, in this study, a solution upon GNN is designed for key information extraction. As illustrated in Fig. 1, our proposed solution includes 3 major steps: text detection, text recognition, and key information extraction. Additionally, it is required a pre-processing step to remove background noise and correct the orientation of the receipt images. Given a mobile-captured receipt image, our solution can draw out pre-required information including SELLER, ADDRESS, TIMESTAMP, and TOTAL_COST.

A. Pre-processing

Mobile-captured receipt images often contain undesired backgrounds and vary in orientations which can make the OCR processing more challenging. Therefore, removing the noisy background and correcting orientation are mandatory steps that aid the text recognition step, hence improving the overall performance of our proposed solution. Most salient object detection methods often suffer from blurry boundaries because of using cross-entropy as the training loss. However, using

¹<https://github.com/pbcquoc/vietocr>

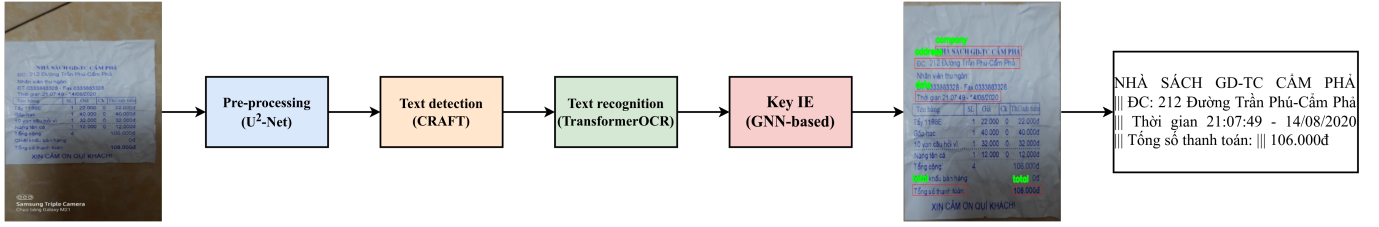


Fig. 1. Proposed framework.



Fig. 2. Pre-processing step.

BASNet [24] could overcome this challenge by employing a hybrid loss that combined binary cross-entropy, structural similarity, and intersection over union (IoU) losses to take into account the information on pixel-level, patch-level, and map-level, respectively. Moreover, an upgrade version namely U²-Net [5] was also designed for salient object detection that includes two nested U-structure. Besides that, it does not require any backbones to train the U²-Net architecture. For this reason, we use the U²-Net to segment the receipt region from an image as the salient object to obtain an accurate boundary in this study. Furthermore, to obtain the best orientation of the receipt image, an unsupervised feature learning method proposed by Gidaris et al. [25] is applied to predict the 2D rotations of the given image. The predicted rotations should belong to the following degrees: 0, 90, 180, and 270. After that, we find the rotated box which fits the receipt best and rotate it to correct the orientation, as shown in Fig. 2.

B. Extracting text

Text from pre-processed receipt image is then detected using CRAFT [6]. CRAFT uses VGG-16-BN as backbone architecture in the encoding part while there are several skip connections in the decoding part that forms a U-Net structure. Given a pre-processed receipt image as an input, CRAFT can return *region score* and *affinity score* that are used to localize individual character regions and joint all detected characters into a single instance.

For recognizing text, we used VietOCR² framework implemented based on TransformerOCR for scene text recognition. Two main parts in TransformerOCR are feature extractor

and transformer modules. In which, the feature extractor uses ResNet-101 [26] as backbone architecture while the transformer module uses only the decoder of the Transformer model. Moreover, VietOCR is very powerful and generalized because it supports Vietnamese hand-writing and printed texts. As a result, given an instance of detected text output by CRAFT, text can be significantly recognized by VietOCR. Finally, the output of this step is a set of rectangle boxes and their associated texts.

C. Key information extraction with graph neural network

1) *Problem statement*: In this study, we designed a GNN-based model by combining RG-GCN [7] and GN [8] for KIE from the mobile-captured Vietnamese receipt images. In which, RG-GCN includes multiple gated GCN layers are formed as ResNet [26] structure, while GN is applied to take into account both local and global structures on the graph by optimizing the weighted combination of different normalization techniques. As a result, our GNN-based approach takes the text boxes detected from the previous step as inputs and labels them with one of 5 pre-defined classes: SELLER, ADDRESS, TIMESTAMP, TOTAL_COST, and OTHERS. Each receipt image is defined as a graph $G(V, E)$ where each text box is a node V and edges E represent the geometric distances between them. Node and edge embedding features are defined below.

2) Node and edge features:

a) *Node features*: Node features are aggregated from text box locations and texts recognized by the OCR module. For each text box, a vector $L = (x_i, y_i) \mid i \in [1, 4]$ (x_i, y_i are the coordinates of a box corner) is fed to a fully connected layer to create an embedding vector of 512, named location encoding. Its text is also encoded by an LSTM [27] to obtain an embedding vector, named semantic encoding. The final feature of a node is formed by an element-wise addition to its location encoding and semantic encoding.

b) *Edge features*: Edge is a connection between two nodes but does not exist for each pair of nodes. Before connecting nodes together, the nodes are sorted according to their associated text box locations. For each node v , we find its neighbors \mathcal{N} by connecting each $v_j \in \mathcal{N}$ to v with an edge if:

$$d(v, v_j) = |(v_y - v_{j,y})| < 3 * h_v \quad (1)$$

where v_y and $v_{j,y}$ are y -coordinates of boxes' centers and h_v is the box height of node v . The connecting scheme of

²<https://github.com/pbcquoc/vietocr>

this graph is designed based on an assumption that texts in a receipt are arranged in a top-down structure. This assumption is trustworthy in most cases where receipt images were properly corrected in orientation after pre-processing step. Edge features are formulated based on the distance between 2 nodes (Eq. 2) fed into a linear layer to get the embedding vector in the same way as the text box location.

$$d(v_i, v_j) = |(v_{i,x} - v_{j,x}, v_{i,y} - v_{j,y})| \quad (2)$$

3) *Network architecture*: After embedding node and edge features, they are fed into RG-GCN [7] architecture with GN [8], where the values of the successive layer are computed as follow:

$$h = x + \left(Ax + \sum_{v_j \rightarrow v} \eta(e_j) \odot Bx_j \right)^+, \quad (3)$$

where $(\cdot)^+$ denotes the rectified linear unit (ReLU) function, \odot denotes the element-wise product, x and h are the input and hidden vector representations of the current node v , x_j is the input of neighbor node v_j linking with v . The weight η of each neighbor node of v is obtained as below:

$$\eta(e_j) = \sigma(e_j) \left(\sum_{v_k \rightarrow v} \sigma(e_k) \right)^{-1}, \quad (4)$$

where σ is the sigmoid function, e_j , e_k are the features of edges connecting current node v with its neighbour nodes v_j and v_k respectively. In its own turn, e_j is calculated as follow:

$$e_j = Ce_j^x + Dx_j + Ex, \quad (5)$$

$$e_j^h = e_j^x + (e_j)^+, \quad (6)$$

where e_j^x and e_j^h are the input and hidden vectors representing feature of edge e_j connecting current node v and node v_j . In equations (3) and (5) above, Ax , Bx_j , Ce_j^x , Dx_j , and Ex are rotations of the input features learned from training process.

In each layer, the node and edge features are normalized using the graph-aware normalization method described in ref. [8]. After stacking L layers ($L = 8$) of RG-GCN [7], each node feature is fed into a dense layer with shared weights to all nodes. Then each node is classified into 5 groups with the cross-entropy loss function.

4) *Dataset and Augmentation*: As shown in Fig. 3, the MC-OCR Challenge 2021 dataset contains 1,155 training receipt images with labels of 4 classes: SELLER, ADDRESS, TIMESTAMP, and TOTAL_COST. We use text detection on receipt images in this dataset to get text boxes. For the text boxes output by our model but not belonging to the aforementioned classes, we create a *pseudo labels* named "OTHERS" class. Any text box obtained from the detection step, which has IoU with all boxes of 4 pre-defined classes less than 0.2, is labeled as "OTHERS". We use both CTPN [28] and CRAFT [6] for detecting box in training phase. Using multiple text detection algorithms may induce different sets of text boxes for each image. This can be considered as an augmentation method. We also augment the dataset by alternating values from the ground truth of SELLER and ADDRESS

classes, and replacing TIMESTAMP, TOTAL_COST classes with random values.

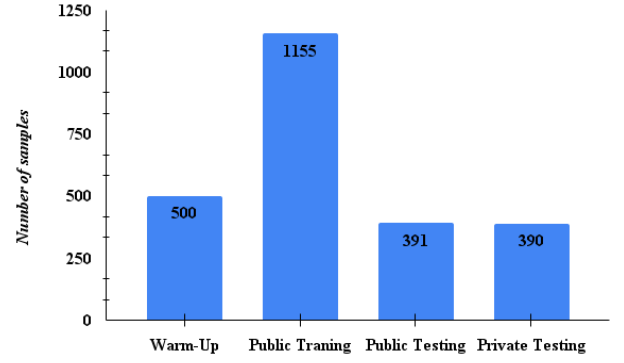


Fig. 3. The distribution of the MC-OCR Challenge 2021 dataset.

D. Post-processing

To improve the accuracy of our proposed solution, we use the regular expression to extract the TIMESTAMP class when the graph model is failed. Furthermore, we create a dictionary based on spelling correction to automatically correct the extracted classes from the graph model. Fig. 4 shows the output of the post-processing step.


Input	Output (1 line, 4 fields, separated by).
	SCTC CO THO 104 TRAN PHU - CAM 104 Trần Phú - phường Cẩm Tây - Thành phố Cẩm Ngày 09/08/2020 (7:44 CH 7:44CH) Tổng thanh toán 115.000d

Fig. 4. The input and required output of TASK 2 - OCR recognition [4].

IV. EXPERIMENTAL RESULTS

A. Evaluation metric

The character error rate (CER) is an evaluation metric in the MC-OCR challenge 2021. Given the number of characters nCs , the CER is obtained as follows:

$$CER = \frac{ins + sub + del}{nCs}, \quad (7)$$

where ins , sub , del are the number of character insertions, substitutions, and deletions respectively to transform the referenced text into the OCR output.

B. Training parameters

Our solution is designed and deployed on an NVIDIA 1080Ti. The proposed solution is trained using Adam optimizer [29] with a learning rate of 0.001, a batch size of 32, and epochs of 10. The amount of augmented data is 10,000 samples and the train/test ratio is 80:20.

C. Results

Table I shows the accuracy of each class using the graph model, in which the SELLER class achieves the highest accuracy at 0.946 while the TIMESTAMP one gains the lowest accuracy at 0.874. Generally speaking, the average accuracy of the graph model is 0.922.

TABLE I
THE ACCURACY OF GRAPH MODEL FOR EACH CLASS

Class	SELLER	ADDRESS	TIMESTAMP	TOTAL COST
Accuracy	0.946	0.938	0.874	0.874

Several other experimental results of the proposed solution are shown in Fig. 5: given the mobile-captured images of the receipts, our proposed solution can automatically extract all key information of 4 classes SELLER, ADDRESS, TIMESTAMP, and TOTAL_COST. In Fig. 5, three columns present the inputs, the outputs of the graph model, and the final outputs of the proposed solution, respectively.



Fig. 5. Experimental results of the proposed solution.

Table II shows an extremely competitive result among other solutions reported in the MC-OCR Challenge 2021. Our proposed solution achieves the CER of 0.25 on the private test set of the challenge. Results in Table II prove that using the GNN-based model serves an important role and can be

TABLE II
COMPARISON OF OUR PROPOSED SOLUTION AND OTHER SOLUTIONS OF TASK 2 OF THE MC-OCR CHALLENGE 2021

Task 2: OCR recognition		
Solution	Method	CER
[19]	YOLO-V5 + TransformerOCR	0.22
[23]	PaddleOCR + MobileNet-V3 + TransformerOCR + PICK	0.23
Ours	U ² -Net + CRAFT + TransformerOCR + GNN-based	0.25
[21]	CRAFT + TransformerOCR + (SMV & PhoBERT & Rule-based)	0.26
[20]	PAN + TransformerOCR + Rule-based	0.30
[18]	Faster R-CNN + Transformer	0.32
[22]	CRAFT + TransformerOCR + Rule-based	0.39

extremely potential to extract key information from structured and semi-structured receipt images. Moreover, the use of the implemented TransformerOCR is also very efficient in most solutions.

D. Discussion

Based on observation, some factors that affect the final results are listed as follows:

- The quality of the mobile-captured receipt images;
- The accuracy of the text detection and recognition modules;
- There are some incorrect labels between TIMESTAMP and TOTAL_COST classes that affect text detection and recognition. Hence, the accuracy of KIE is also reduced.

Hence, an end-to-end training model or joint task learning can be applied to improve the performance of the desired system.

V. CONCLUSION

In this paper, a solution upon GNN is employed for key information extraction by combining the RG-GCN and GN. Besides, we have applied both pre/post-processing to improve the performance of the proposed solution. Furthermore, we have also used the data augmentation method to deal with the lack of data and labels. Our experiments achieve the competitive and potential result for extracting the needed information from mobile-captured Vietnamese receipt images among the solutions in the MC-OCR Challenge 2021. Although using the GNN-based approach is extremely potential for key information extraction for structured and semi-structured documents, some aspects need to be deeply investigated to improve the performance of the pipeline solution in the future: the first one is the quality of the image, the second one is text detection, and the third one is text recognition.

ACKNOWLEDGMENT

The authors would like to thank the Vietnam National Foundation for Science and Technology Development (NAFOS-TED) under grant number 107.01-2019.328.

REFERENCES

- [1] D. Lohani, A. Belaïd, and Y. Belaïd, "An invoice reading system using a graph convolutional network," in *Computer Vision - ACCV 2018 Workshops - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers*, ser. Lecture Notes in Computer Science, G. Carneiro and S. You, Eds., vol. 11367. Springer, 2018, pp. 144–158. [Online]. Available: https://doi.org/10.1007/978-3-030-21074-8_12

- [2] B. Oral, E. Emekligil, S. Arslan, and G. Eryigit, "Information extraction from text intensive and visually rich banking documents," *Information Processing & Management*, vol. 57, no. 6, p. 102361, 2020.
- [3] V. Bhatia, P. Rawat, A. Kumar, and R. R. Shah, "End-to-end resume parsing and finding candidates for a job description using bert," *arXiv preprint arXiv:1910.03089*, 2019.
- [4] X. Vu, Q. Bui, N. Nguyen, T. T. H. Nguyen, and T. Vu, "MC-OCR challenge: Mobile-captured image document recognition for vietnamese receipts," in *RIVF International Conference on Computing and Communication Technologies, RIVF 2021, Hanoi, Vietnam, August 19-21, 2021*. IEEE, 2021, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/RIVF51545.2021.9642077>
- [5] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107404, 2020.
- [6] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 9365–9374. [Online]. Available: http://openaccess.thecvf.com/content/_CVPR/_2019/html/Baek_Character_Region_Awareness_for_Text_Detection_CVPR_2019_paper.html
- [7] X. Bresson and T. Laurent, "Residual gated graph convnets," *arXiv preprint arXiv:1711.07553*, 2017.
- [8] Y. Chen, X. Tang, X. Qi, C.-G. Li, and R. Xiao, "Learning graph normalization for graph neural networks," *arXiv preprint arXiv:2009.11746*, 2020.
- [9] P. Dhakal, M. Munikar, and B. Dahal, "One-shot template matching for automatic document data capture," in *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, vol. 1. IEEE, 2019, pp. 1–6.
- [10] M. A. Valenzuela-Escárcega, G. Hahn-Powell, and D. Bell, "Odinson: A fast rule-based information extraction framework," in *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. European Language Resources Association, 2020, pp. 2183–2191. [Online]. Available: <https://aclanthology.org/2020.lrec-1.267/>
- [11] N. T. Pham, D. N. M. Dang, and S. D. Nguyen, "A method upon deep learning for speech emotion recognition," *Journal of Advanced Engineering and Computation*, vol. 4, no. 4, pp. 273–285, 2020.
- [12] L. H. Nguyen, N. T. Pham, V. H. Do, L. T. Nguyen, T. T. Nguyen, V. D. Do, H. Nguyen, and N. D. Nguyen, "Fruit-cov: An efficient vision-based framework for speedy detection and diagnosis of sars-cov-2 infections through recorded cough sounds," *arXiv preprint arXiv:2109.03219*, 2021.
- [13] N. T. Pham, D. N. M. Dang, and S. D. Nguyen, "Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition," *arXiv preprint arXiv:2109.09026*, 2021.
- [14] T. Tran, N. T. Pham, and J. Lundgren, "A deep learning approach for detecting drill bit failures from a small sound dataset," *Scientific Reports*, vol. 12, no. 1, pp. 1–13, 2022.
- [15] R. Karim, M. Islam, S. R. Simanto, S. A. Chowdhury, K. Roy, A. Al Neon, M. Hasan, A. Firoze, R. M. Rahman *et al.*, "A step towards information extraction: Named entity recognition in bangla using deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 37, no. 6, pp. 7401–7413, 2019.
- [16] Y. Qian, E. Santus, Z. Jin, J. Guo, and R. Barzilay, "Graphie: A graph-based framework for information extraction," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 751–761. [Online]. Available: <https://doi.org/10.18653/v1/n19-1082>
- [17] W. Yu, N. Lu, X. Qi, P. Gong, and R. Xiao, "PICK: processing key information extraction from documents using improved graph learning-convolutional networks," in *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*. IEEE, 2020, pp. 4363–4370. [Online]. Available: <https://doi.org/10.1109/ICPR48806.2021.9412927>
- [18] D. C. Bui, D. Truong, N. D. Vo, and K. Nguyen, "MC-OCR challenge 2021: Deep learning approach for vietnamese receipts OCR," in *RIVF International Conference on Computing and Communication Technologies, RIVF 2021, Hanoi, Vietnam, August 19-21, 2021*. IEEE, 2021, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/RIVF51545.2021.9642128>
- [19] C. M. Nguyen, V. V. Ngo, and D. D. Nguyen, "MC-OCR challenge 2021: Simple approach for receipt information extraction and quality evaluation," in *RIVF International Conference on Computing and Communication Technologies, RIVF 2021, Hanoi, Vietnam, August 19-21, 2021*. IEEE, 2021, pp. 1–4. [Online]. Available: <https://doi.org/10.1109/RIVF51545.2021.9642150>
- [20] H. Le, H. To, H. An, K. Ho, K. Nguyen, T. Nguyen, T. Do, T. D. Ngo, and D. Le, "MC-OCR challenge 2021: An end-to-end recognition framework for vietnamese receipts," in *RIVF International Conference on Computing and Communication Technologies, RIVF 2021, Hanoi, Vietnam, August 19-21, 2021*. IEEE, 2021, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/RIVF51545.2021.9642121>
- [21] H. V. Nguyen, L. D. Bao, H. V. Trinh, H. H. Phan, and T. M. Thanh, "MC-OCR challenge 2021: Towards document understanding for unconstrained mobile-captured vietnamese receipts," in *RIVF International Conference on Computing and Communication Technologies, RIVF 2021, Hanoi, Vietnam, August 19-21, 2021*. IEEE, 2021, pp. 1–5. [Online]. Available: <https://doi.org/10.1109/RIVF51545.2021.9642126>
- [22] B. H. Tran, D. V. Hoang, N. M. Hiep, P. N. B. Anh, H. G. Bao, N. D. Anh, B. H. Phong, T. Nguyen, P. L. Nguyen, and T. Le, "MC-OCR challenge 2021: A multi-modal approach for mobile-captured vietnamese receipts recognition," in *RIVF International Conference on Computing and Communication Technologies, RIVF 2021, Hanoi, Vietnam, August 19-21, 2021*. IEEE, 2021, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/RIVF51545.2021.9642088>
- [23] D. Nguyen, T. Nguyen, and X. Nguyen, "MC-OCR challenge 2021: End-to-end system to extract key information from vietnamese receipts," in *RIVF International Conference on Computing and Communication Technologies, RIVF 2021, Hanoi, Vietnam, August 19-21, 2021*. IEEE, 2021, pp. 1–5. [Online]. Available: <https://doi.org/10.1109/RIVF51545.2021.9642083>
- [24] X. Qin, Z. V. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jägersand, "Basnet: Boundary-aware salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 7479–7489. [Online]. Available: http://openaccess.thecvf.com/content/_CVPR/_2019/html/Qin_BASNet_Boundary-Aware_Salient_Object_Detection_CVPR_2019_paper.html
- [25] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: <https://openreview.net/forum?id=S1v4N2l0->
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: <https://doi.org/10.1109/CVPR.2016.90>
- [27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [28] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9912. Springer, 2016, pp. 56–72. [Online]. Available: https://doi.org/10.1007/978-3-319-46484-8_4
- [29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: <http://arxiv.org/abs/1412.6980>