2021 RIVF International Conference on Computing and Communication Technologies (RIVF) | 978-1-6654-0435-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/RIVF51545.2021.9642126

MC-OCR Challenge 2021: Towards Document Understanding for Unconstrained Mobile-Captured Vietnamese Receipts

Hoai Viet Nguyen*, Linh Bao Doan*, Hoang Viet Trinh*, Hoang Huy Phan
R&D Lab
Sun Asterisk Inc.
{nguyen.viet.hoai, doan.bao.linh, trinh.viet.hoang, phan.huy.hoang}@sun-asterisk.com

Ta Minh Thanh†
Le Quy Don Technical University, 236 Hoang Quoc Viet, Bac Tu Liem, Ha Noi
thanhtm@lqdtu.edu.vn

* Equal contributors † Corresponding author

***Abstract*—The Mobile capture receipts Optical Character Recognition (MC-OCR) [14] challenge deliver two tasks: Receipt Image Quality Evaluation and Key Information Extraction. In the first task, we introduce a regression model to map various inputs, for instance the probability of the output OCR, cropped text boxes, images to actual label. In the second task, we propose a stacked multi-model as a solution to solve this problem. The robust models are incorporated by image segmentation, image classification, text detection, text recognition, and text classification. Follow this solution, we can get vital tackle various noise receipt types: horizontal, skew, and blur receipt.**

***Index Terms*—Information Extraction, Optical Character Recognition - OCR, Image Quality Assessment**

I. Introduction

A. Overview

MC-OCR task is the challenge of recognizing text from structured and semi-structured receipts, invoices, bills captured by mobile devices. In general, receipts carry the information needed for various intention and purposes in various areas, for example, financial, accounting, taxation. The process of extracting information from those documents plays an important role in the streamlining of document-intensive processes and office automation in digital transformation. Our solution can be found below[1].

Mobile-Captured Image Document Recognition for Vietnamese Receipts is the process of extracting main information from receipts image captured by mobile devices. Throughout the challenge, we faced various challenging issues from the dataset, the main difficulties of those tasks are as follows:

- With the given dataset, after performing Exploratory Data Analysis (EDA), we noticed that the dataset contains a lot of noise which are images captured in various condition. In addition, the wrong annotations in the training dataset is unavoidable, including: wrong annotated polygon, missing annotations in captured image, some bounding boxes contain multiple text lines or incorrect labels for the information extraction task, and so on. All of these issues are the main reasons that making this competition more challenging.
- MC-OCR is the first competition for designing a receipts reading system which can extract information from unconstrained Vietnamese documents, captured in complex condition. To resolve this task, we propose an implementation of a pipeline that combines multiple submodels, which including image segmentation, image classification, text detection, text recognition, and text classification. This approaches is highly dependence as result of one model is affected by the output of those previous models. For example, the last step is heavily depended on the output of the previous text recognition model when training text classification model. Additionally, because of the relating information from multiple sources (both images and texts) it's challenging to develop a pipeline that can combine those sources to employ both visual feature extraction and texture information to reach a competitive result.

B. Our contributions

In this paper, we propose our methods for resolving these above problems. In summary, the main contributions of this study are described as follows:

- We propose an efficient method for data preprocessing step. This operation are designed to support clean the data and get ready for the successive steps.
- Taking advantage of various pre-trained model, we perform specific preparation for the dataset, training and

[1]https://github.com/hoainv99/mc-ocr

978-1-6654-0435-8/21/$31.00 ©2021 IEEE

fine-tuning on those data for better representation and got a higher result on related steps.
- We design an end-to-end pipeline for receipt document understanding which incorporate both image processing and natural language processing. Our solution is simple and understandable, in which can be easily extended and adapted for unseen documents.

C. Roadmap

This paper is organized as follows. Section II gives surveys of the related works. Section III introduces our proposed methodology. Section IV presents the results of the experiments and Section V concludes our paper.

II. Related Works

In this section, we review a brief introduction to related works including image segmentation, image classification, text detection, text recognition, and text classification methods that combine both.

A. Image Segmentation

Image Segmentation is a highly active research area with various efficient techniques. This including Mask R-CNN [5], which adopts the same two-stage in Faster R-CNN [10], withhold region proposal network as first stage. Mask R-CNN output parallel binary mask in the later state for each Region of Interest (RoI). Mask R-CNN is also known for the success of instance-level segmentation due to the fact of using fundamental skip connections. Our work chooses segmentation to separate main object from background to reduce complexity of following process.

B. Text Detection

Recently, a majority of text detection methods consider the text as a composition of characters. These methods prove that using a machine based on character-level text detection gained outperforms the state-of-the-art detectors. CRAFT [1] is proposed as a process with the objective is to localize each character in natural images precisely, which utilises a deep neural network to predict character regions and their affinity. With the target is focus on character-level, CRAFT is useful for data-set with low quality like low contrast-image, or curved image. Hence, we select CRAFT for Text Detection process in this work.

C. Text Recognition

Many methods of text recognition have two parts: Extractor Network and Sequence to Sequence Network. Extractor networks frequently used CNN network, for instance, EfficientNet [12], ResNet [6] to extract information from images. After that, using a sequence to sequence model as a mapping between feature extractor and target text. Method [15] presented a method based on Convolution Neural Network, Recurrent Neural Network, and a novel attention mechanism. [8] proposed a technique based on Convolution Neural Network, 2D Positional Encoding, and Transformer [13]. In this work, we used the VietOCR[2] library for the Vietnamese language, which used Transformer for the sequence to sequence model: these methods gained up to 0.88 precision for the full line of Vietnamese data-set.

D. Text Classification

Text classification is the final step to determine output label. Many approaches have been proposed with state-of-the-art performance like XLNet [17], BERT [4], PhoBERT [9], and so on. However, due to the simplicity context of the actual label, we attempt to experiment with basic classifier. For text classification procedure, we combined traditional machine learning method Support Vector Machine [3], PhoBERT and rule-based to classify the desired label. PhoBERT can understand the long context as ADDRESS and SELLER while SVM is better in recognize the total cost.

III. Methodology

A. Dataset

Dataset provided by MC-OCR organizer for this task includes nearly 2,000 images split into three-set: training set, validation set, and test set. Determined labels are quality of captured receipt and required fields need to be recognized in form box coordinate and actual texts. These fields containing information about SELLER, ADDRESS, TIMESTAMP and the real world TOTAL COST of invoices.

Through the EDA process, we confirmed that receipts, invoices, bills are captured in numerous complex condition. Taken images are not in even shape; to be specific, the dataset contains both vertical and horizontal images. Most of the pictures are taken in the vertical direction, in both downward and upward position. Furthermore, vertical images occasionally include skew, unstructured image. These can greatly reduce the performance of our selected method as mention above.

B. Data processing

To overcome the main difficulties of the dataset, we propose a heavy set of preprocessing and post-processing data to achieve the most accurate capabilities for text recognition and text classification. The major goal of data preparation is to transform input images into a standard type.

Our data process stated as below

C. Task 2 - OCR Recognition

1) Data preprocessing:

- Many images contain texts in the background which can lead to excessive/wrong focused text. We employ a segmentation process with Detectron2[3] to keep main objects segregate from unimportant context.
- A simple idea to resolve horizontal and upside-down images is calculating the ratio of width and height of those

[2]https://github.com/pbcquoc/vietocr

[3]https://github.com/facebookresearch/detectron2

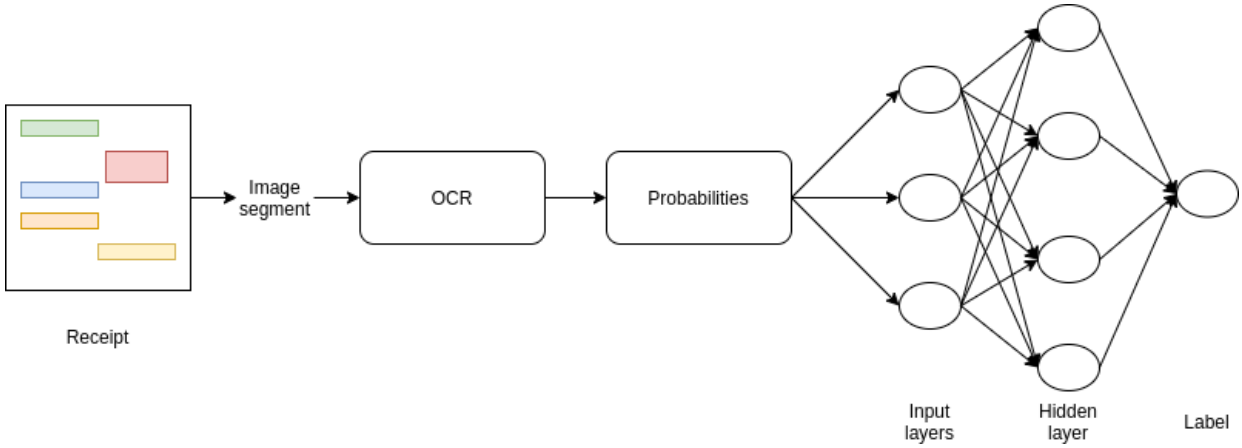

Fig. 1: Task 1 pipeline - OCR probability

images and rotate 90 degrees from horizontal images to a vertical direction. However, this solution cannot settle some edge cases when a vertical image is taken in a horizontal direction or inversely. Consequently, we came up with a more precise method by taking advantage of text detection process. In general, vertical images will carry horizontal detected box while horizontal images have the opposite type. Base on the ratio of these boxes instead of only the image shape, we can determine respectively the direction of receipts for rotation.

- We build a neural network with EfficientNet [12] to classify the upside-down images. These images then will be rotated back to an upward position for later processes. Our experiment also attempted to use text recognition from both the upside and downside of original images then compared the probabilities of each outputs. However, this approach is very time consuming and has poor performance compared to image classification.
- Last step of preprocessing is receipts image alignment. In our method, skew receipts can affect the determination of total_cost label money. Therefore, images are straightened up due to the angle of the detected text box edges to vertical lines.

2) External Data for Information Extraction:

- Information extraction data preparation. To increase the performance of extracting information in Task 2, we apply custom processed data from the original dataset. Fine-tuning on domain data can provide a robust result, hence we re-train VietOCR on the provided input dataset.
- Key Information Extraction (K.I.E) task will include text classification and construct data input. Utilizing previous text detection, we generate external label OTHER among all non-labelled text box. To be specific, every detected box will be calculated with IoU score concerning the true label. This process aims to remain the annotation while generating new OTHER labels.

3) Data post processing:

- For post-processing operation, text correction is applied due to the instability of single character in VietOCR output.

D. Task 1 - Receipt Image Quality Evaluation

The first task is to evaluate the quality of the invoice through image quality. Receipt image quality is measured by the ratio of text lines associated with the “clear”. The quality ranges from 0 to 1, in which a score of 1 means the highest quality and a score of 0 means the lowest quality. Further investigation, we realized that the probability of model OCR prediction was closely related to image quality. After processing the image, given a receipt with N image segments that output of CRAFT, its representation is denoted by $S = [s_1, s_2, ..., s_N]$, where s_i is image cropped contain texts. Then, we forward its throw model OCR and receipt probability of each sentences, given a input as $P = [p_1, p_2, ..., p_N]$. We observe that the amount of probability that has a greater value than 0.9 is so many. This can potentially make noisy input, therefore we sort the ascending order value of probabilities and take the first 100 elements as an input of the regression model to map with the label.

E. Task 2 - Optical Character Recognition

The second task involves recognizing required fields of the receipt, including TOTAL COST (both text label and money cost number), SELLER, ADDRESS and TIMESTAMP. These information will be recognized from extracted text box output from CRAFT to classify the destined fields.

In the initial attempt of text classification, SVM is chosen to be the classifier. We experiment TF-IDF [11] for word and character vectorization. However, the empirical experiment of SVM shows limited in classifying the specific type of text; for instance, misunderstanding date-time type (TIMESTAMP) or wrong SELLER label (which containing address). Therefore, we employ multiple model architectures to reach better performance.

Figure 2 illustrate our overall pipeline for MC-OCR task 2. For SELLER and ADDRESS type, we survey the capabilities of PhoBERT and prove to be more efficient with this type

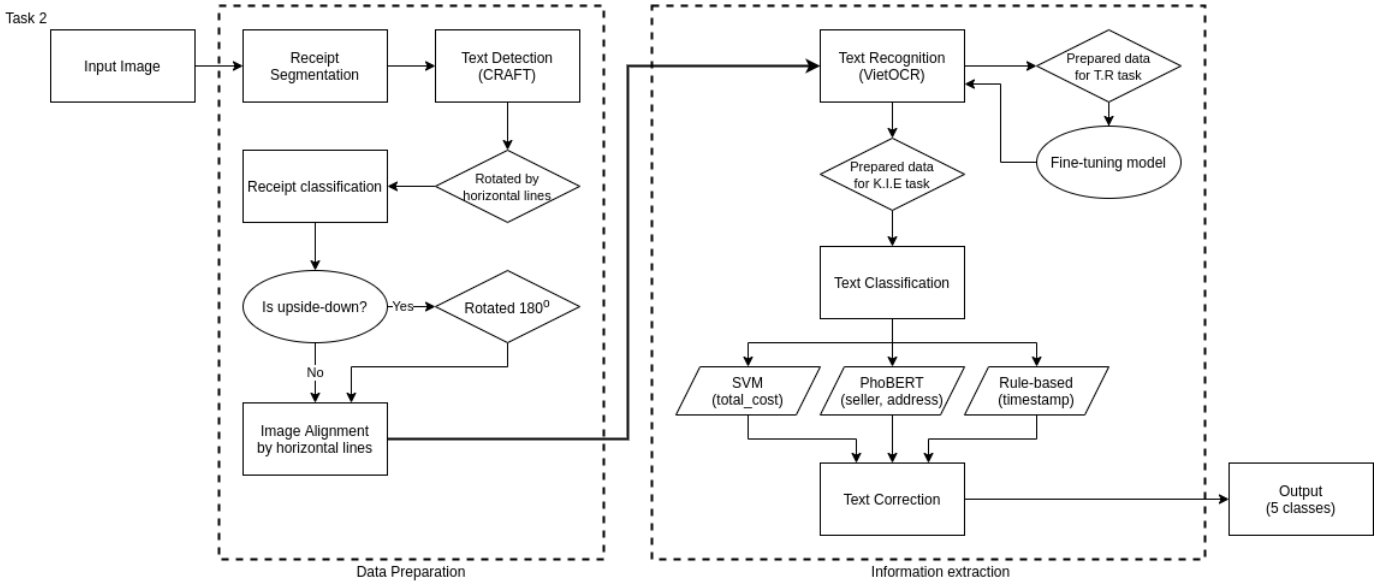

Fig. 2: Task 2 pipeline

of data. With TIMESTAMP data, this contains specific data type, therefore rule-based (Regular Expression) is chosen for ultimate performance. We use Regular Expression to find date time type from output of VietOCR by validating date format ddmmyy (separated by or other character type) and time format. Ultimately, the SVM is used to classify TOTAL COST and OTHER label, which we generated with the previous data preparation. In TOTAL COST label, we only keep the text and remove the number, which is the total cost money, for better classification. This number can be determined based on the position of invoices after we successfully discover the total cost text.

IV. Experiment

In this section, we describe different experimental results of our method.

A. Analysis

In the first task, we use multi-layer perceptron as main method. Given an input as $P = [p_1, ..., p_N]$, where p_i is the probability of each text-box that our model recognized.

In the second task, after our data process, we receive clean and standard images. Then, we forward it through CRAFT to get bounding boxes of text-line in the receipt. Next, we fed cropped-boxes into domain fine-tuned VietOCR. Finally, we used PhoBERT and SVM models to extract key information include seller, address, time-stamp, and total-cost from outputs of VietOCR.

As mentioned above in EDA process, we realize that each class has specific distribution position. These information can greatly impact the performance of overall pipeline. Therefore, we deploy each best methods based on the selected regions. With SELLER and ADDRESS generally would appear on top of the receipts, we use PhoBERT on first 10 detected text boxes to for classification (SELLER, ADDRESS and OTHER).

Finally, TOTAL COST (text type) are classified with SVM classifier. The corresponding money of TOTAL COST are determined base on position of center of boxes respect to TOTAL COST text.

B. System Configuration

Our experiments are conducted on a computer with Intel Core i7 9700K Turbo 4.9GHz, 32GB of RAM, GPU GeForce GTX 2080Ti, and 1TB SSD hard disk.

Training/Testing Data[4] are provided 1,155 training examples with the respective annotated information. The testing set consists of 391 examples without annotations.

C. Results

Comparison of our experimented results are illustrated in Table I and Table II.

TABLE I: Result table

Method	Public Test	Private Test
CNN-base	0.142	0.156
CNN + OCR probability	0.139	0.152
OCR probability	**0.127**	**0.147**

TABLE II: Result table

Text Classification	Public Test	Private Test
SVM	0.37	NA
PhoBERT	0.35	NA
SVM + PhoBERT + Rule base	**0.29**	**0.259**

We employed many different approaches to explore the best method. The final results of Task 1 are illustrated in Table

[4]http://bit.ly/mcocr2021_public_traintest

I. Theoretically, we expect CNN-base can perform well due to the input image type and taking advantage of regression input. However, in practical, OCR-based method shows greater result.

For the second task, we only compared the outputs of the best three trial methods in the public test score. Both SVM and PhoBERT have their own advantages in the specific type of data. By combining all the trial methods, we end up a score of 0.29 on public test and the final model reached 0.259 on the private test.

D. Final Score

With our proposed method, we finished MC-OCR Challenge 2021 with respectable result. For task 1, we ranked 3rd on public test set and 6th on private test set. With task 2, our solution manage to reach 2nd place on public test set and 5th position on private test set.

V. Conclusion

A. Summary

In this paper, we have presented a pipeline solution for MC-OCR tasks challenges. The proposed technique demonstrates the ability to overcome the main difficulties of the competition and possible issues in real-world data type. However, our method still contains has some limitations. Text correction only works for selected domain data, hence, in the real-world problem would decrease accuracy. Our solution also not make use of specific informative details (for example: receipt overall layout) for Task 2. Despite these limitations, our work points towards the two tasks as an intriguing, useful method for this challenge.

B. Future work

Receipts character recognition is an interesting, challenging and potential task. We expect to expand further our work in multiple research direction including improvement of text classification; Post processing will not help if facing new domain data, therefore, we wish to experiment more effective method.

Other promising approaches namely Graph Convolutional Network [7], Graph Neural Network [2] or extracted by layout [16] are considered for further investigation.

Acknowledgment

This work is partially supported by ***Sun-Asterisk Inc***. We would like to thank our colleagues at ***Sun-Asterisk Inc*** for their advice and expertise. Without their support, this experiment would not have been accomplished.

References

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection, 2019.

[2] Yihao Chen, Xin Tang, Xianbiao Qi, Chun-Guang Li, and Rong Xiao. Learning graph normalization for graph neural networks, 2020.

[3] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[7] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2016.

[8] Junyeop Lee, Sungrae Park, Jeonghun Baek, Seong Joon Oh, Seonghyeon Kim, and Hwalsuk Lee. On recognizing texts of arbitrary shapes with 2d self-attention, 2019.

[9] Dat Quoc Nguyen and Anh Tuan Nguyen. Phobert: Pre-trained language models for vietnamese, 2020.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015.

[11] Claude Sammut and Geoffrey I. Webb, editors. *TF–IDF*, pages 986–987. Springer US, Boston, MA, 2010.

[12] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2019.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[14] Xuan-Son Vu, Quang-Anh Bui, Nhu-Van Nguyen, Thi-Tuyet-Hai Nguyen, and Thanh Vu. Mc-ocr challenge: Mobile-captured image document recognition for vietnamese receipts. In *Proceedings of the 15th IEEE-RIVF International Conference on Computing and Communication Technologies*, RIVF '21. IEEE, 2021.

[15] Zbigniew Wojna, Alex Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz. Attention-based extraction of structured information from street view imagery, 2017.

[16] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding, 2020.

[17] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2019.