MC-OCR Challenge: Mobile-Captured Image Document Recognition for Vietnamese Receipts

Xuan-Son Vu*, Quang-Anh Bui†, Nhu-Van Nguyen§, Thi Tuyet Hai Nguyen‡, Thanh Vu**
*Department of Computing Science, Umeå University, Sweden. Email: sonvx@cs.umu.se
†PurchEase company, Paris, France. Email: anh@PurchEase.com
‡La Rochelle University, France. Email: hai.nguyen@univ-lr.fr
§INSA-Lyon, France. Email: vincent.nguyen@insa-lyon.fr
**Oracle Corporation, Melbourne, Australia. Email: thanh.v.vu@oracle.com

***Abstract*—The paper describes the organisation of the "Mobile Captured Receipt Recognition Challenge" (MC-OCR) task at the RIVF conference 2021 [1] on recognizing the fine-grained information in Vietnamese receipts captured using mobile devices. The task is organized as a multi-tasking model on a dataset containing 2,436 Vietnamese receipts. The participants were challenged to build a model that is capable of (1) predicting receipt's quality based on readable information, and (2) recognizing textual information of four required information (i.e., "SELLER", "SELLER_ADDRESS", "TIMESTAMP", and "TOTAL_COST") in the receipts. MC-OCR challenge happened in one month and top winners of each task will present their solutions at RIVF 2021. Participants were competing on CodaLab.Org from 05th December 2020 to 23rd January 2021. All participants with valid submitted results were encouraged to submit their papers. Within one month, the challenge has attracted 105 participants and recorded about 1,285 submission entries.**

I. Introduction

Mobile captured receipt recognition (MC-OCR) is a process of recognizing text from structured and semi-structured receipts and invoices in general captured by mobile devices. This process plays a critical role in the streamlining of document-intensive processes and office automation in many financial, accounting and taxation areas. However, MC-OCR faces big challenges due to the complexity of mobile captured images. First, receipts might be crumpled or the content might be blurred. Second, the quality of photos taken with mobile devices is very diverse because of the light condition and the dynamic environment (e.g., in-door, out-door, complex background, etc.), where the receipts are captured. These issues result in a low quality of recognized information. To address the problem, in this challenge, we target at two tasks including (1) image quality assessment (IQA) of the captured receipt, and (2) key information extraction (KIE) of required fields. Figure 1 briefly shows an example of two tasks. The shared task consists of three phases namely *Warm Up*, *Public Test*, *Private Test*, which was hosted on Codalab from Dec 05, 2020 to Jan 23rd, 2021.

[1]https://rivf2021-mc-ocr.vietnlp.com/, see *dataset* tab for the download information.

This shared task has the following main contributions. First, this shared task provides an evaluation framework for quality evaluation and key information extraction tasks of Vietnamese receipts, which were captured by mobile devices. Thanks to the benchmark dataset, all participants could leverage and compare their innovative models on the same dataset. The insights from different proposed methods may help improve the digitalization process of Vietnamese documents. Second, it is very valuable that MC-OCR challenge provides a novel dataset for both tasks of mobile captured devices. There was a well-known data challenge called SROIE [2], but participants were given scanned receipts and there was no image quality assessment task. The MC-OCR dataset is built based on a novel data annotation process with the use of both human and model-base methods to produce 2,436 mobile captured receipts in Vietnamese. We hope this dataset will be a useful benchmark for further research in related fields. In this shared task, RMSE and CER are utilized as evaluation metrics of the IQA and KIE tasks, respectively.

The remainder of the paper is organized as follows. The next section describes the data collection and annotation methodologies. The shared task description and evaluation are summarized in Section 3. Section 4 describes the competition, approaches and respective results. Finally, Section 5 concludes the paper following with discussions on potential applications for future studies and challenges.

II. Data Collection and Annotation

Figure 2 describes the whole process of raw data collection, annotation, and final data preparation for the MC-OCR data challenge.

TABLE I: Simple statistics of the MC-OCR dataset with the number of receipt images for each corresponding phase.

	Warm-Up	Public Training	Public Testing	Private Testing
#	500	1,155	391	390

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

[2]https://rrc.cvc.uab.es/?ch=13

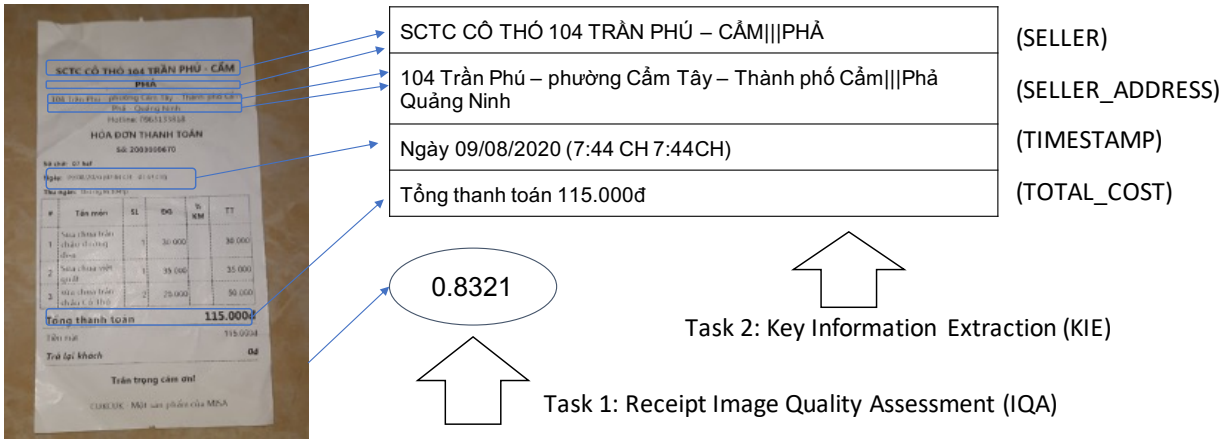


Fig. 1: An example of the receipt recognition process. The challenge consists of two tasks: receipt image quality assessment (IQA) task and key information extraction (KIE) task. It is noted that three symbols "|||" are used as a line separator and they are only used for pre-possessing step before performing the evaluation process.

A. Data Collection

2,436 receipts were contributed by nearly 50 active data collectors over two months. The data collectors were instructed to ask their friends or to see if there were readable receipts. Then they took a photo of the receipt using their mobile phone. To get the original resolution of all receipt images, we deployed an uploading service and asked all data collectors to submit their images.

B. Data Annotation

Annotation Process. The annotation process consists of two phases: receipt image quality phase (phase 1) and key information extraction phase (phase 2). Phase 1 has three main steps: detect textline images (step 1.1) for manual annotation (step 1.2). Then annotated texts of textline images are used for training a transformer-based OCR model, to later be used to infer quality score of receipts (step 1.3). For the phase 2, all steps are manually annotated by two groups of annotators, who are all between the ages of 20 and 40. First group of annotators was asked to annotate polygon regions of the four required fields. The second group was asked to check every annotated region on item-by-item basic to annotate textual information of each region.

Task 1: Image Quality Assessment (IQA). The receipt's image quality is measured by the ratio of clear text lines over the total number of text lines evaluated by a semi-automatic method using both human annotators and an advanced neural model. The quality ranges from 0 to 1, in which 1 means the highest quality and 0 means the lowest quality.

Task 2: Key Information Extraction (KIE). At maximum, one receipt image has 4 key fields provided by human annotators. Based on different receipt's formats, the number of text lines might be different as some receipts do not have all fields. For instance, the SELLER_ADDRESS field might not exist in the receipt or simply, because the line is not readable.

C. Data Pre-processing and Data Format

Raw receipts and annotated texts are provided in *as-is* manner, no pre-processing was performed. Regarding data format, we provide a folder of raw receipts and a '.csv file' containing meta-data information. Meta-data information consists of annotated text fields and the image quality score.

D. Result Submission

Participants are required to submit predicted results in the same order as the testing set in the following format:

```
img_id, anno_texts, anno_image_quality
id1, <text>, <image_quality>
id2, <text>, <image_quality>
...
```

III. Shared Task Description and Evaluation

In this shared task, participants are challenged to build two models or a multi-model to (1) quantify receipt's quality and (2) recognize required text lines.

(Task 1 - IQA). Evaluation metric of IQA task is based on root-mean-square-error (RSME) metric:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

where $\hat{y}_i$ is predicted value, y_i is expected value of the receipt i^{th}, N is the total number of receipt in the test set.

(Task 2 - KIE). Evaluation metric of KIE task is formulated based on *character error rate*. First, *Levenshtein distance* [1] of all fields are computed. Then, the normalized score of *the Levenshtein distance* of all key information is calculated as the final score for the test set. In details, the CER score is calculated as follows:

$$Character\ Error\ Rate\ (CER) = \frac{1}{L}\sum_{i=1}^{N}(i + s + d)$$

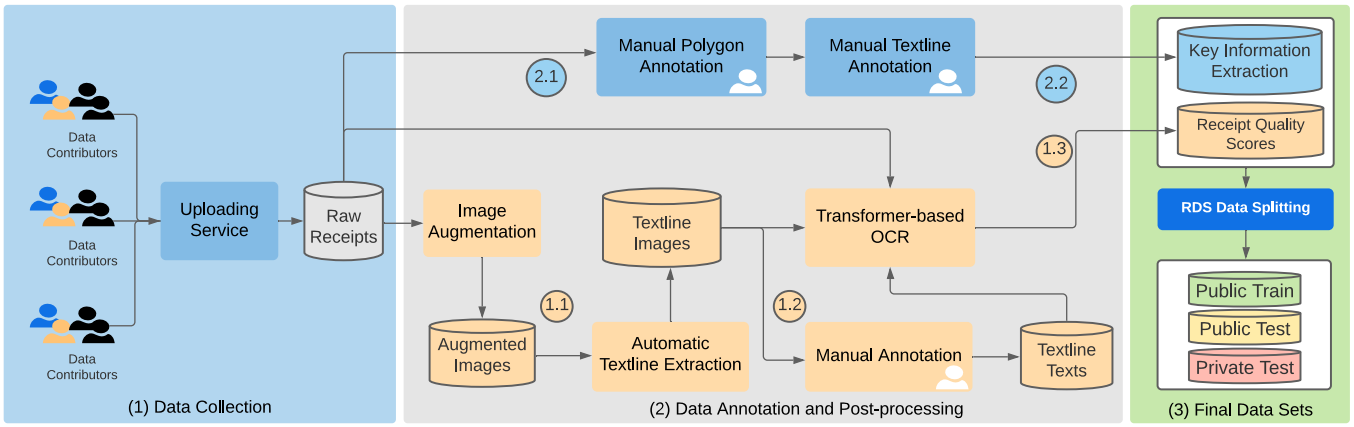


Fig. 2: Data Collection, Annotation Process, and Data Splitting Approach of MC-OCR Data Challenge for two tasks: image quality assessment - IQA (task1), and key information extraction - KIE (task2).

where $L = \sum_{i=1}^{N}(l_i)$ is the total length of all reference texts of the test set, l_i is the length of i^{th} document, N is the total number of test samples. And $(i + s + d)$ is the *Levenshtein distance*, in which, i, s, and d correspond to the minimal number of character insertions, substitutions, deletions required to transform the reference text into the OCR output.

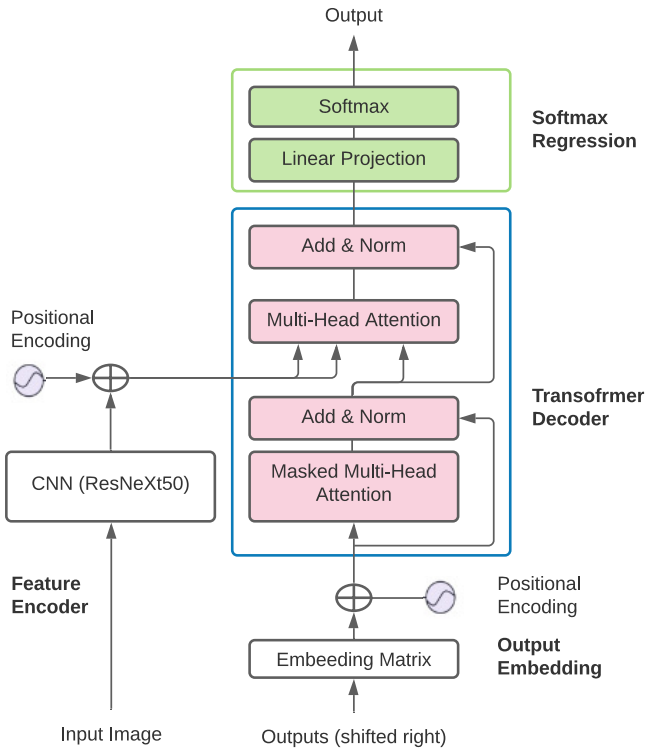


Fig. 3: Transformer-based OCR model of Li et al. [2] for learning on augmented receipts, and be used to infer quality assessment score on raw receipts for Task 1.

A. Semi-automatic annotation for IQA task based on Transformer-OCR

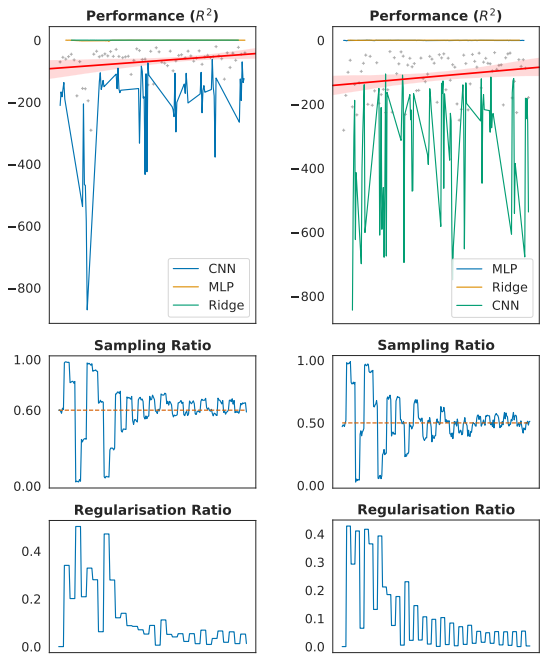
Given the fact that the IQA task is a regression task, it is not practical to simply ask for a direct assessment from annotators. For instance, different annotators might have different eyesight, or concentration, and they will give subjective labeling information during the annotation process. Moreover, *ambiguous evidence* [3] is yet another concern due to different annotators have different domain knowledge, which will affect their justification on whether a certain image is of high quality or low quality. Therefore, we introduce a novel method to produce the quality score for the IQA task in a systematic way with human-in-the-loop. First, we produce another set of receipt images by applying different synthetic approaches (e.g., scale, changing light, colors) on the raw images. Then, these images are used for line detection (step 1.1 in Figure 2). All these line images are used for manual annotations (step 1.2 in Figure 2). We then train a transformer-based OCR (see Figure 3) for recognizing text at line-image level. Finally, the trained OCR model is used to recognize raw receipts, to produce the quality score of receipts based on recognition's confidence of all textlines.

B. Data Splitting

Data splitting for data challenge is a difficult process due to the potential issues of evidence ambiguity [3] and concept drifting [5]. These are the main causes of the unstable ranking issue in data challenges [4, 6].

1) Baselines: To apply RDS [4] for the data splitting process, it requires to have baseline learning models to obtain rewards for the reinforced process. Similar to [4], following baselines are used as base-learners:

1) **CNN** (convolutional neural network) is used most commonly in analyzing visual imagery. The CNN network here consists of 1 embedding layer, 1 Conv2D layer, 1 MaxPool2D layer, and an output layer of one neuron unit. No activation function is used for the output layer



(a) Training Set (b) Val & Test Sets

Fig. 4: Learning Dynamics for splitting data into 3 sets (public training, public testing, and private testing) using RDS Stochastic Choice Reward Mechanism [4].

to form the regression task, and the model predicts numerical values directly without transformation.

2) **MLP** is a *vanilla* neural networks. The MLP network here consists of 3 fully connected layers with the number of units as 512, 256, 128, respectively. The activation function was ReLU [7]. Similar to the above CNN model, the output layer has one neuron unit, and no activation function is used for the output layer to form the regression task.
3) **Ridge Regression** [8] is a method of regularization of ill-posed problems. It estimates the coefficients of multiple-regression models in scenarios where independent variables are highly correlated. Ridge regression is chosen as a representative of the conventional feature-based machine learning approach to increase the diversity among base-learners.

It is noted that representation of all receipts is extracted based on textual information of raw receipts. As described in the Figure 2, all textline images are annotated manually by human annotators. These textual information is used to extract representation for each receipt using Fasttext-Sent2Vec [9] with the pre-trained model provided in ETNLP[3] [10].

Since RDS only supports splitting one dataset into two subsets at a time. We have to run the splitting process two times. First, the full dataset is feed into RDS to split the data into two subsets with a ratio of 6:4. In which, 60% of the data will be used for the public training set. Next, 40% of the full data is feed into RDS to split into two subsets with the ratio of 50%, to form the public testing and private testing sets. Figure 4 shows the learning dynamic of RDS for data splitting process.

[3]https://github.com/vietnlp/etnlp

IV. Participants and Results

A. Participation

Table II summaries the approach and the results obtained by the 6 teams submitted papers describing their approach for the final assessment. In particular, main approaches of the teams with the paper reports are summarized as follows:

- **DataMining VC** focused on Task 2 (KIE). Instead of using the normal pipeline (i.e., text detection, text recognition, information extraction), they proposed to use an information detection step followed by an OCR step. The information detection step includes text-block localization and text-block classification. This approach is well suited for Task 2 and they reached the *first* rank in the challenge with a CNER score of 0.22 on Task 2 private test data.
- **SDSV_AICR** implemented three approaches to Task 1 including detecting the blurry of an image, averaging the confidence scores returned by a text detector (PaddleOCR [11]), and training a regression model with depth-wise separable convolution (DSC). The DSC-based regression model achieved a RMSE score of 0.12 on Task 1 private test data. For Task 2, the authors proposed a pipeline system consisting of five components: a text detector using PaddleOCR, a rotation corrector (Mobilenet v3 [12]), a text line rotator, a text recognizer (VietOCR), and a key information extractor (a GCN-based model, PICK [13]). The pipeline system produced a CER score of 0.23 on Task 2 private test data.
- **SUN-AI** proposed a regression model based on multi-layer perceptron to handle Task 1. Specifically, a receipt image is first segmented using CRAFT [14], then text is recognized using VietOCR [15]. The 100 lowest confidence scores produced by VietOCR are used as the input features to train the regression model. The approach produced a RMSE socre of 0.15 on Task 1 private test data. Regarding Task 2, using the text extracted from Task 1, the authors employed SVM for TOTAL_COST, PhoBERT [16] for SELLER and ADDRESS, and a rule-based approach for TIMESTAMP, achieving a CER score of 0.26 on Task 2 private test data.
- **Tung-nguyen** employed Faster R-CNN [17] to detect the information location, and then utilized the TransformerOCR, a combination of CNN and Transformer to recognize text. With the proposed approach, the authors achieved the CER score of 0.32 on Task 2 private test data.
- **UIT_CS_AIClub** used the EfficientNet [18] architecture to train the quality model. Receipt recognition is based on four steps: preprocessing, text detection (PAN model [19]), text recognition (VietOCR) and structured

TABLE II: Top 6 teams on private test data with submitted papers describing their final approaches. The table presents main approaches of each team and the RMSE (Task 1) and CER (Task 2) scores on the private test data (in random order).

ID	Team	Method	Task 1 (RMSE)	Task 2 (CER)
52	**DataMining VC**	- Task 1: Patch Sifting - Task2: Yolov5 + VietOCR	0.15	0.22
83	**SDSV_AICR**	- Task 1: Mobilenet V1 - Task 2: PaddleOCR + Mobilenet v3 + VietOCR + PICK	0.12	0.23
58	**SUN-AI**	- Task 1: CRAFT + OCR prob - Task 2: Craft + VietOCR + SVM + PhoBERT + Rule base	0.15	0.26
36	**Tung-nguyen**	- Task 2: Faster R-CNN + TransformerOCR	-	0.32
50	**UIT_CS_AIClub**	- Task 1: EfficientNet - Task 2: Faster R-CNN + EfficientNet-B4 + PAN + VietOCR	0.10	0.30
72	**BK_OCR**	- Task 1: VGG-16 - Task 2: CRAFT + VietOCR + NLP	0.11	0.39

information extraction. For structured information extraction, they utilized a rule-based method to classify the OCR text into one of the target classes. The approach produced a RMSE score of 0.10 and a CER of 0.30 on Task 1 and Task 2 private test data, respectively.

- **BK_OCR** utilized VGG-16 [20] to train a regression model to predict the quality of the captured receipts achieving a RMSE score of 0.11 on Task 1 private test data. For Task 2, the authors proposed a pipeline system in which the text location was detected using CRAFT, the text content was recognized using VietOCR, and predefined information was extracted using a rule-based approach. The proposed pipeline system produced a CER score of 0.39 on Task 2 private test data.

B. Outcomes

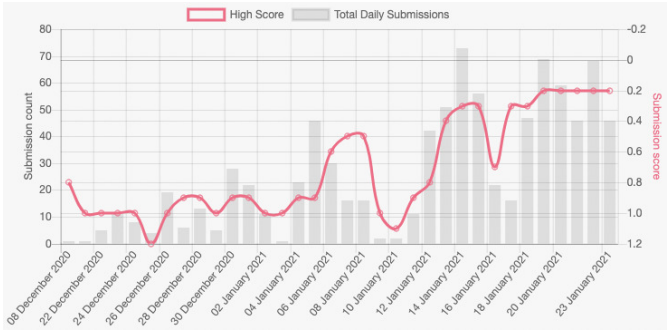


Fig. 5: Number of submissions during the challenge time of MC-OCR 2021 on Codalab.Org.

Table III shows main statistics of the challenge. In total, there were 105 registered participants, of which some participants jointly formed 9 teams. Meanwhile, the rest competed as individuals. It is noted that, only 16 participants signed the corpus user agreement to officially compete for the prizes. Other participants joined with valid submissions for practicing purposes only, and they did not compete for the prizes.

Regarding results on two tasks, there were 1,285 successful entries on both public and private test data. The main statistics of submissions and scores during the challenge time are reported in Figure 5. The best performance on the public and private test data are respectively 0.10 (RMSE) and 0.22 (CER) for Task 1 and Task 2.

Tables IV and V summarize results of task 1 and task 2, respectively. From these two tables, we can see following insights. First, it is clear that the use of RDS data splitting approach helps to balance the data distribution accross different sets. For both tasks, there are no significant different gaps between the best score on the public test and private test. Second, Table V shows that KIE task for Vietnamese receipts is not an easy task. The overall CER score is around 22%, which means that the research community would need to investigate more into this problem, to better facilitate the digital transformation of Vietnam.

TABLE III: Participation summary of the MC-OCR challenge

Metric	Value
Number of registered participants	105
Number of signed agreements	16
Number of submitted papers	6

TABLE IV: Results summary of Task 1 (IQA)

	Public Test	Private Test	Overall
Total Entries	640	161	801
Best RMSE	0.10	0.10	0.10
Mean RMSE	0.22	0.17	0.20
Std. RMSE	0.21	0.07	0.17

V. Conclusions

In this paper, we have presented an overview of the MC-OCR challenge "Mobile-captured image document recognition for Vietnamese receipts": (1) provide details of the tasks, data preparation process and the task organization, and (2)

TABLE V: Results summary of Task 2 (KIE)

	Public Test	**Private Test**	**Overall**
Total Entries	645	161	806
Best CER	0.24	0.22	0.22
Mean CER	0.73	0.34	0.56
Std. CER	0.29	0.18	0.32

report the results obtained by the top participating teams and their adopted approaches. Digital transformation is a very important process in the development of any country. It plays a vital role in transforming the business process into a smart integration of digital services. This data challenge introduces a novel dataset that covers a great number of receipt images, which are the key to the automation process in document process and accounting. The annotation process with a mixture of both systematic model-based approaches with human-in-the-loop helps to establish an important dataset for future research in automatic document processing in Vietnam. We believe the dataset will encourage researchers and machine learning practitioners to contribute their knowledge for the image document recognition task in Vietnam.

ACKNOWLEDGMENT

The authors would like to thank the DopikAI Technology Company (http://DoPik.AI), the AiViVN team, and the fourteen annotators for their hard work to support the shared task. Without their support, the task would not have been possible.

REFERENCES

[1] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, Feb. 1966.

[2] B. Li, X. Tang, X. Qi, Y. Chen, and R. Xiao, "Hamming ocr: A locality sensitive hashing neural network for scene text recognition," 2020.

[3] S. Lee, S. Purushwalkam, M. Cogswell, V. Ranjan, D. Crandall, and D. Batra, "Stochastic multiple choice learning for training diverse deep ensembles," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 2127–2135.

[4] H. D. Nguyen, X.-S. Vu, Q.-T. Truong, and D.-T. Le, "Reinforced data sampling for model diversification," 2020.

[5] S. Bach and M. Maloof, "A bayesian approach to concept drift," in *Advances in Neural Information Processing Systems*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., vol. 23. Curran Associates, Inc., 2010.

[6] D.-T. Le, X.-S. Vu, N.-D. To, H.-Q. Nguyen, T.-T. Nguyen, T. K.-L. Le, A.-T. Nguyen, M.-D. Hoang, N. Le, H. Nguyen, and H. D. Nguyen, "Reintel: A multimodal data challenge for responsible information identification on social network sites," in *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*. Hanoi, Vietnam: Association for Computational Lingustics, December 2020, pp. 84–91.

[7] A. F. Agarap, "Deep learning using rectified linear units (relu)," *CoRR*, vol. abs/1803.08375, 2018.

[8] D. E. Hilt and D. W. Seegrist, "Ridge: a computer program for calculating ridge regression estimates," 1977.

[9] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features," in *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

[10] X.-S. Vu, T. Vu, S. N. Tran, and L. Jiang, "Etnlp: A visual-aided systematic approach to select pre-trained embeddings for a downstream task," in *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, 2019.

[11] Y. Du, C. Li, R. Guo, X. Yin, W. Liu, J. Zhou, Y. Bai, Z. Yu, Y. Yang, Q. Dang, and H. Wang, "Pp-ocr: A practical ultra lightweight ocr system," 2020.

[12] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for mobilenetv3," *CoRR*, vol. abs/1905.02244, 2019.

[13] W. Yu, N. Lu, X. Qi, P. Gong, and R. Xiao, "Pick: Processing key information extraction from documents using improved graph learning-convolutional networks," 2020.

[14] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," *CoRR*, vol. abs/1904.01941, 2019.

[15] C. Q. B. Pham, "Vietocr," https://github.com/pbcquoc/vietocr, 2020.

[16] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1037–1042.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[18] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114.

[19] W. Wang, E. Xie, X. Song, Y. Zang, W. Wang, T. Lu, G. Yu, and C. Shen, "Efficient and accurate arbitrary-shaped text detection with pixel aggregation network," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8439–8448.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.