12th International Conference on Computer and Knowledge Engineering (ICCKE 2022), November 17-18, 2022, Ferdowsi University of Mashhad, Iran

MultiPath ViT OCR: A Lightweight Visual Transformer-based License Plate Optical Character Recognition

Alireza Azadbakht, Saeed Reza Kheradpisheh, Hadi Farahani
Department of Computer and Data Sciences, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran
al.azadbakht@mail.sbu.ac.ir, {s_kheradpisheh, h_farahani}@sbu.ac.ir

2022 12th International Conference on Computer and Knowledge Engineering (ICCKE) | 978-1-6654-7613-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCKE57176.2022.9960026

***Abstract*—Because of the natural conditions of license plate images, the Optical Character Recognition (OCR) of these images is generally a challenging problem. OCR systems are utilized in edge devices with limited computation power. Despite the considerable progress of deep neural networks, state-of-the-art models are not always an excellent solution to this problem. Most models have many parameters, and in practice, they need many resources to train, maintain and implement on edge devices. We propose a lightweight model based on Visual Transformer architecture and achieve competitive results against traditional CRNN models. Due to the lack of a rich and large-scale dataset for Persian license plates, we gathered and annotated 1.3M images of license plates in various natural conditions from different points of view and different cameras. We call this dataset as LicenseNet. Our proposed model achieves 77.25% accuracy against CNN models with 75.18% accuracy and embedded OCR models in cameras with 60.37% accuracy on the LicenseNet test set. Furthermore, we achieved better accuracy with 3.21 times fewer training parameters than previously proposed models.**

***Index Terms*—Vision, Visual Transformer, Optical Character Recognition, License Plate OCR.**

I. Introduction

Optical Character Recognition (OCR) of the license plate is a crucial task used in traffic management, digital security surveillance, vehicle tracking, car ticketing, and parking management in modern cities. Developing an OCR system is not a simple task, and the system should be robust to many factors such as blurry images, poor lighting conditions, variability of license plates colors and special characters, physical impacts, possible frauds, and weather conditions such as rainy and snowy weathers [1].

A robust OCR system needs to cope with various environments while maintaining a high level of accuracy. In other words, this system should work well in natural conditions [1], [2].

As the field of computer vision grows and advanced deep learning architectures are introduced, the accuracy and confidence of deep learning models for tasks like image classification, object detection, and text recognition are improving daily. However, the downside of these advances is that deep learning models usually need advanced computation resources and memory units to perform inference in real time. They can not be implemented on small chips such as Raspberry PI hardware. These limitations prevent the companies and researchers from using the full potential of state-of-the-art deep learning models in practice [3].

Visual Transformers (ViT) is a novel deep learning model in the computer vision research community. The Transformer architecture was initially introduced in Natural Language Processing. After achieving competitive results in comparison with previous models like Recurrent Neural Networks, it gained much attention in the research community [4],[5],[6],[7]. After Dosovitskiy et al. [8] introduced a novel model based on Transformers for computer vision and achieving state-of-the-art accuracy in many vision tasks and large-scale datasets, many researchers tended to work on this model to find the capability of Transformers and self-attention mechanism in computer vision tasks [9],[10].

The ViT model is a simpler variation of Transformers that only uses encoder modules and one multilayer perceptron on the top of a stack of encoders. This model uses the power of the self-attention mechanism [7] to replicate some properties like translational invariance of Convolution Neural Networks (CNN). The self-attention mechanism in ViT architecture handles translation and permutation invariance in images and helps the final multilayer perceptron to have an even better and more accurate view of the input image in comparison with convolutional layers' features and finally helps the whole model to achieve better decision boundary for the given task [11]. The full potential of ViT unlocks with large-scale datasets [8].

In this paper, due to the lack of a rich, large-scale license plate dataset, we first build the LicenseNet dataset, which is a rich and large-scale license plate dataset consisting of 1,300,000 images. After that, we focus on training a small ViT model on a large number of data points and exploit the permutation invariance properties of this architecture to account for various problems in captured license plate images, such as different angles of license plates in image and deformations and frauds in images that CNN models have difficulties in addressing, and in the meantime we keep the model lightweight so we could use this model on edge devices with limited computation power.

Our proposed model has 3.21 times fewer training parameters than previously proposed CNN-based models, and

it achieves 77.25% accuracy on the LicenseNet test set. For comparison, we chose Convolutional Recurrent Neural Network (CRNN) model [12] and trained it on two subsets of the LicenseNet dataset. The CRNN model is extensively used in industrial use cases. Even today, this architecture is used as a solution to many vision tasks that tend to extract information from images in an iterative manner, like extracting text from natural images and OCR of scanned documents. The 100k subset CRNN model achieves 65.31% accuracy, and MultiPath ViT achieves less accuracy of 63.54%, but on the 1M dataset, MultiPath ViT outperforms the CRNN model with 2.07% better accuracy.

The structure of this paper is as follows, in section II, we introduce the MultiPath ViT model and LicenseNet dataset, and we explain different modules in detail. In section III, we explain implementation details and training hyperparameters and compare the model's performance against the previously employed models.

II. Method

This section proposes a new way of thinking in OCR systems and a novel deep learning model based on ViT architecture for license plate OCR.

Previous OCR deep models generally use a set of convolution layers to extract useful features from license plate images and a set of recurrent layers to generate characters in an iterative manner [12]. This kind of architecture uses many computation resources. Usually, it has many parameters, and even inference time is so high that they are not useable in real time. Usually, these models can not be implemented on edge devices.

Another problem with this way of thinking is that when we use iteration to generate characters, it is possible to generate fewer or more characters than the standard format, and due to the gradient vanishing property of these kinds of recurrent layers, it is not feasible to make this part bigger [13]. So generally, these models are very biased to location, angle, and properties of license plates in the training dataset input images. In practice, for example, changing one surveillance camera could dramatically decrease recognition accuracy. In order to resolve this problem, we can use the online learning framework. However, due to a large number of parameters, we need more computation, which is not appropriate for edge devices.

Generally speaking, the number of parameters in a deep learning model directly relates to the model's computation, so we can say that if a model has a small number of parameters, it is more lightweight and needs less computation power. It is much more suitable for being implemented on edge devices and small kits. It is worth mentioning that on some small kits or embedded systems, managing the overall power consumption of a device is a challenging problem, and lightweight models have less power consumption, so naturally, They can be helpful in these kinds of situations as well.

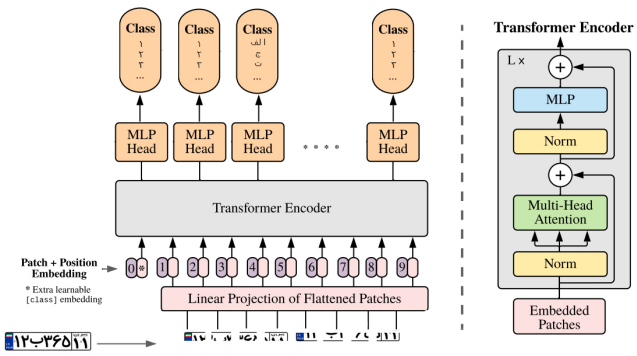

Fig. 1. MultiPath Visual Transformer OCR architecture.

A. MultiPath Visual Transformer

In this section, a ViT model for Persian license plate OCR will be thoroughly introduced. MultiPath ViT OCR, like standard ViT, consists of two modules. First, a stack of transformer's encoders tries to extract rich features from patches of the input image, and second on top of encoder modules, a MultiPath module with N number of multilayer perceptron with a different number of layers tries to perform a classification task of finding each character of license plate. N is the number of characters in the standard license plate; in standard Persian license plates, there are eight characters (7 digits and one Persian character). Our model has eight paths on top of the encoder's stack (see Fig. 1).

In order to bypass the iterative generation of characters and avoid recurrent layers, each path in the MultiPath module is responsible for recognizing one specific position of the character in a license plate. For example, the third character in a Persian license plate has 26 distinct Persian characters, so the third path of the MultiPath module does the task of classification on 26 possible characters. In other words, the last layer of the multilayer perceptron has 26 neurons. Each path computes its classification loss, and the summation of each path's loss backpropagates on the whole network.

B. LicenseNet Dataset

This study aims to train a lightweight OCR; naturally, ViTs are data-hungry models. Official and most computer vision benchmarks like the ImageNet dataset have approximately 1 million images in them [14]. Due to the lack of a rich and large-scale license plate dataset, we gather a rich and large-scale Persian license plate dataset. We collect and annotate 1,300,000 license plate images (1 million images as train set and 100,000 images as test set, and 200,000 images as the validation set).

As of our knowledge, there is only one publicly available dataset for Persian license plates that only consists of 83,000 images [15]. Unfortunately, this dataset has many biases toward natural conditions and the distribution of characters in different positions of the license plate number. Training an

TABLE I
ACCURACY OF PROPOSED METHOD AND COMPARISON WITH PREVIOUS MODELS AND BASELINE EMBEDDED MODEL IN CAMERAS.

Model	Test Accuracy(%)	Val Accuracy(%)	Number of Parameters	Train Dataset
ViT OCR	77.25	77.14	4,949,888	1M
CRNN OCR [12]	75.18	75.11	15,937,987	1M
ViT OCR	63.54	63.20	4,949,888	100K
CRNN OCR [12]	65.31	65.28	15,937,987	100K
Camera's Embedded OCR	60.37	60.25	-	-

Fig. 2. Sample of LicenseNet dataset.

OCR model on this dataset does not give good practice and industrial usage accuracy. LicenseNet dataset is approximately 16 times larger than other datasets. Because of the large number of data points, it is applicable for training novel models like ViTs or developing robust, powerful OCR models without biases to characters and natural conditions.

In order to create an unbiased dataset, images in the LicenseNet dataset were captured with various cameras from different brands in different conditions like night or day, different weather, and different shooting angles (see Fig. 2).

We used the Irapardaz crowdsourcing company's infrastructure to annotate each license plate. In order to annotate each license plate, we use the pipeline as follows. First, we show each image to a human user and ask him to write the license plate number. Second, we ask another human user to do the same, and if the answers of these two users are the same, the annotation is complete. Otherwise, we employ another user to do the recognition task, and this loop continues until five nonequal answers are collected or two answers are the same. If five nonequal answers are collected, we consider the license plate unreadable and remove it from the dataset.

We gathered and annotated around 30,000,000 images and after that, we carefully chose 1,300,000 images so that any position of characters has a uniform distribution, and also day and night images are the same, and we would have various license plate angles in the dataset.

III. EXPERIMENTAL RESULTS

A. Implementation Details

The experiments were conducted in the Google Colaboratory environment. The PyTorch framework [16] is applied to implement the proposed method. The accuracy of our approach is evaluated on the LicenseNet dataset, we train our model for 200 epochs with Adam Optimizer [17], and as the loss function of each path, we used Cross-Entropy loss with the learning rate 5e-4. The size of input images was resized to 32×160, and the patch size of the ViT was 8, so each image splits into 80 patches. In encoders, each patch transforms to the embedding of size 160, and each self-attention module consists of 16 self-attention heads, each inner encoder's multilayer perceptron map to a vector of size 512. In the model, six encoders were stacked together, and each path only acts as a readout layer and is a single linear layer with the output shape of possible path values.

B. Results

We used an OCR system in cameras to collect our dataset, so as a baseline, we used the accuracy of these embedded models. For comparison with CNN-based models, we tested a CRNN model [12] on the LicenseNet dataset. The CRNN model consists of 7 CNN layers and two bidirectional RNN layers, and the whole model trains with Connectionist Temporal Classification (CTC) loss with standard training settings.

As we can see in TABLE I, the 100K subset MultiPath ViT has a competitive result, and on the whole dataset with a tiny number of parameters, it gains better performance. CRNN and MultiPath ViT models achieve better accuracy than the camera's embedded OCR. On the full LicenseNet dataset, MultiPath ViT outperformed its CRNN counterpart with a margin of 2.07% accuracy. In the meantime, MultiPath ViT is 3.21 times smaller than the CRNN model in trainable parameters. So not only MultiPath ViT has better accuracy, and it is much easier to train and employ on edge devices and small kits.

IV. CONCLUSION

We showed that with a large number of training datasets, we could achieve a competitive result in the license plate OCR challenge compared to previous popular models, with a lightweight and a tiny number of parameters. It is worth mentioning that the CRNN model was initially trained on a small subset of datasets consisting of 100,000 data points. However, the computation limitations made it impossible to train this model with the whole dataset on Google Colaboratory GPUs, so we trained it on a more powerful server. Therefore, applying Visual Transformer architecture in edge devices or computation-limited situations is helpful to keep the

model lightweight when we have a large amount of training data.

ACKNOWLEDGMENT

Collection of the LicenseNet dataset would not be possible without the help of the Irapardaz company, so a special thanks to the Irapardaz team.

REFERENCES

[1] S.-L. Chang, L.-S. Chen, Y.-C. Chung, and S.-W. Chen, "Automatic license plate recognition," *IEEE transactions on intelligent transportation systems*, vol. 5, no. 1, pp. 42–53, 2004.

[2] S. Zherzdev and A. Gruzdev, "Lprnet: License plate recognition via deep neural networks," *arXiv preprint arXiv:1806.10447*, 2018.

[3] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, "The computational limits of deep learning," *arXiv preprint arXiv:2007.05558*, 2020.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[10] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[11] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 116–12 128, 2021.

[12] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[13] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 02, pp. 107–116, 1998.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[15] A. Tourani, S. Soroori, A. Shahbahrami, and A. Akoushideh, "Iranis: A large-scale dataset of iranian vehicles license plate characters," in *2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA)*. IEEE, 2021, pp. 1–5.

[16] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.