Proceedings of the 2nd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/19/20230458

License plate Chinese character recognition based on ViT model

Xiaoyu Zhang[1]

[1] Department of Computer Science and Technology, South China University of Technology, Guangzhou, 510006, China

cszhangxy@mail.scut.edu.cn

Abstract. Transformer applications have been widely used in the computer vision field. Many related literatures show that the advantages of the model such as increased receptive field and globality are gradually emerging in image processing. However, with the popularity of the transformer, whether it can compete with the convolutional neural network (CNN) in terms of performance is still questionable and remains to be further studied. This paper will use the most basic structural model in the visual transformer (ViT) to classify and identify Chinese characters that are frequently used in the field of transportation and logistics and compare them with two classical CNN models. The results demonstrate that the performance of the transformer is obviously better than that of the traditional CNN structure, and the final accuracy of character recognition is higher than that of CNN, up to 98.66 %. It fully shows the infinite potential and excellent performance of the transformer in the area of computer vision and has high reliability and generalization ability.

Keywords: Chinese characters, vision transformer, convolutional neural network.

1. Introduction

For a long time, image text and character recognition has been a hot research subject in the computer vision field [1]. So far, this research field has achieved great success. Nowadays, image text recognition technology has penetrated all aspects of society and has an irreplaceable important position in the fields of machine automation, automatic driving, license plate recognition, and logistics recognition [2]. Among them, 31 Chinese provincial characters are very representative in terms of Chinese characters and are one of the processing objects that cannot be ignored. They are widely used in the important identification characters of logistics industry, transportation industry and other industries. Using more efficient identification and classification methods is extremely conducive to improving industrial production efficiency, logistics distribution efficiency, etc., and greatly saving human resources.

In recent years, the use of convolutional neural network (CNN) to realize image character recognition has gradually matured and achieved major breakthroughs, replacing traditional character recognition methods, and widely used in practical production applications. However, CNN still has defects such as limited receptive field and full connection redundancy. In order to further seek a more efficient model, this paper will use the shorthand visual converter ViT model to identify the Chinese character data sets of 31 Chinese provinces. The custom data sets used in the study were collected and marked by the author himself. At the same time, the recognition results of the ViT model will be compared with the classic CNN models AlexNet [3] and ResNet [4].

2. Related works

Since Rosenblatt proposed the perceptron in the 1950s [5], this simple linear classification model has defined the foundation for the coming development of deep learning. Compared with Western characters, Chinese characters have their unique structure [6], and there are new challenges and problems in the field of character recognition. With the birth of Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM), Reference [7] combined this training framework to realize the end-to-end Chinese character recognition for the first time. After that, the first convolutional neural network architecture LeNet-5 [8] was born, and the relevant research teams in China carried out in-depth research to achieve the great success of LeNet-5 framework in license plate recognition [9]. With the rise of convolutional networks such as AlexNet, VGGNet [10], and ResNet, Chinese character recognition has made further breakthroughs. For example, Reference [11] realized the application of AlexNet in license plate recognition system.

However, the CNN model consumes a lot of memory in practical applications. When faced with a large amount of data, parameter adjustment is very complicated. Transformer model based on attention mechanism was born to solve the above problems, is also rapidly changing the field of computer vision. Among them, Vision Transformer is undoubtedly an exciting research result. Up to now, the research on Chinese character recognition based on ViT model is still relatively rare. Therefore, in order to make up for the lack of relevant data in this field, the experiment in this paper will evaluate the performance of ViT structure through character recognition accuracy and compare it with two classical CNN structure frameworks for comprehensive analysis.

3. Methods

Transformer and CNN have been two hot frameworks for Chinese character recognition in recent years. This experiment will use the basic structure ViT model in Transformer to realize the recognition of key Chinese characters with certain practical significance. At the same time, it will be compared with the classical convolutional network AlexNet and residual neural network ResNet in CNN. Figure 1 shows the basic structure of the ViT model. The ViT model mainly includes four steps for image processing. Step 1 first divides the input image into patches, and then linearly maps the flattened patches. Step 2 uses linear embedding and position embedding to produce embedded patches. Step 3 is to input the patches obtained in the previous step into Transformer Encoder for coding and feature extraction. Step 4 is to use multi-layer perceptron Head for classification to output results.

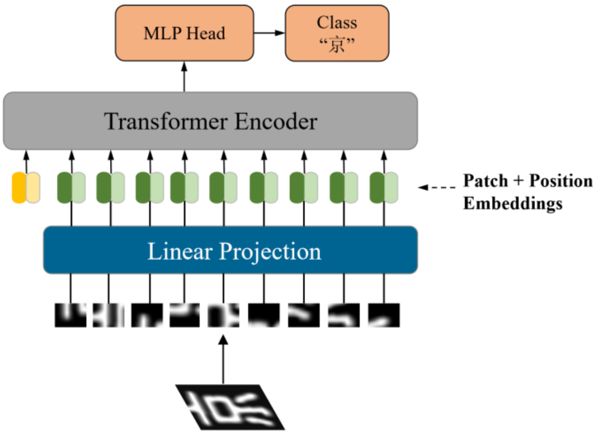

Figure 1. Vision Transformer model overview.

In addition to the ViT model used in this experiment, AlexNet and ResNet are also used as the model of this experiment to compare with the method proposed in this paper. Figure 2 shows the structure model of AlexNet. The AlexNet model has eight layers of network. Specifically, the first five are convolutional layers and the last three are fully connected layers. The activation function used in the first layer is ReLU function, which requires maximum pooling and LRN processing. The second layer is similar to the first layer. The 3rd, 4th and 5th convolutional layers are connected to each other, and the pooling layer is not connected in the middle. Another CNN model used in this experiment is the residual neural network ResNet structure model, and its core idea residual learning unit is shown in Figure 3.

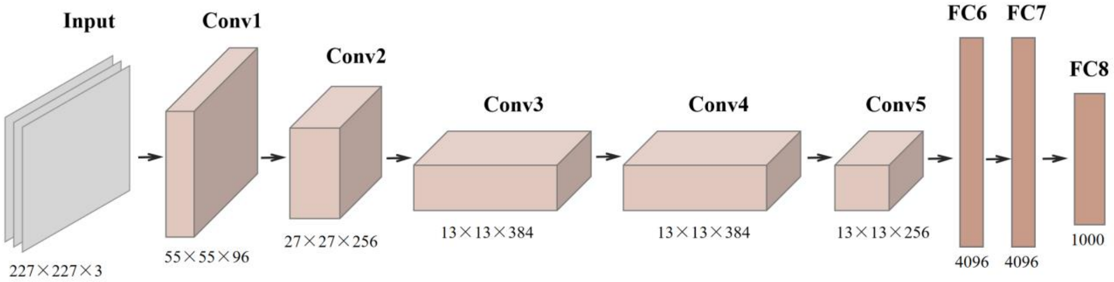

Figure 2. Structure of AlexNet model.

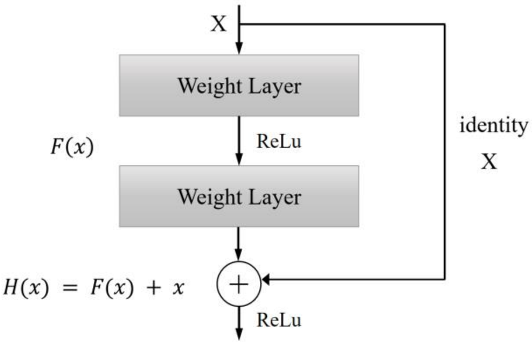

Figure 3. A building block of ResNet.

4. Experiments

4.1. Dataset

Based on the license plate recognition scene, this experiment extracts 31 Chinese characters from the license plate to form the dataset. Figure 4 shows some images of our dataset. For these 31 kinds of Chinese characters, the difficulty of recognition is to distinguish the blurred images accurately and distinguish the characters with similar structures. The dataset contains 6000 Chinese character images that have been extracted and cut, of which 4800 are used as training sets and 1200 are used as test sets, with a ratio of 8:2. In addition, 10% of the training set will be used as the validation set.

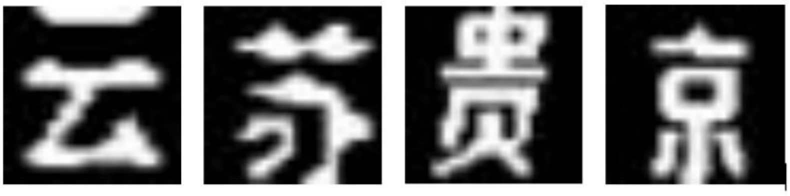

Figure 4. Dataset pictures.

4.2. Experimental environment and parameter setting

The CPU used in this experiment is 11th Gen Intel(R) Core(TM) i5-1135G7; operating system is Windows 10(64bit); build models for training using the open source machine learning framework Pytorch with numpy1.20.3, python3.9.7, Jupyter notebook6.4.5.

The parameter setting of this experiment has some changes compared with the setting of ViT model in Reference [2]. For each input image with a size of 224×224, the image is divided into 49 patches with a resolution of 32×32. In the character selection module, set the hyperparameter K to 64. During training, set the learning rate (LR) to 3e-5, use the Adam optimizer with the cross entropy loss function, and set the batch size to 32.

4.3. Experimental results and analysis

Table 1 shows the relationship between training times and sample accuracy changes when using the ViT model. It can be seen from the table that as the number of training iterations increases, the accuracy of the test sample set and the training sample set increases, and the loss rate of both decreases. This demonstrates the good performance of the ViT model in Transformer. When the epoch is 700, the accuracy of the sample test set can reach 98.66 %, which has a very good performance.

Table 1. Accuracy and Loss Value of ViT Model Using Different Training Times.

Epochs	Accuracy	Loss Value
100	96.43%	0.1617
300	97.51%	0.0766
500	97.71%	0.0621
700	98.66%	0.0251

For the purpose of further testing the performance of the ViT model, this experiment compares the accuracy value of the ViT model with the accuracy of AlexNet and ResNet recognition. In this experiment, the learning rate of AlexNet is set at 0.01, batch size is set as 128; ResNet adopts 18 layers of weight layer, and the learning rate of this experiment is set to 1e-2. Both CNN models use Adam optimizer. Table 2 displays the different performance of the three models using the same dataset. It can be seen from Table 2 that compared with AlexNet and ResNet models, ViT model has significantly higher recognition accuracy for license plate Chinese character data sets.

Table 2. Recognition accuracy of three models on our dataset.

Model	Accuracy
AlexNet	92.56%
ResNet	94.29%
ViT	98.66%

5. Conclusion

Focusing on the problem of Chinese character recognition, this paper proposes a method based on ViT model, which fills the research gap in this field to some extent. The experimental results demonstrated that the ViT model in Transformer has a much higher accuracy rate than the traditional CNN model in Chinese character recognition, showing excellent performance. In the future, the experimental results can be further analyzed and evaluated by increasing the type and quantity of Chinese character data, and the ViT parameter settings can be partially optimized and adjusted to improve the overall recognition effect. At the same time, try other models in Transformer to further prosper Chinese character recognition research in license plates and other scenarios.

References

[1] DosoViTskiy A, Beyer L, Kolesnikov A, et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning

Representations.

[2] Xiangping Wu. (2021). "Research on key technologies of image text recognition." Harbin Institute of Technology.

[3] Technicolor T, Related S , Technicolor T , et al. (2017). " ImageNet Classification with Deep Convolutional Neural Networks [50]." Communications of the ACM, 60(6), 84-90.

[4] K. He, X. Zhang, S. Ren and J. Sun, (2016) "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778.

[5] Rosenblatt, F. (1957). "The perceptron, a perceiving and recognizing automaton Project Para." Cornell Aeronautical Laboratory.

[6] Ruwei Dai,Chenglin Liu and Baihua Xiao. (2007). "Chinese character recognition: history, status and prospects." Frontiers of Computer Science in China(2).

[7] R. Messina and J. Louradour, (2015). "Segmentation-free handwritten Chinese text recognition with LSTM-RNN," 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 171-175.

[8] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, (November, 1998). "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324.

[9] Zeng-qiang, M. (2010). "License Plate Character Recognition Based on Convolutional Neural Network LeNet-5." Computer Simulation.

[10] Karen, Simonyan., Andrew, Zisserman. (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition."

[11] Zhengqiang Liu. (2016). Application of deep learning algorithm in license plate recognition system.University of Electronic Science and Technology of China,MA thesis.