Handwritten Chinese Character Recognition Based on Swin Transformer Network

Yang YU
Shenyang institute of Computing Technology,Chinese Academy of Sciences;University of Chinese Academy of Sciences
Beijing,China
Email:yuyang20@mails.ucas.ac.cn

Jing Li
Shenyang institute of Computing Technology,Chinese Academy of Sciences;University of Chinese Academy of Sciences
Beijing,China
*Corresponding author:771891882@qq.com

***Abstract*—After 2020, the accuracy of convolutional neural network model to recognize handwritten Chinese characters has encountered a bottleneck.In this paper, the HWDB1.1 handwritten Chinese character dataset is used to fine-tune the Swin Transformer model based on the method of Swin Transformer, and three groups of comparative experiments are done.Based on the recognition results of handwritten Chinese characters by Swin Transformer model, the experimental results are compared with the current convolutional neural network models (ResNet50, MobilenetV2, EfficientnetV2 etc) which have excellent performance in the field of image classification, and the simplest convolutional neural network model,and compared with human eyes, three groups of experiments are conducted to recognize handwritten Chinese characters.The above experimental results show that the Swin Transformer network model has the highest recognition accuracy of handwritten Chinese characters.**

Keywords-Deep learning;Image classification;

Swin Transformer;

I.Introduction

The earliest writing may have been hieroglyphic symbols that appeared on stone blocks to represent items or concepts. With the development of human civilization, these pictographs have gradually been abstracted into an important tool for information expression and communication with a simpler structure. Writing plays a very important role in the inheritance and communication of human culture. Nowadays, writing is widely used in People's Daily life and work, and handwriting is also a skill that everyone must learn in social production. However, with the rapid development of science and technology, the appearance of computers has changed people's habit of handwriting, and many handwritten documents in daily life and work have been replaced by electronic documents. With the widespread use of computers in life and work, how to make effective interaction between humans and computers has become a problem to be solved. The invention of the mouse and keyboard enabled people to accurately and efficiently input words or commands into the computer, and they are still widely used in all kinds of computers today. However, the way you use a mouse and keyboard is very different from the way you write, which is naturally the most efficient way you interact. Therefore, the advanced human-computer interaction based on handwriting has aroused great research interest. In order to realize this kind of natural human-computer interaction, how to make the computer correctly understand and recognize the handwritten text is the core of the problem, so the handwriting recognition problem arises.

Chinese characters are one of the most important communication carriers in Chinese society, along with the connection between the Internet and human society is becoming more and more close, handwritten Chinese character recognition Handwritten Chinese character recognition plays an increasingly important role in people's life.Handwritten Chinese Character Recognition is mainly divided into online handwritten Chinese character recognition and offline handwritten Chinese character recognition.The former uses related equipment to record the data of the writing track, and uses The stroke sequence information was used for text recognition. The latter uses the image acquisition device prepare to acquire handwritten Chinese character images, through the learning of image and Chinese character encoding the mapping between the codes is used to identify Chinese characters.

Handwritten Chinese character recognition has been continuously developed since the 1980s, transmission the unified method has gradually formed the process of "preprocessing, feature extraction and classification".To recognize handwritten Chinese characters, and obtained a good recognition effect fruit.But in practice, more complex handwriting styles and recognition molds equation makes the

text recognition rate decrease, and it is difficult for users to obtain the best performance experience.Since 2012, some researchers began to use of deep learning method for handwritten Chinese character recognition, with the depth of the is given priority to with convolution neural network learning method to obtain the recognition rate of the far beyond the traditional method, and shows great potential in the field of deep learning in handwritten Chinese character recognition.

At present, the accuracy of convolutional neural network model in handwritten Chinese character recognition has encountered a bottleneck.To this end, this paper uses Swin Transformer as the backbone network for offline handwritten Chinese character recognition, which improves the recognition accuracy.

II.Materials and Methods

A. Source of the dataset

This article uses the CASIA-HWDB off-line single character database the HWDB 1.1 dataset of CASIA-HWDB is used for experiments.This data set was obtained from China Branch Construction of the National Laboratory of Pattern Recognition at the Institute of Automation of the College.Among them, HWDB1.1 data set contains 3755 Chinese characters in the GB2312-80 character set.Due to the inconsistent size of the samples, they were all processed into a size of 64×64 before being input into the network.The training set of each Chinese character contains 240 samples, and the test set contains 60 samples. The number of samples in the whole data set is 1126500.Some sample examples in the data set are shown in Figure 1.

Figure 1 Dataset Examples

B. Enhance the data

Handwritten Chinese characters are more difficult to identify than printed ones because of their variety of writing styles, the existence of many similar characters and the large variety of characters. In order to improve the accuracy of handwritten Chinese character recognition, in addition to improving the structure of the network model, good performance is also inseparable from the support of large-scale data sets. Generally speaking, the neural network model with high accuracy and strong generalization ability needs huge and diverse data sets for training. Small datasets may cause a risk of overfitting the network model. In this paper, the method of data enhancement is adopted to create more samples and improve the robustness of the model by performing position processing such as flipping, translation and rotation for handwritten Chinese characters. The practice is divided into two steps, first of all, the original picture according to a certain Angle of rotation, to prevent the emergence of improper Angle of the picture recognition. Move the data up and down and left and right, and scale up or down the image. Secondly, the data are distorted to different degrees to adapt to different people's handwriting. The data are stretched to different degrees to increase the generalization ability. And do projection changes to the text, to identify different positions.

To binarize a dataset, the original three-channel RGB ima ge in the dataset is converted into a single-channel gray image, and the gray value of the pixels on the conversion is set to 0 o r 255.For common image classification tasks, the binarization process is a very serious process of image information loss.Ho wever, for the recognition of handwritten Chinese characters, t he data itself is white background and black Chinese character s.Therefore, the binarization of the image will not affect the re cognition effect,but also remove a lot of information in the RG B image that we do not need when recognizing handwritten Chinese characters, and greatly reduce the amount of calculati on when recognizing handwritten Chinese characters.

C. Swin Transformer network model

Swin Transformer's innovative ideas are mainly reflected in Hierarchy, Locality and Translation invariance. Hierarchy Layers similar to CNN are used in the feature extraction stage The input image was 4x and 8x, respectively and 16 times of downsampling to get the multi-scale feature map; Locality is mainly reflected in the process of self-attention calculation, which computes the constraints in the partitioned local non-overlapping window, so that the algorithm the complexity has changed from being a function of the square of the image size.The linear relation greatly reduces the computation amount and improves the efficiency of the algorithm rate; In the two layers of Transformer modules, there is no overlap.
The window partition is shifted by half a window compared to the previous layer, this allows the information in the upper and lower Windows to interact effectively, compared with the sliding window design seen in CNN, it maintains Translation invariance while not causing accuracy reduction.

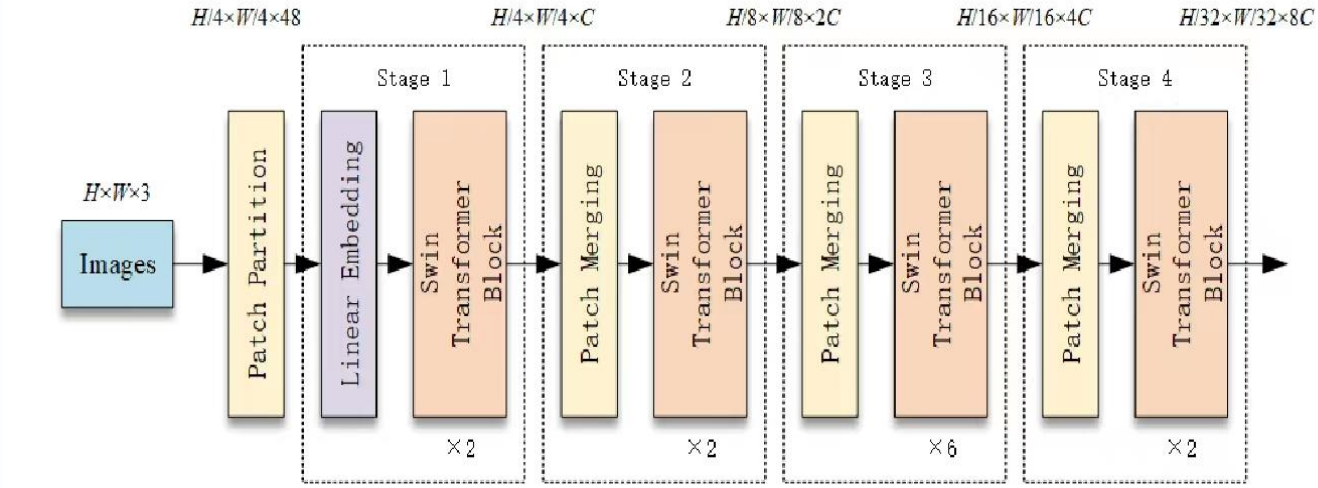

Figure 2 Architecture of Swin Transformer

The algorithm flow of Swin Transformer is roughly as follows.The input image is an RGB three-channel image of the size of H×W.The Patch Partition module is used for Partition operation: Divided into 4×4 patches, followed by edges flatten the three channel directions and flatten the size after flattening 4×4×3=48, which is implemented by Conv2D within the Linear Embeding module, as shown in Figure 1. Then the data enters the Swin Transformer module, and after normalization, enters the Windows multi-head self-attention, further divides the image block into non-overlapping areas, and calculates the self-attention within the area. Since in W-MSA, only the self-attention calculation is carried out in each window, there is no need to send information between Windows. Therefore, the negated Windows multi-head self-attention module, SW-MSA has been introduced. After the non-overlapping windows are divided at layer L, the window is redivided by half the window distance in layer L+1, so that the information of some Windows in different layers can be exchanged. The above operation can be summed up as Stage 1.

In order to obtain multi-scale feature information, construction is needed a hierarchical structure designed in Swin Transformer network the Patch Merging module, implements classes like the pooling operation in CNN, the width and height of the image after Mosaic half the size, double the number of channels, and then through Swin Transformer modules, such structures are connected in series connect, the second through fourth phases in Figure 1.

Swin Transformer has been implemented to meet different requirements the serialization of network structure, the network from simple to complex in turn is swin-t (Tiny), swin-s (Small), swin-b (Baes) and Swin-L (Large), the more complex the network the stronger, the greater the computational overhead, as shown in Figure 2 Swin-t for example.

III. Results & Discussion

A. Experimental Environment

The experimental conditions: the processor CPU is Intel I7 8700K, the GPU is Nvidia RTX3090 16GB, and Pytorch is selected as the deep learning framework.

B. Evaluation Metrics

The experiment uses Accuracy as an evaluation metric, which is the ratio of the number of correctly classified samples to the total number of samples.The target task of this paper is a multi-class classification problem (the number of classes is N =3755). The binary classification is extended to multi-class classification, and the accuracy calculation formula is as follows in Figure 3.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Figure 3 Accuracy Calculation Formula

C. Experimental process and conclusions

In order to illustrate the accuracy and effectiveness of the Swin Transformer network model in recognizing handwritten Chinese characters, three groups of comparative experiments are designed in this paper.

1) Experiment 1

In Experiment I, the recognition accuracy of Swin Transformer is compared with other convolutional neural network models (ResNet50, MobilenetV2, EfficientnetV2 etc) that perform well in the field of image classification.From the experimental results, it can be seen that Swin Transformer has a significant improvement in accuracy compared with other convolutional neural network models.

Tab.1 Comparison results of various network models

Network Model	Accuracy	Params
Swin Transformer	96.31%	29M
ResNet50	92.46%	25.5M
EfficientNet V1	95.21%	66M
Mobilenet V2	91.42%	6.9M
AlexNet	82.91%	60M
Vgg19	84.76%	144M
Inception V3	90.23%	23.8M

2) Experiment 2

Experiment II is based on the recognition accuracy of Swin Transformer and compared with the convolutional neural network model with the simplest structure.From the experimental results, it can be seen that the simplest convolutional neural network model, Swin Transformer improves the accuracy greatly, which verifies the necessity of complex model structure to improve the recognition accuracy of handwritten Chinese characters.

Tab.2 Comparison results of simple network models

Network Model	Accuracy	Params
Swin Transformer	96.31%	29M
Two layers of Convolution plus One layer of Full Connection	60.21%	2M

3) Experiment 3

Experiment 3 was based on the recognition accuracy of Swin Transformer and compared with the recognition accuracy of handwritten Chinese characters by human eyes.It can be seen from the experimental results that Swin Transformer has a significant improvement in accuracy compared with human eye recognition.It is verified that the accuracy of the computer has surpassed that of the human in recognizing handwritten Chinese characters.

Tab.3 Comparison results of Human Eyes

Network Model	Accuracy	Params
Swin Transformer	96.31%	29M
Human Eyes	95%	None

4) Experimental Conclusions

Through three groups of comparative experiments, it can be verified that Swin Transformer has the highest accuracy compared with other convolutional neural network models and human eyes.The accuracy of Swin Transformer in recognizing handwritten Chinese characters is 96.31%.

IV. Conclusions

Aiming at the problem that the accuracy of convolutional neural network model in recognizing handwritten Chinese characters has encountered a bottleneck, this paper proposes a stripping method based on Swin Transformer handwritten Chinese character recognition network model.The main research works are as follows:

Firstly, the HWDB 1.1 data set is preprocessed and the three-channel RGB images are binarized into single-channel grayscale images.

The Swin Transformer network model is fine-tuned, mainly adjusting the first and last layer of Swin Transformer, and training and testing are carried out on the handwritten Chinese character data set processed in the previous step.

Three groups of comparative experiments are designed to compare the training results of Swin Transformer model with other convolutional neural network models and the accuracy of human eye recognition.The experimental results show that the Swin Transformer algorithm proposed in this paper has the highest accuracy in handwritten Chinese character recognition.The accuracy of Swin Transformer in recognizing handwritten Chinese characters is 96.31%.

References

[1] Epshtein B, Ofek E, Wexler Y.Detecting text in natural scenes with stroke width transform (2010), in IEEE Computer Vision and Pattern Recognition (CVPR)

[2] Su B,Lu S.Accurate scene text recognition based on recurrent neural network[C]. Asian Conference on ComputerVision. Springer, Cham, Singapore, 2014: 35-48.

[3] Wang K, Babenko B, Belongie S. End-to-end scene text recognition[C].IEEE, Barcelo-NA, 2011: 1457-1464. [C].

[4] Krizhevsky A , Sutskever I , Hinton G E . ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, Six (6) : 84-90.

[5] C. M. Bishop. Neural networks for pattern recognition. Oxford university press, 1995.

[6] W. L. Briggs, S. F. McCormick, et al. A Multigrid Tutorial. Siam, 2000.

[7] K. Chatfifield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In BMVC, 2011.

[8] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, And A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. IJCV, Pages 303 -- 338, 2010.

[9] S. Gidaris and N. Komodakis. Object detection via a multi-region & semantic segmentation-aware cnn model. In ICCV, 2015.

[10] CTPNTian Z, Huang W, He T, et al. (2016) Detecting Text in Natu-ral Image with Connectionist Text Proposal
Network. In: Leibe B., Matas J., Sebe N., Welling M. (EDS) Computer Vision -- EC-CV 2016. ECCV 2016. Lecture Notes in Computer Science, 9912. Springer, Cham.