Vision Transformer Based on Knowledge Distillation in TCM Image Classification

Ge Yuyao[1], Cheng Yiting[2], Wang Jia[3], Zhou Hanlin[4], Chen Lizhe[5]
[12345] North China University of Technology
[12345] Beijing 100144, China
e-mail: 2912899504@qq.com, 994502781@qq.com, 930359879@qq.com, xeon9931@foxmail.com, orichen@qq.com

***Abstract*—In order to improve the ViT model accuracy of image classification task in Chinese medicine, this paper proposes a sharpening image preprocessing method of coupling residual algorithm, the image preprocessing method can make deep learning network makes it easier to extract the image edge character. In this paper, through a series of experiments to compare the algorithm under different parameters in traditional Chinese medicine classification accuracy of the data sets. Improved the vision Transformer structure of knowledge distillation and proposed the way of overlapping image blocks in PatchEmbeding operation to extract more information of the original image. A series of experiments were carried out on the traditional Chinese medicine data set. It is proved that the accuracy of the model is about 2% higher than that of the original knowledge distillation ViT structure.**

Keywords-attention mechanism; transformer; knowledge of distillation; deep learning

I. INTRODUCTION

At present, the classification of Traditional Chinese medicine in China mainly adopts the method of manual screening, which is low in efficiency and high in cost. In order to meet the urgent needs of the society for Traditional Chinese medicine, image technology applied in the field of traditional Chinese medicine has been developed successively.

In 2020, a Google team published ICLR2021 paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" [1] applies transformer structure to the field of vision and proposes vision Transformer (hereinafter referred to as ViT) [2]. In this paper, the author points out that if ViT is trained on a large data set, its performance is more outstanding than that of traditional convolutional neural network under sufficient computing power. Despite ViT's impressive performance on large data sets, training ViT is not easy.

To solve this dilemma, the Facebook team proposed a ViT model of knowledge distillation structure in ICML 2021, which published "Training data-effificient image transformers & distillation through attention" [3]. The introduction of knowledge distillation greatly improves the efficiency of model training and makes it affordable for ordinary scholars to train the ViT of knowledge distillation structure.

II. KNOWLEDGE DISTILLATION VIT

The ViT model of knowledge distillation can be disassembled in three parts. They are respectively PatchEmbeding part, Transformer encoder part and knowledge distillation part.

A. PatchEmbeding

Since ViT's design comes from the field of natural language understanding, ViT follows NLP's transformer thinking and sets ViT inputs as word vectors. So, the PatchEmbeding step before Transformer considers how to map images into word vectors into Transformer.

ViT author pointed out that firstly, a (C, H, W) image was divided into multiple non-overlapping image blocks of (C, P, P) size. After segmentation, HW/P can be obtained. At this time, the matrix shape becomes (HW/P^2, C, P, P), then flatten the last three dimensions of the image, that is, perform the flatten operation, and the matrix shape becomes (HW/P^2, C * P * P). Finally, the matrix is mapped into a matrix with the shape of (HW/P^2, Embed_dim) through the linear transformation layer.

B. Encoder

Since ViT does not have a Decoder, a one-dimensional class token needs to be added at the beginning of the Transformer input for final classification. Because knowledge distillation ViT added knowledge distillation on the basis of ViT, we also needed to add a one-dimensional random vector called distillation token to make the output value of knowledge distillation as close as possible to that of teacher model. Finally, all the word vectors and position vectors are spliced along the feature direction to get 2+ HW/P^2 vectors of words of length T, T=C*P*P.

Transformer Encoder is composed of several Encoder, and each Encode Encoder is mainly composed of three parts, which are layer normalization layer, multi-attention mechanism layer and multi-perceptron layer respectively. The input word vector first passes through the layer normalization layer, then passes through the multi-attention mechanism layer, and then the resulting word vector is added to the original input word vector, and then passes through the layer normalization layer and the multi-perceptron layer, and then the output result is added to the output result of the

multi-attention mechanism layer. As shown in Figure 1. Encoder structure.

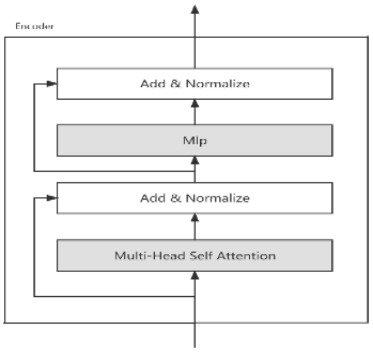

Figure 1. Encoder structure

C. Multi-head Self-attention

The most important structure in Transformer is multi-head self-attention. Multi-head self-attention structure with two heads is shown in Figure 2. Schematic diagram of attention mechanism.

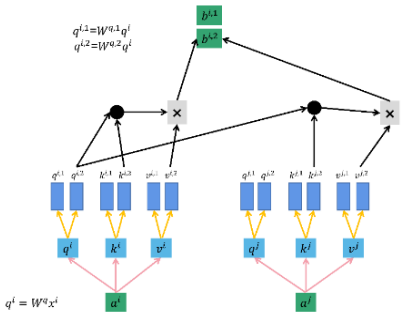

Figure 2. Schematic diagram of attention mechanism

The weighted inner product is calculated as follows:

$$a(1,i) = q^1 * k^i / \sqrt{d} \qquad (1)$$

Where, d is the dimension of q and K, because the value of Q * k increases with the increase of the dimension, so the value divided by $\sqrt{d}$ is equivalent to the effect of normalization.

Next, take the calculated α^i as softmax operation, and then multiply it by $v^{(i,j)}$ [5].

D. Knowledge Distillation

In knowledge distillation ViT, the author's team uses CONv-based architecture as teacher model network, introduces local assumptions into Transformer through distillation, and achieves the effect of improving training efficiency. As shown in Figure 3. Schematic diagram of hard distillation. The authors' team proposed two distillation methods:

(1) Soft Distillation: make the model self-iterate by minimizing the KL divergence of teacher network and student network output values, as shown in Equation (2).

$$L_{global} = (1-\lambda)L_{CE}(\psi(Z_s), y) + \lambda\tau^2 KL(\psi(Z_s/\tau), \psi(Z_t/\tau)) \qquad (2)$$

Where, Z_s 、 Z_t respectively represent the output of student network and teacher network.

(2) Hard-label Distillation: A variation of knowledge Distillation proposed in this paper, in which the output of teacher networks is regarded as Hard label.

$$y_t = \text{argmax}_c Z_t(c) \qquad (3)$$

$$L_{global}^{hardDistill} = (1-\alpha)L_{CE}(\psi(Z_s), y) + \alpha L_{CE}(\psi(Z_s), y_t) \qquad (4)$$

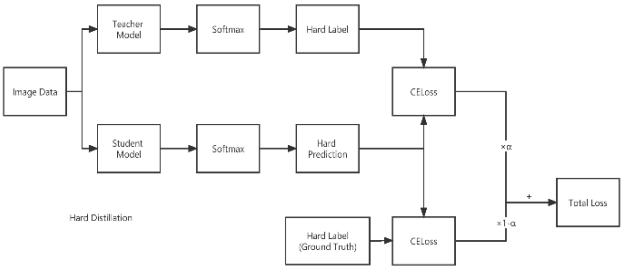

Figure 3. Schematic diagram of hard distillation

III. VERLAP_VIT

A. Sharpening the Coupling Residual Algorithm

In order to enhance the effective information in the image, we design a sharpening coupling residual algorithm. By sharpening the image processed by the coupling residual algorithm, its edge information is effectively strengthened, so that the edge information can be better maintained in the following multiple Encoder structures. As shown in Figure 4. Comparison of different sharpening methods.

$$Output = (1-\lambda)Sharpen(Sharpen(Input)) + \lambda Input \qquad (5)$$

In this structure, how to properly adjust the proportion of the original image and the sharpened image is our key concern.

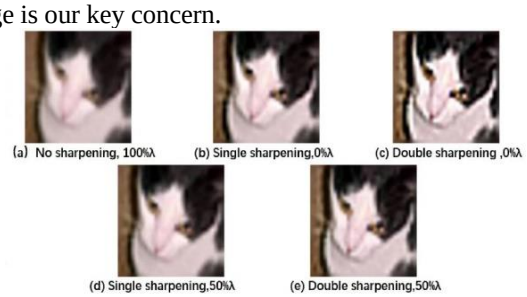

Figure 4. Comparison of different sharpening methods

We designed the following experiment. In order to explore which processing method has better effect between single sharpening and double sharpening, we designed the following four experiments: Input the single sharpened image directly, input the double sharpened image directly, input the single sharpened image after 1:1 fusion with the original image, input the double sharpened image after 1:1 fusion with the original image, and input the original image without processing to compare the Top1 accuracy. As shown in Figure 5. Schematic diagram of sharpening coupling residual algorithm, (*) is executed during double sharpening operation.

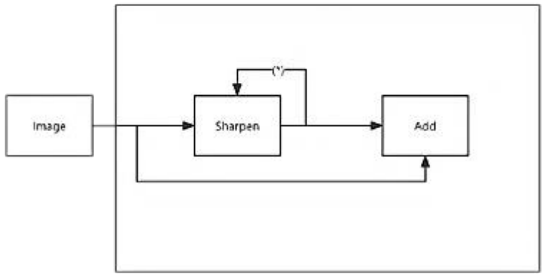

Figure 5. Schematic diagram of sharpening coupling residual algorithm, (*) is executed during double sharpening operation

B. PatchEmbeding Method of Overlapping Image Blocks

In "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale"[1], the author adopts the method of dividing a (C,H,W) image into multiple non-overlapping image blocks of (C,P,P) size and mapping them into word directions After measuring, input multiple word vectors into the attention mechanism layer. We found that this method focused on the relationship between word vectors and word vectors, And the relationship between word vectors is missing, which resulted in the information loss of the original image, resulting in the decline of accuracy. Therefore, we decided to present in the original paper PatchEmbeding is improved on the basis of PatchEmbeding. By overlapping image blocks, there is a certain overlap between image blocks. In this way, the ability of extracting original image information of the model can be improved to the maximum extent within the time controllable and acceptable range of equipment, so as to improve the accuracy of model recognition. As shown in Figure 6. Schematic diagram of overlapping PatchEmbeding.

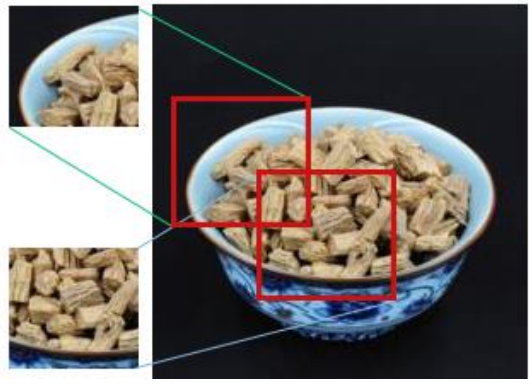

Figure 6. Schematic diagram of overlapping PatchEmbeding

IV. The Data Set

A. Data Sources

This case data set comes from Baidu flying paddle AI Studio platform.

B. Introduction to Data Sets

The dataset is named the Identification dataset of Chinese Medicinal Materials. As shown in Figure 7. Example of Traditional Chinese medicine image. Among them, we label codonopsis pilosula, Honeysuckle, lily, locust flower and wolfberry as' 0 ', '1', '2' and '3' respectively.

TABLE I. Data set display table of Chinese medicinal materials1

	codonopsis pilosula	honeysuckle	lily	locust flower	wolfberry
quantity	190	180	180	167	185
proportion	21.06%	19.95%	19.95%	18.51%	20.51%
grade	0	1	2	3	4

Figure 7. Example of Traditional Chinese medicine image

V. Experimental Design

The experiment was divided into two experimental groups. The first experimental group compared the accuracy of sharpening coupling residual algorithm with different parameters under the two data sets, and selected the pretreatment method with the highest accuracy to enter the second experimental group to continue the experiment.

The first experimental group:

In the first experimental group, two sharpening methods were set up, namely single sharpening and double sharpening. Two different λ values were set for the single sharpening mode, respectively 0% (Experiment 2) and 50% (Experiment 4). Two different λ values were also set for the double sharpening mode, respectively 0% (experiment 3) and 50% (experiment 5). The first experimental group is composed of five groups of experiments by combining these four groups of experiments with the direct input of the original image (Experiment 1). Are shown in table 2.

The second experimental group:

In the second experimental group, the best condition in experimental group 1 was used as the image preprocessing method, and the processed images were respectively input into knowledge distillation ViT (Experiment 4) and Overlap_ViT (Experiment 6). Compared with the original image input knowledge distillation ViT (Experiment 1), VGG19 (Experiment 7) and Resnet50 (Experiment 8), the performance of the image processed by sharpening coupling residual algorithm in the Overlap_ViT model compared with other models was obtained. See Table 3 and Table 4

TABLE II. EXPERIMENTAL PARAMETERS OF THE FIRST GROUP

	experiment 1	experiment 2	experiment 3	experiment 4	experiment 5
λ	100%	0%	0%	50%	50%
sharpen mode	No sharpening	Single sharpening	Double sharpening	Single sharpening	Double sharpening
image_size	224×224				
patch_size	16				
stride	16				
embed_dim	768				
mlp_ratio	4.0				
depth	12				
num_ratio	12				
num_epochs	300				
warmup_epochs	5				
optimizer	AdamW				
batch_size	32				
weight_decay	1e-4				
base_lr	6e-4				
warmup_strart_lr	1e-6				
end_lr	5e-4				
distillation_alpha	0.5				
remark	Knowledge distillation VIT	The input image undergoes a single sharpened knowledge distillation VIT	The input image has been sharpened twice by knowledge distillation VIT	The input images are distilled by the sharpened coupling residuals algorithm knowledge VIT	The input images are distilled by the sharpened coupling residuals algorithm knowledge VIT

TABLE III. EXPERIMENTAL PARAMETERS OF THE SECOND GROUP (1)

	experiment1	experiment4	experiment6
λ	100%	50%	50%
sharpen mode	No sharpening	Single sharpening	Single sharpening
image_size	224×224		
patch_size	16	16	14
stride	16		
embed_dim	768		
mlp_ratio	4.0		
depth	12		
num_ratio	12		
num_epochs	300		
warmup_epochs	5		
optimizer	AdamW		
batch_size	32		
weight_decay	1e-4		
base_lr	6e-4		
warmup_strart_lr	1e-6		
end_lr	5e-4		
distillation_alpha	0.5		
remark	Knowledge distillationVIT	The input images are distilled by the sharpened coupling residuals algorithm knowledgeVIT	The sharpened coupling residuals algorithm is used Overlap_VIT

Explanation of some parameters:

λ: λ is the residual ratio of the original image in the sharpening coupling residual algorithm

Image_size: In the pre-processing, the image will be scaled to Image_size and Image_size by bilinear interpolation algorithm

Patch_size: Size of image blocks in overlapping PatchEmbeding operation

Stride:In the overlapping PatchEmbeding operation, take the step length of the image block

Distillation_alpha: The proportion of the teacher's model in a knowledge distillation operation that is transmitted back

TABLE IV. EXPERIMENTAL PARAMETERS OF THE SECOND GROUP (2)

	experiment7	experiment8
λ	100%	100%
sharpen mode	not	not
image_size	224×224	
optimizer	Adam	
learning_rate	1e-4	
batch_size	32	
remark	Vgg19	Resnet50

VI. DATA ANALYSIS

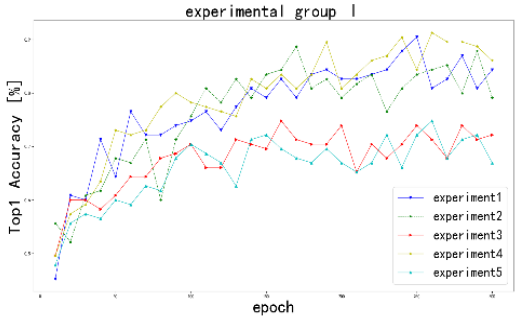

Figure 8. Comparison of accuracy in the first experimental group

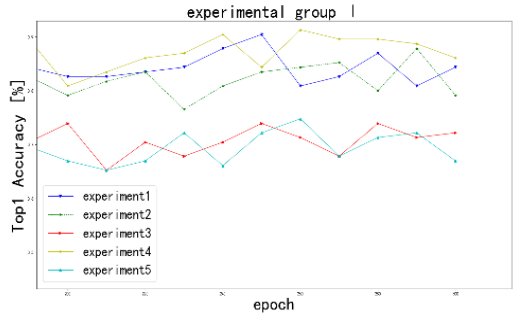

Figure 9. Comparison of the accuracy of 200~300 epochs in the first group

Finally, by observing the conditions of the next 100 epochs, in 71% of epochs, the Top1 accuracy of the single sharpening image and the 1:1 fusion of the original image (Experiment 4) was higher than that of all other processing methods. We believe that the reasons are as follows: As the edge features of the image without sharpening are not obvious, redundant information will appear in the structure of multiple Encoder stacks, leading to the misjudgment of the model, thus affecting the accuracy. However, the double-sharpened structure lost a lot of original image features because the edge features were too prominent due to the double-sharpened structure, so that the accuracy rate could not be improved.

After single sharpening and 1:1 fusion, the output can not only ensure that the edge of the image is strengthened, but also retain the original information of the image due to the addition of residual of the source image. That is, on the premise of ensuring that the original information of the image is not lost, the original information can be retained in the structure of multiple Encoder stacking by strengthening the edge. Figure 8. Comparison of accuracy in the first experimental group and Figure 9. Comparison of the accuracy of 200~300 epochs in the first group.

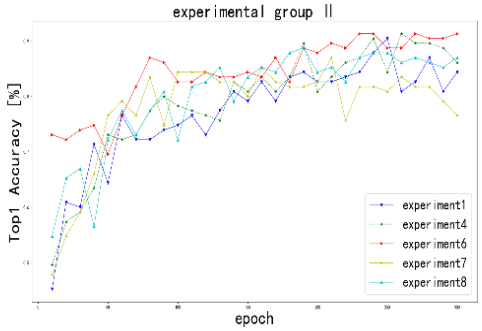

Figure 10. Comparison of accuracy of the second experimental group

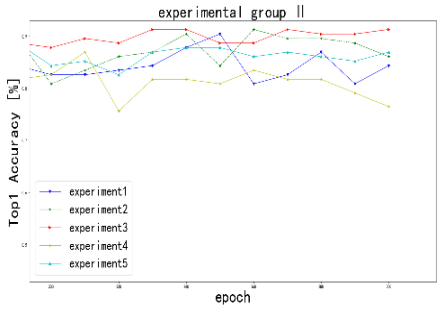

Figure 11. Comparison of the accuracy of the second experimental group 200-300 epoch

By comparing experiment 6 with Experiment 7 and Experiment 8, Overlap_ViT with single sharpening coupling residual algorithm has higher accuracy than traditional convolutional neural network. Figure 10. Comparison of accuracy of the second experimental group. Figure 11. Comparison of the accuracy of the second experimental group 200-300 epoch .After analysis, it can be concluded that there are three reasons for this result:

1) Due to the multi-head attention mechanism of ViT, it can make ViT pay attention to the global information and obtain the global receptive field. Compared with the local receptive field of traditional convolutional neural network, this global attention mechanism can make ViT more information from the image, thus improving the accuracy.

2) Due to the sharpening coupling residual algorithm, Overlap_ViT can obtain enhanced edge information when reading images in the training set, which makes Overlap_ViT more familiar with edge features of images when facing similar images in the test set, thus improving accuracy.

3) Since Overlap_ViT's teacher model adopts the result model of Experiment 8, Overlap_ViT can obtain the information of teacher model for iteration when the accuracy is lower than that of Resnet50 model in Experiment 8. When the number of training rounds is constant, the accuracy of Overlap_ViT cannot be lower than that of the teacher model.When the accuracy of Overlap_ViT is higher than that of the teacher model, the information from ground truth can be used for iteration to make its accuracy higher than that of the teacher model.

Compared with Experiment 4, the accuracy of Overlap_ViT with overlapping Patch_Embding operation was higher than that of knowledge distillation ViT.The reason can be summarized as that knowledge distillation ViT adopts the way of non-overlapping image block, which leads to the loss of some information inside the image block.However, Overlap_ViT's overlapping Patch_Embding operation alleviates this situation to a certain extent. As image blocks are overlapped, the same information can be found in two image blocks, and more information of their own image blocks can be noticed in the multi-attention mechanism layer, thus improving the accuracy.

By comparing experiment 6 with Experiment 1, it is proved that Overlap_ViT with single sharpening coupling residual algorithm has higher accuracy than knowledge distillation ViT in this task.

VII. Conclusion

In view of the five TCM classification tasks in this experiment, visual Transformer model is adopted to accelerate the training of the model through knowledge distillation, so that ViT which originally needs training on large data sets can get better performance on small data sets. The new overlapping PatchEmbeding and single sharpening coupling residual algorithm proposed in this paper can effectively further improve the accuracy of the model and achieve better results.

In the next work, more experimental parameters can be changed to obtain better experimental results.

References

[1] Dosovitskiy A,Beyer L,Kolesnikov A,et al.An image is worth 16x16 words: Transformers for image recognition at scale[J].arXiv preprint arXiv: 2010. 11929,2020.

[2] Liu Wenting, Lu Xinming. Progress of Transformer Research Based on Computer Vision [J]. Computer Engineering and Applications, 202, 58(06):1-16.

[3] Touvron H, Cord M，Douze M, et al.Training data-efficient image transformers& distillation through attention[J]. Arxiv,2020:2012.12877 .

[4] Isaman Sangbamrung, Panchalee Praneetpholkrang, and Sarunya Kanjanawattana, "A Novel Automatic Method for Cassava Disease Classification Using Deep Learning," Journal of Advances in Information Technology, Vol. 11, No. 4, pp. 241-248, November 2020. doi: 10.12720/jait.11.4.241-248

[5] Sun Qiang, LI Yiquan, Yu Zhanjiang, LI Chengchao, XU Jinkai. Research on Wear Prediction of Micromilling Cutter for Inception-ViT Model [J]. Tool Engineering,202,56(01):3-8.

[6] Yuan L, Chen Y, Wang T, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet[J]. arXiv pre-print arXiv:2101.11986, 2021.

[7] Duan yang. Based on the deep study of lymphatic metastasis breast cancer classification method research [D]. Hangzhou university of electronic science and technology, 2020. The DOI: 10.27075 / d. cnki. ghzdc. 2020.001106.

[8] S. Bunrit, N. Kerdprasop, and K. Kerdprasop, "Improving the Representation of CNN Based Features by Autoencoder for a Task of Construction Material Image Classification," Journal of Advances in Information Technology, Vol. 11, No. 4, pp. 192-199, November 2020. doi: 10.12720/jait.11.4.192-199

[9] Jolitte A. Villaruz, "Deep Convolutional Neural Network Feature Extraction for Berry Trees Classification," Journal of Advances in Information Technology, Vol. 12, No. 3, pp. 226-233, August 2021. doi: 10.12720/jait.12.3.226-233

[10] Ying Chen, Weiwei Du, Xiaojie Duan, Yanhe Ma, and Hong Zhang, "Squeeze-and-Excitation Convolutional Neural Network for Classification of Malignant and Benign Lung Nodules," Journal of Advances in Information Technology, Vol. 12, No. 2, pp. 153-158, May 2021. doi: 10.12720/jait.12.2.153-158