Researches Advanced in Deep Learning based Image Classification

Siying LI*

Sydney Institute of Intelligent Technology, Northeastern University Qinhuangdao, China.

* Corresponding Author Email: 202019230@stu.neuq.edu

Abstract. Image recognition has always been a popular research topic in computer vision, whose basic task is to learn a model to predict the category of a given image. Early image classification algorithms mainly relied on handcrafted features, while their classification results often failed to meet practical application requirements due to the limitation of handcrafted features expressiveness ability. Thanks to the rapid development of deep learning, image recognition algorithms based on convolutional neural networks have achieved great success. Generally, stacking network layers can improve the prediction accuracy, while increasing the network depth can also lead to problems such as gradient disappearance, gradient explosion, and degradation. In recent years, due to its powerful representation ability, Transformer-based image classification algorithms have achieved new breakthroughs in recognition accuracy. This paper first introduces the classic deep learning algorithms in the field of image classification, including networks such as AlexNet, GoogLeNet, VGG, and ResNet. Meanwhile, the visual transformer (ViT) and the data-efficient image transformer are further introduced to handle image classification tasks. In addition, this paper analyzes the application and development of these two models in image classification, classifies the different models, and analyzes their advantages and disadvantages.

Keywords: Image classification, deep learning, CNN, Transformer.

1. Introduction

With the rapid development of Internet technology, multimedia technology and computers, images have become a common form of expressing and storing information. However, the problem of image information disorder is becoming more and more prominent. Therefore, in the face of massive image data, how to use the computer to intelligently and efficiently process it and classify and identify it, so as to extract and organize the required data information, has become a highly-demanding item in the field of machine learning. subject of attention. Numerous researchers have carried out a lot of related research in the field of image classification, which to a certain extent has made artificial intelligence a big step forward. However, today's image classification technology is still far below our expectations, because of the complexity and variability of the image itself and the limitations of theoretical and technical development.

Image classification is one of the basic tasks in the field of computer vision. Given an image or a group of pictures, the computer uses the corresponding algorithm to identify and classify what category it belongs to. Image classification can be applied in a wide range of fields. For example, in the field of vehicle autonomous driving, image features are often used to identify trees, people, animals, traffic lights, etc. around traffic during driving. In the medical service domain, image classification can identify local shadows and highlights in medical images [1].

In the process of image classification, the most important is image feature extraction for image classification, as well as image preprocessing and classifier classification [2]. Early image recognition algorithms were mainly based on manual features. However, limited by the insufficient ability of hand-crafted features to express the scene, the performance of traditional image recognition algorithms based on machine learning cannot meet the actual application requirements. Thanks to the rapid development of convolutional neural networks and powerful feature expression capabilities, image recognition algorithms based on deep learning have become the current mainstream framework. The increase of network depth is conducive to the degree of nonlinear change in feature learning, while it also often brings about the problem of gradient disappearance, gradient explosion, and

degradation. In recent years, transformers that have achieved excellent performance in natural language processing tasks have been extended to computer vision tasks, greatly improving the performance of image classification.

Focusing on the framework of deep convolutional neural network, this paper systematically introduces the research progress of image recognition task through a detailed literature analysis. Specifically, this paper firstly introduces the development of classical classification algorithms based on convolutional neural networks in detail; secondly, it introduces the representative algorithms of transformers in the field of image recognition, and analyzes and summarizes the advantages and disadvantages of these algorithms. In addition, this paper quantitatively compares the performance of different representative classification algorithms on classical image classification datasets. Finally, this paper also summarizes the existing problems in the current image recognition research field and predicts its future development direction.

2. Image classification based on convolutional neural network

2.1. Basic network structure

Convolutional neural networks have always been a hot topic in pattern recognition and representation learning. For any kind of convolutional neural network, each neuron is only connected to its upper local neuron, and the network parameters are greatly reduced, which makes it very powerful in extracting local image features [1]. Traditional convolutional neural networks are usually composed of the following layers: convolutional layer, linear rectification layer, pooling layer and full connection layer.

(1) Convolution layer.In the convolutional neural network, multiple convolutional units constitute the convolutional layer, and the parameters of the convolutional unit are optimized through the back propagation algorithm. The more layers of the convolutional network, the more complex features are extractedPooling layer. CNN further reduces the amount of computation through pooling. Pooling is the process of reducing the input image to retain only important information. The size of the pooled area is usually 2 x 2, which is converted to a certain value according to the corresponding rules.

(2) Full connection layer. The full connection layer and the convolution layer can be converted to each other, that is, for any convolution layer, the weights only need to be converted into a huge matrix, in which the local perception is mostly 0 except for some specific blocks, and the weights of many blocks are the same, because the weights are shared. Conversely, any fully connected layer can also be a convolution layer.

2.2. Representation convolution neural network

The emergence of convolutional neural networks is marked by lenet-5 model proposed by LeCun in 1998. Convolutional neural network consists of one or more convolutional layers, full connection layers and pooling layers. This structure enables convolutional neural network to take advantage of the two-dimensional structure of input data, which can be trained using back propagation algorithm and requires fewer parameters, especially compared with deep neural network and feedforward neural network. Therefore, convolutional neural network is more suitable for image classification and speech recognition.

2.2.1 AlexNet

The first development of convolutional neural networks was the proposal of AlexNet network. Krizhevsky and Hinton won the ILSVRC (ImageNet Large-scale Visual Recognition Challenge) in 2012 with AlexNet by 10.9 percentage points. At this point, the birth of AlexNet aroused the enthusiasm of many experts and scholars to further study the field of deep learning. The structure of this network is very similar to LeNet, but the dual GPU network structure is adopted, so that a larger and deeper network can be designed. Krizhevsky et al. have introduced several key concepts in AlexNet. One is ReLU, which prevents the problem of gradient extinction, and the other is Dropout,

which prevents overfitting and allows neurons in each layer to turn on and off randomly. However, AlexNet also has defects due to various conditions at that time, that is, because it uses GTX580 with only 3GB of video memory, but the algorithm is constantly improving, which leads to its inability to keep up with the further development of the algorithm.

2.2.2 GoogLeNet

After AlexNet, a large number of network structures have been improved for image classification tasks, mainly from the perspective of increasing the network scale, including the improvement of network depth and width. However, directly increasing the scale of the network will bring the following two problems: it is easy to lead to overfitting and increase the amount of calculation. To solve both of these problems, GoogleNet is proposed with a novelty Inception mechanism, which enables multi-scale processing of images. A typical Inception module structure is shown in Figure 1.

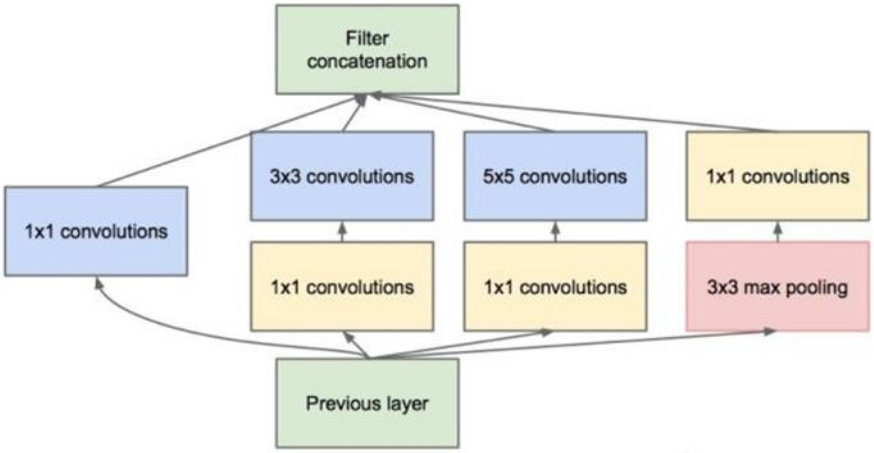

Figure 1 The structure of typical Inception module

The Inception module consists of three sets of convolutional nuclei and a pool cell that collectively accept input images from the previous layer. GoogleNet makes two improvements over Alex Net. The first is to replace the full connection layer with the full dozen average pool, because the number of parameters in the full connection layer leads to overfitting. Second, the Inception module in Google Net improves parameter utilization and further increases the branching structure. The Inception module has the advantage of integrating many convolutional kernels and pooling layers of different sizes, which greatly reduces the number of model parameters. In addition, GoogLeNet adds complexity to the network by extending its width. This method not only reduces the parameters, but also greatly improves the adaptability of the network to multi-scale.

2.2.3 VGGNet

In 2014, Simonyan and Zisserman proposed a series of VGG models, and in ImageNet Challenge in the same year, VGG network won the second place in classification task and the first place in localization task. At that time, VGG was already a very deep network with a depth of 19 layers, which was a great breakthrough in the field of convolutional neural networks, because the fitting ability of neural network models usually increases with the scale of model network layers. At the same time, the scalability of VGGNet is very strong, which makes it migrate to other image data generalization is very good, greatly expanding the scope of application. The entire network structure of VGGNet adopts convolution kernel size(2*2) and pooling size(3*3). Even now network computing resources and computing power is very developed, VGGNet is still often used to extract image features, is widely used in various tasks in the field of vision.

The main innovation of VGG network is the adoption of small-size convolution kernel. All convolution layers use the convolution kernel, and the step of convolution is 1. In order to ensure that the size of the image remains unchanged after convolution, 1 pixel is filled on each side of the image. All pooling layers have a core of step 2. The full connection layer consists of 3 layers, including 4096,

4096 and 1000 nodes respectively. All layers adopt ReLU activation function except the last fully connected layer. In fact, VGG has some similarities with AlexNet, they are both composed of 5 convolution layers and the superposition part of activation function and 3 fully connected layers. However, the difference is that VGG deepens the superposition of the previous 5 convolution layers and activation function, so that each part is not composed of a convolution layer and an activation function. Instead, multiple such combinations form parts, each of which is brought together.

Moreover, unlike other convolution neural network, VGG uses a smaller consecutive 2×2 convolution kernels to simulate larger convolution kernels, such as the layer 2 consecutive 2×2 convolution layer can reach a layer of 5×5 convolution the feeling of the wild, but the quantity is less, some 18 two 2×2 convolution kernel parameters, and a 5×5 convolution kernels has 25 parameters. As for the advantages of this, the author points out the following two points: reducing the number of network parameters; As the number of parameters is greatly reduced, the previous convolutional layer with a large receptive field can be replaced by multiple convolutional layers with small receptive fields, thus increasing the nonlinear expression ability of the network. Subsequent residual networks have continued this feature.

2.2.4 ResNet

In 2015, He et al. proposed ResNet, which not only solved the degradation problem of neural network, but also solved the problems of poor fitting ability and high training/testing error when deep neural network reaches a certain depth compared with shallow neural network. ResNet won first place in ILSVRC and COCO tasks for classification, location, detection and segmentation. Residual network solves the problem that deep network is difficult to train by fitting residual terms through cross-layer connections, and makes the number of layers of network reach an unprecedented scale. The ImageNet test set, which won the first place in the ILSVRC2015 classification task, achieved a top 5 error rate of 3.57%, and the 100 - and 1000-layer residual networks were analyzed on the CIFAR-10 dataset.

Previous experience has proved that increasing the number of layers of the network will improve the performance of the network, but after increasing to a certain extent, with the increase of layers, the training error and test error of the neural network will increase, which is different from overfitting, which only has a large error in the test set, and this problem is called degradation. Based on this problem, He et al. proposed the residual structure, changing the original fitting function $H(x)$ to $F(x)$, where $F(x) = H(x) - x$. The principle of this structure is shown in Figure 2.

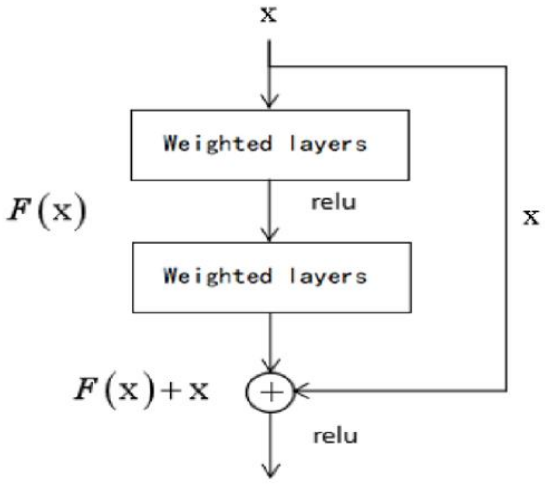

Figure 2 The structure of residual module

Resnet solves the problem of information leakage between traditional convolutional neural network and connected network, and constructs deep network residual by building blocks. It only needs to study the difference between input and output parts of the whole network.

2.3. Improvement based on mainstream CNN

The above representative convolutional neural networks have greatly improved the performance of various vision tasks. In recent years, many experts and scholars have carried out a lot of research on how to further improve these networks and applied them to image classification tasks in various fields.

Shaojuan Li et al. [3] improved the image classification algorithm of AlexNet by adding a deconvolution layer on the basis of traditional AlexNet and classifying images through the full connection layer. Honghai Hoang et al. [4] studied that the long-hop connection has stronger convergence ability and better performance than the corresponding model when performing the image classification task, and the long-hop connection is applicable to any model. Jiyeon Kim et al. [5]proposed an automatic classification method for tourist photos according to tourist attractions. The clustering was extracted by hierarchical clustering of noisy applications and density-based spatial clustering, and then deep learning was applied to remove the noise in the clustering, and then the photos were classified according to categories. Kan Wu et al. [6]proposed a relative position encoding method for 2D images, called iRPE, which can improve DeIT-S and Detr-Resnet 50 by 1.5% and 1.3% in ImageNet and COCO, respectively, by inserting a self-focus layer.

3. Image classification based on Transformer

3.1. Transformer

Transformer Is a classic model for NLP applications introduced by the Google team in June 2017 in Attention Is All You Need. Transformer's advantage over CNN and RNN is that it uses only encoders, decoders and attention mechanisms, enabling parallel training and efficient global analysis. Transformer has greatly changed the situation that the former SOTA model is only based on RNN, LSTM and other recursive neural networks. Transformer does not use serial mode but uses parallel language processing, completely changing the serial mode and allowing all text to be analyzed at the same time. This parallelism is entirely dependent on the attention mechanism, which allows Transformer to consider the relationship between any two words in order to analyze which word gets more attention.

(1) Self - Attention mechanism. Self-attention is a self-attention mechanism used in Transformer, which integrates the whole context into each word, helping to learn the dependence of words within sentences and the internal structure of sentences. The self attention mechanism in Transformer is very similar to the attention mechanism in CV. Both of them establish a connection between a set of input data to determine which part of data is more important.

(2) Multi-Head Attention. Multi-head Attention maintains a separate query, key, and value weight matrix for each header. Each of these sets is randomly initialized, which extends the model's ability to focus on different locations and gives multiple representation subspaces of the focus layer.

3.2. Transformer-based Image classificationr

Although Transformer structures are widely used in the NLP arena, their use in the visual arena is limited. In the visual world, attention is used either in conjunction with CNN networks or in place of specific components in CNN.

The application of Attention mechanism in image classification domain to achieve better performance. Yehui Tang et al. [7]propose a new enhancement shortcut scheme to improve the feature diversity of visual transformers (FIG. 1). In addition to traditional identity shortcuts, parallelizing the MSA module with multiple parameterized shortcuts provides more alternative paths to bypass the attention mechanism. Ben Graham et al. [8] propose LeViT, which is a hybrid neural network for rapid inference image classification that is significantly superior to existing convnets and Vision transformers in terms of speed/precision trade-offs. LeViT, for example, was five times faster on CPU than EfficientNet with 80% ImageNet top-1 accuracy.

Vision Transformer(ViT) is a model proposed by Google in 2020 that directly applies Transformer to image classification. ViT is a pure converter that performs image classification well when applied directly to image block sequences. They follow the original transformer design as closely as possible. The input to the model is a plane pixel vector extracted from a block of pixel size PxP. Each input pixel is fed into the linear projection layer, producing what is called patch embedding. Notice that at the beginning of the sequence, the model attaches an additional learnable embed. The embedding method is used to predict the type of input image after self-attention update. Each embed also adds a learnable location embed. This classification simply places the MLP header on top of the Transformer structure, where it is inserted as an additional learnable embedding location added to the sequence. After Embedding, a D-dimensional Embedding class is added in the front of the sequence, which is similar to the class mark in BERT, as the input of the transformer encoder after position coding. In the output part, VIT removes the decoder part, and sends the features obtained by the encoder directly to the MLP head, which is finally classified by Softmax. Instead of selecting projected image blocks, it uses ResNet's earlier feature maps as input to Transformer. Through end-to-end training of Transformer model and CNN backbone, the model can obtain the best image classification results.

Data-efficient Image Transformer (DeiT) is an Image converter based on knowledge distillation proposed by Facebook in a 2021 study. Compared to VIT, DeiT requires less data and fewer computing resources to generate a high-performance image classification model. The DeiT model was trained on an 8-GPU server for 3 days and the method achieved 84.2% top-1 accuracy in the ImageNet benchmark test. In the training stage, no external data is used, and the results are comparable to those of the top-level convolutional neural network.3.3 Comparison between Transformer and CNN

As Transformer is introduced into the field of computer vision, more and more researchers begin to study the advantages and disadvantages of Transformer and CNN, and further give a variety of related applications. The visual converter (ViT) model provides an alternative design paradigm for convolutional neural networks (CNN). So much so that the inductive bias inherent in evolution towards local processing is replaced by global processing performed by the ViT with multi-headed self-care[9]. Yutong Bai et al.[10]found that Transformer outperformed CNN when measuring the robustness of countermeasures, and proposed a method to improve CNN by adopting transformer's training method so that CNN could be as powerful as Transformer in defending against hostile attacks. The key to this training approach is the self-focused architecture of Transformer. The work of Daquan Zhou et al. [11] demonstrates an important benefit of attentional representation in robust generalization and is consistent with recent studies looking at robustness in ViTs.

4. Performance comparison

This image classification performance comparison experiment is divided into three parts. The first part is based on the comparison of error rates of CNN image classification model in ImageNet data set. The second part is to study the time complexity and spatial complexity of CNN image classification model to improve model efficiency. The third part is more complex image multi-classification task simulation experiment to observe the change of loss value in CNN model training.

4.1. Comparison of error rates of ImageNet data sets based on CNN image classification model

Table 1 [12]shows the classification performance of different CNN models on the imagenet dataset. Top-1 error rate represents the ratio that the category with the largest prediction probability among the learned labels is not the correct category. Top-5 error rate represents the ratio that the five categories with the largest prediction probability among the learned labels do not contain the correct category. The fourth column is the error rate of test sets. All are the error rates of integrated network test set submitted in ILSVRC competition of the same year. It can be seen that convolutional neural network has developed particularly rapidly in the field of image classification, and there is still room

for breakthrough. Therefore, subsequent in-depth research on convolutional neural network is very important in the field of image classification.

Table 1. Error rate comparison of ImageNet data set based on CNN image classification model
Note: Validation with val is short for validation and test is short for test

CNN model	Top - 1 error rate (val)	Top - 5 error rate (val)	Top - 5 error rate (test)
Alex Net	36.70%	15.40%	15.30%
GoogLeNet	-	8.43%	7.32%
VGGNet	-	7.89%	6.66%
ResNet-152	19.38%	4.49%	3.57%
Attention-92	19.50%	4.80%	-

4.2. FLOPS and the number of parameters

Note: FLOPS stands for number of floating point operations per second.

Time complexity determines the number of operations required for model training and prediction, which can be measured by FLOPS. The spatial complexity determines the number of parameters, which can be measured by the number of model parameters. Table 2 [12] shows the ImageNet data sets different CNN FLOPS and model parameters of the model, a LeNet respectively, AlexNet, VGGNet, GoogLeNet, ResNet and Attention. In recent years, CNN has improved the image classification performance in terms of structure and depth. However, with the deepening of network depth, the corresponding time complexity and spatial complexity keep increasing, resulting in the reduction of network efficiency. Time complexity of the model with the development of hardware and improve, but the space complexity greatly increases because the model dimension, and will result in an increase in model time complexity, so it can deduce that as long as the reduced model of the whole space complexity, guarantee the model performance, can improve the efficiency of study, reduce the time complexity, lightweight model.

Table 2. Time complexity and space complexity of CNN image classification model

CNN model	FLOPS	Parameter Quantity/number
LeNet-5	4*10^5	6*10^4
Alex Net	7*10^9	6*10^7
VGGNet-16	1.5*10^11	1.38*10^8
GoogLeNet	1.5*10^10	5*10^6
ResNet-50	3.86*10^9	2.5*10^7
Attention-92	1.04*10^10	5.13*10^7

4.3. Complex image multi-classification task simulation experiment

In order to better analyze the performance of convolutional neural network models and show the training and evaluation methods of models in recent years, popular classification models (Alexnet-8, VGGNET-16, Googlenet-22, RESNET-50, RESNET-10) were trained on ciFRA10 data set. The training results are shown in Figure 3[13]. In this experiment, a single Tesla M40GPU with 12g video memory was used for training through tensorflow, and the training process was visualized through tensorboard. In order to facilitate the comparison between various models, the optimization algorithm RMSprop was used for all models in this training, and the optimizer parameters and training skills were the same. This experiment only considers the performance difference caused by the model itself. The change of model loss value in the training process is shown in the figure. The five models use the same optimization algorithm, corresponding parameters and the same degradation learning rate, so the change process of loss value is mainly determined by the model structure. As can be seen from the figure, except alexnet-8, the loss value is almost zero after

100,000 iterations, while the loss value fluctuation of other models is relatively stable between 200,000 and 300,000 times, which can be considered as model training convergence. The loss value reflects the accuracy of model prediction of training samples, that is, the greater the loss value, the lower the accuracy of model description of training samples. Therefore, when the model converges, Alexnet-8 is the most accurate description of the training set sample, and the loss of GoogLeNet is slightly greater than that of other models.\

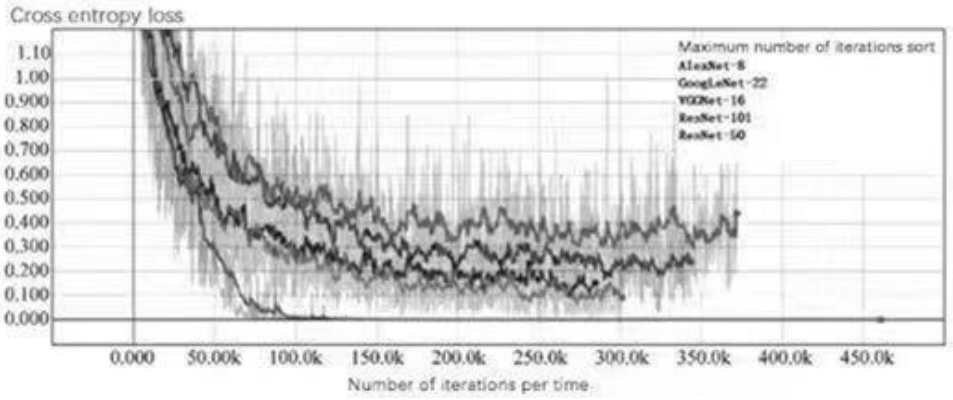

Figure 3 The change of loss value during the training of CNN model

5. Discussion

Although algorithms based on convolutional neural networks and transformers have greatly improved the accuracy and speed of image classification tasks, there are still some problems to be solved.

(1) Image classification by deep convolutional neural network is supervised learning, that is, it is necessary to label each image and tell the model the category of the image, so as to learn the features of the image of this category, and then use the features learned to judge the image of the unknown category. While the human brain is capable of thinking and is more flexible, it is more unsupervised learning, and the human visual system does not need to be told the categories of thousands of images before it judges an image. Therefore, how to make the model capable of thinking like the human brain may be the direction of future research, and unsupervised learning is the future development prospect.

(2) Since Transformer is lightweight, processing ultra HD graphics is a challenge. Because transformer models are typically large, the computational cost is higher than a lightweight CNN model, such as a ViT model that requires 18B FLOPs to process the image. By comparison, the lightweight CNN model GhostNet only needs about 600M failures to achieve similar performance. This means that using transformer models is still expensive at this stage. Although several methods for compressing converters have been proposed, their complexity is still considerable. Thus, an efficient Transformer model is one where agents deploy visual Transformers on devices with limited resources.

Although many Transformer based models have been proposed for computer vision tasks, these efforts are suggestive and early solutions that have a lot of room for improvement. For example, the Transformer architecture in ViT fully follows Transformer's standard converters in the NLP domain, and many of these features are also not applicable in the computer vision domain, so an improved version of Transformer designed specifically for CV needs to be further explored.

6. Conclusion

This paper summarizes the development of deep learning algorithms in the field of image classification, focusing on the development of convolutional neural networks in different stages of the field of image classification, from CNN, AlexNet, GoogLeNet, VGGNet to ResNet, as well as transformer series models, which are very popular in the field of image classification recently.

Compared with THE convolutional neural network CNN, which is very important in the field of image classification, it shows outstanding advantages. The differences between CNN and Transformer are summarized as follows: Although Transformer belongs to machine learning, it does not have operations such as convolution, pooling or circulation. Transformer lacks some inductive biases inherent in CNN, such as translation equivalence and location; Transformer makes good use of the correlation between each line of data, and the mechanism is highly explanatory and is more suitable for NLP; Image classification is introduced in ViT, based on Transformer, but with changes. CNN focuses on the correlation between two-dimensional local data. With the deepening of layers, the area of concern will be wider and more suitable for image processing. However, with the self-attention mechanism, Transformer has a great range of applications for image classification. In the future, Transformer in the field of computer vision will be used for more tasks.

Reference

[1] Kaur, P., Singh, S. K., Singh, I., & Kumar, S. (2021, December). Exploring Convolutional Neural Network in Computer Vision-based Image Classification. In International Conference on Smart Systems and Advanced Computing (Syscom-2021).

[2] Zhang Ke,Feng Xiaohan and Guo Yurong. A review of Deep Convolutional Neural Network models for Image Classification. Journal of Image and Graphics,26(10):2305-2325 [DOI: 10.11834 / JIG.200302].

[3] Li, S., Wang, L., Li, J., & Yao, Y. (2021, February). Image classification algorithm based on improved AlexNet. In Journal of Physics: Conference Series (Vol. 1813, No. 1, p. 012051). IOP Publishing.

[4] Hoang, H. H., & Trinh, H. H. (2021). Improvement for convolutional neural networks in image classification using long skip connection. Applied Sciences, 11(5), 2092.

[5] Kim, J., & Kang, Y. (2022). Automatic Classification of Photos by Tourist Attractions Using Deep Learning Model and Image Feature Vector Clustering. ISPRS International Journal of Geo-Information, 11(4), 245.

[6] Wu, K., Peng, H., Chen, M., Fu, J., & Chao, H. (2021). Rethinking and improving relative position encoding for vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10033-10041).

[7] Tang, Y., Han, K., Xu, C., Xiao, A., Deng, Y., Xu, C., & Wang, Y. (2021). Augmented shortcuts for vision transformers. Advances in Neural Information Processing Systems, 34, 15316-15327.

[8] Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., & Douze, M. (2021). Levit: a vision transformer in convnet's clothing for faster inference. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 12259-12269).

[9] Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., & Girshick, R. (2021). Early convolutions help transformers see better. Advances in Neural Information Processing Systems, 34, 30392-30400.

[10] Bai, Y., Mei, J., Yuille, A. L., & Xie, C. (2021). Are Transformers more robust than CNNs?. Advances in Neural Information Processing Systems, 34, 26831-26843.

[11] Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., & Alvarez, J. M. (2022, June). Understanding the robustness in vision transformers. In International Conference on Machine Learning (pp. 27378-27394). PMLR.

[12] Su Fu, Lu Qin & Luo Renze.(2019). A survey of image classification based on deep learning. Telecommunications Science (11),58-74. [21]Boyu Chen , Peixia Li, Chuming Li , Baopu Li , Lei Bai , Chen Lin ,Ming Sun , Junjie Yan , Wanli Ouyang. GLiT: Neural Architecture Search for Global and Local Image Transformer.

[13] Yang Zhenzhen, Kuang Nan, Fan Lu & Kang Bin.(2018). A survey of image classification algorithms based on convolutional neural networks. Signal Processing (12),1474-1489. doi:10.16798/j.issn.1003-0530.2018.12.009.