Journal of Computers Vol. 35 No. 1, February 2024, pp. 17-31
doi: 10.53106/199115992024023501002

ACANet: A Fine-grained Image Classification Optimization Method Based on Convolution and Attention Fusion

Zhi Tan*, Zi-Hao Xu

School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, 102616, Beijing, China

tanzhi@bucea.edu.cn, 2108550021035@stu.bucea.edu.cn

Received 10 April 2023; Revised 26 July 2023; Accepted 10 October 2023

Abstract. The key to solve the problem of fine-grained image classification is to find the differentiation regions related to fine-grained features. In this paper, we try to add new network components to the original network and adjust various parameters to try to propose a new fine-grained image classification network. We propose a fine-grained image classification network based on the fusion of asymmetric convolution, convolution and self-attention mechanisms. Firstly, an enhanced module using asymmetric convolution to assist classical convolution proposed to help convolution learn deep features. Secondly, according to the common points of convolution and self-attention mechanism, we invented a fusion module of convolution and self-attention mechanism to improve the learning ability of the network.We integrate these two modules into the residual network and invent a new residual network .Finally, according to the experience, we design a new downsampling layer to adapt to the new component of the attention mechanism and improve the performance of the model. The experiment test on three publicly available datasets, and three methods for comparison. The results show that the new structure can effectively complete the task of fine-grained image classification, and the classification accuracy of different methods and different datasets are significantly improved.

Keywords: attention mechanism, asymmetric convolution, fine-grained image classification

1 Introduction

Fine-grained image classification (FGIC) refers to the classification task that is further subdivided from the classification of the main classes of common objects (such as birds, cars, etc.). It has a broad prospect and high demand in many fields such as academic research and practical application. Fine-grained image classification task is more challenging than ordinary image classification task because it has the characteristics of large differences among different individuals within the same category and small differences among different individuals of different categories.

Earlier work mainly relied on supervised learning methods for manually labeled images. For example, the local one-to-one feature learning method proposed by Berg [3] et al., and the double cross entropy loss function proposed by Li [18] et al., by adding regularization terms to the cross entropy loss function. Ma et al. [4] proposed the channel maximum pooling (CMP) method by inserting a new layer between the convolution layer and the fully connected layer, which has achieved excellent results. However, the drawback that such a method cannot be extended effectively and widely becomes increasingly prominent soon. Therefore, most of the latest research methods are unsupervised learning. For example, Cong [19] et al propose a global-and-local collaborative learning architecture, which includes a global correspondence modeling (GCM) and a local correspondence modeling (LCM) to capture comprehensive inter-image corresponding relationship among different images from the global and local perspectives., and Hoang [5] et al proposed a novel framework, namely Deep Cross-modality Spectral Hashing (DCSH), to tackle the unsupervised learning problem of binary hash codes for efficient cross-modal retrieval. Unsupervised learning has been shown to be more effective than supervised learning because it can uncover missing or unidentifiable parts of human data. What they all have in common is a focus on finding the most distinctive parts. In order to make up for the lack of labeled data, the latest fine-grained classification efforts are mainly in the direction of adding better network components and learning more unique features.

* Corresponding Author

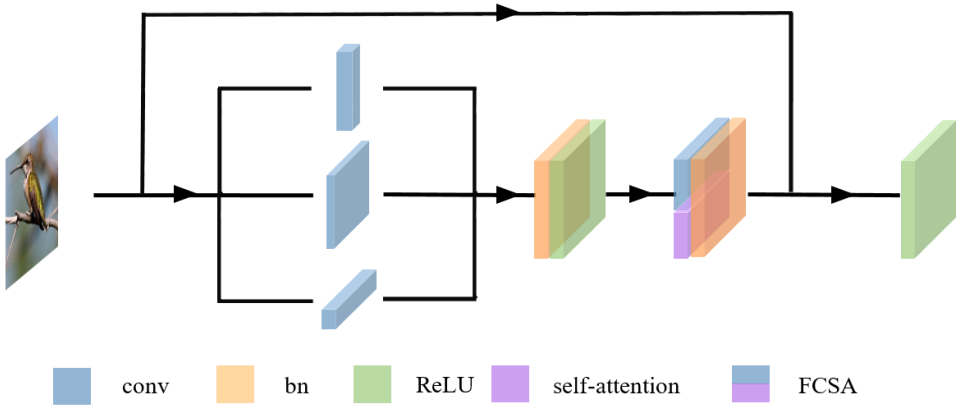


Fig. 1. A new residual structure based on the self-attention and convolution fusion module and based on asymmetric convolution enhancement techniques

In this paper, we follow the motivation outlined above to address the unique challenges of fine-grained image classification. We try to add new network component in the network to improve the ability of the network to extract detailed features in the image. At the same time, while improving the performance of the network, we also try to control the number of network parameters and calculation time to take care of practical applications.

We try to add new network modules to enhance the ability of the network to learn features. Importantly, a new residual structure is proposed in this paper on the basis of the original convolutional network structure. By using the new residual module, the performance of the network is significantly improved under several different loss functions and several different datasets. Compared with the existing technology, the new module proposed in this paper has some significant advantages: (1) The attention mechanism is introduced at a small cost and the model performance is improved without significantly increasing the network parameters and extending the inference time. (2) Theoretically, it can be easily applied to any existing or future network architecture.

More specifically, we use the additivity of convolution operation to design the enhancement module and design the fusion module of self-attention and convolution according to the common properties of self-attention and convolution . A new downsampling layer is designed based on the vision transformer (ViT) experience. In this paper, the new network can have the advantages of both convolutional network and ViT model. The schematic diagram is shown in Fig. 1.

In this paper, a large number of experiments were carried out on three commonly used fine-grained classification datasets, CUB-200-2011 [11], Flower-102 [6] and Stanford Cars [7], and the results show that the proposed method has achieved remarkable results. On this basis, we carried out ablation experiments to further study the actual effects of each component.

The specific works of this paper are as follows:

(1) We strengthen the convolution kernel with a set of asymmetric convolution, and adjust the size and position for the asymmetric convolution combination.

(2) We integrated the attention mechanism with convolution, and adjusted the parameters to help the fusion module adapt to fine-grained image classification tasks.

(3) Learning from the experience of other models, we design a new downsampling layer for the convolution fusion self-attention module to help it adapt to fine-grained image classification tasks.

(4) After many experiments, we combine the above three modules and other auxiliary components into a new residual block, and based on this, we propose a new fine-grained image classification network.

The remaining sections of this paper are organized as follows. In Section II , we first introduce some cases and techniques of fine-grained image classification and network components separately in the Related Works section. In Section III , in the Methods section, we introduce the details and principles of the method used in this paper and show the effect. In this part, we introduce Asymmetric Convolution and Convolution Block (ACCB), Fusion of Convolution and Self-Attention (FCSA) and Downsampling Layer technology respectively by using formulas and images,and the results are presented at the end. In Section IV, We present the results of our proposed method on different data sets and analyze the results and experimental procedures. In Section V, in the Ablation Study, we verified the effect of each module by disassembling several modules one by one, and showed the results with

images and figures. In Section VI, we discuss and summarize in the Conclusions section and make an outlook on the development of this technology in the future.

2 Related Works

2.1 Fined-Grained Image Classification

In the field of fine-grained image classification, there are two main methods for feature extraction, broadly classified as object-part-based method and attention-based method.

The object part-based approach aims to find the local area of the object for recognition by using the model to generate candidate regions and then extracting discriminant features from them. Zheng et al [21]. trained both localization and classification accuracy by clustering feature maps into object sections. This unsupervised classification enhances feature learning by dividing patterns into object parts. This method allows learning features and locations at the same time. In addition, Hu et al. [22] added data by clipping local extreme values to discover other discriminant features. On the other hand, attention-mechanism-based approaches are mainly used to enhance feature learning and locate object details. The approach proposed by Sun et al. [23] works by generating multiple sets of features that are enhanced by attention mechanisms. Luo et al. [24] used attention graphs from multiple excitation models to learn features from different classes. The method proposed by Zhuang et al. [25] and Zhang et al. [26] enhances the discriminant representation by using two images as inputs to calculate the attention between feature maps. Behera et al. [27] calculated the self-attention graph of output features to express the relationship between feature pixels. Bera et al. [28] used graph convolutional neural networks to describe the relationship between features. Rao et al. [29] proposed to add a counterfactual intervention to the attention diagram to predict categories. With the development of the field of computer vision, many excellent workers have proposed improved architectures, such as Wang et al. [30], Sun et al. [31], He et al. [32], Zhang et al. [33] , who use self-attention graphs at the converter layer to enhance feature learning and locate object details.

2.2 Network Component

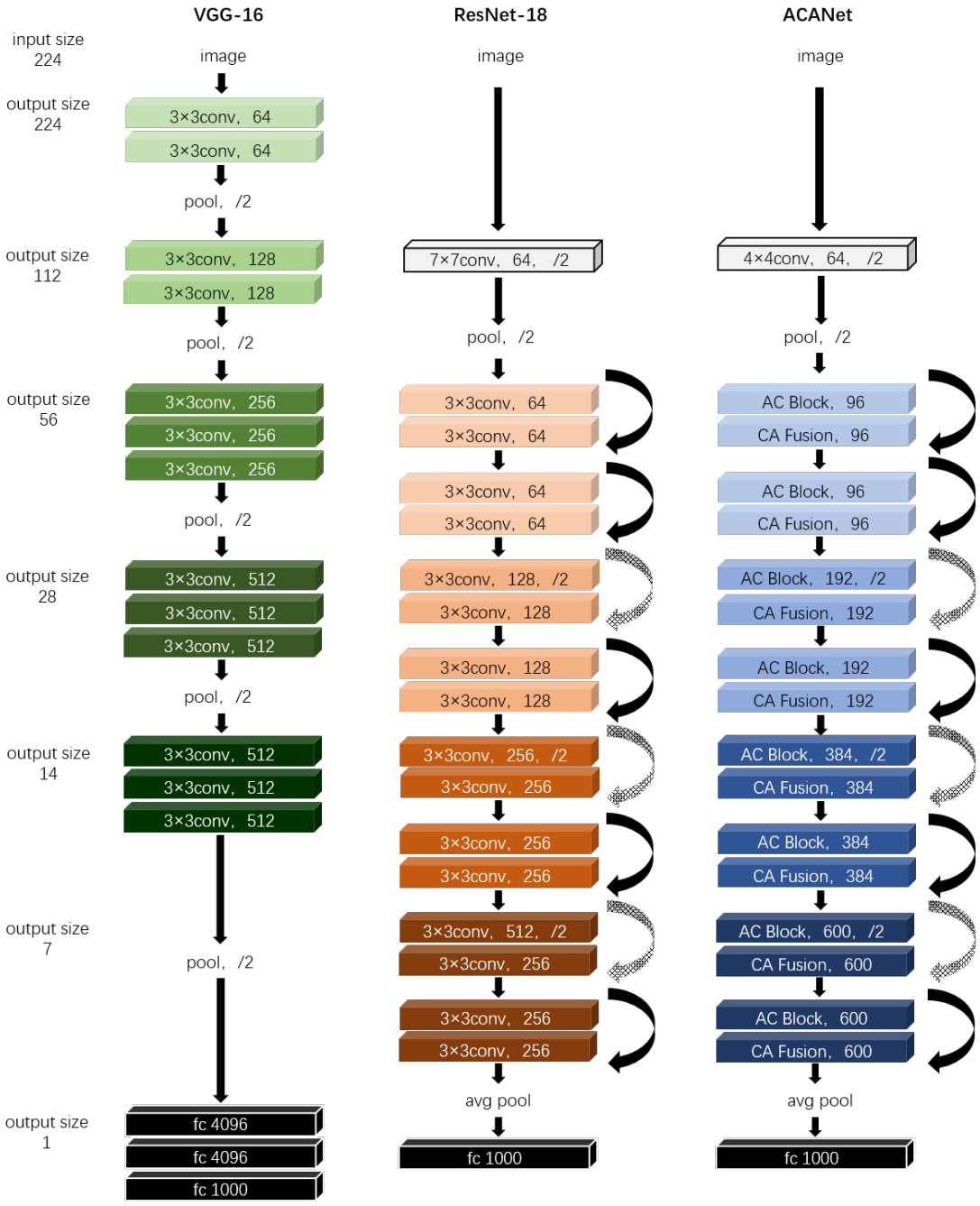
CNN has made a lot of achievements in image recognition. Since the accuracy of AlexNet [38] on ImageNet [39] greatly exceeds that of traditional methods, various convolutional architectures have been proposed to improve the accuracy of recognition. For example, ResNet [2] and DenseNet [37] have performed quite well on various recognition tasks using deep network structures with fast connections, and their pre-trained models on ImageNet have also been successfully transferred to various computer vision tasks. EfficientNet [40] further studied the depth, width, and input resolution of the convolutional network to optimize the network structure. The method successfully maintains high precision with a small number of parameters.

Since the publication of Transformer [41], the self-attention mechanism has been widely used in the field of natural language processing and computer vision, and its performance is quite good. For example, Vit [1] achieved 90.7% accuracy on ImageNet, which convert the patch of the image into a token through linear transformation, and complete the global information exchange through the self-attention mechanism. However, its disadvantage is that there is no hierarchical expression ability for the features of local regions. In SwinTransformer [31], the extraction of local region features at different scales is done through a multi-layer self-attention structure. Because of the success of the above architectures, we will use these models as the backbone network to test the capabilities of the proposed modules.

Therefore, this paper tries to combine the advantages of CNN model and ViT model, and invents a new component module. Then, this component module is successfully applied to ResNet, which makes it significantly improved.

Asymmetric convolution is often used to approximate the replacement of the existing square kernel convolution layer for compression and acceleration. Some previous work, such as Ding [8] et al., present a simple but powerful architecture of convolutional neural network, which has a VGG-like inferencetime body composed of nothing but a stack of 3 × 3 convolution and ReLU, while the training-time model has a multi-branch topology; DBB [14] is a universal building block of Convolutional Neural Network to improve the performance without any inference-time costs. They also showed that the traditional standard d×d convolution can be decomposed into a combination of 1 × d and d × 1 convolution to reduce the amount of computation required. The principle

behind this is that if a 2D convolution kernel has rank 1, then the operation can be equivalent to a series of 1D convolutions. However, due to the distributed eigenvalues in the deep network, its internal rank is often not equal to 1 in practice, so the direct decomposition of the convolution kernel will cause significant information loss.



(a) The overall structure diagram of VGG-16 network, where pool represents the pooling layer, and reduce the image size halved. (b) The ResNet-18 network structure diagram. In particular, the curve on the right side represents residual connection, and the virtual curve represents reduction of the feature map size through the downsampling layer. (c) The method of ACANet in this paper. It should be noted that the downsampling layer used in the virtual curve on the right of the figure is the one using 2×2 convolution.

Fig. 2. For the three main network structure diagrams related to the work in this paper, input size and output size are respectively the input and output sizes of this layer

Therefore, we do not directly replace the original layer with a one-dimensional asymmetric convolution as the decomposed convolution layer, but as part of the architectural design. This way we can enrich the feature space while training and then incorporate what we learn from it into parallel convolution layers.

Although convolutional neural networks, which use convolutional kernels to obtain local feature information, have become the most ubiquitous technology in various computer vision tasks. At the same time, however, the self-attention mechanism also shows its excellence in a wide range of natural languages processing tasks, such as BERT and GPT3. According to relevant theoretical analysis, when the memory capacity is strong enough, the self-attention mechanism can completely replace convolution. Therefore, some recent studies have explored the possibility of using self-attention mechanisms in computer vision tasks, such as the work of Alexey [1] et al. and Jie Hu et al. [9]. There are two main technical routes, one directly uses the self-attention mechanism as a component in the network, and the other is about the integration with the self-attention mechanism and convolution. The approach used in this article is more of the second.

However, although the above methods have achieved very remarkable results. But, most methods simply add the self-attention mechanism directly into the convolutional neural network. These methods have some problems, such as long reasoning time and heavy burden on hardware. So, we devise a whole new network by fusing convolution and self-attention mechanisms.

3 Methods

We use asymmetric convolution to enhance the data of traditional convolution, so that convolution can better extract the features of input images. We call it Asymmetric Convolution and Convolution Block (ACCB). Then, we propose a new fundamental building block by combining the similarities between convolution and self-attention mechanisms. We call it Fusion of Convolution and Self-Attention (FCSA). After that, we integrate ACCB and FCSA into the residuals block of ResNet to propose a new residuals module. We call this network of new residual module Asymmetric Convolutional Attention Networks (ACANet). At last, we design a new VIT-like downsampling layer for ACANet to help the convolutional neural network adapt to the self-attention mechanism. The overall structure of the network can be shown in Fig. 2.

3.1 Asymmetric Convolution and Convolution Block (ACCB)

Asymmetric convolution has been proposed for a long time. It is used to improve model performance by replacing the existing traditional square convolution of $k \times k$ size. Some previous work has shown that the traditional $k \times k$ convolution layer can be split into two $1\times k$ and $k \times 1$ convolution layers to reduce the model parameters and the amount of computation required to improve the model. Its theoretical basis is: if the rank of a two-dimensional matrix is 1, it can be equivalently converted into a series of one-dimensional matrices. Similar studies, such as Max Jaderberg [20], successfully learned horizontal convolution kernel and vertical convolution kernel by minimizing L-2 reconstruction error. Different from the former, in this paper, asymmetric convolution is regarded as a supplement to the symmetric convolution of the trunk, and the lightweight advantage of asymmetric convolution is utilized to try to improve the model performance at a small cost. The network structure is shown in Fig. 3.

We note that convolution has a property that if two or more sid-compatible two-dimensional convolution kernels compute the same input feature graph at the same step size, producing an output of the same resolution, then their outputs can be added. This property inspired us to add his output so that multiple convolution nuclei of different sizes are equivalent to one convolution kernel. This process can be expressed as:

$$I \times K_1 + I \times K_2 = I \times (K_1 + K_2). \tag{1}$$

Where I is the input feature graph, and respectively represent two $2D$ convolution kernels with the same compatible size. Therefore, it can be seen from the above equation that a symmetric convolution can be enhanced by multiple asymmetric convolutions with the same compatible size and then merged into the same convolution.

In the training stage, each convolution kernel is trained through the input feature graph; in the reasoning stage, the structure reparameter technology is used to fuse multiple convolution nuclei into a traditional symmetric convolution to initialize the network parameters. Therefore, in the reasoning stage, the network structure is exactly

the same as the original network. However, the network parameter adopts the parameter with stronger feature extraction ability, namely the convolution kernel parameter after fusion, so the calculation amount will not be increased in the inference stage.

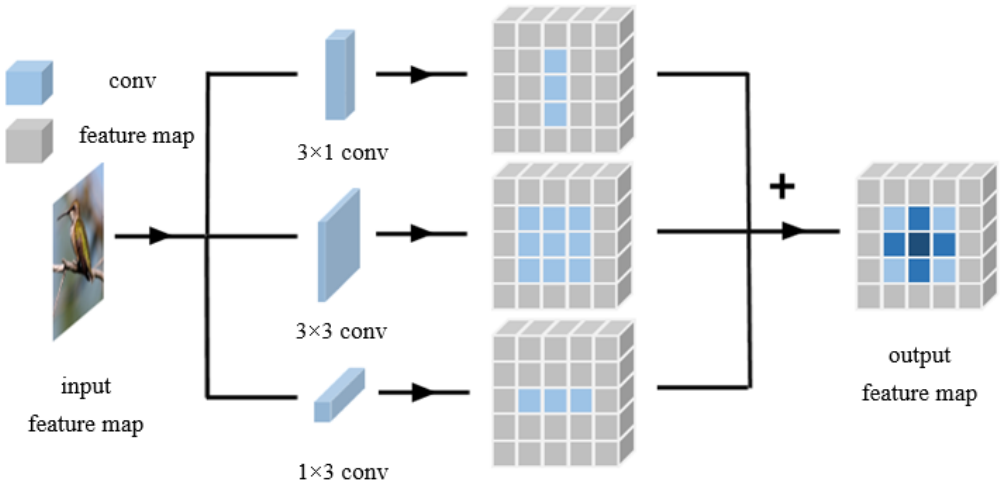


Fig. 3. A schematic diagram of using asymmetric convolution as the asymmetric convolution module enhanced by trunk symmetric convolution data in the training stage

The branch fusion process in this paper refers to the fusion of the asymmetric convolution kernel into the corresponding position of the square symmetric convolution through structural reparameter technology, so as to improve the performance of the model without increasing the number of parameters. In practice, this technique is implemented by first building the original structure of the network and using the fused weights for initialization. The fusion process can be shown in equation (2).

$$X'^{(i)} = \frac{\alpha_i}{\beta_i} X^i + \frac{\alpha_{(1)i}}{\beta_{(1)i}} X^{(i)}_{(1)} + \frac{\alpha_{(2)i}}{\beta_{(2)i}} X^{(i)}_{(2)}. \tag{2}$$

Where X^i represents the symmetric convolution kernel of the square, and represent the asymmetric convolution kernel of 1 × 3 and 3 × 1, respectively, and x' represents the new convolution kernel after merger. The offset term bi can be shown in equation (3).

$$b_i = -\frac{\gamma_i \alpha_i}{\beta_i} - \frac{\gamma_{(1)i} \alpha_{(1)i}}{\beta_{(1)i}} - \frac{\gamma_{(2)i} \alpha_{(2)i}}{\beta_{(2)i}} + \delta_i + \delta_{(1)i} + \delta_{(2)i}. \tag{3}$$

Therefore, it is easy to conclude that for any filter X_i , the asymmetric convolution branch fusion process can be shown in equation (4).

$$Y_i + Y_{(1)i} + Y_{(2)i} = \sum_{k=1}^{N} Z_k \times X'^{(i)}_k + b_j. \tag{4}$$

Where Y_i, $Y_{(1)i}$ and $Y_{(2)i}$ are the outputs of the original 3 × 3 symmetric convolution, 1 × 3 and 3 × 1 asymmetric convolution, respectively. In particular, although this module can be converted into the standard layer equivalent, the equivalence is only valid when reasoning, and because there are different situations in the training process, it will show different performance in the reasoning process. The non-equivalence of the training process is caused by the random initialization of the kernel weights and the gradient derived from the different computational flows in which they participate.

3.2 Fusion of Convolution and Self-Attention (FCSA)

As an important part of the convolutional neural network, the convolution is assumed to be a 3 × 3 convolution with step size 1, and its specific process is shown in Fig. 4. Consider a standard convolution with a kernel of $Z \in R^{C_{in} \times C_{out} \times k^2}$, where k is the kernel size and C_{in} and C_{out} represent the number of input and output channels, respectively. Given that the tensors $X \in R^{C_{in} \times H \times W}$ and $Y \in R^{C_{out} \times H \times W}$ are feature maps of input and output, where H and W represent height and width, we express $X_{ij} \in R^{C_{in}}$ and $Y_{ij} \in R^{C_{out}}$ as feature tensors of pixels (i, j) corresponding to X and Y, respectively. So standard convolution can be shown in formula (5).

$$Y_{ij} = \sum_{a,b} Z_{a,b} f_{i+a-\frac{k}{2}, j+b-\frac{k}{2}}. \tag{5}$$

Where $Z_{a,b} \in R^{C_{in} \times C_{out}}$, $a, b \in \{0, 1, \ldots, k-1\}$, Represents the weight of the kernel location (a, b).

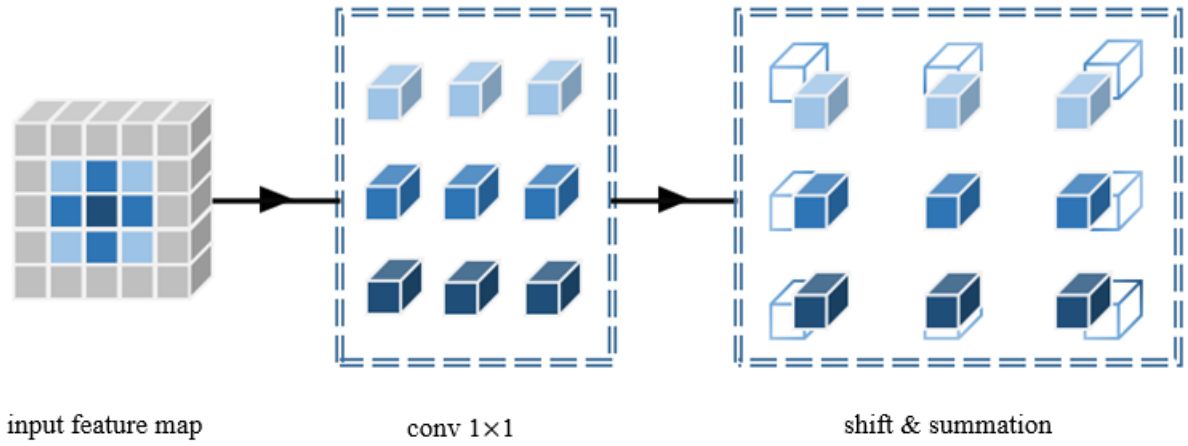


Fig. 4. Taking 3 × 3 convolution as an example, independent 1 × 1 convolution is performed on the input feature graph and then the convolution result is obtained by shifting and merging operations. In this process, the independent convolution calculation assumes a large part of the computational power consumption.

So, the standard convolution can be summarized as two stages:

$$\text{STAGE I: } Y'^{(a,b)}_{i,j} = Z_{a,b} \times f_{i,j}. \tag{6}$$

$$\text{STAGE II: } Y^{(a,b)}_{i,j} = Shift(Y'^{(a,b)}_{ij}, a - \frac{k}{2}, b - \frac{k}{2}). \tag{7}$$

$$Y_{ij} = \sum\nolimits_{a,b} Y^{(a,b)}_{ij}. \tag{8}$$

In the first stage, which is the single convolution calculation stage, the input feature map is linearly projected from a position onto the weight of the convolution kernel. This is almost identical to the classical 1 × 1 convolution. However, in the second stage, which is the shift and aggregation stage, the projected feature map moves according to the position of the convolution kernel and aggregates together. According to the careful calculation in this paper, the main computational burden in the process of convolution operation comes from the first stage, while the second stage is relatively easy.

Due to the proposal of ViT model, attention mechanism has become another important component in addition to convolution. Compared with traditional convolution, attention mechanism can make the model pay more at-

tention to a larger range of image information. The operation process of attention mechanism can be shown in Fig. 5.

If a multi-head attention mechanism with multiple heads has the number of heads N, $X \in R^{C_{in} \times H \times W}$, $Y \in R^{C_{out} \times H \times W}$ represents input and output features, and $x_{ij} \in R^{C_{in}}$, $y_{ij} \in R^{C_{out}}$, represents the corresponding tensor corresponding to a particular point (i, j) in the image. Thus, a single head in the multi-head attention mechanism can be shown in Equation (9).

$$y_{ij}^{(N_1)} = \sum_{a,b \in N_k(i,j)} A(Z_q^{(l)} x_{ij}, Z_k^{(l)} x_{ab}) Z_v^{(l)} x_{ab}. \tag{9}$$

Where Z_q, Z_k, Z_v is the projection matrix corresponding to Q, K, V, and $N_k(i, j)$ represents the local region of space range k with (i, j) as the center pixel. And $A(W_q^{(l)} x_{ij}, W_k^{(l)} x_{ab})$ is about $N_k(i, j)$ characteristics of the corresponding matrix.

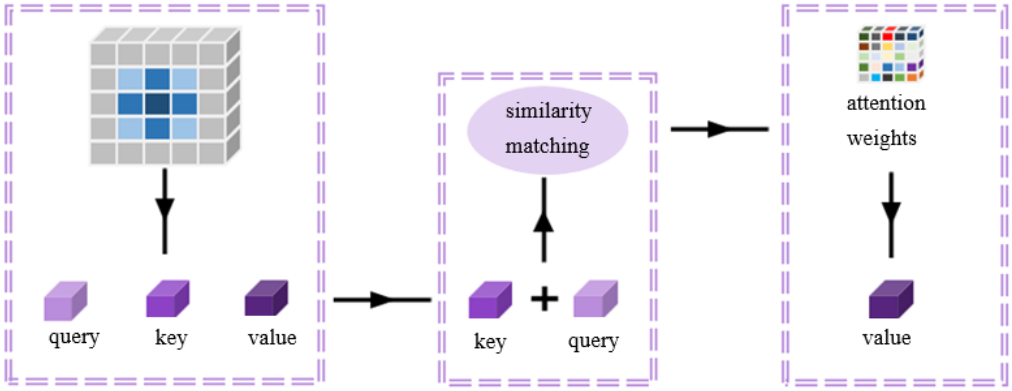


Fig. 5. The input feature map is first projected into three values, Q, K and V, which are similar to the 1 × 1 convolution process, and the weight of attention is calculated by the values of Q and K, which is used to aggregate the values to form the final result.

Also, multi-head self-sttention can be decomposed into two stages, and summarized as:

$$\text{STAGE I: } q_{ij}^{(l)} = Z_q^{(l)} x_{ij}, k_{ij}^{(l)} = Z_k^{(l)} x_{ij}, v_{ij}^{(l)} = Z_v^{(l)} x_{ij}. \tag{10}$$

$$\text{STAGE II: } y_{ij} = \Big\|_{l=1}^{N} \Big(\sum_{a,b \in N_k(i,j)} A(q_{ij}^{(l)}, k_{ab}^{(l)} v_{ab}^{(l)}\Big). \tag{11}$$

Similar to the traditional convolution, 1 × 1 convolutions are first conducted in stage I to project the input feature as query, key and value. On the other hand, Stage II comprises the calculation of the attention weights and aggregation of the value matrices, which refers to gathering local features. The corresponding computational cost is also proved to be minor comparing to Stage I, following the same pattern as convolution.The fusion of convolution and attention mechanisms is shown in Fig. 6.

3.3 Downsampling Layer

Generally speaking, the downsampling layer is concerned with how to process the input image data for later operation. Due to the widespread natural redundancy in images, the downsampling layer in ConvNets and Transformer tends to actively downsample the input image to an appropriate size for subsequent operations.

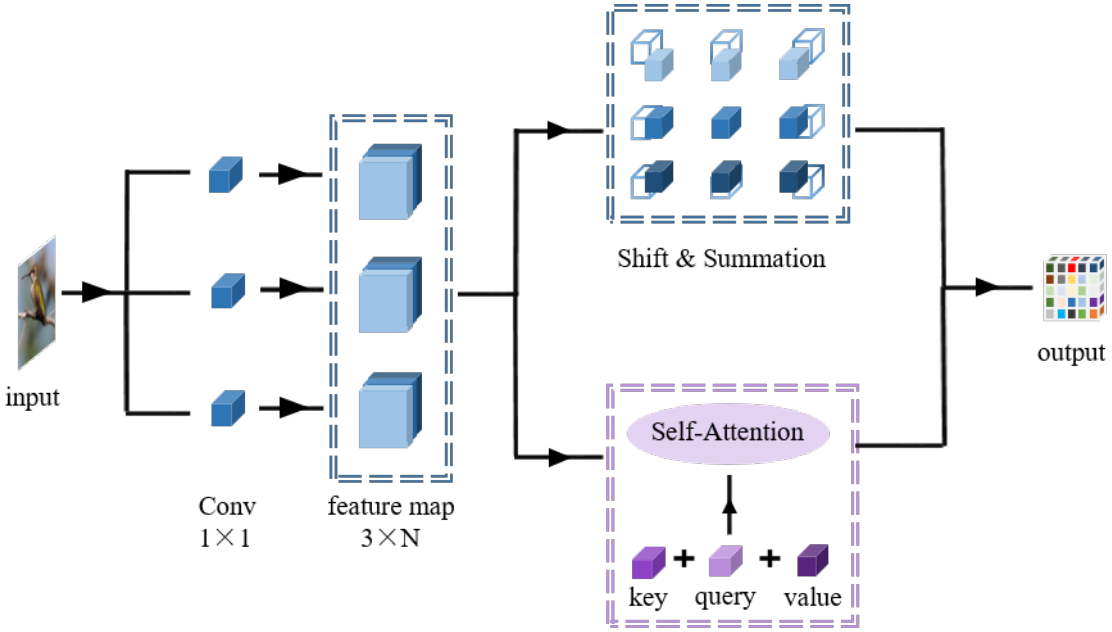


Fig. 6. The input images are first convolved by three groups of multiple 1 × 1 convolution to obtain multiple groups of intermediate feature graphs, and then are combined into unified feature graphs by shift aggregation of convolution operations and concatenation after attention processes. This process achieves an efficient self-attentional mechanism by merging the common parts of convolution and self-attentional mechanisms.

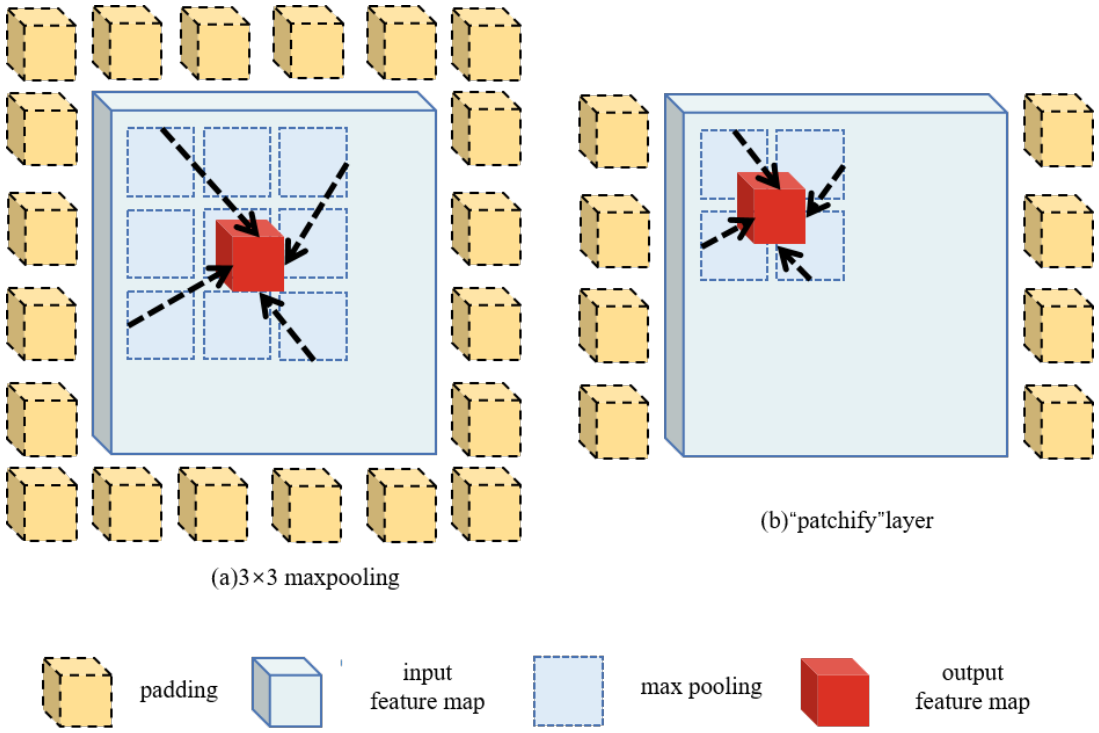


Fig. 7. Diagram of the downsampling layer, wherein (a) The diagram of the classic 3 × 3 maxpooling layer. In order to maintain the same size of the input feature graph and the output feature graph, the periphery of the original feature graph is filled for one week before the pooling operation. (b) The 2 × 2 downsampling layer similar to the patch layer of the ViT model proposed in this paper. In the figure, due to the particularity of 2 × 2 convolution, only two sides of the input feature graph are filled.

In the classical residual network ResNet, a 7 × 7 convolution layer with step size of 2 is first used to convolve the input image, and then the maximum pooling layer is used to further reduce the input image and remove the redundant information. ResNet uses this set of operations as the initial downsampling layer of the network, and the input image passing through this layer will be reduced to a quarter of its original size.

In particular, the Swin Transformer model uses a more specific downsampling strategy. They used a "patchify" layer of size 4 to neatly divide the entire input image into multiple non-overlapping blocks of the same size as the convolution kernel for downsampling operations.Therefore, this paper tries to apply this method to the classical residual network, so that the classical residual network can better play the role of attention mechanism and realize the integration of convolution and attention.

The initial downsampling layer downsample the natural image input to the network to a quarter size. In this paper, convolution kernel with step size 2 is used for the different stages of the residual network. This option allows us to reduce the input image to half of the original image and gradually focus the key information, thus achieving excellent results with limited computational resources. The structure of the downsampling layer can be shown in Fig. 7.

4 Experimental Results and Comparative Analysis

4.1 Experimental Results

We evaluate our method on three widely recognized publicly available fine-grained image classification datasets: CaltechUCSD Birds (CUB200-2011) [11], Standford Cars [7] and Flower-102 [6]. A detailed summary of the datasets is provided in Table 1. In order to be consistent with the other datasets, we only divide them into training sets and test sets.

Table 1. Statistics of datasets

Dataset	Category	Train	Test
CUB-200-2011	200	5994	5794
Standford Cars	196	8144	8041
Flowers-102	102	2040	6149

Based on the experimental data shown in Table 2 and Table 3, we tested our approach on three publicly available fine-grained image classification data sets. We selected three representative methods from existing methods, CE Loss, Focal Loss and MC Loss, in order to reach a more comprehensive conclusion. As you can see, our approach is a much better improvement than the traditional ResNet18 network. For CE Loss, although the three public data sets all have some improvement, the improvement is not obvious, which may be because CE Loss itself has poor adaptability to fine-grained image classification operations, and has good adaptability to the new network structure proposed in this paper. Therefore, although it shows a relatively stable improvement effect, the overall improvement is not obvious. Focal Loss, produces a good effect in some datasets, but poor performance in some datasets. This paper believes that Focal Loss can better adapt to the task of fine-grained image classification compared with CE Loss. Therefore, it shows that the method proposed in this paper can be well amplified in some tasks, but sometimes it also magnifies the shortcomings.

Based on the experimental data shown in Table 2, we tested our approach on three publicly available fine-grained image classification data sets. We selected three representative methods from existing methods, CE Loss, Focal Loss and MC Loss, in order to reach a more comprehensive conclusion. As you can see, our approach is a much better improvement than the traditional ResNet18 network. For CE Loss, although the three public data sets all have some improvement, the improvement is not obvious, which may be because CE Loss itself has poor adaptability to fine-grained image classification operations, and has good adaptability to the new network structure proposed in this paper. Therefore, although it shows a relatively stable improvement effect, the overall improvement is not obvious. Focal Loss, produces a good effect in some datasets, but poor performance in some datasets. This paper believes that Focal Loss can better adapt to the task of fine-grained image classification compared with CE Loss. Therefore, it shows that the method proposed in this paper can be well amplified in some tasks, but sometimes it also magnifies the shortcomings.

For MC Loss, although it can be well adapted to fine-grained image classification tasks, there may be some conflicts with the method proposed in this paper, so it shows acceptable improvement in some datasets, but certain fluctuations in more datasets.

Table 2. Results of different methods on the CUB-200-2011, Stanford Cars and Flower-102 datasets

Model	Method	Dataset		
		CUB-200-2011	Stanford Cars	Flower-102
ResNet 18	CE Loss	52.50	83.39	52.23
	Focal Loss [14]	51.45	80.03	50.27
	MC Loss [10]	59.95	86.73	57.11
	Center Loss [12]	50.26	81.84	49.51
	A-softmax Loss [13]	49.67	82.15	50.56
	COCO Loss [15]	46.01	72.38	56.76
	LGM Loss [16]	44.91	74.37	56.84
	LMCL Loss [17]	46.01	71.17	57.12
VGG 16	CE Loss	28.53	76.59	40.90
	Focal Loss	31.12	77.02	48.19
	MC Loss	42.57	82.48	55.33
	Center Loss	51.38	85.27	52.53
	A-softmax Loss	60.79	85.71	52.34
	COCO Loss	48.31	67.27	53.31
	LGM Loss	28.14	71.27	57.18
	LMCL Loss	41.11	49.57	56.43
ACANet (ours)	CE Loss	52.73	84.80	53.37
	Focal Loss	55.06	78.82	51.46
	MC Loss	62.74	86.75	57.40

For the CUB-200-2011 datasets and Flower-102 datasets, the method in this paper can be well adapted to this dataset and complete the task of fine-grained image classification. However, for the dataset Stanford Cars, the method proposed in this paper cannot solve the fine-grained image classification task to some extent, but the effect is not ideal.

As can be seen from the curve data in Fig. 8, compared with ResNet18, the method proposed in this paper produce a big fluctuation when the learning rate decreases to 0.001 in about the 400th epoch. This may be due to the entirely new module of FCSA and ACCB fluctuate when the learning rate decreases because they cannot adapt to each other. This also happens when the learning rate decreases from 0.1 to 0.01 in the 200th epoch.As the change is relatively small, this fluctuation is also relatively small compared with that in the 400th epoch.

For CE Loss, the method in this paper can reduce its amplitude of shock in the plateau period to a certain extent, and mostly increase upward, making the training process of the network smoother, faster and more excellent. For Focal Loss, the method proposed in this paper can also reduce the amplitude of oscillation when it is in the plateau period to a certain extent.

Table 3. Results of different models on the CUB-200-2011, Stanford Cars and Flower-102 datasets

Model	Method	Dataset		
		CUB-200-2011	Stanford Cars	Flower-102
ResNet 18	CE Loss	52.50	83.39	52.23
VGG 16	CE Loss	28.53	76.59	40.90
GoogleNet [34]	CE Loss	29.32	71.21	49.74
MobileNetV3 [35]	CE Loss	41.57	81.72	50.94
DenseNet [37]	CE Loss	43.68	81.29	51.71
ConvNeXt [36]	CE Loss	52.08	83.68	53.05
ACANet (ours)	CE Loss	52.73	84.80	53.37

Although the transient instability occurs when the learning rate is changed in the later period, it is acceptable for the improvement of the overall effect. For MC Loss, although the method presented in this paper performs

slowly before entering the first plateau period, the plateau period is extended to a certain extent, and certain fluctuations will occur when the learning rate is changed in the later period. Considering the increase in the end result, this phenomenon is acceptable.

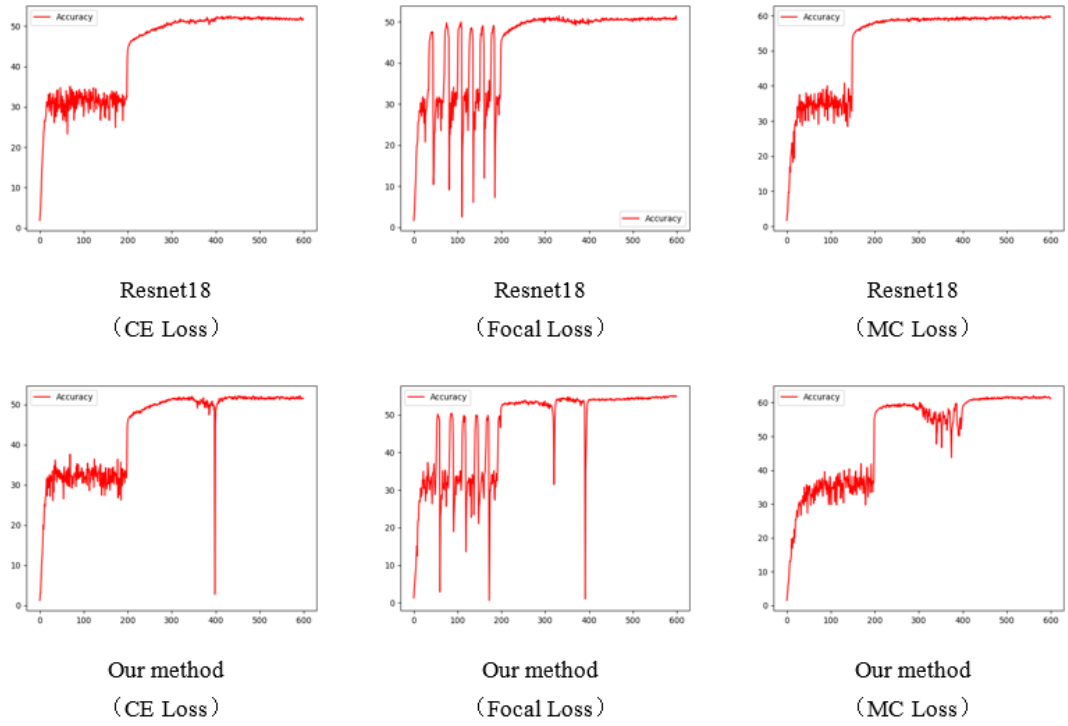


Fig. 8. Test set accuracy curve based on the CUB-200-2011 dataset

(It can be seen that the method proposed in this paper can provide better stability in the early stage of training, while it will show certain fluctuations in the late stage of training.)

4.2 Comparative Analysis

In the ACCB block, we use the asymmetric convolution set to enhance the convolution capability, and use the reparameterization technique to integrate them after the completion of the training phase. With this module, we can improve the capability of the network without increasing the network parameters. However, in the process of experiments and tests, we realized that simply transferring this module to the network for training can not solve the fine-grained image classification task. Therefore, through a large number of experimental studies, we determined that the balance between network capability and computational burden can be achieved when the convolution size is set to 3×3. We also found that asymmetric convolution performs very well when flipping images. Therefore, we believe that adding asymmetric convolution to the network can enhance the anti-flipping performance of the network to help the network better adapt to the diversified situations in the real scene.

We combine the common parts of self-attention and convolution to propose a new FCSA module. This module allows us to use both self-attention and convolution modules at a fraction of the cost to improve network capabilities. However, during the experiment, we also found that different parameter Settings would have a great impact on the experimental results and calculation burden. In addition, different parameters will also produce different effects for different loss functions. Therefore, after a lot of experiments, we finally set different parameter Settings for the above three recognized loss functions, and designed a special downsampling layer for FCSA. Our proposed downsampling layer not only mimics the data form of self-attention habit but also caters to the form of convolution operation. We believe that this downsampling layer can be used as a downsampling auxiliary component of FCSA to help it better deploy to other networks.

5 Ablation Study

For the ablation experiment, we use ResNet18 as the backbone architecture and the size of each input image adjust to 224 × 224. The initial learning rate of the whole network is uniformly set at 0.1, and the learning rate is multiplied by 0.1 in the 200th and 400th epoch, for a total of 600 epoch. We use CE Loss and MC Loss based on the CUB-200-2011 dataset, ACCB and FCSA is replaced by 3 × 3 convolution successively mechanisms are used to study their respective roles and actual roles in the network.

Table 4. Ablation experiments using CE Loss and MC Loss on CUB-200-2011 dataset

Method	Base model	Loss	Accuracy
Complete structure	ACANet	CE Loss	52.73
Minus ACCB	ACANet	CE Loss	52.310.42
Minus (FCSA+ACCB)	ACANet	CE Loss	48.983.75
Complete structure	ACANet	MC Loss	62.74
Minus ACCB	ACANet	MC Loss	60.612.13
Minus (FCSA+ACCB)	ACANet	MC Loss	60.392.35

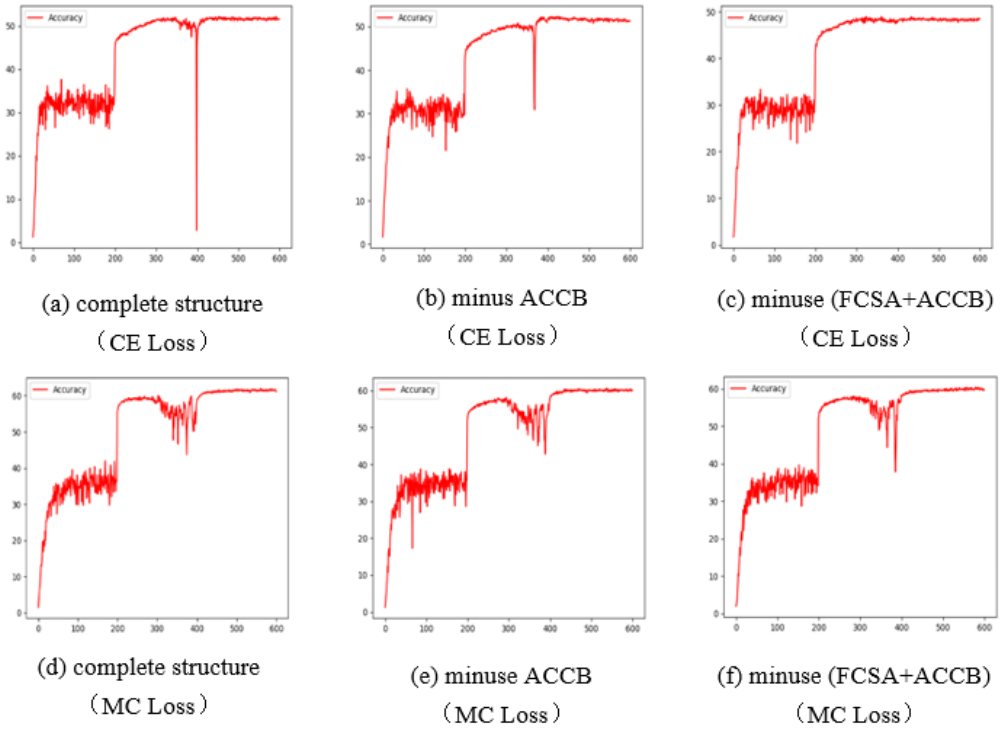


Fig. 9. Accuracy curves of the ablation experimental test set under multiple conditions

(As can be seen with CELoss, the new modules produced some fluctuations in the late training period, but not with MCLoss.)

As can be seen from the Table 4, the ACCB and the FCSA indeed play an important role in the network. When the overall structure of the proposed method is complete in the CUB-200-2011 dataset, the best results are obtained under the two loss functions. However, when the ACCB are removed, MC Loss fluctuates significantly compared with CE Loss. This paper believes that this may be because MC Loss itself is a multi-channel loss function designed on the basis of convolution operation, so when the number of convolution kernels is reduced, it will produce large fluctuations. So, in contrast, the fluctuation in removing the attention mechanism is relatively small. In particular, CE Loss fluctuates greatly when 3 × 3 convolution is used instead of FCSA. Looking at

this change in isolation, it is not clear that the main reason is the attention mechanism or the new downsampling layer. Therefore, combined with the comprehensive analysis of the results of MC Loss under this change, this paper believes that it can be concluded that, compared with other Loss functions, the self-attention mechanism has a more obvious influence on CE Loss.

According to the accuracy curve shown in Fig. 9, when the accuracy switches from 0.01 to 0.001 for more detailed learning at about the 400th epoch, CE Loss has a large fluctuation and this fluctuation will gradually decrease with the ablation experiment, while this does not exist for MC Loss. This paper argues that This may be because CE Loss cannot adapt well to the ACCB and FCSA There are multiple paths of self-attention mechanisms module, and the design logic of MC Loss is unique, so it can better adaptto this change. Therefore, we should pay attention to the choice of loss function when using these two new modules.

6 Conclusions

In this paper, a fine-grained image classification method using convolution fusion attention technique is proposed; It can work effectively using both convolution and attention mechanisms while using fewer network parameters. An improved convolution structure ACCB is proposed for the detailed features in fine-grained image classification. To help the attention mechanism and convolution fusion, we propose a new downsampling structure similar to the VIT model. We evaluated the proposed method on multiple datasets with promising results. Therefore, our proposed fine-grained image classification method has good universality while integrating convolution and attention mechanism.

The current experimental scenario is relatively homogeneous compared with the actual complex application environment. In the actual application scenario, it will have a very complex impact due to the Angle, illumination, distance and other issues of the acquisition device. In future work, we plan to add more complex experimental scenarios or artificially add disturbing information that can be used to train generic and more robust learning models. In order to ensure that the fine-grained image classification and recognition system can maintain a high recognition speed and accuracy in the complex environment.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: Proc. 2021 International Conference on Learning Representations, 2021.

[2] K.-M. He, X.-Y. Zhang, S.-Q. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[3] T. Berg, P.N. Belhumeur, Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation, in: Proc. 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013.

[4] Z.-Y. Ma, D.-L. Chang, J.-Y. Xie, Y.-F. Ding, S.-G. Wen, X.-X. Li, Z.-W. Si, J. Guo, Fine-Grained Vehicle Classification With Channel Max Pooling Modified CNNs, IEEE Transactions on Vehicular Technology 68(4)(2019) 3224-3233.

[5] T.T. Hoang, T. Do, T.V. Nguyen, N.M. Cheung, Unsupervised Deep Cross-modality Spectral Hashing, IEEE Transactions on Image Processing 29(2020) 8391-8406.

[6] M.-E. Nilsback, A. Zisserman, Automated Flower Classification over a Large Number of Classes, in: Proc. 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing, 2008.

[7] J. Krause, J. Gebru, J. Deng, L.-J. Li, F.-F. Li, Learning Features and Parts for Fine-Grained Recognition, in: Proc. 2014 22nd International Conference on Pattern Recognition, 2014.

[8] X.-H. Ding, X.-Y. Zhang, N.-N. Ma, J.-G. Han, G.-G. Ding, J. Sun, RepVGG: Making VGG-style ConvNets Great Again, in: Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[9] J. Hu, L. Sheng, G. Sun, Squeeze-and-Excitation Networks, in: Proc. 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[10] D.-L. Chang, Y.-F. Ding, J.-Y. Xie, A.K. Bhunia, X.-X. Li, Z.-Y. Ma, M. Wu, J. Guo, Y.-Z. Song, The Devil is in the Channels: Mutual-Channel Loss for Fine-Grained Image Classification, IEEE Transactions on Image Processing 29(2020) 4683-4695.

[11] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset, california institute of technology, 2011.

[12] Y.-D. Wen, K.-P. Zhang, Z.-F. Li, Y. Qiao, A Discriminative Feature Learning Approach for Deep Face Recognition, in: Proc. 2016 European Conference on Computer Vision, 2016.

[13] W.-Y. Liu, Y.-D. Wen, Z.-D. Yu, M. Li, B. Raj, L. Song, SphereFace: Deep Hypersphere Embedding for Face Recognition, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[14] X.-H. Ding, X.-Y. Zhang, J.-G. Han, G.-G. Ding, Diverse Branch Block: Building a Convolution as an Inception-like Unit, in: Proc. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
[15] Y. Liu, H.-Y. Li, X.-G. Wang, Rethinking Feature Discrimination and Polymerization for Large-scale Recognition, in: Proc. 2017 Computer Vision and Pattern Recognition, 2017.
[16] W.-T. Wang, Y.-Y. Zhong, T.-P. Li, J.-S. Chen, Rethinking Feature Distribution for Loss Functions in Image Classification, in: Proc. 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[17] H. Wang, Y.-T. Wang, Z. Zhuo, X. Ji, D.-H. Gong, J.-C. Zhuo, Z.-F. Li, W. Liu, CosFace: Large Margin Cosine Loss for Deep Face Recognition, in: Proc. 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[18] X.-X. Li, L.-Y. Yu, D.-L. Chang, Z.-Y. Ma, J. Cao, Dual Cross-Entropy Loss for Small-Sample Fine-Grained Vehicle Classification, IEEE Transactions on Vehicular Technology 68(5)(2019) 4204-4212.
[19] R. Cong, N. Yang, C.-Y. Li, H.-Z. Fu, Y. Zhao, Q.-M. Huang, S. Kwong, Global-and-Local Collaborative Learning for Co-Salient Object Detection, IEEE Transactions on Cybernetics 53(3)(2023) 1920-1931.
[20] M. Jaderberg, A. Vedaldi, A. Zisserman, Speeding up Convolutional Neural Networks with Low Rank Expansions, in: Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[21] H.-L. Zheng, J.-L. Fu, T. Mei, J.-B. Luo, Learning multi-attention convolutional neural network for fine-grained image recognition, in: Proc. 2017 IEEE International Conference on Computer Vision, 2017.
[22] T. Hu, H.-G. Qi, See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification, in: Proc. 2019 IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[23] M. Sun, Y.-C. Yuan, F. Zhuo, E.-R. Ding, Multi-attention multi-class constraint for fine-grained image recognition, in: Proc. 2018 European Conference on Computer Vision, 2018.
[24] W. Luo, X.-T. Yang, X.-J. Mo, Y.-H. Lu, L.S. Davis, J. Li, J. Yang, S.N. Lim, Cross-x learning for fine-grained visual categorization, in: Proc. 2019 IEEE/CVF International Conference on Computer Vision, 2019.
[25] P.-Q. Zhuang, Y.-L. Wang, Y. Qiao, Learning attentive pairwise interaction for fine-grained classification, in: Proc. 2020 AAAI conference on artificial intelligence, 2020.
[26] T. Zhang, D.-L. Chang, Z.-Y. Ma, J. Guo, Progressive co-attention network for fine-grained visual classification, in: Proc. 2021 International Conference on Visual Communications and Image Processing, 2021.
[27] A. Behera, Z. Wharton, P. Hewage, A. Bera, Context-aware attentional pooling (cap) for fine-grained visual classification, in: Proc. 2021 AAAI conference on artificial intelligence, 2021.
[28] A. Bera, Z. Wharton, Y.-H. Liu, N. Bessis, A. Behera, Sr-gnn: Spatial relation-aware graph neural network for fine-grained image categorization, IEEE Transactions on Image Processing 31(2022) 6017-6031.
[29] Y.-M. Rao, G.-Y. Chen, J.-W. Lu, J. Zhou, Counterfactual attention learning for fine-grained visual categorization and re-identification, in: Proc. 2021 IEEE/CVF International Conference on Computer Vision, 2021.
[30] J. Wang, X.-H. Yu, Y.-S. Gao, Feature fusion vision transformer for fine-grained visual categorization, in: Proc. 2021 32nd British Machine Vision Conference, 2021.
[31] H.-B. Sun, X.-T. He, Y.-X. Peng, Sim-trans: Structure information modeling transformer for fine-grained visual categorization, in: Proc. 2022 30th ACM International Conference on Multimedia, 2022.
[32] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y.-T. Bai, C.-H. Wang, Transfg: A transformer architecture for fine-grained recognition, in: Proc. 2022 AAAI Conference on Artificial Intelligence, 2022.
[33] Y. Zhang, J. Cao, L. Zhang, X.-C. Liu, Z.-Y. Wang, F. Ling, W.-Q. Chen, A free lunch from vit: adaptive attention multi-scale fusion transformer for fine-grained visual recognition, in: Proc. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.
[34] C. Szegedy, W. Liu, Y.-Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going Deeper With Convolutions, in: Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[35] H. Andrew, S. Mark, C. Grace, L.-C. Chen, B. Chen, M.-X. Tan, W.-J. Wang, Y.-K. Zhu, R.-M. Pang, V. Vijay, Q.-V. Le, H. Adam, Searching for MobileNetV3, in: Proc. 2019 IEEE/CVF International Conference on Computer Vision, 2019.
[36] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
[37] G. Huang, Z. Liu, L.V.D. Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, in: Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[38] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Communications of the ACM 60(6)(2017) 84-90.
[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, F.-F. Li, Imagenet: A large-scale hierarchical image database, in: Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[40] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: Proc. 2019 36th International Conference on Machine Learning, 2019.
[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proc. 2017 Advances in Neural Information Processing Systems, 2017.
[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proc. 2021 IEEE/CVF International Conference on Computer Vision, 2021.