Applied Intelligence (2022) 52:2872–2883
https://doi.org/10.1007/s10489-021-02573-2

Multilayer feature fusion with parallel convolutional block for fine-grained image classification

Lei Wang[1] · **Kai He**[1] · **Xu Feng**[1] · **Xitao Ma**[1]

Accepted: 27 May 2021 / Published online: 24 June 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Fine-grained image classification aims at classifying the image subclass under a certain category. It is a challenging task due to the similar features, different gestures and background interference of the images. A key issue in fine-grained image classification is to extract the discriminative regions of images accurately. This paper proposed a multilayer feature fusion (MFF) network with parallel convolutional block (PCB) mechanism to solve this problem. We use the bilinear matrix product to mix different layers' feature matrixes and then add them to the fully connection layer and the softmax function. In addition, the original convolutional blocks are replaced by the proposed PCB, which has more effective residual connection ability in extracting the region of interest (ROI) and the parallel convolutions with different sizes of kernels. Experimental results on three international available fine-grained datasets demonstrate the effectiveness of the proposed model. Quantitative and visualized experimental results show that our model has higher classification precision compared with the state-of-the-arts ones. Our classification accuracy reaches 87.1%, 91.4% and 93.4% on the dataset CUB-200-2011, FGVC Aircraft and Stanford Cars, respectively.

Keywords Multilayer feature fusion · Parallel convolutional blocks · Fine-grained classification · Deep learning

1 Introduction

Fine-grained image classification is to identify the subclass of an image under a specific category. Although deep neural networks (DNN) [1–6] perform well in general image classification owning to their effective information extraction ability, they always fail in fine-grained image classification due to the light or shading in the images, which largely raises the difficulty of discriminative feature extraction. Furthermore, inter-class similarity and intra-class difference among the fine-grained images also affect the classification result greatly. So it is quite difficult to realize the accurate fine-grained image classification only by broadening or deepening CNN network. In this case, extracting the accurate region of interest (ROI) has become the key pointinfine-grained image classification. By now, many methods were proposed to solve this challenging problem. In order to get the accurate location of ROI, some scholars [7–10] applied expensive human annotations to extract the salient parts of an image. However, using manual information is unrealistic in practice. In this case, more and more approaches [13, 16–18, 20] focused on using the image itself and its corresponding labels to extract discriminative informationin recent years.

In this paper, we improve the original convolutional block with the proposed parallel convolutional block (PCB) by extracting the salient features as the output information. In this process, a PCB of different kernel sizes is proposed to improve the residual connection and prevent the feature information missing. Considering that higher layers tend to extract more discriminative features and lower layers have more global information, we proposed a new model called multilayer feature fusion (MFF), which uses the matrix multiplication on the last hierarchy to combine the different layers in pairs. In this way, our model can effectively improve the ability of inter-layer information interaction. Multiple experiments were conducted on the international datasets using the proposed method.

The main contributions are summarized as follows:

✉ Kai He
hekai@tju.edu.cn

[1] Tianjin University, Tianjin, 300072, China

1. We propose a MFF network, which uses bilinear matrix product to combine the features in multilayer for fine-grained image classification. In this way, the global information and the local feature can be adequately considered as a whole.
2. We propose an improved convolutional blocks (PCB). Compared with the original convolutional blocks (CB), the proposed PCB is helpful to extract more salient features due to using the parallel convolutions with different sizes of kernels and their effective residual connection.
3. We compare the proposed model with state-of-the-art ones on three internationally available fine-grained datasets. Visualized and quantitative experimental results show the superiority of the proposed model.

The rest of this paper is organized as follows. Section 2 introduces the related works. Section 3 illustrates the proposed model. Section 4 shows the experiment results on three datasets and Section 5 concludes the whole paper.

2 Related work

This part introduces the previous works related to ours from two viewpoints, including the fine-grained image classification and the fine-grained feature extraction.

2.1 Fine-grained image classification

Fine-grained image classification, also known as subcategory image classification, is a branch of general image classification. The goal of fine-grained image classification is to distinguish the subclass images from those in the same basic class. Compared to the common image classification, fine-grained one is more difficult due to the similar features, large intra class differences but small inter class differences, as well as the interference from different perspectives and backgrounds. In the original fine-grained image classification researches, most common methods are based on the strong supervised learning, which utilize the provided border and label information to improve the classification accuracy. For example, Zhang et al. [7] utilized the bounding box annotations to learn the geometric constraints between the whole object and part detectors. Berg et al. [9] proposed a part-based one-to-one feature method for fine-grained categorization. However, these methods are always unrealistic due to the requirement of a large number of human and material resources.

Compared with the strong supervised learning algorithms, the weak supervised learning ones [11–15] usually have good feature expression ability. These algorithms have been widely used recently due to only using the images themselves and their corresponding tags. For example, Zhang et al. [12] picked the deep filter responses after computing the response of each candidate patch. Besides, Xiao et al. [11] applied a two level attention model, which contained an object level filter net and a part level attention for fine-grained image classification.

2.2 Fine-grained feature extraction

In order to improve the robustness of fine-grained image classification, most researches tend to enhance the interconnection between the neural networks in different layers by changing the network structure. For example, Bilinear Convolutional Neural Network (BCNN) [13] combined CNNA and CNNB with matrix outer product and used average pool to obtain bilinear feature representation. In this way, the two networks can cooperate with each other to conduct the class detection and target feature removal, and then better complete fine-grained image recognition. Considering the huge amount of parameters and computation of BCNN [13], the methods in [16, 17, 20] further reduced the computational complexity while keeping the accuracy. Hierarchical bilinear pool (HBP) [18] improved the feature representation of network structure by enhancing the interaction between different layers. In addition, boost CNN [19] improved the usability by only using the category tags to achieve fine-grained image classification. HIHCA [22] improved the classification accuracy by fusing high-order multi-level convolution features. Discriminative Filter Learning (DFL) [30] proved that the feature learning can be enhanced in CNN network structure, and designed a new asymmetric multi-stream structure based on hierarchical information and global appearance. Navigator-Teacher-Scrutinizer (NTS) [31] network, which is composed of navigator, teacher and scrutinizer agents, can be regarded as a kind of multi-agent cooperation. In which different agents benefit from each other to provide more accurate fine-grained classification in the reasoning process. Besides, MOMN [35] proposed a multi-objective matrix normalization method, which uses three methods to normalize the bilinear representation.

In recent years, more and more scholars used attention model [21, 24, 25, 27] to extract ROI. For example, RA-CNN [24] proposed a new cyclic attention convolutional neural network, which optimized the intra scale classification and the inter scale ranking loss to learn more accurate regional attention and fine-grained representation. MA-CNN [25] used a multiple attention convolutional neural network for detection by learning the location and fine-grained features simultaneously. OPAM [26] used object part two-level attention model to promote multi-view and multi-scale feature learning, and achieved good

performance. HBPASK [29] proposed a new hierarchical bilinear pool framework with mask mechanism inspired by SCDA [41], to capture the cross-layer interactions of local characteristics. Bilinear attention network (BAN) [34] supervised the learning process of attention map by proposing attention center loss and attention dropout. In addition, Zheng et al. [36] proposed a three line attention mechanism to locate the details, and extract and optimize them as the classification basis.

In addition, some other scholars improved the classification accuracy by accurately locating the salient regions of images [28, 33, 40]. For example, Chen et al. [32] used the "broken reconstruction" learning method to solve the problem of fine-grained image recognition. In which, the image was decomposed into several local blocks and was randomly scrambled, and then the original and the reconstructed images were distinguished using anti loss. Zheng et al. [40] proposed a progressive attention network to locate the discriminating part on multiple scales. Besides, [33] used the random local model to locate ROI.

3 Proposed approach

This paper proposes novel network architecture to extract more discriminative features of fine-grained images. Inspired by the previous works [13, 18], we attempt to realize the interaction of different layers with the same dimensions. We adopt ResNet34 as the basic model and extract the feature maps in the middle layers. Besides, we replace the original blocks in the deep layer network with the proposed PCB, which is a two stream parallel structure with an effective residual connection to prevent the information missing and to mine ROI. In addition, we propose a multilayer feature fusion (MFF) network to enhance the information interaction ability between different network layers as well as the the feature description ability of model structure. Different from HBP [18], which uses multiplication by elements in the last layer, our proposed MFF uses bilinear matrix for the output of different series parallel convolution blocks. The bilinear operators in different layers are beneficial to strengthen the feature expression ability of the proposed network structure.

Compared with the existing feature fusion algorithms, ours is multi-layer feature fusion with parallel convolution blocks. We also propose an improved parallel convolution block for high-level feature extraction and use the proposed MFF to fuse the high-level feature matrix with the same dimension information obtained from the parallel convolution block. Thus, a new and complete classification model is established to realize the fine-grained image classification.

3.1 Parallel Convolutional Block (PCB)

The original and the improved convolution blocks are shown in Fig. 1, where Fig. 1a represents the original convolution block CB, and Fig. 1b is our proposed parallel convolution block (PCB). As shown in Fig. 1a, the layer 4 module in the ResNet34 network (CB) is composed of three 3×3 convolution layers. It is used to reduce the size of feature map and expand the number of channels. However, due to only using the output of the last layer in high-level feature extractor and sending it to the full connection layer as the classification basis, CB tends to ignore the key feature information in the shallow layer network, thus the final classification effect will be reduced.

The deep structure in the last layer network usually has strong discriminative feature extraction ability, while the shallow structure has effective global feature information. Both of them are conducive to locate the whole target. Therefore we proposed the PCB, as shown in Fig. 1b, to make full use of the relevant characteristics of the three convolution layers in the original layer4 module. Our proposed PCB uses a two stream structure with different convolutional kernels to extract abundant features. In addition, the proposed PCB is helpful to prevent information losing in the residual connection by changing the convolution with a larger one and setting stride to 2 for down-sampling. The parallel convolution operators with different kernel sizes are essential for extracting ROI. So this paper uses different convolutions to obtain two representative feature maps, as shown in layer4.1 in Fig. 1b. In which, we use concat operation to splice the different feature representation matrixes and expand the channel dimension, thus the diversity of features are increased.

We use the convolution kernels with the size of 1×3 and 3×1 for further feature extraction. Considering that $Conv_{1\times1}$ tends to loss some feature information, we replace it with $Conv_{3\times3}$ as the residual connection to match dimension information. Besides, we replace $Conv_{3\times3}$ with $Conv_{1\times3}$ and $Conv_{3\times1}$ to further reduce the computational complexity.

It is well known that pooling layer can reduce the output eigenvector of convolution layer. However, the feature information dimensions of the deep and shallow structures in layer 4 must be consistent. So we change the concat operation to direct addition, and remove the activation function and the pooling layer in the middle of layer 4.2. In addition, in order to match the number of channels and reduce the complexity of computation, we cancel the convolution layer in residual connection of layer 4.2. Similarly, the structure of layer 4.3 is consistent with that of layer 4.2.

In addition, in order to prevent the gradient disappearing problem, we replace the ReLU activation function, defined

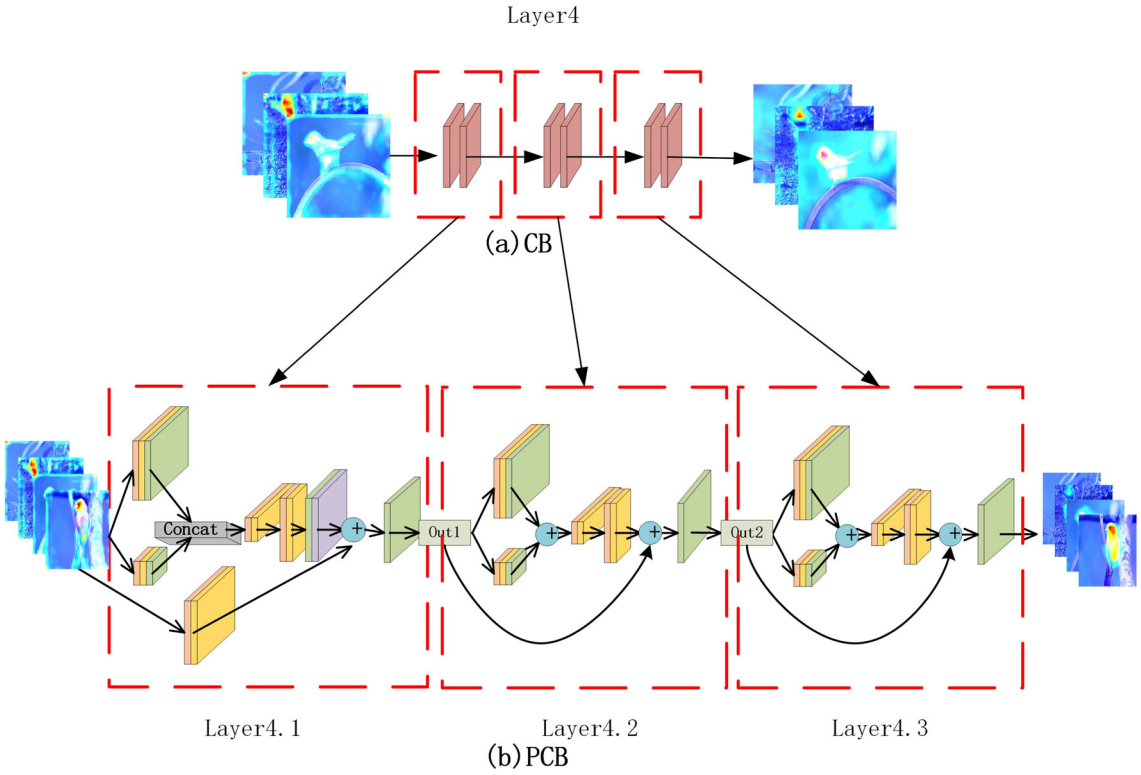


Fig. 1 Parallel convolutional blocks (PCB): **a** the original convolution block CB **b** the proposed PCB blocks. Where Pink, yellow, green and purple blocks represent the convolution, batchnorm, leakyRelu, and maxpool, respectively. Concat represents splicing on the channel dimension and + means direct addition. In layer4.1, the channel dimension is extended using concat operation. In the jump connection, the feature extraction filter with large receptive field is used to sample the feature map. In layer 4.2 and 4.3, the direct addition is used as dimension processing to maintain the dimension information obtained from layer 4.1

as formula (1) in CB, with the Leaky ReLU, defined as formula (2). Where a represents the gradient of negative interval. After multiple attempts, we take the parameter a as 0.2 to achieve the best fitting effect of nonlinear function.

$$f(x) = max(0, x) \tag{1}$$

$$f(x) = max(ax, x)(0 < a < 1) \tag{2}$$

3.2 Multilayers Feature Fusion (MFF)

To extract more discriminative features, we replace the original CB with the proposed PCB as the feature extractor in high layer. In addition, to strengthen the inter-layer information interaction and enhance the feature expression ability of network structure, we fuse the deep-layer feature output with the shallow-layer feature matrix through the matrix outer product and normalization operation. In this way, the discriminative regions can be accurately obtained, and then the feature expression ability is improved. Suppose $X \in R^{h_x \times w_x \times c_x}$, $Y \in R^{h_y \times w_y \times c_y}$ and $Z \in R^{h_z \times w_z \times c_{xz}}$ represent the multi-layer feature matrix extracted from the serial parallel convolution block. Our model is defined as follows:

$$O_{bp} = \sigma(N(\beta(X, Y)) + N(\beta(X, Z)) + N(\beta(Y, Z))) \tag{3}$$

where σ is the softmax function, N is the normalization operation, and β is the bilinear operator in each feature matrix.

$$X = M(Conv_{3\times1}(Conv_{1\times3}(concat(Conv_{1\times1}(F), Conv_{3\times3}(F))))) + Conv_{3\times3}(F) \tag{4}$$

$$Y = Conv_{3\times1}(Conv_{1\times3}(Conv_{1\times1}(X) + Conv_{3\times3}(X))) + X \tag{5}$$

$$Z = Conv_{3\times1}(Conv_{1\times3}(Conv_{1\times1}(Y) + Conv_{3\times3}(Y))) + Y \tag{6}$$

where $Conv$ includes convolution, batch normalization and activation layers. F represents the feature maps in the former layer. $Conv_{3\times3}$, $Conv_{1\times1}$, $Conv_{1\times3}$, $Conv_{3\times1}$ represents the convolutional kernel with the size of 3×3, 1×3, 3×1, 1×1, respectively. In order to reduce the size of feature maps, max pooling is used for down-sampling

and preserving the regions with the largest response value. To extract the specific features, different convolutional blocks do not share parameters. The output feature matrix is defined as

$$O = X^T \otimes Y + X^T \otimes Z + Y^T \otimes Z \tag{7}$$

where $\otimes$ represents the bilinear operator between the different layers in a mutual way. The proposed MFF network structure is shown in Fig. 2.

Inspired by HBP [18], which considers the interaction of different layers in the same hierarchy by projecting the feature maps to higher dimension and combining them with the hadamard product on each channel, we use the matrix outer product and the normalization to improve the interaction of feature maps in different layers, and project the feature information from low dimension to high dimension. The last three layers' outputs are fused using matrix product and added to the fully connection layer and softmax function.

As shown in Fig. 2, we obtain the feature matrixes X, Y, Z using the layer 4.1, 4.2 and 4.3, respectively. The dimensions of all feature matrixes are (B, C, H, W), where B is the batch size, C is the number of channels or feature maps, and $H \times W$ is the size of each feature map. For the features' interaction, we resize the feature matrix X, Y into $(B, C, H \times W)$ and $(B, H \times W, C)$, respectively. The essence of multi-layer feature fusion is matrix product. After normalization, the dimension of $X^T \otimes Y$ is resized from (B, C, C) to $(B, C \times C)$. Performing the above operations each other for X, Y, Z, we can obtain the input of fully connection layer by adding the three results together. In this way, the features of different layers can be effectively fused. Since the interactions of different layers are beneficial for extracting discriminative features, bilinear matrix product is initially applied into different layers to enhance the features expression in this paper.

Compared with other feature fusion methods, like HBP [18], our MFF doesn't need other convolution layer parameters to project the feature map to higher feature dimension. Besides, our method considers different layers interaction in a each other way rather than only uses two layers feature information, like BCNN [13].

4 Experiments

In this section, we conduct the experiments on three standard international fine-grained datasets to demonstrate the effectiveness and the accuracy of the proposed model. Moreover, quantitative and visualized experiments were conducted to evaluate our results against several state-of-the-arts ones.

4.1 Datasets

Experiments were conducted on three internationally used datasets, i.e., CUB-200-2011 [37], FGVC [38] and Stanford Cars [39]. To ensure the practicality of the algorithm, only the images and their corresponding labels were used while any bounding box or part annotations were discarded.

CUB-200-2011 [37] Published by California institute of technology, including over 10,000 images of birds in 200 categories. Among them, 5,994 were used for training and 5,794 were used for testing.

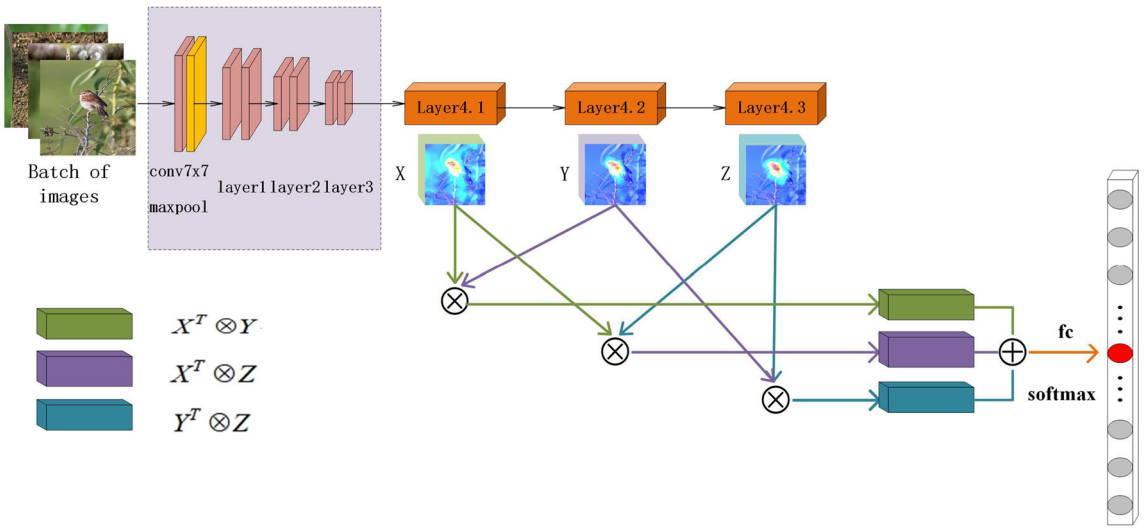


Fig. 2 Proposed multilayer feature fusion (MFF) model. Among them, layer 1, layer 2 and layer 3 are the first three layers in the basic ResNet34 model. There are three 3×3 convolution layers in layer 1, four 3×3 convolution layers in layer 2 and six 3×3 convolution layers in layer 3

FGVC Aircraft [38] Contains 100 aircraft types and a total of 10,000 images. Among them, the training and testing set were divided by a ratio of 2:1.

Stanford Cars [39] Published by Stanford University, including 196 categories and a total of 16,185 images of cars. Among them, 8144 images were used for training and 8041 were used for testing. Each category was classified according to the year, manufacturer, or model.

Some sample images in CUB-200-2011 dataset are shown in Fig. 3, where the images in the same row belong to the same category. The images from top to bottom are California Gull, Glaucous winged Gull and Herring Gull, respectively. We can see that the difference between each category is quite small while the difference within the same category is relatively large due to the influence of light or posture.

4.2 Implemention details

Our experiments use pytorch to implement algorithms. Due to the limited number of fine-grained image samples in each category, performing training directly tends to produce overfitting. To solve this problem, the weight transfer learning was adopted and the training dataset was enhanced using the methods of randomly cropping and flipping horizontally. Moreover, the weighted parameters trained on the ImageNet dataset were used as the initialization values to achieve the rapid convergence.

In all experiments, we resize the images to the single standard 448 × 448 and train the model in two stages. The first stage is to fine-tune the fully-connection layer and the additional parameters that do not belong to ResNet34, while the second stage is to train all the parameters in the network. We set momentum to 0.9, weight decay to 1e-5, and use the stochastic gradient descent method as the network optimizer. Our model is saved as the .pth file and tested on the corresponding dataset to obtain the final results. The batch size is set to 8 for the dataset CUB-200-2011 Birds and 16 for the dataset FGVC Aircraft and Stanford Cars.

4.3 Quantitative evaluation results

To demonstrate the effectiveness of the proposed MFF and PCB, we perform the relative comparison experiments with the original network structures. The quantitative results are shown in Table 1. where CB represents the original convolution block and PCB represents our proposed parallel convolution block.

Since most previous works [13, 18] were completed on VGG [2] network structure while ours was performed on ResNet34, this paper conducted BCNN-RNet classification test on relative datasets and adopted the results of HBP-RNet in [29] for comparison. From Table 1, we can see that after replacing the original BCNN and HBP with our proposed MFF, the results are obviously improved on all datasets. This means that our proposed MFF is superior to the original BCNN and HBP. The results demonstrate that it is helpful to improve the feature expression ability using the matrix outer product to enhance the interaction

Fig. 3 Examples of CUB-200-2011 Dataset

Table 1 Classification accuracy (%) comparison of different models on three fine-grained datasets

Dataset	CUB-200-2011	FGVC-Aircraft	Stanford-Cars
BCNN+CB	84.2	89.6	91.5
HBP+CB	85.8	90.2	92.2
MFF+CB	85.4	90.6	92.4
MFF+PCB(ours)	87.1	91.4	93.4

among different layers. Similarly, after further replacing the original CB with the proposed PCB, our network structure MFF+PCB achieves the highest accuracy on all datasets, meaning that the PCB with parallel structure and effective shortcut is able to extract more salient features. The results demonstrate the effectiveness of the proposed network. The visualization results are shown in the visualization results of Section 4.3.

To further demonstrate the accuracy of the proposed algorithm quantitatively, we compare our algorithm to several state-of-the-art ones, which are divided into three parts for more clear comparison. The approaches in the first part were performed by adjusting network structure or using pooling, including LRBP [17], CBP [16], Boost-CNN [19], Improved B-CNN [20], HIHCA [22] and KP [23]. The approaches in the second part were performed by adding attention mechanism and mask, including RA-CNN [24], MA-CNN [25], DT-RAM [21], OPAM [26], A3M [27], WARN [42], HBPASK [29], SCAB [43] and Chen [44]. While the approaches in the third part aims at locating discriminative parts, including WSDL [28], RP-CNN [33] and CNL [45].

The comparison results are shown in Table 2. Among them, the results of [29] were performed using the codes or pre-trained models provided by their authors, while other results came from the relative references. The "-" in Table 2 represents that the relative method has not been performed on the relative dataset. We can see in the first part of Table 2 that our algorithm is superior to all the bilinear pooling-based ones without any additional annotations. It shows that our method has better feature expression power than others. The methods in the second part of Table 2 are based on attention and mask network to extract salient information. We can see that our algorithm outperforms most methods on three fine-grained datasets. It illustrates that our algorithm can better concentrate on the salient parts and ignores the unrelated parts. Although the methods in the last part of Table 2 can locate discriminative regions accurately, our

Table 2 Classification accuracy (%) comparison of different approaches on three fine-grained datasets. Bold represents the best results, and underline represents the second best results

Methods	Year	CUB-200-2011	Stanford-Cars	FGVC-Aircraft
LRBP [17]	2017	84.2	90.9	87.3
CBP [16]	2016	84.0	–	–
Boost-CNN [19]	2016	85.6	92.1	88.5
Improved B-CNN [20]	2017	85.8	92.0	88.5
HIHCA [22]	2017	85.3	91.7	88.3
KP [23]	2017	86.2	92.4	86.9
RA-CNN [24]	2017	85.3	92.5	88.2
MA-CNN [25]	2017	86.5	92.8	89.9
DT-RAM [21]	2017	86.0	93.1	–
OPAM [26]	2018	85.8	92.2	–
A3M [27]	2018	86.2	–	–
WARN [42]	2020	85.6	90.0	–
HBPASK [29]	2019	86.8	93.8	91.3
SCAB [43]	2019	84.7	91.7	88.3
Chen [44]	2020	85.1	–	84.2
WSDL [28]	2019	85.7	92.3	–
RP-CNN [33]	2019	84.5	93.0	89.9
CNL [45]	2020	86.7	93.1	–
Ours	–	87.1	93.4	91.4

method surpasses all of them on three datasets. Compared with the latest method [45], our method also surpasses 0.4% and 0.3% on the datasets CUB-200-2011 and Stanford-Cars, respectively. Besides that, our algorithm exceeds 2.4%, 1.7%, 3.1% on CUB-200-2011, Stanford-Cars and FGVC-Aircraft, respectively, against SCAB [43]. On the whole, our method outperforms most state-of-the-art ones on three fine-grained datasets. It illustrates that our model has better feature representation ability than others.

4.4 Visualization results

As illustrated in Section 3.1, our convolution block tends to have better performance in extracting discriminative regions. This performance is further demonstrated in Fig. 4 using the hot images obtained by different convolution blocks on different fine-grained datasets. where Fig. 4a, b, and c represents the results on CUB-200-2011 Bird, FGVC Aircraft, and Stanford Cars dataset, respectively. The images on the top row show the results of CB, while those on the bottom row are obtained by the proposed PCB. From left to right are the original image and the hot images on different channels, respectively. From Fig. 4, we can see that the information extracted by CB convolution block usually contains background and other interference information. However, the information extracted by our proposed PCB convolution block mainly focused on the discriminative features of the image, which have strong expression ability and are crucial to solve the problem of small inter differences in fine-grained image classification.

The visual results of confusion matrix on Stanford Cars are shown in Fig. 5, and the images from left to right correspond to the training and testing confusion matrixes, respectively.

Some misclassified examples are shown in Fig. 6. where Fig. 6a and b represents the results on the dataset CUB-200-2011 and FGVC Aircraft, respectively. The sub-image on the left is the test image with corresponding true label, while the sub-images on the right are the samples of its wrongly predicted labels. From Fig. 6, we can see that the test images are quite similar with the wrongly predicted ones which are difficult to be distinguished even for human beings.

It is essential to extract salient features accurately in fine-grained image classification. Figure 7a-d are the

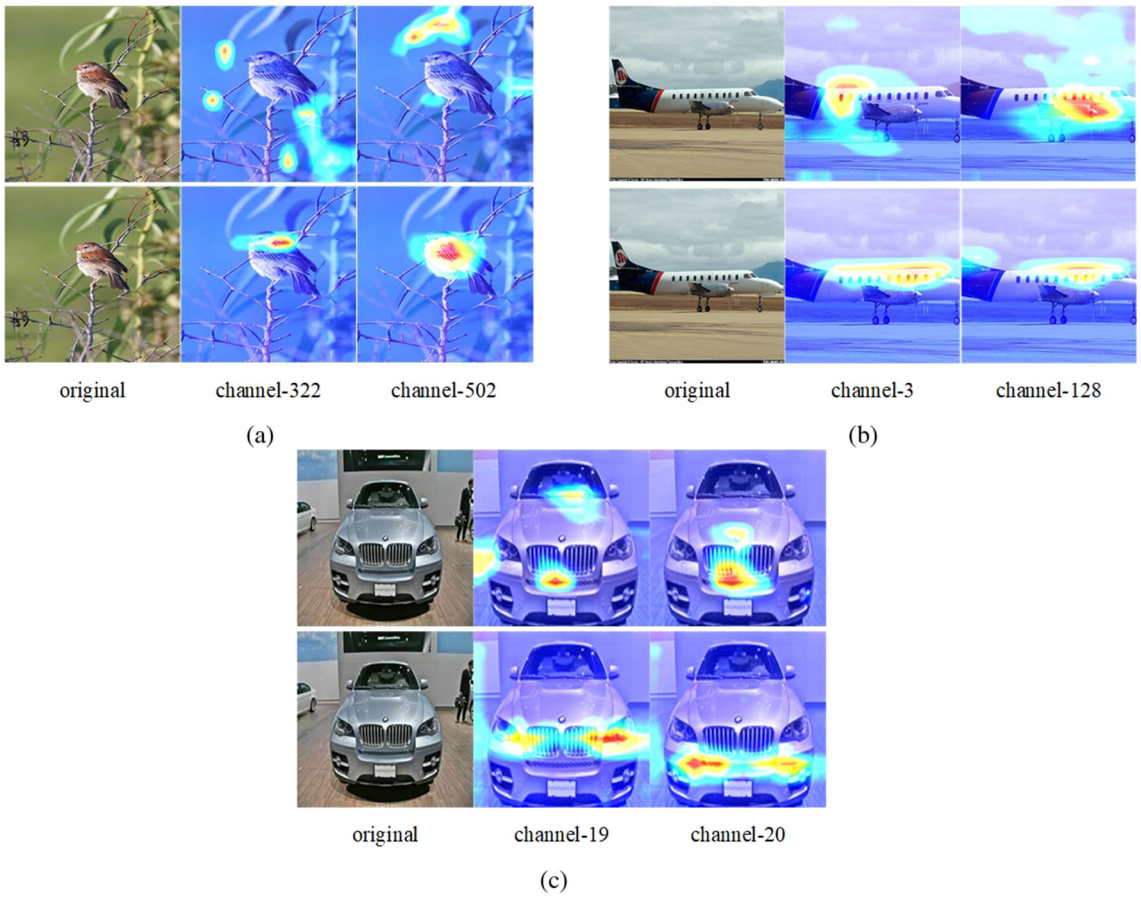


Fig. 4 Hot images on different channels using different convolution blocks. (**a**)-(**c**) represent the sample images in different datasets. From top to bottom are the results of standard CB and the proposed PCB. From left to right are the original images and the hot images on different channels

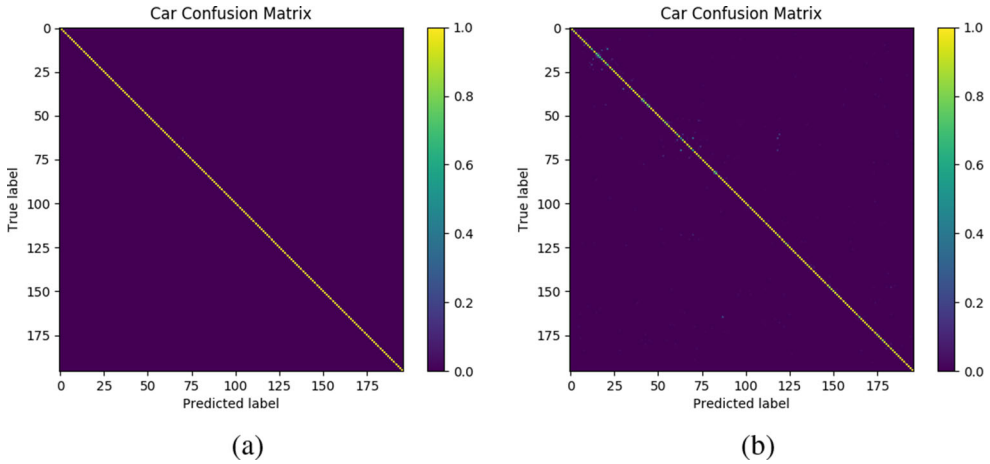


Fig. 5 Confusion matrixes on dataset "Stanford Cars". The images from left to right are the training and testing confusion matrixes, respectively

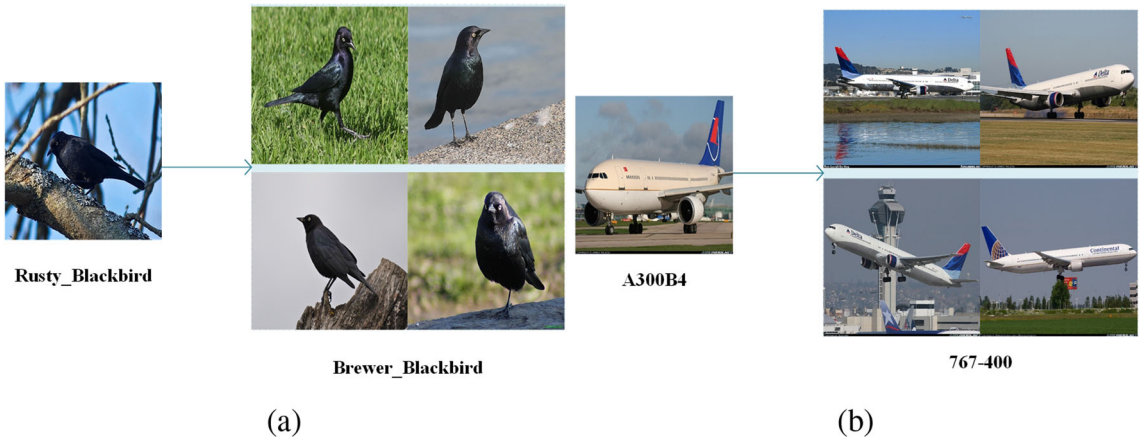


Fig. 6 Some misclassified examples on different datasets. The left sub-image is the test image with its corresponding true label, while the right sub-images are the samples with its wrongly predicted samples

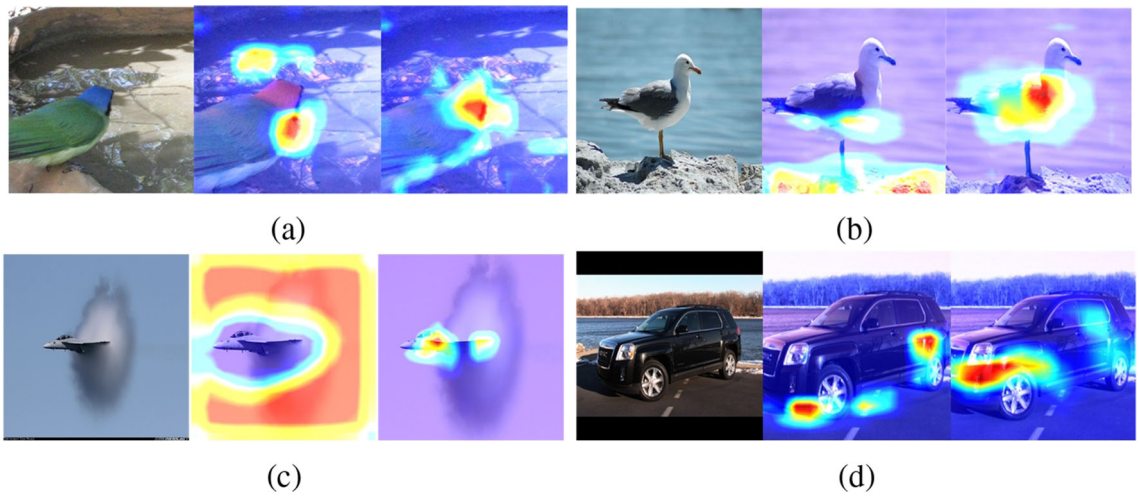


Fig. 7 Comparison of salient features images using CB and PCB. (**a**)-(**d**) represent the sample images in different datasets. From left to right are the original images, and the hot images on a certain channel using CB and PCB for salient feature extraction

salient features extraction results of different datasets. The leftmost column is the original image, and the last two columns are the hot map obtained by CB and PCB on same channel. From Fig. 7 we can see that compared to the CB, our PCB module extracts features mostly in the significant regions of image, while ignores the background information. Therefore, the improved convolution block PCB is more accurate in extracting salient image feature, which is helpful to improve the accuracy of fine-grained image classification.

5 Conclusions

In this paper, we propose a multilayer feature fusion with parallel convolutional block approach for fine-grained image classification. A PCB mechanism is proposed to extract the discriminative features with two different convolutional kernel sizes in a two stream way. Besides, the proposed MFF uses bilinear matrix multiplication to enhance their interaction ability. In the training process, the two-step training method is utilized to obtain better weight parameters. Experimental results demonstrate that our method can achieve the accuracy of 87.1%, 91.4% and 93.4% on the dataset CUB-200-2011, FGVC-Aircraft and Stanford-Cars, respectively. Qualitative and quantitative experimental results show that our method has higher precision against the state-of-the-arts ones.

In the future, we attempt to use weakly supervised methods to localize salient regions, like [32, 33, 40], or utilize the attention network, like [24, 25], to extract discriminative information.

References

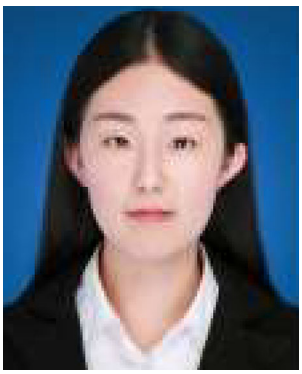
1. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp 1097–1105
2. Simonyan K (2014) A Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
3. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning, pp 448–456
4. He KM, Zhang XY, Ren SQ, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
5. Howard A-G, Zhu M, Kalenichenko B-D, Wang W, Weyand T, Andreetto M, AdamChen H (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861
6. Huang G, Liu Z, Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2261–2269
7. Zhang N, Donahue J, Girshick R, Darrell T (2014) Part-based r-CNNs for fine-grained category detection. In: Proceedings of the European Conference on Computer Vision, pp 834–849
8. Branson S, Van Horn G, Belongie S, Perona P (2014) Bird species categorization using pose normalized deep convolutional nets. In: Proceedings of the BMVC 2014—British Machine Vision Conference
9. Berg T, Belhumeur P-N (2013) POOF: Part-based One-vs-one features for fine-grained categorization, face verification, and attribute estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 955–962
10. Xie L, Tian Q, Hong R, Yan S, Zhang B (2013) Hierarchical part matching for fine-grained visual categorization. In: Proceedings of the IEEE Conference on International Conference on Computer Vision, pp 1641–1648
11. Xiao T, Xu Y, Yang K, Zhang J, Peng Y, Zhang Z (2015) The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , pp 842–850
12. Zhang X, Xiong H, Zhou W, Lin W, Tian Q (2016) Picking deep filter responses for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1134–1142
13. Lin T-Y, RoyChowdhury A, Maji S (2015) Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1449–1457
14. Jaderberg M, Simonyan K, Zisserman A (2015) Spatial transformer networks. Advances in Neural information Processing Systems, pp 2017–2025
15. Ji Z., Zhao K., Zhang S., Li M (2019) Classification of fine-grained fish images based on spatial transformation bilinear networks. Journal of TianJin University 52:475–482
16. Gao Y, Beijbom O, Zhang N, Darrell T (2016) Compact bilinear pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 317–326
17. Kong S, Fowlkes C (2017) Low-rank bilinear pooling for fine-grained classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 365–374
18. Yu C, Zhao X, Zheng Q, Zhang P, You X (2018) Hierarchical bilinear pooling for fine-grained visual recognition. In: Proceedings of the IEEE Conference on European Conference, pp 595–610
19. Moghimi M, Belongie SJ, Saberian MJ, Yang J, Vasconcelos N, Li L-J (2016) Boosted convolutional neural networks. In: Proceedings of the British Machine Vision Conference
20. Lin TY, Maji S (2017) Improved bilinear pooling with CNNs. In: Proceedings of British Machine Vision Conference, pp 395.1–395.12
21. Li Z, Yang Y, Liu X, Zhou F, Wen S, Xu W (2017) Dynamic computational time for visual attention. In: Proceedings of International Conference on Computer Vision Workshops, pp 1199–1209
22. Cai S, Zuo W, Zhang L (2017) Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In: Proceedings of the IEEE International Conference on Computer Vision, pp 511–520
23. Cui Y, Zhou F, Wang J, Liu X, Lin Y, Belongie S (2017) Kernel pooling for convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2921–2930
24. Fu J, Zheng H, Mei T (2017) Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 4438–4446

25. Zheng H, Fu J, Mei T, Luo J (2017) Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp 5209–5217
26. Peng Y, He X, Zhao J (2018) Object-part attention model for fine-grained image classification. IEEE Trans Image Process 27:1487–1500
27. Han K, Guo J, Zhang C, Zhu M (2018) Attribute-aware attention model for fine-grained representation learning. In: Proceedings of the Multimedia Conference on Multimedia Conference, pp 2040–2048
28. He X, Peng Y, Zhao J (2019) Fast Fine-Grained image classification via weakly supervised discriminative localization. IEEE Trans Circuits Syst Video Technol 29:1394–1407
29. Tan M, Wang G, Zhou J, Peng Z, Zheng M (2019) Fine-grained classification via hierarchical bilinear pooling with aggregated slack mask. IEEE Access 7:117944–117953
30. Wang YM, Morariu VI, Davis LS (2018) Learning a discriminative filter bank within a CNN for fine-grained recognition. In: Proceedings of the IEEE Computer Vision and Pattern Recognition, pp 5209–5217
31. Yang Z, Luo T, Wang D, Hu Z, Gao J, Wang L (2018) Learning to navigate for fine-grained classification. In: Proceedings of the European Conference, pp 420–435
32. Chen Y, Bai Y, Zhang W, Mei T (2019) Destruction and construction learning for fine-grained image recognition. In: Proceedings of Computer Vision and Pattern Recognition, pp 5157–5166
33. Xin Q, Lv T, Gao H (2019) Random part localization model for fine grained image classification. In: Proceedings of International Conference on Image Processing, pp 420–424
34. Hu T, Xu J, Huang C, Qi H, Huang Q, Lu Y (2018) Weakly Supervised Bilinear Attention Network for Fine-Grained Visual Classification. arXiv:1808.02152
35. Min S, Yao H, Xie H, Zha ZJ, Zhang Y (2020) Multi-objective matrix normalization for fine-grained visual recognition. IEEE Trans Image Process 29:4996–5009
36. Zheng H, Fu J, Zha Z. J., Luo J (2019) Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In: Proceedings of Computer Vision and Pattern Recognition, pp 5012–5021
37. Wah C, Branson S, Welinder P, Perona P, Belongie S (2011) The Caltech-UCSD birds-200-2011 dataset, Comput Neural Syst California Inst Technol
38. Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A (2013) Fine-grained visual classification of aircraft. arXiv:1306,5151
39. Krause J, Stark M, Deng J, Fei-Fei L (2013) 3D object representations for fine-grained categorization. In: Proc IEEE Int Conf Comput Vis Workshops, pp 554–561
40. Zheng H, Fu J, Zha ZJ, Luo J, Mei T (2020) Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. IEEE Trans Image Process 29:476–488
41. Wei XS, Luo JH, Wu J, Zhou ZH (2017) Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Trans Image Process 26:2868–2881
42. Rodríguez P, Velazquez D, Cucurull G, Gonfaus JM, Roca FX, Gonzàlez J (2020) Pay attention to the activations: a modular attention mechanism for Fine-Grained image recognition. IEEE Trans Image Process 22:502–514
43. Wang W, Zhang J, Wang F (2019) Attention bilinear pooling for fine-grained classification. Symmetry 11:1033
44. Chen F, Huang G, Lan J, Wu Y, Pun C, Ling WK, Cheng L (2020) Weakly supervised Fine-Grained image classification via salient region localization and different layer feature fusion. Appl Sci 10:4652
45. Ye Z, Hu F, Liu Y, Xia Z, Lyu F (2020) Pengqing Liu:Associating Multi-Scale Receptive Fields For Fine-Grained Recognition. ICIP: 1851–1855

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Lei Wang received the B.E. degree from the College of Electrical Engineering, Sichuan University, China, in 2020. She is currently pursuing the degree with the school of Electrical and Information Engineering, Tianjin University. Her research interests include few shot learning, image classification, and machine learning (deep learning).

Kai He received the Ph.D. degree in communication and information system in 2003 from Jilin University, Changchun, China. He was a postdoctoral researcher in Peking University, Peking, China from 2004 to 2006. He is currently an associate professor with the School of Electrical and Information Engineering, Tianjin University, China. He was a visiting scholar from 2014 to 2015 in the School of Computing Science, University of Glasgow, Glasgow, UK. His current research interests are in the areas of digital image/video processing and computer vision. He has published over 70 scientific papers in the international journals and conferences on image processing and computer vision.



Xu Feng received the B.E. degree from the College of Electronic Information Engineering and Automation, Tianjin Technology University, China, in 2018. She is currently pursuing the degree with the school of Electrical and Information Engineering, Tianjin University. Her research interests include computer vision, image classification, and machine learning (deep learning).

Xitao Ma received the B.E. degree from the College of Electronical and Information Engineering, Hebei University of Technology, China, in 2019. He is currently pursuing the degree with the school of Electrical and Information Engineering, Tianjin University. His research interests include computer vision, scene text recognition, and machine learning (deep learning).