RESEARCH ARTICLE

Fine-grained classification based on multi-scale pyramid convolution networks

Gaihua Wang^{1,2}, Lei Cheng^{1*}, Jinheng Lin¹, Yingying Dai¹, Tianlun Zhang¹

1 School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan, China, **2** Hubei University of Technology Cooperative Innovation Center of Hubei Province for Efficient Use of Solar Energy, Wuhan, China

* 1774437797@qq.com

OPEN ACCESS

Citation: Wang G, Cheng L, Lin J, Dai Y, Zhang T (2021) Fine-grained classification based on multi-scale pyramid convolution networks. PLoS ONE 16(7): e0254054. <https://doi.org/10.1371/journal.pone.0254054>

Editor: Thippa Reddy Gadekallu, Vellore Institute of Technology: VIT University, INDIA

Received: April 30, 2021

Accepted: June 21, 2021

Published: July 9, 2021

Copyright: © 2021 Wang et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: We uploaded the data to Figshare, which made up our minimal dataset. The minimal dataset can be obtained by link: (<https://figshare.com/s/3647e47f6b6ea4492fb6>).

Funding: This work was supported by The National Key R&D Program of China under Grant 2017YFB1302400. The author, Gaihua Wang, has received financial support from this fund, which has provided financial help for our design. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of

Abstract

The large intra-class variance and small inter-class variance are the key factor affecting fine-grained image classification. Recently, some algorithms have been more accurate and efficient. However, these methods ignore the multi-scale information of the network, resulting in insufficient ability to capture subtle changes. To solve this problem, a weakly supervised fine-grained classification network based on multi-scale pyramid is proposed in this paper. It uses pyramid convolution kernel to replace ordinary convolution kernel in residual network, which can expand the receptive field of the convolution kernel and use complementary information of different scales. Meanwhile, the weakly supervised data augmentation network (WS-DAN) is used to prevent over fitting and improve the performance of the model. In addition, a new attention module, which includes spatial attention and channel attention, is introduced to pay more attention to the object part in the image. The comprehensive experiments are carried out on three public benchmarks. It shows that the proposed method can extract subtle feature and achieve classification effectively.

1 Introduction

Fine grained image classification is a more and more concerned subject in the field of computer vision. It has been widely studied in new retail, automatic driving, ecological protection. Different from traditional image classification, fine-grained image classification aims to divide the same species into different subclasses, such as Shiba Inu and Akita Inu. Because most of the subtle differences between classes can only be distinguished effectively by region location and discriminative feature learning, fine-grained image recognition is regarded as a more challenging task.

Early works on localization-based methods usually use strong supervision to annotate the part information of the image, and then extract the features of the parts for fine-grained classification [1–3]. However, the methods of strong supervision rely heavily on manual object annotation, which is too expensive to be widely used in practice. Weak supervision using attentional mechanisms [4–8] have been more popular in subsequent researches. In [5], a recurrent attention convolutional neural network that locates the region of interest is proposed. By using the mutual promotion and enhancement of region detection and feature extraction, it gradually

the manuscript. There is no conflict of interest with the funders. The author would like to thank the national key R & D program of China under Grant 2017yfb1302400 for its financial support.

Competing interests: The authors have declared that no competing interests exist.

locates and identifies the region to complete the feature extraction from coarse-grained to fine-grained. In [6], it proposes to use pairwise interaction to distinguish differences and find the key areas of each image by comparing a pair of fine-grained images.

These methods all adopt the idea of location recognition, and strengthen the attention and discrimination to the fine features through the weak supervision method. However, they ignore the multi-scale features of the network. For fine-grained classification tasks, multi-scale features are crucial, because target parts have multifarious sizes and shapes in images [9–11], and the extracted feature of single convolution is insufficient.

This paper proposes a fine-grained classification network based on multi-scale pyramid convolution kernel, which uses pyramid convolution kernel with multiple sizes to extract multi-scale features. And weakly supervised data augmentation network (WS-DAN) is used to enhance the network data. To reduce the interference of image background, a new attention mechanism is also added to extract subtle features. The main contributions of this paper are as follows: (1) The pyramid convolution kernel is introduced to extract multi-scale features without increasing the computational cost, and improve the ability to capture fine features by WS-DAN. (2) A new lightweight multi-attention module, including spatial attention and channel attention, is designed to retain important information and suppress the interference caused by background noise. The attention module can be seamlessly integrated into any convolutional neural networks (CNN) architecture effectively (3) the proposed method achieves state-of-the-art performance on three datasets, including the CUB-200-2011 [12] dataset, Stanford Dogs [13] dataset, and Stanford Cars [14] dataset.

The rest of this paper is organized as follows: in Section 2, we introduce the Weakly Supervised Data Augmentation Network. Section 3 introduces the multi-scale pyramid convolution kernel and multi-attention module. Section 4 gives our experiment result, including the introduction about CUB-200-2011, Stanford Dogs, and Stanford Cars dataset, and the experiment settings. Finally, we conclude in Section 5.

2 Related work

2.1 Weakly supervised data augmentation network

Deep learning is an emerging technology in the field of machine learning, which has attracted the attention of many researchers [15–17]. Data enhancement is a common strategy in deep learning task. It can increase the number of training data by introducing more data variance in the way of random cropping. However, due to random sampling of clipping area, a large part of the area contains a lot of background noise, which may affect the quality of feature extraction and offset its advantages. Therefore, we use WS-DAN instead of traditional data enhancement.

WS-DAN uses attention learning to generate attention maps to represent the spatial distribution of the discriminating object parts, and then enhance the data of object. After generating the attention graph, the design is divided into two parts: the first part includes Attention cropping and Attention dropping. Attention cropping is the acquisition of subtle feature to enhance the representation of local features. Attention dropping removes subtle feature from the image randomly to learn new detail features. In the second part, the attention map is used to locate the whole target accurately, which enlarges the target to get more attention, and suppresses the interference of irrelevant noise.

2.2 Visual attention

Visual attention is widely used in various deep learning tasks [18–20]. It can be employed to discover the subtle inter-class differences in fine-grained image categorization. For instance, in

[21], it proposes a cross layer non local module based on visual attention. By establishing a query layer and a response layer, the deep and shallow characteristics of the network are correlated to improve the representation ability of the network. The paper [8] uses the method of soft attention, which imposes a soft mask on the feature maps to generate the attention maps and guide the enhanced area of the image. The paper [22] uses a self-attention method to discover the complementary information related to channels through the interaction between channels.

3 The proposed method

The overall architecture of the network is shown in Fig 1. The backbone network is used to extract feature information, the multi-attention module is used to integrate spatial and channel information, and the weakly supervised data augmentation network (WS-DAN) is used to enhance image data and improve model performance.

The process of data passing through the network is as follows: first, the input image is sent into the backbone network to obtain feature maps. Then, attention maps are obtained through the convolution operation by feature maps, at the same time, feature maps are input into the multi-attention module. Output of attention maps are divided into two branches. One branch goes through the weakly supervised data augmentation network (WS-DAN) and feeds the results back to the input data; the other branch is fused with the output of the multi-attention module. Finally, the fused results are classified.

3.1 Multi-scale pyramid convolution

According to Convolutional Neural Networks (CNN) theory, the convolution operator can be transformed as follows: $T: X \rightarrow Y, X \in R^{h \times w \times C}, Y \in R^{h' \times w' \times C}$, where $h \times w$ represents the spatial dimension and C represents the number of channels. Compared with ordinary convolution kernel, pyramid convolution kernel contains multiple scale convolution kernel. It can extract multi-scale features by using the convolution kernel of different sizes. Fig 2 is the structure of multi-scale Pyramid convolution (Pyconv).

In Fig 2, it contains several convolution kernels of different sizes ($\text{kernel}_1, \text{kernel}_2, \dots, \text{kernel}_n$). The number of their output channels is $C_1, C_1, \dots, C_n$, and the number of final channel is

Fig 1. The overall architecture of the network.

<https://doi.org/10.1371/journal.pone.0254054.g001>

Fig 2. Multi-scale pyramid convolution kernel.

<https://doi.org/10.1371/journal.pone.0254054.g002>

C_{out} :

$$C_{out} = C_1 + C_2 + \dots + C_n \quad (1)$$

Pyconv combines these features in channel dimensions to complete feature fusion, which helps the network obtain richer semantic information and locate the key areas of the image accurately.

In order to use different depth kernel in each level of Pyconv, the input feature maps are divided into different groups. As shown in Fig 3, they are three different groups: the numbers of groups are one group, two groups, and four groups. And the kernel is applied independently for each group, which is called block convolution. When the number of groups increases, the number of parameters and the calculation cost of convolution decrease.

The block convolution can reduce the parameters of the network. The total number of parameters of a convolutional layer is:

$$T(conv) = C \times H \times W \times K \quad (2)$$

Where C represents the number of channels, $H \times W$ represents the size of convolution kernel, and K represents the number of cores. In Table 1, the number of parameters in the first stage

Fig 3. Grouped convolution.

<https://doi.org/10.1371/journal.pone.0254054.g003>

Table 1. PyconvResNet50 convolution kernel size information.

Stage	Output	ResNet50(basic)	PyconvResNet50
	112×112	$7 \times 7, t = 64, s = 2$ $3 \times 3, \text{maxpool}, s = 2$	$7 \times 7, t = 64, s = 2$
1	56×56	$\begin{bmatrix} 1 \times 1, t = 64 \\ 3 \times 3, t = 64 \\ 1 \times 1, t = 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, t = 64 \\ \text{Pyconv4}, t = 64 \\ \begin{bmatrix} 9 \times 9, t = 16, G = 16 \\ 7 \times 7, t = 16, G = 8 \\ 5 \times 5, t = 16, G = 4 \\ 3 \times 3, t = 16, G = 1 \end{bmatrix} \\ 1 \times 1, t = 256 \end{bmatrix} \times 3$
2	28×28	$\begin{bmatrix} 1 \times 1, t = 128 \\ 3 \times 3, t = 128 \\ 1 \times 1, t = 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, t = 128 \\ \text{Pyconv3}, t = 128 \\ \begin{bmatrix} 7 \times 7, t = 64, G = 8 \\ 5 \times 5, t = 32, G = 4 \\ 3 \times 3, t = 32, G = 1 \end{bmatrix} \\ 1 \times 1, t = 512 \end{bmatrix} \times 4$
3	14×14	$\begin{bmatrix} 1 \times 1, t = 256 \\ 3 \times 3, t = 256 \\ 1 \times 1, t = 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, t = 256 \\ \text{Pyconv2}, t = 256 \\ \begin{bmatrix} 5 \times 5, t = 128, G = 4 \\ 3 \times 3, t = 128, G = 1 \end{bmatrix} \\ 1 \times 1, t = 1024 \end{bmatrix} \times 6$
4	7×7	$\begin{bmatrix} 1 \times 1, t = 512 \\ 3 \times 3, t = 512 \\ 1 \times 1, t = 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, t = 512 \\ \text{Pyconv1}, t = 512 \\ \begin{bmatrix} 3 \times 3, t = 512, G = 1 \end{bmatrix} \\ 1 \times 1, t = 2048 \end{bmatrix} \times 3$
	1×1	global avgpool fc	global avgpool fc
Params		25.56×10^6	24.85×10^6
FLOPs		4.14×10^9	3.88×10^9

<https://doi.org/10.1371/journal.pone.0254054.t001>

can be obtained through the following formula:

$$T(\text{basic}) = (64 \times 3 \times 3 \times 64) \times 256 \times 3 = 36864 \times 256 \times 3 \quad (3)$$

$$\begin{aligned} T(\text{PyConv4}) &= ((64 \div 16 \times 9 \times 9 \times 16) + (64 \div 8 \times 7 \times 7 \times 16) + (64 \div 4 \times 5 \times 5 \times 16) \\ &\quad + (64 \div 1 \times 3 \times 3 \times 16)) \times 256 \times 3 \\ &= 27072 \times 256 \times 3 \end{aligned} \quad (4)$$

Compared with ordinary convolution kernel, Pyconv kernel can reduce the number of parameters. we can get PyConvresnet50 by replacing the ordinary convolution kernel in the ResNet50 with multi-scale Pyconv kernel. The information of the convolution kernel of pyconvresnet50 is shown in Table 1, s represents the step size, t represents the number of channels, and G represents the channel group. The steps of PyConvresnet50 are as follows:

- Step 1: The ordinary convolution is replaced by Pyconv4, which contains $9 \times 9, 7 \times 7, 5 \times 5$ and 3×3 kernels. The input channel of each group is $G = 16, G = 8, G = 4$ and $G = 1$, and the output channel of all kernels is 16.
- Step 2: The ordinary convolution is replaced by Pyconv3, which contains $7 \times 7, 5 \times 5$ and 3×3 kernels. The input channel of each group is $G = 8, G = 4$ and $G = 1$, and the output channel is 64, 64 and 32;
- Step 3: The ordinary convolution is replaced by Pyconv2, which contains 5×5 and 3×3 kernels, The input channel of each group is $G = 4$ and $G = 1$, and the output channel of all kernels is 128;
- Step 4: Replace the original ordinary convolution kernel with Pyconv1 that only contains a convolution kernel 3×3 . The number of output channels is 512, and the input channel is $G = 1$.

3.2 Multi-attention module

To better extract the subtle features between different categories, a multi-attention mechanism, which includes channel attention and spatial attention, is designed. The structure of multi-attention module is shown in Fig 4.

The input X is paralleled through the channel and spatial attention to obtain the channel and spatial weight respectively. The network can learn the location information of the key area by multiplying with the channel and spatial weight to remove the interference of irrelevant background. Then the attention result is combined with the input feature X . It can be described as follows:

$$F = F_s * (F_c * X) + X \quad (5)$$

where X is input feature, F is output, F_c and F_s are output of channel attention and spatial attention respectively.

Channel attention module. Channel attention can effectively capture the context information between channel information. Fig 5 shows the specific method of the channel attention module. Firstly, global maximum pooling and global average pooling are used to map input features from space (H, W, C) to $(1, 1, C)$. Then, the results of the two pooling methods are spliced to get the feature map with dimension of $(1, 1, 2C)$. Because the channel number of the

Fig 4. Multi-attention module.

<https://doi.org/10.1371/journal.pone.0254054.g004>

Fig 5. Channel attention module.

<https://doi.org/10.1371/journal.pone.0254054.g005>

original input feature graph is C , we need to go through two convolution kernels with the size of 1×1 to reduce the dimension of the channel number to further extract the channel features. R represents the channel compression ratio, in this experiment, $R = 16$, the action process can be expressed as:

$$F_C = \text{Conv}_{1 \times 1}(\text{ReLU}(\text{BN}(\text{Conv}_{1 \times 1}(\text{concat}[\text{maxpool}(X), \text{avgpool}(X)])))) \quad (6)$$

where F_C is the channel attention output, Conv , ReLU , BN , maxpool and avgpool represents convolution operation, activation function, batch normalization, global maximum pooling and global average pooling.

Spatial attention module. Spatial attention focuses on the location information of the image, and removes the interference of background noise. For example, CBAM [23] adopts pooling channel compression in spatial branch. And Bottleneck Attention Module (BAM) [24] adopts serial convolution and hole convolution channel compression. In order to obtain more abundant spatial information, this paper uses different sizes of parallel convolution when compressing channels, as shown in Fig 6. Convolution kernels of 1×1 and 3×3 are used respectively to extract features to obtain rich feature information, in which convolution kernels of 3×3 are decomposed into 1×3 and 3×1 . The maximum pooling and average pooling are used to aggregate the channel information on the two branches respectively. The channel number is compressed to 1 by convolution, and the information of the two branches is fused. The process of spatial attention module can be described as:

$$F_s = \text{Conv}_{1 \times 1}(\text{ReLU}(\text{concat}[\text{maxpool}(\text{Conv}_{3 \times 1}(\text{Conv}_{1 \times 3}(X))), \text{avgpool}(\text{Conv}_{3 \times 1}(\text{Conv}_{1 \times 3}(X)))])) \times \text{Conv}_{1 \times 1}(\text{ReLU}(\text{concat}[\text{maxpool}(\text{Conv}_{1 \times 1}(X)), \text{avgpool}(\text{Conv}_{1 \times 1}(X))])) \quad (7)$$

where F_s is the output of spatial attention module, Conv , ReLU , BN , maxpool and avgpool

Fig 6. Spatial attention module.

<https://doi.org/10.1371/journal.pone.0254054.g006>

represents convolution operation, activation function, batch normalization, global maximum pooling and global average pooling.

4 Experiment

In this section, all kinds of the experimental settings will be introduced and the classification results of all related methods will be analyzed.

4.1 Datasets and training settings

We conducted experiments on three challenging fine-grained image classification datasets, namely CUB-200-2011 [12], Stanford dogs [13] and Stanford cars [14]. Table 2 summarizes the detailed statistics of datasets.

The hardware configuration of the experiments: Intel Xeon e5-2683 V3 CPU, 32g running memory, single NVIDIA GTX 1080ti graphics card, 11g video memory. In win10 system, Pytorch framework is used as the experimental platform.

The input image is resized to 448×448 . Each dataset is trained with 80 epochs and tested at the end of each epoch. The batch size is 8, the learning rate is 0.001, the momentum is 0.9, and the weight attenuation is set to 0.00001. SGD is used to train the loss function of the model. The pretrained parameters on ImageNet are used as the initial weight.

Classification accuracy is used to measure the performance of networks. It is expressed by:

$$A_{accuracy} = \frac{I_a}{I} \quad (8)$$

Where: $A_{accuracy}$ represents the classification accuracy, I_a and I represent the number of correct classification and the total number of test images.

4.2 Experimental analysis

In order to verify that the proposed method can effectively improve the classification accuracy of fine-grained images, we conducted ablation studies and comparative analysis.

4.2.1 Ablation studies. In order to explore the influence of different scale in multi-attention module, and try to find the best convolution combination, we use convolution combinations of different sizes to construct spatial attention module, including 1×1 and 3×3 , 3×3 and 5×5 , 5×5 and 7×7 . The constructed attention module is tested on CUB-200-2011 data set, and the results are shown in Table 3.

From the Table 3, the convolution group with 1×1 and 3×3 can get the highest classification accuracy. We use the combination in the final multi-attention module.

To explore the effect of multi-attention module and pyramid convolution, WS-DAN is used as a benchmark. We conducted the following experiments on CUB-200-2011: 1) WS-DAN network with resnet50 as the backbone; 2) add the multi-attention module at the last layer of backbone; 3) add pyramid convolution to the backbone; 4) fuse the multi-attention with pyramid convolution. Table 4 shows the results of networks in different configurations. Compared with the baseline, for WS-DAN, using multi-attention module has 0.59%

Table 2. Details of the three datasets.

Datasets	#Category	#Total	#Train	#Test
CUB-200-2011 [12]	200	11788	5994	5794
Stanford dogs [13]	120	20580	12000	8580
Stanford cars [14]	196	16185	8144	8041

<https://doi.org/10.1371/journal.pone.0254054.t002>

Table 3. Top1 accuracy (%) of convolution combinations with different sizes on CUB-200-2011.

Convolution combination	Backbone	Top1(%)
1×1 and 3×3	ResNet50	84.91
3×3 and 5×5	ResNet50	84.68
5×5 and 7×7	ResNet50	84.35

<https://doi.org/10.1371/journal.pone.0254054.t003>

improvement. Using pyramid convolution can be improved by 1.31%. Fused module provides 1.60% improvement. In addition, using the channel group, our model has less computation cost than WS-DAN, which saves 0.08×10^7 parameters and 0.27×10^9 calculated numbers.

4.2.2 Comparative analysis. 1) *Comparison with WS-DAN model.* We improve the WS-DAN model based on the weakly supervised fine-grained algorithm, replacing the ordinary convolution kernel in resnet50 network with pyramid convolution kernel, and introducing the attention mechanism into the last layer of the network. The last layer of the original model and the proposed model is visualized respectively, in which Fig 7 is the back propagation saliency map, and Fig 8 is the visualization of attention.

Fig 7(A) is the original image. Fig 7(B) and 7(C) are the backpropagation saliency map of WS-DAN and the proposed method respectively. Compared Fig 7(B) and 7(C), WS-DAN has more background noise in feature extraction. The proposed method uses pyramid convolution kernel and multi-attention module to extract features and suppress the background noise effectively.

The visualization of attention is shown in Fig 8. Fig 8(A) is the original image, and Fig 8(B) and 8(C) are the visualization of the attention of WS-DAN and the proposed method. Different colors represent the different attention. And the deeper the red is, the higher the attention is. Comparing Fig 8(B) and 8(C), our method can locate the regions with distinguishing features more accurately, such as higher attention to the head of birds in the second column of Fig 8, which makes more computing resources incline to these key areas in the training and testing.

Fig 9 is the curve of accuracy and loss. From Fig 9, we can see that the train loss of the proposed method on three datasets can be reduced steadily, which proves the rationality and universality of the proposed method. Because of the pretrained model, the test accuracy curve is rapidly improved in the first 10 epochs. Meanwhile, the test accuracy curve does not decline, which proves that there is no fitting phenomenon in this process.

2) *Comparison of different fine-grained algorithms.* We select resnet50 [25], BCNN [4], RA-CNN [5], MA-CNN [26], WS-DAN [8], and PMG [27] to compare with the proposed method, and test them on three public fine-grained data sets. The experimental results are shown in Tables 5 and 6. “1-Stage” indicates whether the training can be done in one stage. For one stage, the whole process of training and prediction is completed in the model, which does not need to be divided into multiple stages.

It can be seen from Tables 5 and 6 that the accuracy of the proposed method on CUB-200-2011, Stanford Dogs and Stanford Cars data sets reaches 85.92%, 85.82% and 93.64%. Compared with WS-DAN, the accuracy of these three datasets is improved by 1.60%, 1.98% and 1.13%.

Table 4. Experimental results of ablation. Top1 accuracy (%) on CUB-200-2011.

Model	Backbone	Top1(%)	param	FLOPs
WS-DAN (baseline)	ResNet50	84.32	3.67×10^7	4.13×10^9
WS-DAN (att)	ResNet50+attention	84.91	3.79×10^7	4.15×10^9
WS-DAN(Py)	ResNet50+Pyconv	85.63	3.48×10^7	3.85×10^9
Ours (att + Py)	ResNet50+ attention +Pyconv	85.92	3.59×10^7	3.86×10^9

<https://doi.org/10.1371/journal.pone.0254054.t004>

Fig 7. Back propagation saliency map.

<https://doi.org/10.1371/journal.pone.0254054.g007>

Compared with other methods, the proposed method also achieves the highest classification accuracy which further proves the effectiveness and universality of the proposed method.

5 Conclusion

In this paper, we design a new fine-grained classification network based on multi-scale pyramid convolution kernel. It adds pyramid convolution kernel into the main network. Through

Fig 8. Visualization of attention.

<https://doi.org/10.1371/journal.pone.0254054.g008>

Fig 9. The test accuracy curve and train loss curve on CUB-200-2011, Stanford Cars and Stanford Dogs datasets.

<https://doi.org/10.1371/journal.pone.0254054.g009>

Table 5. Comparison results on CUB-200-2011.

Method	Backbone	1-Stage	Top1(%)
ResNet50 [25]	resnet50	✓	73.01
BCNN [4]	vgg16	×	80.31
RA-CNN [5]	vgg19	×	83.27
MA-CNN [26]	vgg19	✓	82.54
WS-DAN [8]	resnet50	✓	84.32
PMG [27]	resnet50	✓	85.31
Ours	pyconvresnet50	✓	85.92

<https://doi.org/10.1371/journal.pone.0254054.t005>

Table 6. Comparison results on Stanford Dogs and Stanford Cars.

Method	Top1(%)	
	Stanford Dogs	Stanford Cars
ResNet50	63.27	84.62
BCNN	79.36	88.29
RA-CNN	82.52	91.33
MA-CNN	81.73	90.57
WS-DAN	83.84	92.51
PMG	-	93.48
Ours	85.82	93.64

<https://doi.org/10.1371/journal.pone.0254054.t006>

experiments, we find that pyramid convolution kernel has a better performance in extracting subtle features. In addition, a multi-attention module is designed, which includes spatial and channel attention. It has the advantages of lightweight and can be seamlessly connected in any CNN architecture. The experimental results show that the proposed method is better than other existing methods. However, the proposed method has still a large number of parameters. In the future, we will combine knowledge distillation to study the lightweight of fine-grained classification algorithm.

Supporting information

S1 File.

(DOCX)

Acknowledgments

The author would like to thank Catherine Wah, Aditya Khosla, Jonathan Krause, et al. for making the CUB-200-2011, Stanford dogs, Stanford cars dataset publicly available. The authors would like to acknowledge Professor ShuJun from the School of electrical and electronic engineering of Hubei University of technology and his team for providing GPU computing resources.

Author Contributions

Conceptualization: Gaihua Wang, Lei Cheng.

Methodology: Gaihua Wang, Lei Cheng, Jinheng Lin.

Software: Lei Cheng.

Validation: Yingying Dai.

Writing – original draft: Gaihua Wang.

Writing – review & editing: Lei Cheng, Tianlun Zhang.

References

1. Zhang N, Donahue J, Girshick R, et al. Part-Based R-CNNs for Fine-Grained Category Detection. European Conference on Computer Vision, 2014: 834–849.
2. Huang S, Xu Z, Tao D, et al. Part-Stacked CNN for Fine-Grained Visual Categorization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE 2016: 1173–1182.
3. Wei X, Xie C, Wu J, et al. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recognition the Journal of the Pattern Recognition Society, 2018: 1073–1082.
4. Lin T Y, Roychowdhury A, Maji S. Bilinear CNNs for Fine-grained Visual Recognition. 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 1449–1457.
5. Fu J, Zheng H, Mei T. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition. IEEE Conference on Computer Vision & Pattern Recognition. IEEE 2017: 4476–4484.
6. Zhuang P, Wang Y, Qiao Y. Learning Attentive Pairwise Interaction for Fine-Grained Classification. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7):13130–13137.
7. Ye Z, Hu F, Liu Y, et al. Associating Multi-Scale Receptive Fields For Fine-Grained Recognition. IEEE International Conference on Image Processing (ICIP)2020: 1851–1855. <https://doi.org/10.1109/icip40778.2020.9191018>
8. Tao H, Qi H. See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification, arXiv:1901.09891, 2019.
9. Zhao B, Wu X, Feng J, et al. Diversified visual attention networks for fine-grained object classification. IEEE Transactions on Multimedia, 2017, 19(6): 1245–1256.
10. Luo W, Yang X, Mo X, et al. Cross-X learning for fine-grained visual categorization. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8242–8251.
11. He K., et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence 37. 9(2014):1904–16. <https://doi.org/10.1109/TPAMI.2015.2389824> PMID: 26353135
12. Wah Catherine, Branson Steve, Welinder Peter, Perona Pietro, and Belongie Serge. The caltech-ucsd birds200-2011 dataset, 2011.
13. Khosla Aditya, Jayadevaprakash Nityananda, Yao Bangpeng, and Li Fei-Fei. Novel dataset for fine-grained image categorization: Stanford dogs. CVPR Workshops, 2011.
14. Krause J, Deng J, Stark M, et al. Collecting a large-scale dataset of fine-grained cars. 2013.
15. Alazab A, Venkatraman S, Abawajy J, et al. An optimal transportation routing approach using GIS-based dynamic traffic flows. ICMTA 2010: Proceedings of the International Conference on Management Technology and Applications. Research Publishing Services, 2010: 172–178.

16. Gadekallu T R, Alazab M, Kaluri R, et al. Hand gesture classification using a novel CNN-crow search algorithm. *Complex & Intelligent Systems*, 2021: 1–14.
17. Gadekallu T R, Rajput D S, Reddy M P K, et al. A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. *Journal of Real-Time Image Processing*, 2020: 1–14.
18. Javed A R, Usman M, Rehman S U, et al. Anomaly detection in automated vehicles using multistage attention-based convolutional neural network. *IEEE Transactions on Intelligent Transportation Systems*, 2020.
19. Rehman A, Rehman S U, Khan M, et al. Canintelliids: Detecting in-vehicle intrusion attacks on a controller area network using cnn and attention-based gru. *IEEE Transactions on Network Science and Engineering*, 2021. <https://doi.org/10.1109/tNSE.2020.3032117> PMID: 33997094
20. Gaihua W, Tianlun Z, Yingying D, et al. A Serial-parallel Self-attention Network Joint with Multi-scale Dilated Convolution[J]. *IEEE Access*, 2021.
21. Ye Z, Hu F, Liu Y, et al. Associating Multi-Scale Receptive Fields For Fine-Grained Recognition. *IEEE International Conference on Image Processing (ICIP)2020*: 1851–1855.
22. Gao Y, Han X, Wang X, et al. Channel Interaction Networks for Fine-Grained Image Categorization. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020, 34(07): 10818–10825.
23. Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module. *Proceedings of the European conference on computer vision (ECCV)*. 2018: 3–19.
24. Park J, Woo S, Lee J Y, et al. Bam: Bottleneck attention module. *arXiv:1807.06514*, 2018.
25. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*, 770–778.
26. Zheng H, Fu J, Mei T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition. *Proceedings of the IEEE international conference on computer vision*. 2017: 5209–5217.
27. Du R, Chang D, Bhunia A K, et al. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. *European Conference on Computer Vision*. Springer, Cham, 2020: 153–168.