ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Shanghai, 19-21 June, 2020, pp. 37-41
doi: 10.18178/wcse.2020.06.007

Fine-grained Classification Using Multi-channel ResNet

Di Zang [1+], Yiqing Yan [1], Jun Chen [2, 3] and Yang Li [2]

[1] Department of Computer Science and Technology, Tongji University, Shanghai, China
[2] College of Civil Engineering, Tongji University, Shanghai, China
[3] State Key Laboratory of Disaster Reduction in Civil Engineering, Tongji University, Shanghai, China

Abstract. At present, fine-grained classification has attracted extensive attention. The task of fine-grained classification is difficult due to the challenge of accurately locating the key regions with resolution and extracting valid features from the detected key regions for classification. In this paper, we propose a new convolutional neural network (Multi-channel ResNet). Multi-channel ResNet uses Mask R-CNN for foreground extraction to reduce the interference of image background on fine-grained classification results. In addition, the four-channel ResNet module is used to learn fine-grained features at multiple scales, and Gaussian blur processing and crop processing are used to learn details and contours, all and local features, so as to improve the accuracy of fine-grained classification. The model does not require bounding box/part annotations. We experiment with the CUB_200_2011 dataset, and the results show that Multi-channel ResNet has an improvement in fine-grained classification tasks on the baseline of no pre-trained ResNet-18.

Keywords: Fine-grained Classification, Multi-channel ResNet model, Convolutional Neural Networks

1. Introduction

In recent years, fine-grained classification has a wide range of research needs and application scenarios in industry and academia. It is of significance to use computer vision to realize fine-grained classification.

Fine-grained classification is challenging because there are very few differences between categories. The process framework of previous fine-grained image classification algorithms usually consists of two steps: find the foreground object and its local area in an unsupervised manner or by using the supervised bounding box/local annotations, and then extract the features of these areas respectively which are used to complete the training and prediction of the classifier.

We propose a new fine-grained classification model using no bounding box/part annotations. The model consists of two parts: the foreground is extracted by Mask R-CNN, and multi-scale learning is conducted by Multi-channel ResNet to obtain more fine-grained features. In the Multi-channel ResNet module, we process different channels in different ways (Gaussian blur processing and crop processing). These modules enable the model to eliminate the interference of the background of the images to the fine-grained classification and learn the fine-grained features of multi-scale. We perform an experiment with CUB_200_2011 and it shows that the performance of the model has been improved greatly.

2. Related Work

2.1. Discriminative Feature Learning

Fine-grained image classification mainly studies discriminative feature learning and sophisticated part localization. Discriminative feature learning is the key to fine-grained image classification. The deep residual network [1] expands the number of network layers to 152 based on residual functions. This model obtained

[+] Corresponding author. Tel.: + 021-69589867; fax: + 021-69589867.
E-mail address: zangdi@tongji.edu.cn

an error rate of only 3.75% on the ImageNet test dataset [2]. CNNs are combined with Generative Adversarial Nets (GANs) [3] to improve classification accuracy. Another method proposes to combine the spatially weighted representation of Fisher Vector with CNN [4], and this model is superior to the existing methods in CUB_200_2011 [5] and Stanford Dogs datasets [6].

2.2. Sophisticated Part Localization

Some previous methods use extra annotations of bounding box and part annotations which take a lot of effort and time. In this case, some works use unsupervised approaches. Deep filter response picking [7] proposes to find distinctive filters responding to specific patterns significantly and consistently, and learn a set of part detectors by iteratively and alternately learning. A multiple granularity framework [8] is proposed to encode informative and discriminative features by constructing multi-grained descriptors. On this basis, a new recurrent attention convolutional neural network (RA-CNN) [9] is proposed and it recursively learns discriminative regional attention and region-based feature representation on multiple scales.

3. Proposed Model

The overall structure of the proposed model Multi-channel ResNet is shown in figure 1. In order to make the fine-grained classification results more accurate, on the one hand, we used the Mask R-CNN module to extract the image foreground for eliminating the interference of the background to the classification results. On the other hand, we refer to the residual net module and design a four-channel residual net module to learn multi-scale information.

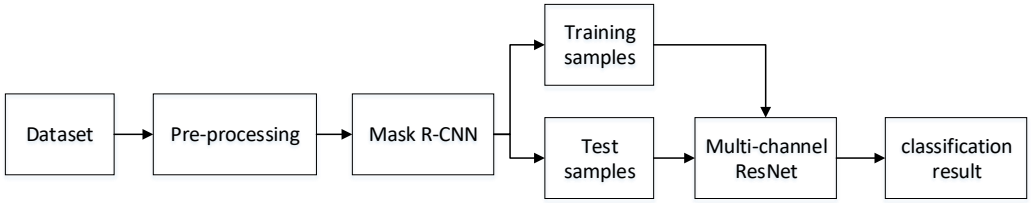

Fig. 1: The process of the Multi-channel ResNet model.

3.1. Mask R-CNN

We used Mask R-CNN to extract the suitable qualified foreground. First, we extract the foreground of the dataset. The foreground detected as bird in the dataset is extracted, and the background is filled with black pixels. Second, we select the first extracted foreground image of each dataset image to be saved as the result. The experimental results show that a number of eligible foreground images may be detected in the images of the dataset, and the small size of some foreground images may have an influence on subsequent experiments. And Mask R-CNN usually detects the most eligible foreground first. Finally, we fill the image background to make it square. Because there may be a big gap between the width and height of the extracted images and it has a certain impact on the subsequent experiments.

3.2. Multi-channel ResNet

The multi-channel ResNet module is the key part of this model. The module learns the idea of ResNet model [1] and extends the single channel to four channels. We use Gaussian blur processing and crop processing to process the input image in different channels. We weaken the details of the image by Gaussian blur processing with different radii so that the model could learn more contour information of different scales. We find the part region of the picture with the most foreground information by crop processing so that the model could learn the characteristics of the parts.

First, process the dataset. We resize the input images to the specified size (224*224), and normalize the RGB value. Second, as shown in Figure 2, the input image is processed with Gaussian blur (r=3, r=11) and partial interception, and the original image and the processed images are respectively used as the input of the four channels. Third, the original images and the processed images of the first three channels are fed into a convolutional layer, a max pooling layer, LAYER1 and LAYER2 respectively to extract feature images (28*28). At the same time, the processed image (56*56) intercepted by the fourth channel passes through the convolutional layer and the dropout layer to extract the feature map with the same size. Finally, we

concatenate the feature maps of four channels in the dimension of channel number, and put it into a convolutional layer, LAYER3 and LAYER4. LAYER1-4 denote the residual layer in ResNet model. Through the final average pooling layer and the full-connect layer, we get the classification results.

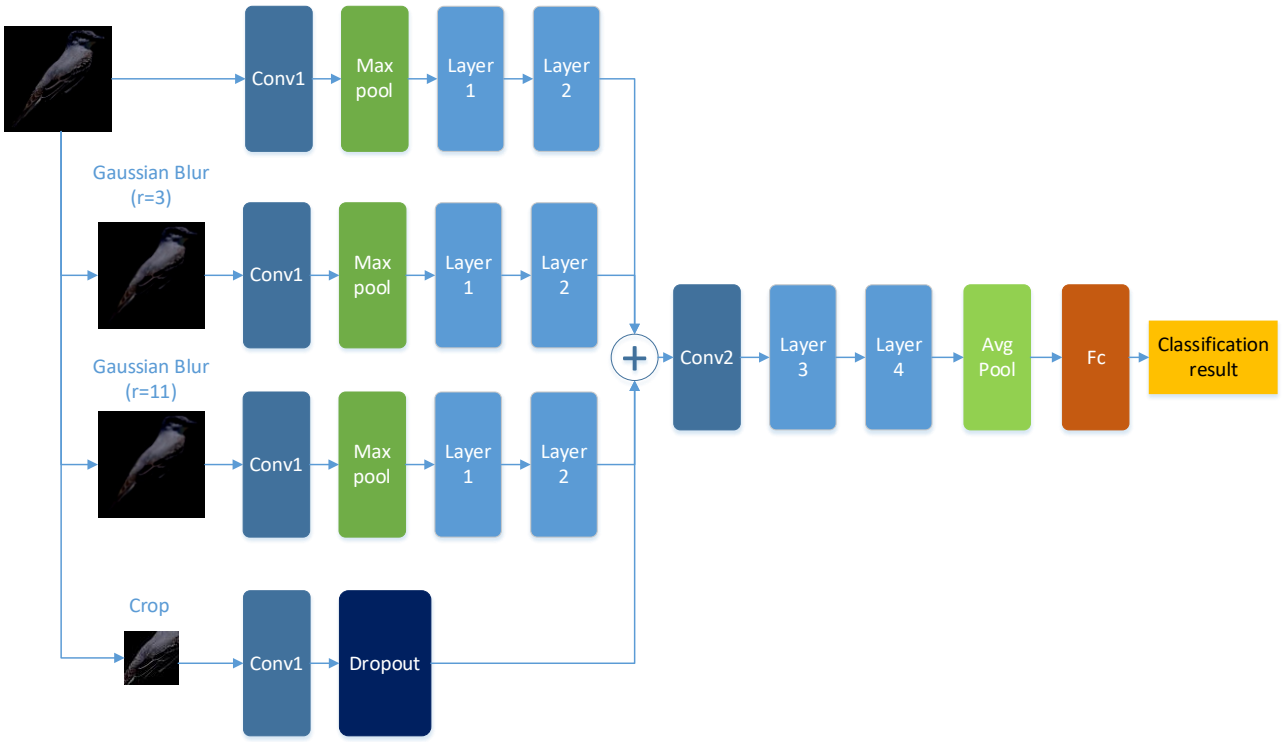

Fig. 2: The structure of the multi-channel ResNet module.

Residual Network. As shown in Figure 3, the first channel is similar to the front part of Resnet-18 model. The original images are fed into a convolutional layer, a max pooling layer, LAYER1 and LAYER2 respectively to extract feature images (28*28). LAYER1 and LAYER2 are made up of two building blocks of the residual network, respectively.

Gaussian Blur Processing. The second and the third channels do different-radius (r=3, r=11) Gaussian blur processing for the input, respectively. Gaussian blur is an image blur filter that uses the normal distribution to calculate the transformation of each pixel in the image.

The coordinate of the center point in the image is set as (0, 0), and the weight matrix can be calculated according to the corresponding coordinates of other pixels. The sum of the weights of all the points in the matrix may not be 1. Therefore, the weight of each point is divided by the sum of the weights, so that the sum of the new weights is equal to 1. We get the final weight matrix. The final weight matrix is taken as the fixed convolution kernel of the convolution layer and convolved with each channel of the RGB three channels respectively. The layers after the Gaussian blur processing layer are the same as for channel 1.

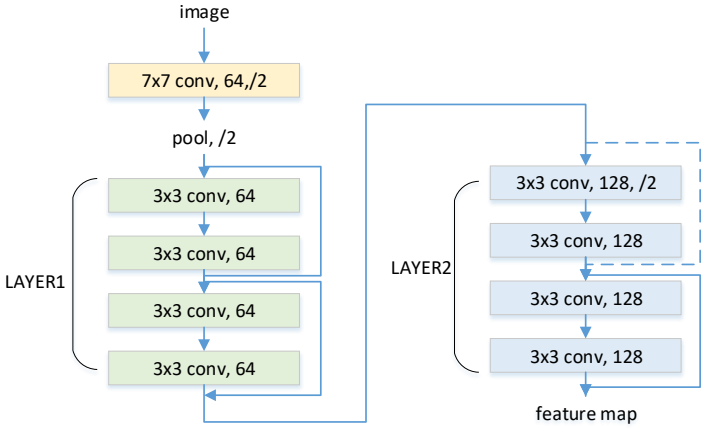

Fig. 3: The first channel.

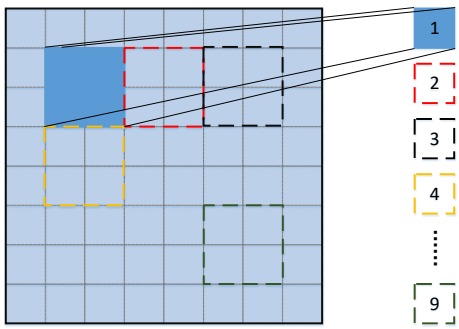

Fig. 4: Candidate regions.

Crop processing. The forth channel selects quarter of the input with the most foreground information.

First, we do the binarization processing for the image matrix. Typically, in the image whose background is black, the background pixel value is 0 and the foreground pixel value is greater than 0. Since the dataset is normalized, 0 cannot be selected as the threshold. Therefore, we set the value of the pixel in the upper left corner of the image as the threshold to distinguish foreground from background, and set the pixel whose value is above the threshold as 1, and vice versa.

Second, we divide the candidate regions and calculate the foreground information value of each candidate region. As shown in figure 4, the nine parts are selected as candidate regions and we calculate the sum of the pixel value in these candidate regions as the foreground information value respectively. Finally, the candidate region with the largest foreground information value is selected as the final region. With crop processing, we get the feature image (56*56). Then the feature image is fed by a convolutional layer and a dropout layer to down-sampling, and we get a feature map with the same size as the first three channels.

4. Experiment and Results

4.1. Dataset Description

We conduct experiments on the challenging fine-grained images recognition dataset Caltech-UCSD Birds (CUB_200_2011). There are 200 classes and a total of 11,788 pictures in this dataset, each containing about 60 pictures. Among them, there are 5994 images of the training dataset and 5794 images of the test dataset.

4.2. Mask R-CNN Results

We use pre-trained Mask R-CNN of the COCO dataset for foreground extraction. Mask R-CNN foreground extraction results are shown in figure 5. We detect the foreground of a bird in 11,621 images in the CUB_200_2011 dataset. These 11,621 images are used as a new dataset for the Multi-channel ResNet module. Split the training dataset and test dataset of the new dataset based on the original splitting of the CUB_200_2011 dataset. Therefore, there are 5911 images of the training dataset and 5710 images of the test dataset.

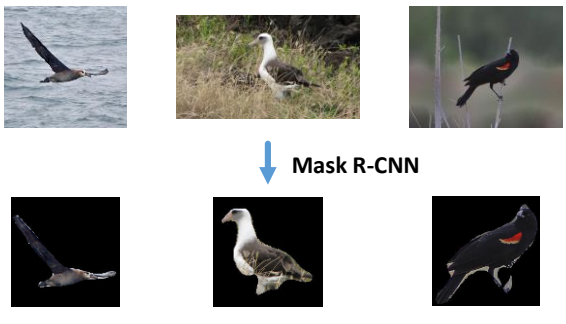

Fig. 5: Mask R-CNN foreground extraction results.

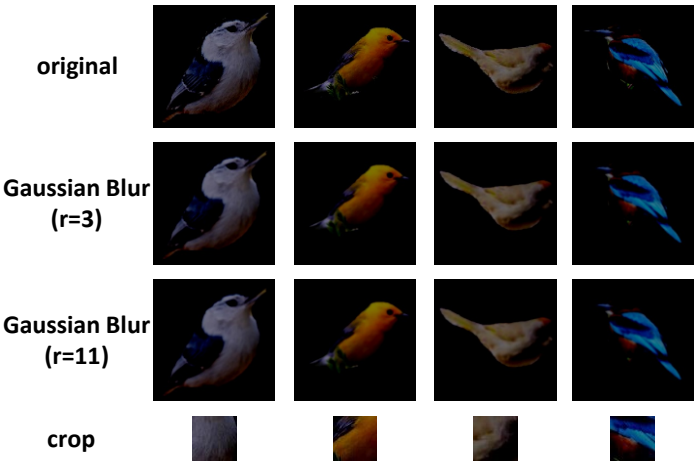

Fig. 6: Gaussian blur and crop processing results.

4.3. Gaussian Blur Processing and Crop Processing Results

The results of Gaussian blur processing and crop processing are shown in figure 6. After Gaussian blur processing, some details in the images are lost, and the model can learn more about the contour. After crop processing, the captured images are basically bird's feathers, and the module can learn details such as the color and pattern of feathers.

4.4. Classification Accuracy

The fine-grained classification accuracy results are shown in Table 1. We use no pre-trained ResNet-18 as the baseline and do not use bounding box/part annotations. As can be seen, the classification accuracy of ResNet is greatly improved by first processing the dataset with Mask R-CNN, the accuracy of the model

with Mask R-CNN is increased by 7.365% compared with the baseline. Moreover, in the Multi-channel ResNet module, each time the channel is added, the accuracy rate will be improved on the original basis. The accuracy of the model with Mask R-CNN and the four-channel module has the best accuracy, which is increased by 13.275% compared with the baseline. It indicates that we can reduce the interference of image background on classification results through Mask R-CNN module. And Through Gaussian blur processing and crop processing, multi-scale and multi-channel learning is added, which can obtain more scale information in the image and improve the accuracy of the model.

Table 1: The accuracy of the fine-grained classification results.

Models	Accuracy (%)
ResNet (channel 1)	50.897
Mask R-CNN + ResNet (channel 1)	58.262
Mask R-CNN + ResNet (channel 2)	57.300
Mask R-CNN + ResNet (channel 1 and 2)	63.053
Mask R-CNN + ResNet (channel 1, 2 and 3)	64.067
Mask R-CNN + ResNet (channel 1, 2, 3 and 4)	64.172

5. Conclusion

In this paper, we propose a multi-channel convolutional neural network (Multi-channel ResNet) based on mask R-CNN and ResNet for fine-grained classification to learn fine-grained characteristics at different scales. The neural network does not require bounding box/part annotations for training and does not use the pre-trained model for training. Experiments show that the performance of the model is improved greatly on the basis of the original model ResNet-18 in the fine-grained classification of bird datasets CUB_200_2011.

6. Acknowledgment

This work is supported by National Natural Science Foundation of China (No. 61876218) and National Natural Science Foundation of China (U1711264).

7. References

[1] K. He, X. Zhang, S. Ren, J. Sun. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 770-778.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in neural information processing systems. 2012, pp. 1097-1105.

[3] E. Xie, G. Li, W. Liu. Improving Fine-Grained Object Classification Using Adversarial Generated Unlabelled Samples. In 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM). 2018, pp. 1-5.

[4] F. Perronnin, D. Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, pp. 3743-3752.

[5] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie. The Caltech-UCSD Birds-200-2011 dataset. 2011.

[6] A. Khosla, N. Jayadevaprakash, B. Yao, F. F. Li. Novel dataset for fine-grained image categorization: Stanford dogs. In Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC). 2011, Vol. 2, No. 1.

[7] X. Zhang, H. Xiong, W. Zhou, W. Lin, Q. Tian. Picking Deep Filter Responses for Fine-grained Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 1134-1142.

[8] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, Z. Zhang. Multiple Granularity Descriptors for Fine-grained Categorization. In Proceedings of the IEEE international conference on computer vision. 2015, pp. 2399-2406.

[9] J. Fu, H. Zheng, T. Mei. 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. 2017, pp. 4438-4446.