LEARNING TO FOCUS AND DISCRIMINATE FOR FINE-GRAINED CLASSIFICATION*

Zhicong Feng, Keren Fu†, Qijun Zhao

National Key Laboratory of Fundamental Science on Synthetic Vision
College of Computer Science, Sichuan University, Chengdu, China

ABSTRACT

Existing state-of-the-art fine-grained classification methods usually use separated networks for discriminative region localization and feature learning/classification, and are thus complicated to implement and optimize. In this paper, we aim to provide a compact solution by deepening the collaboration between the region localization, feature learning and classification modules during the learning process of fine-grained classification. We thus propose a method that can learn to simultaneously localize discriminative regions and extract discriminative features by exploring the localization ability of classification convolutional neural networks and joint optimization of different modules. Our method, while being built upon a single backbone network and trained with only softmax losses, achieves state-of-the-art performance on three benchmark fine-grained datasets, which proves that our method is simple but effective for fine-grained classification.

Index Terms— Fine-grained classification, region proposal, discriminative region localization, attention, convolutional neural networks

1. INTRODUCTION

Fine-grained classification assigns subordinate categories to objects that belong to the same basic-level category, e.g., species of birds and flowers, and models of aircraft and cars. It is obviously a challenging task to fulfill fine-grained classification based purely on the appearance of objects because of intrinsically small inter-class differences and potentially large intra-class differences [1]. In the early days, researchers turned to domain experts for prior knowledge of discriminative features of objects in different subcategories, and relied on manually annotated bounding boxes or object parts [2, 3, 4]. However, acquiring human annotations is expensive, and expert-defined discriminative features might not be optimal for auto-classification by computers.

Recent methods [1, 5, 6, 7, 8, 9, 10, 11, 12] attempt to automatically find discriminative regions in objects and learn effective feature representations without the need of bounding box/part annotations. Some of them [5] treat images as a whole and directly learn discriminative texture features for fine-grained classification. Others [1, 6, 7, 10, 12] first detect discriminative regions in objects and then extract features from the regions to distinguish objects in different subcategories. With assistance of local attention mechanisms, these latter approaches achieve state-of-the-art performance. Yet, they usually require separated sub-networks for generating proposals of discriminative regions and for extracting features and classification. To train the networks, they often have to use specially designed losses. As a result, these methods are mostly complicated to implement and hard to be optimized.

The goal of this paper is to provide a compact approach that can simultaneously detect discriminative regions and extract features for recognizing objects in different subcategories. Our approach is motivated by the work in [13], which shows that convolutional neural networks (CNNs) trained with image-level labels have the ability to localize discriminative image regions. Based on a backbone network for fine-grained classification, we utilize its extracted feature maps to generate candidate discriminative regions, and zoom in on these regions to further extract local discriminative features. Unlike existing methods, our proposed method, while being implemented with a single network and trained by merely softmax losses, can learn to focus on discriminative image regions and meanwhile improve the effectiveness of extracted features in distinguishing objects of different subcategories. Evaluation results on three benchmark databases prove that the proposed method is simple but effective for fine-grained classification.

The rest of the paper is organized as follows. Section 2 briefly reviews related work. Section 3 introduces our method in detail. Section 4 presents and analyses the experimental results, followed by conclusions in Section 5.

2. RELATED WORK

How to localize and represent discriminative regions in images has for a long time been a central issue in fine-grained classification. Many efforts have been made in the past five years to solve this problem. In early methods [1], discriminative region localization and fine-grained feature extraction are

*This work is supported by the National Natural Science Foundation of China (61773270, 61703077), the Fundamental Research Funds for the Central Universities (YJ201755), and the Shenzhen Fundamental Research fund (JCYJ20180305125822769).

†Corresponding Author - Keren Fu (fkrsuper@scu.edu.cn).

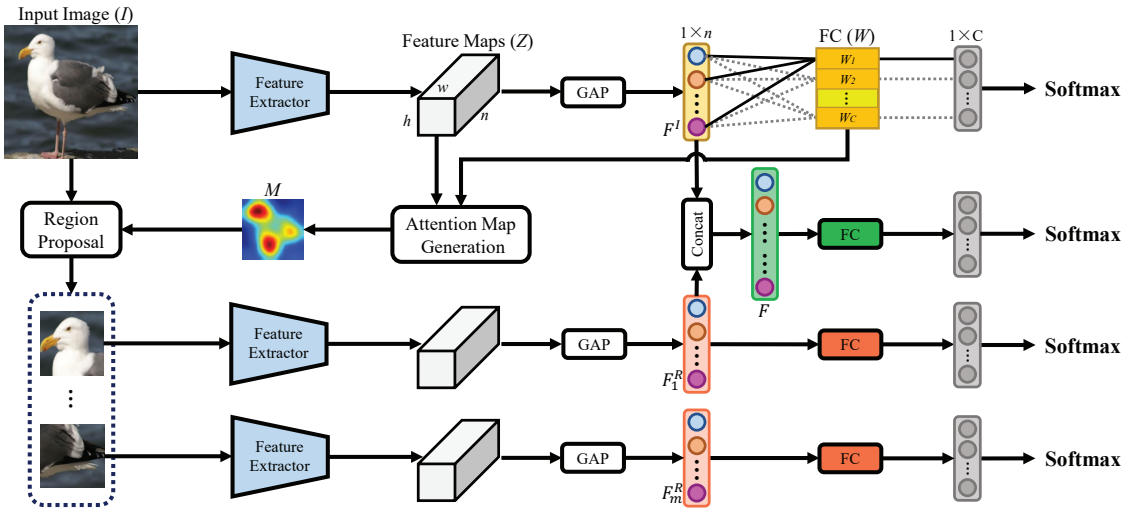

Fig. 1. The overall framework of our method. The input image I is fed into the fully-convolutional feature extractor to generate feature maps Z, which are converted to the condensed *image-level* feature F^I by global average pooling (GAP). The image-level classification module (shown in yellow color) then produces classification scores from F^I via a fully-connected (FC) layer and softmax. Based on the image-level feature maps Z and FC weights W, an attention map M is generated from which m regions of top discriminativeness are proposed and cropped from the input image. *Region-level* features $\{F_1^R, \cdots, F_m^R\}$ are then extracted from the zoomed-in versions of these local regions and classified by the region-level classification module (show in orange color). The final classification result is obtained based on the *hybrid* feature F that is a concatenation of image-level feature and top-l region-level features, and by using the hybrid classification module (shown in green color).

usually separately implemented. Observing the close correlation between region detection and feature learning, Fu et al. [9] and Zheng et al. [10] propose to jointly optimize the attention proposal networks (APN), which aim to detect discriminative local regions, and the classification networks (CN), which aim to learn feature representations and recognize the fine-grained categories of objects. In their methods, APN and CN partially overlap by sharing their feature extraction modules and are trained under multiple supervisions, which enable the mutual boost of region localization and feature learning. Recently, Yang et al. [12] further develop these methods by proposing a more elaborate training paradigm with specially designed navigation losses. Despite the non-negligible improvement achieved by them, the implementation of their method is inevitably complicated.

Revisiting these existing state-of-the-art fine-grained classification methods, we see on the one hand the importance of integrating region localization and feature learning and categorization. On the other hand, we also observe that existing methods share only feature extractors between the modules of region localization and classification. Being inspired by the localization ability of classification CNNs [13], we propose in this paper a deeper collaboration between region localization and classification modules by sharing both feature representation and classification layers. This way, not only does the overall fine-grained classification network become simple, but also more discriminative local regions and feature representations can be obtained and thus higher classification accuracy as well. It is worth mentioning that the authors of [13] apply their method to fine-grained classification, but only for selecting one local region without joint learning with feature extraction and classification.

3. OUR APPROACH

3.1. Overview

Figure 1 shows the overall framework of our proposed method. Given an input image I, a set of n feature maps $\{Z_i \in \mathbb{R}^{h \times w} | i = 1, 2, \cdots, n\}$ are extracted from it via a feature extractor consisting of a number of convolutional layers. Global average pooling (GAP) is then applied to these feature maps, resulting in a condensed *image-level* feature representation $F^I \in \mathbb{R}^{1 \times n}$. This feature is fed into a fully connected layer, which maps the feature to scores measuring the probabilities of the input image belonging to different fine-grained classes as follows,

$$S_k = \sum_{i=1}^{n} W_{ki} F_i^I, k = 1, 2, \cdots, C. \quad (1)$$

Here, C is the total number of fine-grained classes, and $W_k \in \mathbb{R}^{1 \times n}$ denote the stacked weights associated with the k^{th} class. According to these scores, image-level fine-grained classification result is obtained via softmax.

Meanwhile, based on the classification weights W and the feature maps Z, an attention map $M \in \mathbb{R}^{h \times w}$ is generated, which indicates the discriminativeness of local regions with respect to fine-grained classification. According to M, m regions of top discriminativeness are localized on the input image. These regions are cropped and resized to higher resolution. All of them then go through the same feature extraction process as in the image-level, resulting in m condensed *region-level* features $\{F_r^R \in \mathbb{R}^{1 \times n} | r = 1, 2, \cdots, m\}$. From each of these region-level features, region-level fine-grained classification results are obtained. Moreover, top-l region-level features are concatenated with the image-level feature, leading to a *hybrid* feature $F = [F^I F_1^R \cdots F_l^R]$, from which the final fine-grained classification result is obtained via softmax. Note that in real-world deployment, image-level and region-level softmax classification are used during training only.

3.2. Discriminative Region Localization

Next, we introduce in detail how the attention map is generated based on the extracted features and the classification weights, and how the discriminative local regions are determined. Given the image-level feature maps $Z \in \mathbb{R}^{h \times w \times n}$ and the image-level classification weights $W \in \mathbb{R}^{C \times n}$, the attention map is defined by

$$M = \max\{M_k = \sum_{i=1}^{n} W_{ki} Z_i \in \mathbb{R}^{h \times w} | k = 1, 2, \cdots, C\}, \tag{2}$$

where '$\max$' is element-wise max pooling across the C class activation maps (i.e., M_k, which reveals the importance of different local regions on the input images for the recognition of specific object classes [13]). Instead of using single class activation maps (i.e., the one of the ground truth class during training, or the one of the predicted class during testing) as in [13], we max-pool all the class activation maps. This is because we believe that once the image-level classification module predicts wrong class labels, the attention map would be misleading due to the only use of unreliable class activation map of the wrongly predicted class. To further increase the robustness of M to noise, we apply Gaussian smoothing to the attention map. By integrating the class activation maps, our attention map considers both the representativeness of local regions for specific classes and the separability between different classes.

To obtain multiple discriminative regions, we produce a series of pre-defined regions from the attentiona map M inspired by the idea of region proposal network in [14]. Specifically, each pixel in M corresponds to a region in the input image I, and the pixel value measures the discriminativeness of the corresponding region. We sort the regions according to their discriminativeness and adopt non-maximum suppression (NMS) to reduce redundancy. Then we select top-m discriminative regions, crop them from the input image and zoom in on them to increase their resolution such that finer detail can be extracted.

3.3. Training Paradigm

We now introduce how the proposed method is trained. Note that the feature extractor modules are shared between image level and region level, and the classification modules (i.e., the fully connected layers) for different local regions are identical. Therefore, the trainable parameters in our method are owed to the feature extractor module, and the fully connected (FC) layers in the image-level classification module, the region-level classification module and the hybrid classification module. To learn these parameters, we apply softmax losses for all the classification modules involved in our proposed network, and train the entire network in an end-to-end manner. Note that in our method the learning of discriminative region detection, feature representation, and classification is interleaving, and more importantly, is implemented in a more deeply collaborative way by sharing more parameters. This greatly contributes to the superior performance of our proposed method, as being demonstrated by our experimental results that are reported in the next section.

4. EXPERIMENTS

4.1. Datasets and Baselines

We evaluate our proposed method on three benchmark datasets, CUB-200-2011 [15], FGVC Aircraft [16] and the Stanford Cars [17]. These datasets are widely used in the fine-grained image classification literature. CUB-200-2011 contains 200 species of birds with $5,994$ images for training and $5,794$ images for testing. This dataset is considered one of the most competitive datasets for fine-grained classification due to the limited number of images of each species. FGVC Aircraft contains 100 classes of aircraft models with $6,667$ images for training and $3,333$ images for testing. The Stanford Cars dataset includes 196 car models with $8,144$ images for training and $8,041$ images for testing.

We compare our proposed method against the following baseline methods, the Bilinear-CNN [5], RA-CNN [9], The ResNet-50 implemented in [18], Boost-CNN [19], MA-CNN [10], PartNet-Full [11] and NTS-Net[12], all of which do not depend on bounding box/part annotations. For a fair comparison, the accuracy of these methods, if available, is directly cited from their source papers.

4.2. Implementation Details

In our experiments, we adopt the ResNet-50 model in [20] as the backbone network (i.e., the feature extractor module), which has been pre-trained on ILSVRC2012 [21]. The input

images, the localized discriminative regions, and the zoomed-in local regions, respectively, have resolution of 448×448, 160×160, and 224×224 pixels. The NMS threshold is set to 0.25. For a fair comparison, the hyper parameters m and l are set according to NTS-Net [12], i.e., $m = 6$ and $l = 4$. Stochastic Gradient Descent is employed to optimize our model with weight decay of $5e - 4$ and momentum of 0.9. The training process continues for 100 epochs with the base learning rate initialized at 0.01 and multiplied by 0.1 after every 40 epochs. Note that the learning rate for the pre-trained layers in the feature extractor module is 0.1 times the base learning rate.

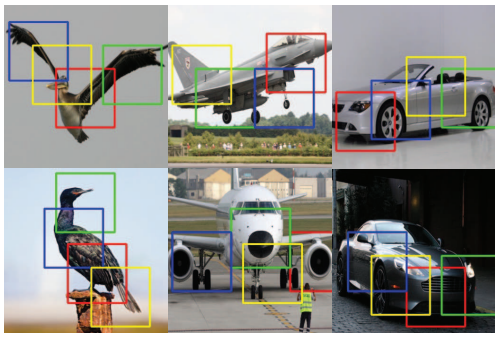

Fig. 2. Four discriminative regions localized by our proposed method on images of birds, aircraft and cars. Top-1 to top-4 regions are shown in red, gree, blue and yellow, respectively.

4.3. Results

The classification accuracy of our method and the counterpart methods on CUB-200-2011 is presented in Table 1, from which we can see that our method achieves the highest accuracy. Note that our method outperforms the method in [18], which is also based on ResNet-50, by 3.1%. We believe that this improvement owes to the more effective learning ability of our method for both discriminative region detection and feature extraction and classification.

Table 1 also shows the results on FGVC Aircraft and Stanford Cars. As can be seen, our method overwhelms the existing top method, NTS-Net, by a clear margin of (0.7%) and establishes a new state-of-the-art result on FGVC Aircraft. Similarly, the accuracy of our method on Stanford Cars (94.1%) is also the best one, compared with the counterpart methods. These results demonstrate the superiority of our proposed method in learning effective feature representations for fine-grained classification. Figure 2 visualizes the top-4 local regions detected by our method on some example images. Obviously, the selected regions appear to coincide with our human beings' perception of these objects (e.g., focus on head and wing of birds, headlight and logo of cars, and head and wing of aircraft).

To further assess the contribution of different components

Method	Accuracy (%) on		
	Birds	Aircraft	Cars
Bilinear-CNN [5]	84.1	84.1	91.3
ResNet-50 [18]	84.5	-	-
RA-CNN [9]	85.3	88.2	92.5
Boost-CNN [19]	85.6	88.5	92.1
MA-CNN [10]	86.5	89.9	92.8
PartNet-Full [11]	87.3	-	-
NTS-Net [12]	87.5	91.4	93.9
Ours	**87.6**	**92.1**	**94.1**

Table 1. Classification accuracy on CUB-200-2011 (Birds), FGVC Aircraft (Aircraft) and Standord Cars (Cars).

in our method, we conduct ablation study on CUB-200-2011 with different settings. First, we remove the hybrid classification module during training and use the image-level classification module during testing to predict the fine-grained classes. According to the results in Table 2, in this case our method achieves an accuracy of 86.7%, which outperforms the ResNet-50 [18] by 2.2% and even outperforms the image-level NTS-Net [12] by 1.4%. Yet, after including the hybrid classification module, the accuracy of our method is further improved to 87.6%.

Second, following the method in [13], we use single class activation maps, rather than max-pooling all the class activation maps, to generate the attention map. Consequently, the accuracy of our method drops from 87.6% to 87.3%, which proves the importance of considering the classification weights of all classes.

Method	Accuracy (%)
Res-Net50 [18]	84.5
NTS-Net (image level only) [12]	85.3
Ours (image level only)	86.7
Ours (without max-pooling)	87.3
Ours (hybrid + max-pooling)	**87.6**

Table 2. Ablation study results on CUB-200-2011.

5. CONCLUSIONS

In this paper, we propose a method for fine-grained classification, which can learn to focus on discriminative local regions and extract more feature representations in a more effective way. Its key idea is to deepen the collaboration between the region localization, feature learning and classification modules in fine-grained classification networks. Evaluation experiments on three benchmark databases demonstrate that the proposed method is simply but effective, and establishes new state-of-the-art accuracy compared with existing methods.

6. REFERENCES

[1] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *CVPR*, 2015.

[2] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell, "Part-based r-cnns for fine-grained category detection," in *ECCV*, 2014.

[3] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas, "Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition," in *CVPR*, 2016.

[4] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang, "Part-stacked cnn for fine-grained visual categorization," in *CVPR*, 2016.

[5] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, "Bilinear cnn models for fine-grained visual recognition," in *ICCV*, 2015.

[6] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue, and Zheng Zhang, "Multiple granularity descriptors for fine-grained categorization," in *ICCV*, 2015.

[7] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian, "Picking deep filter responses for fine-grained image recognition," in *CVPR*, 2016.

[8] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan, "Diversified visual attention networks for fine-grained object classification," *Trans. Multi.*, vol. 19, no. 6, pp. 1245–1256, 2017.

[9] Jianlong Fu, Heliang Zheng, and Tao Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *CVPR*, 2017.

[10] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *ICCV*, 2017.

[11] Yabin Zhang, Kui Jia, and Zhixin Wang, "Part-aware fine-grained object categorization using weakly supervised part detection network," *arXiv preprint arXiv:1806.06198*, 2018.

[12] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang, "Learning to navigate for fine-grained classification," in *ECCV*. 2018.

[13] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[15] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie, "The caltech-ucsd birds-200-2011 dataset," *Tech. rep.*, 2011.

[16] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi, "Fine-grained visual classification of aircraft," *Tech. rep.*, 2013.

[17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei, "3d object representations for fine-grained categorization," in *ICCV Workshops*, 2013.

[18] Zhichao Li, Yi Yang, Xiao Liu, Feng Zhou, Shilei Wen, and Wei Xu, "Dynamic computational time for visual attention," in *ICCV Workshops*, 2017.

[19] Mohammad Moghimi, Serge J Belongie, Mohammad J Saberian, Jian Yang, Nuno Vasconcelos, and Li-Jia Li, "Boosted convolutional neural networks," in *BMVC*, 2016.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.