Fine-grained image classification based on attention-guided image enhancement

Junming Lu¹, Wei Wu²

¹Department of Information Engineering, Wuhan University of Technology, Wuhan, Hubei Province, 430000, China
²Department of Information Engineering, Wuhan University of Technology, Wuhan, Hubei Province, 430000, China

*Corresponding author’s e-mail: wuwei@whut.edu.cn

Abstract. Extracting distinguished fine-grained features is essential for fine-grained image recognition tasks. Many researchers use expensive manual annotations to learn to distinguish part models, which may not be possible in practical applications. Unlike previous strongly supervised fine-grained classification networks that require additional image annotations, weakly supervised fine-grained image classification only requires label annotations. Recently, image enhancement has been increasingly used in network structures, but random enhancement will lead to background noise and filter out irrelevant areas. In this article, we propose a weakly supervised fine-grained image classification network based on attention-guided image enhancement to study the effect of image enhancement on the classification network. In detail, we use the backbone network to generate the feature map of the image, then generate the corresponding attention map through a custom mask, and use the attention map to guide the image enhancement process (including image cropping and image dropping). We conducted experiments on three commonly used fine-grained image classification datasets, and achieved sota effects in CUB, FGVC-Aircraft, and Stanford Cars.

1. Introduction

Compared with traditional coarse-grained image classification which only focuses on distinguishing the basic categories of objects, fine-grained image classification is more challenging. It aims to distinguish subordinate categories with subtle visual differences, such as the species of bird, model of car and type of airplane. Compared with coarse-grained classification, fine-grained visual classification mainly has the following reasons: (1) The difference between classes is small. For example, the color of a part of a bird's feathers can determine which species it belongs to. (2) Training data is limited. Usually the dataset is only labelled with the category it belongs to. Fine-grained image classification has very important research value. For example, in real life, this technology is used to classify different types, such as identifying different goods in unmanned retail supermarkets. Using fine-grained image classification can effectively reduce human and financial resources. It has great research significance for both academic field and industrial field.

The key step of fine-grained image classification is to extract the most distinguishing local features from the picture. If the feature area is manually marked, it will take extra time and economic costs. Therefore, more methods are considered to use weakly supervised learning to achieve classification. Because ordinary neural networks cannot handle the problem of fine-grained classification, some people propose to use data enhancement to filter out the areas that need attention. Data enhancement is a
commonly used method to increase the amount of data training data, which is used to prevent overfitting and improve the performance of the model. Many of the commonly used data enhancement methods use random approaches, such as random image cropping, rotation, enlargement. There is a certain probability that it can crop the target to our needs, but the greater probability is the target that are not needed, such as noisy targets or background information. So we designed a data enhancement module guided by attention, which can achieve better classification results. Our contributions are summarized as follows:

- We propose a FGIC approach based on attention-guided image enhancement that achieves comparable performance to the state-of-the-art methods.
- We propose an attention extraction module using mask generation procedure to get different salient feature areas of the image.

2. Related Work
In the past decade, fine-grained visual classification has been a very active research area. The early solution was a classification algorithm based on strongly supervised algorithm. Researchers have introduced additional bounding boxes or additional annotations to help locate the target object and align the components in the image. Although the performance of these methods was satisfactory, they are not optimal for fine-grained image classification problems, because manual annotation can be very time-consuming and requires professional knowledge. Instead, there are some new methods to solve this problem with weak supervision.

2.1. Strongly supervised Fine-grained Visual Classification
The classification algorithm based on strongly supervised information appeared earlier, and its notable feature is that during network training, not only the category labels of fine-grained images are used, but also the location of the object in the image, the key parts of the object and other artificial labelling information are needed. This additional annotation help the network learn more discriminative features, thereby improving classification accuracy. In order to reduce the difficulty of intra-class differences, researchers focus on capturing the discriminative semantic parts of fine-grained objects, and then construct intermediate representations corresponding to these parts for final classification.

Part-based R-CNN[1] is based on the famous target detection algorithm R-CNN. The main idea is: First, use the selective search method to generate candidate frames that may be objects or key parts of objects in fine-grained images. Then, three detection models are supervised and trained based on the label box and the label information of the key parts of the object, one of which is the object-level detection model, and the other two are the detection models of different key parts of the object. Then, the corresponding geometric position constraints are added to the predicted frames obtained by the three detection models, so as to obtain more ideal detection results of objects and key parts. However, in the Part-based R-CNN algorithm, it is not only necessary to use object position labelling boxes and key parts labelling information during training, but also to require the predicted image to provide these additional artificial labelling information during model prediction. Pose Normalized CNN is an improved version inspired by the Part-based R-CNN algorithm. Compared with the Part-based R-CNN algorithm, which simply adds the corresponding geometric position constraints to the predicted frame obtained by the detection model, Pose Normalized CNN uses a more accurate sub-image pose alignment operation.

2.2. Weakly supervised Fine-grained Visual Classification
The advantage of the above classification algorithm based on strongly supervised method is that with the help of additional manual labelling information, the network can extract refined features more accurate, thereby improving the classification accuracy of the network. The shortcomings are also obvious: due to the need for additional labelling information, it will consume a lot of manpower and time, and is not practical. In addition, it is easy to make mistakes using manually labelling information, which will directly affect the robustness of the algorithm. Therefore, some researchers have proposed
classification algorithms based on weakly supervised method. This type of algorithm does not require any manual labelling information and only use the category information of the object in image.

The Bilinear CNN[2] algorithm relies on the linkage of two feature extractors whose outputs are multiplied using outer product at each location of the image and pooled to obtain an image descriptor. The key is bilinear pooling, which uses second-order statistical information to fuse information from different channels to achieve fine-grained classification. However, the problem with the original bilinear pool is its high dimensionality, so some people have proposed some methods to solve this problem. The CBP[3] proposed by Gao et al. (Compact bilinear pooling) uses Tensor Sketch projection and Random Maclaurin projection to greatly reduce the size without reducing accuracy. Many previous methods ignore the interaction of local features between layers and the learning of fine-grained features are interrelated and can also strengthen each other. The author of WSDAN[4] proposed a bilinear attention pooling mechanism (BAP) and an attention-based data enhancement strategy to improve their representation ability. They propose attention-guided data augmentation to improve the efficiency of data augmentation, including attention cropping and attention dropping.

However, in the WSDAN network, the threshold of cropping and dropping is a fixed value which is artificially designed. Instead, we designed a threshold based on image binarization, so that each picture uses the corresponding threshold range according to its own characteristics. This can effectively improve the classification performance.

3. Approach

In this chapter, we introduce the proposed network structure in detail, including weakly supervised attention map, attention map guided data enhancement and loss function design. For each input image, a weakly supervised learning network is designed to generate a corresponding attention map to represent the most discriminative feature of the target, and then crop or enlarge it through data enhancement methods based on the attention map, and finally integrate these inputs through the network to get the final classification result. The process is shown in Fig.1.

![Figure.1 Overview of our approach](image)

3.1. Weakly supervised attention map

In the training and testing phases, we do not use bounding boxes or key point related information. Instead, we use weakly supervised learning to predict the location distribution of objects by labelling object categories. The specific method is as follows. For the part of weakly supervised learning, we use resnet50 as the backbone network. After extracting the basic feature map for each training image, the corresponding attention map is generated through the designed mask to represent the most salient feature part of the image. By using the attention map, we manage to design a more flexible network structure, and end-to-end training for fine-grained classification can be achieved more easily.

Inspired by the bilinear pooling network, we designed the second part of the network structure. Each attention map represents a specific part of the target image. Each part is obtained by multiplying the obtained feature map and the elements of the attention map. Then use convolution to process the partial feature maps, and finally combine each partial feature map to obtain a feature matrix.
3.2. Attention map guided data enhancement

The efficiency of random data enhancement is very low, especially if the object occupies a relatively small amount of the whole picture, it will lead to more background noise, so using the attention map for data enhancement[4] can achieve better results. Besides this reason, the object occupies different proportions in different image. So instead of using a manually designed value to adjust the threshold, we let the image itself determine its own threshold. An adaptive binarization threshold algorithm is used to obtain the average value of the entire image, which is used as the benchmark. In this way, we can get the most suitable image enhancement for different image. The region which is larger than the benchmark is cropped to obtain the key region, and weakly supervised learning is used on this region to generate an attention map to represent the salient features of the target, which we call cropped attention. In addition, we add another branch, discarding areas with higher benchmark and using the left parts, so that the network can learn some features of non-key areas, and thus generate a secondary focus attention map, which we refer as dropped attention. Combined with these two branches, our image enhancement can effectively extract the most useful regional features in the image.

![Figure 2. comparison of original image, mask-generation image and detail-focused image (from left to right).](image)

The detail-focused images can be seen in Figure 2. The first picture is the original picture, the second picture is the image obtained after designed mask processing, it can be seen that the main part and the background part are distinguished, and the third picture is the image obtained after data enhancement guided by the attention map. The red part is the area with the most obvious feature information.

3.3. Loss function design

Inspired by Focal loss[5], we combine the standard cross-entropy loss with the results of focal loss. By reducing the weight of samples which are easy to classify, the model can focus more on difficult-to-classify samples during training. The network can adapt to more datasets by controlling the different proportions of the two parts. The equation is as below.

\[
\text{focalloss} = \begin{cases} 
-\alpha(1 - y')\log(y') & , y = 1 \\
-(1 - \alpha)y'^\gamma\log(1 - y') & , y = 0 
\end{cases} 
\]  

(1)

At the same time, the center loss[6] function is introduced to our network. A class center is maintained in the feature map for each category of the training set. In the training process, the distance constraint between the sample in the feature map and the class center after network mapping is increased to achieve better classification results.

\[
\text{centerloss} = \frac{1}{2N} \sum_{i=1}^{y} \| x_i - c_i \|^2 
\]

(2)

The proportion of the two loss functions is controlled by the hyper parameter \( \beta \). We design to adjust the \( \beta \) value for different datasets to achieve the best results.

\[
\text{Loss} = \text{focalloss} + \beta \times \text{centerloss} 
\]

(3)
4. Experiments
In this section, we conducted a lot of experiments and comparisons to prove the accuracy of the model. First, we introduce the datasets used in the experiment. Secondly, we introduced the environment in which the experiment was run. Finally, our model is compared with the state-of-the-art methods.

4.1. datasets
Caltech-UCSD Birds-200-2011 (CUB-200-2011) This dataset is the most widely used benchmark for fine-grained image classification. The data set covers 200 species of birds, including 5994 training images and 5794 test images. In addition to the category label, each image will be further annotated with 1 bounding box, 15 part key points and 312 attributes. The Stanford Cars dataset contains a total of 16185 car pictures of different models, of which 8144 are the training set and 8041 are the test set. FGVC-Aircraft Benchmark aircraft data set contains 10,200 aircraft images, of which 102 different aircraft, each with 100 images.

4.2. Implement Details
The experimental environment is based on the Ubuntu 16.04 operating system, the CPU is Intel i7-9700, the GPU is 2 NVIDIA RTX 2080, and the running memory is 16G. This model employs the Resnet50 model pre-trained on ImageNet as the basic network. We removed the original fully connected layer and added a convolutional layer, down-sampling layer and feature extraction layer to extract regions of interest and use the custom mask to generate attention map. The weight decay and momentum are set to $1 \times 10^{-5}$ and 0.9 respectively, the learning rate is initially set to 0.001, and the decay rate is 0.9.

4.3. Comparison
We compare our method with the latest methods on the fine-grained classification dataset described above. The results are shown in the table respectively. It can be seen that our network architecture achieves the best classification effect on all these fine-grained data sets.

| Dataset     | CUB-200-2011 | Stanford-Cars | FGVC-Aircraft |
|-------------|--------------|---------------|---------------|
| BCNN[2]     | 84.1         | 91.3          | 84.1          |
| RA-CNN[7]   | 85.3         | 92.5          | 88.2          |
| MA-CNN[8]   | 86.5         | 92.8          | 89.9          |
| LRB[9]      | 84.2         | 90.1          | 87.3          |
| HBP[10]     | 85.8         | 92.2          | 90.2          |
| NTS-Net[11] | 87.5         | 93.9          | 91.2          |
| CrossX[12]  | 87.7         | 94.5          | 92.6          |
| DCL[13]     | 87.8         | 94.5          | 93.0          |
| WSDAN[4]    | 88.4         | 94.4          | 93.0          |
| Ours        | **88.6**     | **94.5**      | **93.1**      |

The comparison results are shown in Table 1. We can see: 1) Our network structure is better than many strongly supervised methods that use additional image annotations, which shows that our model can achieve excellent classification results without other information; 2) Our network is better than most weakly supervised method. This means that our proposed image enhancement mechanism based on the attention map is effective and can get more accurate ROI area while discarding useless background information.
5. Conclusion
In this article, we propose a weakly supervised fine-grained image classification network based on self-constrained image enhancement. Weakly supervised classification network generates attention map, and self-constraint can get enhanced image, which in turn guides to get better classification results. The two modules interact to allow the network model to obtain more local information and ensure the accuracy of the classification results.

Acknowledgments
Thanks for the assistance of colleagues and mentor, as well as the financial support from the National Natural Science Foundation of China: Research on the visual perception enhancement method of unmanned surface ships in severe navigable environment.

References
[1] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, ‘Part-based R-CNNs for fine-grained category detection’, in European conference on computer vision, 2014, pp. 834–849.
[2] T.-Y. Lin, A. RoyChowdhury, and S. Maji, ‘Bilinear cnn models for fine-grained visual recognition’, in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1449–1457.
[3] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, ‘Compact bilinear pooling’, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 317–326.
[4] T. Hu, H. Qi, Q. Huang, and Y. Lu, ‘See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification’, ArXiv Prepr. ArXiv190109891, 2019.
[5] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, ‘Focal loss for dense object detection’, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
[6] C. Szegedy et al., ‘Going deeper with convolutions’, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[7] J. Fu, H. Zheng, and 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.
[8] H. Zheng, J. Fu, T. Mei, and J. Luo, ‘Learning multi-attention convolutional neural network for fine-grained image recognition’, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5209–5217.
[9] S. Kong and C. Fowlkes, ‘Low-rank bilinear pooling for fine-grained classification’, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 365–374.
[10] C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, ‘Hierarchical bilinear pooling for fine-grained visual recognition’, in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 574–589.
[11] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, ‘Learning to navigate for fine-grained classification’, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 420–435.
[12] W. Luo et al., ‘Cross-X learning for fine-grained visual categorization’, in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8242–8251.
[13] Y. Chen, Y. Bai, W. Zhang, and T. Mei, ‘Destruction and construction learning for fine-grained image recognition’, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5157–5166.