Research on fine-grained pattern recognition based on attention pattern-generated model

Jiayin Zhou¹, Kaiping Feng², a, * and Lihong Luo ², b

¹School of Computer Science, Guangdong University of Technology, Guangzhou Guangdong 510006, China
²School of Art and Design, Guangdong University of Technology, Guangzhou Guangdong 510090, China

*, a Corresponding author email: fengkp@gdut.com, b zhjy1994517huigui@163.com

Abstract. The major difficulty in classifying the type of fine-grained patterns is that the visual differences between subordinate types are probably very subtle, which, however, may be extremely big in the same kind. A fine-grained pattern recognition algorithm is put forward based on the attention pattern-generated model in this article. First, extract high dimensional characteristic pattern from the image via CNN. Then apply several separators to outputting category predictive activation volumes on the characteristic mapping. After that, polymerize these activation volumes to obtain attention mapping. As for the fine-grained recognition, we use the generated attention pattern to tailor and zoom the intentional area so as to carry out multi-scale prediction. This article has been assessed on CUB-200-2011, FGVC-Aircraft, Stanford Cars, and the model mentioned in this article is more superior than baseline multi-scale resnet50 by more than 2%. Attention pattern generated from the model is helpful for other models to increase their recognition accuracy as well.

1. Introduction
The difference and difficulty of the fine-grained image analysis task relative to the general/generic images task is that the granularity of the category to which the image belongs is more granular. To cope with this challenge, the first step in most fine-grained recognition models is to locate the object area. Usually the model is trained with some explicit annotation information, such as bounding boxes, attributes, and so on. However, these annotations need to be applied to domain knowledge and take time and effort. In recent years, visual attention mechanisms have been introduced into many image-related tasks. In fine-grained recognition scenarios, it helps the model to focus on discriminating regions, which is a promising approach.
The idea behind this article is as follows: If the classifier can use local regions to correctly predict the object class, it means that the region is different and requires the attention of our predictor. When the classifier is applied at each location, attention is generated in the classifier activation to obtain a reliability distribution for all categories. This probability distribution allows us to obtain rich information about local regions, rather than just indicating the binary output as or not as a region of attention. If we have multiple classifiers, the aggregated predictions will further allow us to correct the mispredictions for each classifier by backpropagation during training. For example, given a local region, classifier A can predict that it is a medium region with a medium probability, and classifier B can predict that it is one of the fine-grained categories with a high probability. The local area is then used as the discriminating area, and classifier A is trained to perform the correct target fine-grained prediction next time. The core idea of this paper is to use the attention map instead of the label to train each classifier.

2. Related research
Fine-grained recognition models typically include target area positioning. According to how the model learns the target area, we divide it into two categories: a model based on supervised positioning and a visual attention model. Since the models in the second category are more relevant to our work, we will briefly introduce the first category and a more detailed introduction to the second category.

2.1. Supervised positioning model
An easy way to find an object area is to use object labeling to train the model. Branson et al. used the detected target key points to align multiple regions with the prototype model and extracted the regional features using a deep convolutional neural network. Lin et al. proposed a valve linkage function that makes it possible to backpropagate all sub-networks through fine-grained identification, part alignment and positioning. Wang et al. mined the geometrically constrained plaque triples and automatically found the discriminant region, which helps to improve the recognition accuracy.

In the training phase, such models usually require the correct label. However, collecting these tags is time consuming and labor intensive, which makes it difficult to extend the model to large, fine-grained data sets.

2.2. Visual attention model
In order to overcome the shortcomings based on the supervised positioning model, the researchers used a bottom-up or top-down visual attention mechanism.

3. Detailed analysis of the method
The flow chart of the attention generation process in this paper is shown in Figure 1. The method first extracts a high-dimensional feature map from the image through CNN. Multiple classifiers are then applied to the feature map to output the category predictive activation volumes, which are then
aggregated to obtain an attention map. For fine-grained recognition tasks, we use the generated attention map to crop and scale the region of interest to perform multi-scale prediction.

3.1. High dimensional feature extraction
Advanced feature extraction using a Full Convolutional Network (FCN). There are two reasons for using FCN: on the one hand, the FCN output convolution feature map is used to generate the attention map; on the other hand, the FCN can efficiently process different sizes of input images, so there is no need to crop the image during the test. We represent the output of the FCN as f(x), where x is the input image. The size of f(x) is c × h × w, where c is the number of channels, and h and w are the spatial height and width of the feature map.

3.2. Intensive local activation
Before describing the core components of the model in detail, first introduce the concepts that will be used in this article.

In this work, the local classifier is a linear classifier that outputs (L+1) dimension activation. L represents the number of fine-grained categories. An extra one is used to indicate the background category. The input to the classifier is the feature vector along the channel dimension in the convolutional feature map. Intuitively, categorization based on these feature vectors is equivalent to categorization based on partial image segments, since the receptive field of each advanced neuron corresponds to a local region on the input image.

Images are a kind of high-dimensional data, so using a single local classifier is not enough to achieve high recognition accuracy. Therefore, we use n local classifiers to obtain n activation quantities A_i, i = (0, 1, 2, ..., n). In order to aggregate these active volumes, we choose to maximize pooling activation along the classifier dimension, namely:

\[ A = \max \{ A_1, A_2, ..., A_n \} \]

The dimension of A is the same as A_i.

3.3. Attention Pattern-Generation Model
Given A, we only need to maximize the pooling A along the channel dimension (not including the last background category), and we can get the attention map M. The specific steps are shown in Figure 2.

![Figure 2. Flow chart for generating a attention map](image)

Inspired by the research on semantic segmentation, this paper considers each slice of aggregate activation a as a attention map corresponding to a specific category. Ideally, an accurate mask can be used for training. On the target mask, the classifier needs to make the correct fine-grained category predictions. At a location other than the target mask, the classifier should be able to mark it as a background. Here, we use the otsu method [6] to obtain a binary target attention map Mb from M, and replace it with the target fine-grained target mask, which plays the same role as the split mask in semantic segmentation. It should be noted that regardless of whether the resulting regions are activated by errors
or correct predictions, they are considered part of the region of interest as long as the activation probabilities are sufficiently high. This paper establishes the hypothesis that a higher probability means that the local features contain rich information. Although this assumption is not always correct, research shows that it does apply to our scenario.

3.4. Fine-grained identification

For fine-grained recognition tasks, we use different scales of image training models. Multi-scale recognition refers to multiple recursive positioning, cropping and magnification recognition of the attention area. For each size, the classifier is trained on two levels. One is the local classifier detailed above, and the other is the target classifier.

3.4.1. Local level loss function

This paper uses the binary attention map $M_b$ as the target mask to construct the loss function:

$$l_{loc} = \frac{1}{WH}((1-w)l_1 + wl_0)$$

Among them,

$$l_0 = \sum_{(i,j) \in M_b(i,j)=0} -\log p_{(i,j)}^b$$

Loss that represents the classification of the background category,

$$l_1 = \sum_{(i,j) \in M_b(i,j)=1} -\log p_{(i,j)}^l$$

Loss that represents a fine-grained classification.

Usually, the background area is much larger than the attention area, so the weight $w$ is used to prevent $l_0$ from becoming the dominant factor. Then, $w = \frac{1}{WH} \sum_{(i,j)} M_b(i,j)$ is the proportion of the attention area of the high-level feature map to the entire spatial area.

3.4.2. Target level loss function

The local classifier uses only local information and ignores target level information. Therefore, we also trained a linear classifier based on the target level features. For a given attention mask $M_b$, we spatially maximize the $M_b(i,j)$ pool and get a vector as the target level feature. The loss function is:

$$l_{obj} = -\log p^t$$

Where $t$ is the label of the image and $p^t$ is the softmax probability of the $t$th category.

Combining local level loss with target level loss, the damage function for a single scale is expressed as follows:

$$l = l_{loc} + l_{obj}$$

This damage function consists only of Softmax loss, which can quickly implement the proposed model and simplify the training process. In multi-scale training, the models are trained in order, and the objective functions at each scale are independently optimized.
4. Experimental verification

4.1. Data set
This paper evaluates the three data sets of data set CUB-200-2011 and Stanford Cars.

4.2. Implementation detail analysis
The main neural network in this paper uses ResNet50, which is pre-trained on the ImageNet classification dataset. The convolution feature maps before the last pooling layer are extracted as high-level feature representations. All local classifiers and target classifiers use linear classifiers. For each local classifier, the \((L+1) \times 1 \times 1\) convolution kernel is used for end-to-end training of the model. The number of local classifiers is 16, and the fine tuning learning rate is 1e-4.

4.3. Implementation detail analysis
From the results we can see that our model is better than the baseline multi-scale resnet50 performance by more than 2%. Since both models are trained on the same multi-scale image, it can be considered that the performance improvement is brought about by the generated attention, which helps to train the local classifier and target level on each scale. Classifier. At the same time, the recognition accuracy of the multi-scale resnet50 model is higher than that of the single-scale resnet50 model. This can be explained by the fact that the multi-scale resnet50 model is trained on multi-scale images generated by the model, which allows the model to capture more recognition features, which also demonstrates the effectiveness of the attention generated. The attention map generated by the model in this paper also helps other models to improve the recognition accuracy.

| CUB-200-2011 | Structured annotation | Accuracy /% |
|--------------|-----------------------|-------------|
| PA-CNN       | Yes                   | 82.0        |
| MG-CNN       | Yes                   | 82.8        |
| Mask-CNN     | Yes                   | 87.5        |
| HSNet        | Yes                   | 87.6        |
| B-CNN        | No                    | 84.2        |
| ST-CNN       | No                    | 84.3        |
| MA-CNN       | No                    | 86.7        |
| ResNet50(Single scale) | - | 83.9 |
| ResNet50(Multiscale) | - | 85.2 |
| Method of this paper | No | 87.9 |

| Stanford Cars | Structured annotation | Accuracy /% |
|---------------|-----------------------|-------------|
| R-CNN         | Yes                   | 88.3        |
| MDTP          | Yes                   | 91.4        |
| PA-CNN        | Yes                   | 92.9        |
| HSNet         | Yes                   | 94.0        |
| B-CNN         | No                    | 91.4        |
| RA-CNN        | No                    | 92.6        |
| MA-CNN        | No                    | 92.9        |
| ResNet50(Single scale) | - | 91.3 |
| ResNet50 (Multiscale) | - | 92.1 |
| Method of this paper | No | 94.2 |

The results show that even though our model uses a relatively simple objective function, the recognition accuracy is higher than other most advanced methods. The recognition results on the three data sets show that the proposed model is an effective fine-grained identification method.
5. Conclusion
This paper proposes a fine-grained identification method based on local classifier to activate the attention area. The model can be trained only by image-level labeling and SoftMax damage. The method is difficult to implement and the recognition effect is good. The experiments in the three benchmark datasets show that the proposed model has superior performance.

References
[1] S. Branson, G. Van Horn, P. Perona, and S. Belongie. Bird species categorization using pose normalized deep convolutional nets. In Proceedings of the British Machine Vision Conference. BMVA Press, 2014. 1, 2, 6.
[2] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In Conference on Computer Vision and Pattern Recognition, pages 3640 – 3649, 2016. 1.
[3] C. E. Connor, H. E. Egeth, and S. Yantis. Visual attention: bottom-up versus top-down. Current biology, 14(19):R850 – R852, 2004. 1.
[4] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 1.
[5] J. Fu, H. Zheng, and T. Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In Conference on Computer Vision and Pattern Recognition, 2017.
[6] N. Otsu. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 9(1):62 – 66, 1979. 4.