Fine-Grained Visual Classification with Aggregated Object Localization and Salient Feature Suppression

Haicheng Qu¹, Siqi Zhao¹*, Wanjun Liu¹
¹College of Software, Liaoning Technical University, Huludao, China
*email: 18648610655@163.com

Abstract. Fine-grained visual classification (FGVC) is desired to classify sub-classes of objects in the same super-class. For the FGVC tasks, it is necessary to find subtle yet discriminative information from local areas. However, traditional FGVC approaches tended to extract strong discriminative features, and overlook some subtle yet useful features. Besides, current methods ignore the influence of background noises on feature extraction. Therefore, aggregated object localization combined with salient feature suppression are proposed, which is a stacked network. First, the feature maps extracted by the coarse network are fed into aggregated object localization to obtain complete foreground object in an image. Secondly, the refined features obtained through zooming in complete foreground object are fed into fine network. Finally, through finer network processing, the feature maps are fed into salient feature suppression module to find more valuable region discriminative features for classification. Experiment results on two datasets show that our proposed method can get superior result compared with state-of-the-art methods.

1. Introduction

Different from general image classification, fine-grained visual classification (FGVC) aims to differentiate subcategories of a super-category (e.g., species of birds). The task is much more difficult than coarse-grained image classification, since visual differences between subordinate classes are often subtle and these differences are deeply hidden discriminative region areas. As a consequence, FGVC tasks have two challenges, a) large intra-class variance and small inter-class variance; b) images have quite complex backgrounds, which are also along with occlusions and scale variation. Hence, FGVC have constantly been a burdensome task.

Early works[1] generally rely on some extra information (e.g., bounding box/part annotations) to obtain accurate distinguishing regions position and learn the sophisticated feature representations. Albeit with decent results reported, these supervised learning are not best choice for FGVC as manual annotations can be labor-intensive and are often error-prone. Subsequently, weakly supervised FGVC[2,3] have been proposed with only using class label.

Some models improve classification performance by encoding higher-order information. For example, B-CNN[4] make full use of the second-order encoding to dig out more feature information, but the outer product result is high dimensional, which may prone to out of memory. Other methods find the discriminative regions by using attention mechanism. RA-CNN[5] is proposed to extend to its discriminative areas by attention network, but this method only learn a single discriminative feature in the training stage.
Although excellent achievements have been made, a few subtle features are rather overlooked. What’s more, the features of the interesting object are easily be shadowed by background noises. We argue that accurate location of complete foreground object image and the sophisticated feature representation correlate in a mutually reinforced way. Based on this assumption, aggregated object localization combined with salient feature suppression is proposed to handle the problems mentioned above of fine-grained image classification. First, complete foreground object features can be obtained through aggregated object localization, which is can effective eliminate background noises; Then, on the premise of obtaining complete foreground object features, salient feature suppression can weaken salient feature to enforce the network pay more attention to subtle features; Finally, three-branch loss function is used for constraint learning process.

Main contributions is as follows:

(1) We propose a stacked network, which can effectively locate the interesting object and learn sophisticated feature representation correlated in a mutually reinforcement way.

(2) An Aggregated Object Localization is designed, which utilizes multi-layer activation map to generate the object mask. The object mask edge precisely locates the complete foreground object in the image.

(3) On the premise of getting a complete foreground object, we weaken the high response features through Salient Feature Suppression. Hence, the model can learn more useful features to enhance its robustness.

2. Approach

Our method’s structure is described in detail. The method’s structure is a stacked network, using ResNet-50[6] as the backbone. The proposed method’s structure is shown in Figure 1.

Fig. 1 Overall structure of our approach

2.1. Aggregated object localization

Our complete object localization approach is inspired by SCDA[7], which generate a binary matrix to measure the distribution of object response values.

Suppose \( F \in \mathbb{R}^{W \times H \times C} \) represent the last convolutional feature map of ResNet-50. Activation map \( A \) can be generated by adding up feature maps \( F \) along channel direction. Then we set a threshold \( \theta \), which is the mean value of \( A \). In general, if the value of position \( \langle i, j \rangle \) is greater than \( \theta \), It is set to 1, otherwise it is set to 0. As representation in Eq(1).

\[
M_{\langle i, j \rangle} = \begin{cases} 
1, & A_{\langle i, j \rangle} > \theta \\
0, & \text{otherwise}
\end{cases}
\]  

(1)
The distribution of object can be determined by the intensity of each pixel in the activation maps. So, activation maps of different layers can reflect the different distribution of images. In our method, by fusing the multi-layer activation maps, an aggregated mask is obtained to enhance the object localization performance. Suppose layer\(_{4,1}\), layer\(_{4,2}\) and layer\(_{4,3}\) are the three convolution layers of the Block-4 in ResNet-50, respectively, we generate binary mask map \(M_{4,1}\), \(M_{4,2}\) and \(M_{4,3}\) according to Eq(1). The reason why we don’t use the layer\(_{1–3}\) convolutional feature is that layer\(_{1–3}\) texture information is not class-specific, which has little effect on object localization. Then sum operation is performed on the \(M_{4,1}\), \(M_{4,2}\) and \(M_{4,3}\) to obtain an aggregated mask \(M\). The larger the aggregated mask value, the more likely it is to be an object. Hence, the aggregated mask value equals to 1 only when the sum of all the corresponding multi-layer mask values is 3, otherwise equals to 0. Finally, the object is usually in the largest connected component of \(M\), so the smallest bounding box that contains the largest connectivity area as the result of our object location.

2.2. Salient feature suppression

We obtain complete foreground object by multi-layer aggregated mask features, which eliminate irrelevant background information. To achieve optimal feature learning based on the complete object, we design salient feature suppression network. The salient feature suppression module is shown in Figure 2.

![Figure 2: Structure of SFS](image)

We obtain the self-attention map \(M_{att} \in \mathbb{R}^{H \times W}\) by squeezing feature map \(F \in \mathbb{R}^{H \times W \times C}\) using channel-wise average pooling, and zoom in the same size as original image. Then sigmoid operation is performed on the upsampling feature map. As representation in Eq(2)(3).

\[
M_{att} = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} F(i,j)
\]  
\[
A = \text{sigmoid}(\text{up} \_\text{sample}(M_{att}))
\]

We generate the attention feature map \(A\) where the high value regions are discriminative features, while the low values are subtle areas on which the model pay less attention to focus. Therefore, we strengthen the lower-response values in the attention feature map \(A\), where the network can attach importance to the features of subtle regions rather than those discriminative regions. By doing this, all subtle yet discriminative features in the aggregated foreground object can be full of learned to the model.
Specifically, we set a threshold $\beta$, when the pixel value of the feature map $A$ is greater than the $\beta$, we regard it as discriminative regions, otherwise it is less discriminative regions. In addition, A weakening high response factor $\alpha \in (0.5, 0.9)$ is introduced. That is, setting each pixel value $A(i, j)$ is $\alpha$ if it is greater than $\beta$, and 1 if it is smaller. Instead of directly erasing the discriminative features, we replace it with a small value $\alpha$. By doing so, the model can mime more subtle yet discriminative features in the complete object image. As representation in Eq(4).

$$S_{(i,j)} = \begin{cases} 
\alpha, & A_{(i,j)} > \beta \\
1, & \text{otherwise}
\end{cases} \quad (4)$$

Finally, the aggregated foreground object $I$ is multiplied bitwise by the $S_{(i,j)}$ to obtain a new input $I'$ fed into finer network. The high-response values are replaced by a smaller value, while the others values are strengthened.

### 2.3. Classifier
We construct three-branch network, each branch has a classifier that constrains its learning convergence process. We use cross-entropy loss for the whole process to train the network. The total loss is the sum of the three-branch loss.

### 3. Experimental result and discussion

#### 3.1. Experimental Details
We conducted experiments with PyTorch on GTX 1080Ti GPU. During training stage, images were resized to $448 \times 448$, and the input of the three-branch network were enlarged to the same size. We used SGD as the optimizer and BN for regularization, which momentum is set 0.9 and the weight decay is set 0.001. The epoch was set 120, batch size was 6. The learning rate was 0.001 and in 60 epochs, the learning rate is multiplied by 0.1 [10]. After experimental analysis, we set $\alpha = 0.5$, $\beta = 0.5$ for CUB-200-2011 dataset, $\alpha = 0.6$, $\beta = 0.7$ for Stanford Cars.

#### 3.2. Performance Comparison
To demonstrate the efficiency of method, we conduct experiments on two dataset CUB-200-2011[8] and Stanford Cars [9] datasets, respectively. The comparison results are showed in Table 1.

| Method       | Datasets       | CUB-200-2011 | Stanford Cars |
|--------------|----------------|-------------|--------------|
| NTS [11]     |                | 87.5        | 93.9         |
| Cross-X [12] |                | 87.7        | 94.6         |
| DCL [13]     |                | 87.8        | 94.5         |
| SMA-Net [14] |                | 87.7        | 94.4         |
| Ours         |                | 87.8        | 94.9         |

#### 3.3. Ablation Studies
**The effect of aggregated object localization:** As shown in Table 2, when introducing aggregated object localization into ResNet-50, the model gets 85.9% accuracy on the CUB-200-2011 dataset which is 1.8% higher than the ResNet-50 baseline. The result demonstrates our component is effective.

**The effect of salient feature suppression:** As shown in Table 2, when introducing salient feature suppression, our approach obtains 87.8% accuracy, an improvement of 3.7% compared with baseline. The result demonstrates our component is effective as well.
3.4. Object Localization Performance
To further demonstrate the effectiveness of AOL, we compare it with the SCDA. Compared with SCDA, AOL gets 85.9% accuracy, surpass the SCDA 76.8% accuracy, as shown in Table 3. Besides, to further improve our method effective, we visualize the bounding box, where the red rectangles represent the ground-truth, the green rectangles are predicted through methods. As shown in Figure 3. The top is SCDA, the bottom is our aggregated object localization method. Our method can precisely locate the interesting object.

Tab. 3 Object localization on CUB-200-2011

| Method      | Base      | Accuracy |
|-------------|-----------|----------|
| SCDA        | ResNet-50 | 76.8     |
| AOL(Ours)   | ResNet-50 | 85.9     |

Fig. 3 Visualization of object localization on the CUB-200-2011. The top is SCDA, the bottom is our aggregated object localization.

4. Conclusions
In this paper, Aggregate object localization combined with salient feature suppression is proposed, which works to extract pure features of complete foreground object and strengthen subtle feature learning correlated in a mutually reinforcement way. Extensive experiments confirmed that the proposed method improves the accuracy on different FGVC datasets.

Acknowledgments
This work was funded by the National Natural Science Foundation of China (grant number: 42071351) and the Science Foundation of Liaoning Educational Committee (grant number: LJ2020QNL013).
References

[1] Lin D, Shen X Y, Lu C and Jia J 2015 Int. Conf. on Computer Vision and Pattern Recognition (Boston) Deep LAC: Deep localization, alignment and classification for fine-grained recognition vol 1 pp. 74-1666

[2] Liu X Z, Zhang L F, Li T, Wang D J and Wang Z J 2020 Dual attention guided multi-scale CNN for fine-grained image classification Information Sciences 573 pp 37-45

[3] Sun M, Yuan Y C, Zhou F and Ding E 2018 Multi-attention multi-class constraint for fine-grained image recognition Lecture Notes in Computer Science 11220 pp 50-834

[4] Lin T Y, RoyChowdhury A and Maji S 2015 Int. Conf. on Computer Vision (Santiago, Chile) Bilinear CNN Models for fine-grained visual recognition vol 1 pp 57-1449

[5] Fu J L, Zheng H L and Mei T 2017 Int. Conf. on Computer Vision and Pattern Recognition (Honolulu) Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition vol 1 pp 84-4476

[6] He K M, Zhang X Y, Ren S Q and Sun J 2016 Proc. Int. Conf. on Computer Vision and Pattern Recognition (Las Vegas) Deep residual learning for image recognition vol 1 pp 8-770

[7] Wei X S, Luo J H, Wu J X and Zhou Z H 2017 Selective convolutional descriptor aggregation for fine-grained image retrieval IEEE Trans. Image Process 26(6) 81-2868

[8] Wah C, Branson S, Welinder P, Perona P and Belongie S 2011 The caltech-ucsd birds-200-2011 dataset California Institute of Technology

[9] Jonathan K, Michael S, Jia D and Li F F 2013 Proc. Int. Conf. on Computer Vision Workshops (Sydney) 3d object representations for fine-grained categorization vol 1 pp 554–561

[10] Zhang F, Li M, Zhai G S and Liu Y Z 2021 Multi-branch and Multi-scale Attention Learning for Fine-Grained Visual Categorization Lecture Notes in Computer Science 12572 pp 47-136

[11] Yang Z, Luo T G, Wang D, Hu Z Q, Gao J and Wang L W 2018 Proc. Int. Conf. on European Conference on Computer Vision (Munich) Learning to navigate for fine-grained classification 11218 vol 1 pp 54-438

[12] Luo W, Yang X T, Mo X J, Lu Y H, Davis L, Li J, Yang J and Lim S 2019 Int. Conf. on Computer Vision and Pattern Recognition (Seoul) Cross-X learning for fine-grained visual categorization vol 1 pp 51-8242

[13] Chen Y, Bai Y L, Zhang W and Mei T 2019 Int. Conf. on Computer Vision and Pattern Recognition (Long Beach) Destruction and construction learning for fine-grained image recognition vol 1 pp 66-5157

[14] Liu C, Huang L, Wei Z Q and Zhang W F 2021 Subtler mixed attention network on fine-grained image classification Appl Intell 51 pp 16-7903