Part Attentioned Fine-Grained Categorization Using Context Sensitive Flow

Zhiqin Zhang¹, Bo Zhang²,*
School of Computer Science, Wuhan Donghu University, Wuhan 430212, China
*Corresponding author's e-mail: bob.cheung@ovspark.com

Abstract. It is difficult to recognize fine-grained objects (e.g., dogs or cars) because of the challenges of difficult region localization and fine-grained feature learning. Current approaches neglect the fact that local context feature and global feature are mutually correlated and thus it is possible to use the related information from each other. In this paper, we propose a novel multiple scales attention fusing convolutional neural network which can learn region-based feature and discriminative object attention and at multiple scales in a mutually improved way. The learning is composed of several scales which consist of a classification sub-network and an attention fusing module. A $L_{\text{refine}}$ loss was proposed to refine the second sub-network category performance. We do some overall experiments and show that our proposed method achieves the best performance in two fine-grained tasks, with relative mean accuracy gains of 2.2%, on CUB Birds, and Stanford Cars.

1. Introduction
Fine-grained categories technology based on deep learning techniques (e.g., classifying bird species, dogs and flowers [1-3]) has attracted a lot of attention. Recently, a lot of research was made to improve accuracy on an increasing number of object classes [4]. In contrast to general object recognition, it is very challenging to classify the fine-grained categories (e.g., different grebe species) and it can only be recognized by domain experts. Different from traditional recognition, localizing and classification are the basic function of fine-grained image recognition which can different the very subtle visual differences within each sub-categories, and thus can benefit a lot of applications, such as complex image captioning, expert-level image recognition [5], and so on.

Recognizing fine-grained objects can be divided into two steps: locating the discriminative regions and extracting fine-grained feature from the located regions. Previous research has made state of the art progresses by using mixed part-based object features, which include two steps: 1) identifying different object part regions by differing convolutional responses from feature map in an unsupervised method or by using supervised labels, and 2) extracting significant feature representations from each part of object and using fused compact vectors for classification. Although a lot of perfect results have been reported, it is possible to achieve further improvement by eliminating the following limitations. First, it is not optimal for existing unsupervised methods to learn human-defined regions or the regions for machine classification. Second, the subtle visual differences existed in local regions from similar fine-grained categories can be difficult to learn. Third, context features were not used to learn discriminate feature. We found that adjacent parts region feature fusing can be used to improve feature discrimination, global features and local features learning are mutually correlated and thus can enhance each other.
In this paper, we proposed a new research sub direction by introducing and evaluating a new deep part features fusing framework and adding a context sensitive loss to supervised improve different context features relation. Our main contribution is a formulation of the new deep feature fused architecture, called FFCSnet, which consist of two main components: a new part feature fused attention net and a context sensitive loss. We organized the rest of the paper as follows. Section 2 the related state of the art work was reviewed. Section 3 we proposed our method. Section 4 provides the experiments for evaluation and analysis, followed by the conclusion in Section 5.

2. Related Work
Researching on fine-grained object recognition can be divided into two dimensions: learning discriminative feature representations and locating the different prominent parts.

2.1 Discriminative Feature Learning
Learning the powerful convolutional deep features was used by most of the methods, which have shown great improvement than traditional hand-crafted features on both general and fine-grained classification [6]. In [7], a deep residual network consisting of 152 layers CNN was proposed by concatenating residual feature which can learn stronger feature representation, and the error rate was reduced to 3.75% on ImageNet test set. Due to the success of deep learning, a lot of methods was developed for fine-grained object classification [7]. The main thinking of these approaches was to distinguish subtle differences among similar subcategories by finding the different context structure of different object parts existing in different fine-grained classes.

Recently, a bilinear structure was proposed to compute the mixed pair features by using two independent CNN branches to capture context semantic information, learning discriminative features is crucial to recognize fine-grained different objects, which has achieved the perfect results in fine-grained bird classification [8].

2.2 Difficult Part Localization
Previous works mainly focus on using the extra supervised information of bounding box and part annotations to discriminate different regions in fine-grained recognition [9]. However, it is impossible and not practical to get the exactly correct manual annotations large-scale real problems. Recently, a lot of advanced researchers aimed at a more general scenario and proposed to use unsupervised information to mine region attention. Spatial transformer [10] takes one step further and proposes a dynamic mechanism that can actively learn the transformation in an image for improved classification. Whereas, it is still difficult for existing models to exactly localize subtle regions due to their small sizes.

Our approach is related to existing work aimed at deeply fusing different object parts using no supervision labels and context features. For example, informative object parts are learned without needing part annotations by augmenting an existing CNN architecture with a differentiable spatial transformation module. In contrast to these methods, our FFCSnet consist of two main components: a new part feature fused attention net and a context sensitive loss, discriminated features can be achieved by attention fusing and context sensitive loss.

3. Proposed method
In this section, we will introduce the proposed part based attention convolutional neural network for fine-grained image recognition. We consider the network with one whole scale and one part scale as an example in Figure 1, and finer scales can be stacked in a similar way. The inputs are full-size images in whole branch net to part discriminative regions in part branch net, where the part branch takes the input as the attended regions from the whole branch, respectively. First, images at different scales are fed into convolutional layers to extract region-based feature representation. Second, the CNN network proceed to predict both a probability score by fully-connected and softmax layers and a part region selection attention by an attention proposal network. The proposed network is optimized to
convergence by learning a softmax classification loss at each scale and a refined loss across neighboring scales.

3.1 Part Based Attention fusing CNN

![Fig.1](#)

The CNN consists of one whole and one part scale branch

It is important to mutually improve the learning for both localization and classification which can take advantages of the deeply trained networks. In this paper, we propose a self attention study network (SASN) and it is free to compute the region attention, the SASN can be trained end-to-end without supervision. We implemented an element-wise multiplication for the cropping operation between the input image at the whole scale and an attention mask, which can be formed as:

\[
X_{\text{att}} = X^{\text{in}} \odot M
\]

where \( \odot \) represents element-wise multiplication, \( X_{\text{att}} \) denotes the cropped region in figure 1, \( M \) is a mask function and can be learned from end to end, with the specific form as:

\[
M(.) = h(f(x))
\]

Where \( h(.) \) is a softmax function, \( x \) is the feature maps of the whole scale.

Multi-task network structure of SASN can improve the features learning and reduce overfitting. The whole scale feature maps were added element-wised to the part feature maps, the operation can mix low level features with high level features.

3.2 Context Sensitive Module

Inspired by the Inception-ResNet, it is important to take advantage of the gain of deeper and wider network features. We proposed the context-sensitive layer (CSL), see Fig. 1, in which we concatenate the FC layers of context module in the whole branch and the part branch. By using CSM, all the benefits of the multi-task approach can be utilized while remaining rich contextual features from the whole branch. After that, a \( L_{\text{refine}} \) loss was proposed to make \( L_{\text{cls}}^{2} - L_{\text{cls}}^{1} \) bigger than a margin, which can be formed as:

\[
L_{\text{refine}} = -(P_{\text{cls}}^{2} - P_{\text{cls}}^{1} - \gamma)
\]

where \( P_{\text{cls}}^{2} \) and \( P_{\text{cls}}^{1} \) represent the whole branch and the part branch classification probability, respectively. \( \gamma \) is a margin, we set it 0.15 in all our experiments. The total loss function can be formed as:

\[
L = L_{\text{cls}}^{2} + L_{\text{cls}}^{1} + \beta * L_{\text{refine}}
\]

where \( \beta = 0.5 \).

4. Experiments

4.1 Datasets

We conduct experiments on two challenging fine-grained image recognition datasets, including Caltech- UCSD Birds (CUB-200-2011), and Stanford Cars. The detailed statistics with category numbers and data splits are summarized in Table 1
Table 1. The statistics of fine-grained datasets used in this paper.

| Datasets     | Category | Training | Testing |
|--------------|----------|----------|---------|
| CUB-200-2011 | 200      | 5994     | 5794    |
| Stanford Cars| 196      | 8144     | 8041    |

4.2 Baselines

The baseline approaches were divided into two categories: using human-defined bounding box (bbox) of supervision or part annotation of weak supervision. In the following, the first two methods use human supervision, and the last one is based on part learning methods of weak supervision. Since state-of-the-art results were achieved in both categories, it is reasonable to compare with them. We lists all the baselines as follows:

1) LAC-AI: a pose-aligned part-based classification method was proposed for alignment, deep localization, and recognition.
2) PDFR: consisted on the three main ideas including learning part detectors, picking deep filter responses and finding distinctive filters.
3) PN-CNN: proposed a pose normalized CNN which can estimating the object’s pose and compute local features.

4.3 Experiments on CUB-200-2011 and Stanford Cars

During infer time, we only used the part branch category results, which was showed in the table 2. For fair comparison, we used ResNet52 as the baseline network for all the methods, the results showed that the proposed method achieved the best accuracy.

Table 2. Comparison results on CUB-200-2011 and stanford cars dataset.

| Approach | Accuracy |
|----------|----------|
| LAC-AI   | 80.3     |
| PDFR     | 82.6     |
| PN-CNN   | 85.4     |
| OURS     | 87.6     |

5. Conclusion

In this paper, we propose a part based attention fusing convolutional neural network for fine-grained recognition, which learns discriminative region attention and concatenate region-based feature representations at different scale and context regions. Extensive experiments showed that the advanced performance can be achieved on fine-grained recognition tasks such as birds, dogs. In the future, we will continue to research on one main directions, how to fuse local visual cues and global image multiple scale structure and by context attention, to keep promoting the performance.

Acknowledgement

This paper was co-supported by the Foundation for Young Scholars in Wuhan Donghu University under grant 2018dhzk006. Thanks for the support of Wuhan Donghu University.

References

[1] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian. Picking deep filter responses for fine-grained image recognition. In CVPR, pages 1134–1142, 2016.
[2] S. E. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. In CVPR, 2016.
[3] J. Krause, H. Jin, J. Yang, and F.-F. Li. Fine-grained recognition without part annotations. In CVPR, pages 5546–5555, 2015.
[4] M. Simon and E. Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In ICCV, pages 1143–1151, 2015.
[5] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and F.-F. Li. The
unreasonable effectiveness of noisy data for fine-grained recognition. In ECCV, pages 301–316, 2016.

[6] J. Wang, J. Fu, T. Mei, and Y. Xu. Beyond object recognition: Visual sentiment analysis with deep coupled adjective and noun neural networks. In IJCAI, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.

[8] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[9] S. Huang, Z. Xu, D. Tao, and Y. Zhang. Part-stacked CNN for fine-grained visual categorization. In CVPR, pages 1173–1182, 2016.

[10] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In NIPS, pages 2017–2025, 2015.