Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Weakly Supervised Learning of Part Selection Model with Spatial Constraints for Fine-Grained Image Classification

Xiangteng He, Yuxin Peng*
Institute of Computer Science and Technology, Peking University
Beijing 100871, China
pengyuxin@pku.edu.cn

Abstract

Fine-grained image classification is challenging due to the *large* intra-class variance and *small* inter-class variance, aiming at recognizing hundreds of sub-categories belonging to the same basic-level category. Since two different sub-categories is distinguished only by the *subtle* differences in some specific parts, semantic part localization is crucial for fine-grained image classification. Most previous works improve the accuracy by looking for the semantic parts, but rely heavily upon the use of the object or part annotations of images whose labeling are costly. Recently, some researchers begin to focus on recognizing sub-categories via weakly supervised part detection instead of using the expensive annotations. However, these works ignore the spatial relationship between the object and its parts as well as the interaction of the parts, both of them are helpful to promote part selection. Therefore, this paper proposes a weakly supervised part selection method with spatial constraints for fine-grained image classification, which is free of using any bounding box or part annotations. We first learn a whole-object detector automatically to localize the object through jointly using saliency extraction and co-segmentation. Then two spatial constraints are proposed to select the distinguished parts. The first spatial constraint, called box constraint, defines the relationship between the object and its parts, and aims to ensure that the selected parts are definitely located in the object region, and have the largest overlap with the object region. The second spatial constraint, called parts constraint, defines the relationship of the object's parts, is to reduce the parts' overlap with each other to avoid the information redundancy and ensure the selected parts are the most distinguishing parts from other categories. Combining two spatial constraints promotes parts selection significantly as well as achieves a notable improvement on fine-grained image classification. Experimental results on CUB-200-2011 dataset demonstrate the superiority of our method even compared with those methods using expensive annotations.

Introduction

Fine-grained image classification is an extremely challenging task, which aims to distinguish the objects in subordinate classes, such as bird types (Wah et al. 2011), dog species (Khosla et al. 2011), plant breeds (Angelova and Zhu 2013) and aircraft models (Maji et al. 2013) etc. An inexperienced person can easily recognize basic-level categories such as birds, horses and dogs, since they vary a lot in appearance. He may know several kinds of birds, but it would be very difficult to recognize 200 or even more sub-categories. For example, it is extremely hard for an inexperienced person to distinguish between Herring Gull and Slaty-backed Gull whose appearance are very similar, as both of them have the gray back and pink legs. These subordinate classes share the same global appearance, and are often distinguished by the subtle differences in their parts (e.g. Herring Gull and Slaty-backed Gull are distinguished by the color of the back, the latter's is deeper). Therefore, the object and its salient parts are crucial for fine-grained image classification.

Since the discriminative features are mainly localized on the object and its parts, most existing works follow the pipeline: localizing the object or its parts firstly, and then extracting discriminative features for fine-grained image classification. As the fine-grained image classification datasets (e.g. CUB-200-2011 (Wah et al. 2011)) mostly have the detailed annotations like bounding box and part locations, early works directly use the detailed annotations at both training and testing stage. The works of (Chai, Lempitsky, and Zisserman 2013; Yang et al. 2012) use the provided bounding box to learn part detectors in a unsupervised or latent manner. Several methods even use the part annotations (Berg and Belhumeur 2013; Xie et al. 2013). Since the annotations of the testing image are not available in the practical applications, some works use the object or part annotations only at training stage and no knowledge of annotations at testing stage. Bounding box and Part annotations are directly used in training phase to learn a strongly supervised deformable part-based model (Zhang et al. 2013; Azizpour and Laptev 2012) or directly used to fine-tune the pre-trained Convolutional Neural Net (CNN) (Branson et al. 2014). Further more, Krause et al. (Krause et al. 2015) only uses bounding box at training stage to learn the part detectors, then localize the parts automatically in the testing stage. Recently, there are some promising works attempting to learn the part detectors under the weakly supervised condition, i.e. the bounding box and part annotations are not used at training or testing stage. These works make it possible to put the fine-grained image classification into practical applications. Neural Activation Constellations Part Model

*Corresponding author.

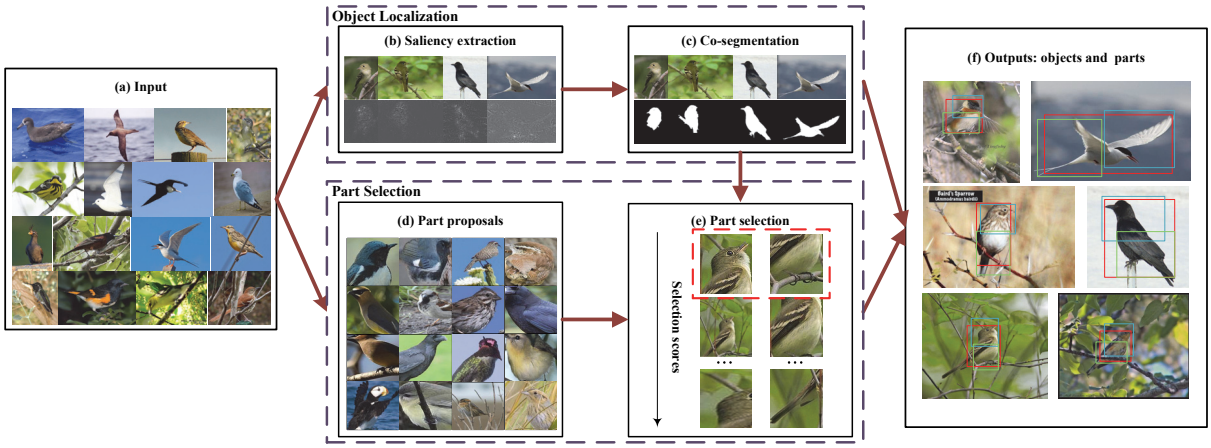

Figure 1: An overview of our proposed method to localize the object and its parts. Our approach consists of two stages. The first stage aims at localizing the object via (b) saliency extraction and (c) co-segmentation for each input image. The second stage is to select the semantic parts for fine-grained image classification. Based on the object location information and the saliency map, parts are selected driven by two spatial constraints: box constraint and parts constraint. The red rectangles in (f) denote the bounding boxes produced by the proposed method automatically, and the green and blue rectangles denote the selected parts.

(NAC) (Simon and Rodner 2015) proposes to localize parts with constellation model, Two Level Attention Model (TL Atten) (Xiao et al. 2015) applies two attention models to select relevant proposals to the object and the discriminative parts, and Picking Deep Filter (PD) (Zhang et al. 2016b) incorporates deep convolutional filters for both part detection and description. However, they all ignore the object localization and the spatial relationship between the object and its parts as well as the interaction of the parts. The object localization can remove the influence of the background noise to obtain the meaningful global feature, and the two spatial relationships is significant for selecting the discriminative and useful parts, both of them can promote significant effects on fine-grained image classification.

Therefore, this paper proposes a framework for fine-grained image classification without using bounding box or part annotations. The main contributions of this paper can be concluded as follows:

Object Localization We learn a whole-object detector without using bounding box, but only class label. First, a saliency map (Simonyan, Vedaldi, and Zisserman 2013) is extracted for each image with the fine-tuned CNN. It can provide the saliency information over the input image, which helps to localize the object preliminarily. But the generated bounding box is typically much bigger than the ground truth bounding box. Then, we leverage co-segmentation (Krause et al. 2015) to make the coarse granularity bounding box become more accurate. To the best of our knowledge, this paper is the first work to localize the object for fine-grained classification only using whole image label, without the expensive labor annotations like bounding box, and the effect of the class label is weakened in our method.

Part Selection In order to find the distinguishing parts, we propose two spatial constraints to guide the part selection process.

- **Box constraint.** One intuitive idea to select the distinguishing parts is to consider the parts inside the object region. However, the previous weakly supervised works ignore the intuitive idea so that the selected parts may be outside the object region, which is a side-effect for classification. Therefore, we apply the box constraint to enforce that the selected parts are definitely located in the region of the object. Additionally, in order to obtain more useful information, we guide the selection process to favor the parts group which has the largest overlap with the object region. With box constraint, the selected parts have these characteristics: definitely located in the object and highly representative.
- **Parts constraint.** Previous works only concern about the parts' responses, but ignore the parts constraint. So that the selected parts of the same object may contain the similar information, which is redundant for fine-grained image classification. From another point of view, it makes some meaningful part left out. Therefore, we conduct the parts constraint on the parts selection process to reduce the overlap with each other and ensure the selected parts are the most distinguishing parts from other categories.

Weakly Supervised Learning of Part Selection Model with Spatial Constraints

In this section, the proposed method is described. It is important to note that only class labels of training images are used in our method. Fig. 1 shows the overview of our proposed method to localize the object and its parts, which consists of two stages. The first stage aims at localizing the object via saliency extraction and co-segmentation. The second stage is to select the distinguishing parts for improving the performance. Based on the object location information and the

saliency map, parts are selected by two spatial constraints: box constraint and parts constraint.

Table 1: Precision of the bounding box produced by the object localization method on CUB-200-2011 dataset. The precision is computed with the ground truth bounding box, defined by the proportion of Intersection-over-Union (IoU) overlap with ground truth bounding box at least 0.5, 0.6 and 0.7 respectively.

Method \ IoU	0.5	0.6	0.7
saliency extraction	64.20%	41.08%	19.31%
+co-segmentation	65.52%	46.16%	28.36%

Object Localization

Existing weakly supervised works focus on the part localization or selection without using the object or part annotations. However, they ignore the object localization which is not only crucial for improving the classification performance but also helpful to select the distinguishing parts. In this section, we propose a new object localization method through jointly using saliency extraction and co-segmentation without using bounding box. Our proposed method consists of two stages: saliency extraction and co-segmentation. The first stage is to localize the object preliminarily with the saliency information produced by the CNN model. The object information obtained from the first stage is not accurate enough, so the second stage is to make the object information more accurate for fine-grained image classification by co-segmentation. Fig. 2 presents the results of each stage on CUB-200-2011. The sub figure (a) shows the original images, (b) shows the saliency maps of the original images and (c) shows the segmentation results of co-segmentation method. In (d) the blue rectangles represent the ground truth bounding boxes of the objects, the red rectangles represent the bounding boxes produced by saliency extraction and the green rectangles represent the bounding boxes produced by co-segmentation on the basis of the red rectangles. We can see that the bounding boxes become more accurate through the co-segmentation process.

Saliency Extraction The previous work (Simonyan, Vedaldi, and Zisserman 2013) generates a saliency map for an input image with a classification CNN model and the class label of the image. However, in testing stage, the class label information is unknown. Fortunately, in fine-grained image classification, all the classes belong to a basic-level category, and they have the similar appearance which allows to get the saliency map without using the class label.

Given an image I (width: n, height: m, channel: c), a class label s (in our experiments, $s = 50$, which affects the results a little) and a classification CNN model which is fine-tuned from the pre-trained CNN model on ImageNet (Deng et al. 2009), the saliency map $M \in R^{m \times n}$ is computed as follows. First, the derivative ω of the class score function is computed by the back-propagation algorithm. Then, to derive a single saliency value for each pixel (i, j) of image I, we take the maximum magnitude of ω across all color channels: $M_{ij} = max_c|\omega_{h(i,j,c)}|$, where $h(i, j, c)$ is the index of the element of ω corresponding to the image pixel in the i-th row, j-th column and c-th channel.

(a) (b) (c) (d)

Figure 2: Illustration of the results our object localization method. (a) shows the original images, (b) shows the saliency maps of the original images and (c) shows the segmentation results of co-segmentation method. In (d) the blue rectangles represent the ground truth bounding boxes of the objects, the red rectangles represent the bounding boxes produced by saliency extraction and the green rectangles represent the bounding boxes produced by co-segmentation on the basis of the red rectangles.

When the saliency map is produced, we conduct the binarization and connected area operations to get the bounding box of the object. But through this process, the bounding boxes are not accurate enough, as the red rectangles shown in Fig. 2 (d), they are much bigger than the ground truth bounding boxes denoted by the blue rectangles.

Co-segmentation In order to get more accurate bounding box, a figure-ground segmentation of each image is established via co-segmentation (as shown in Fig. 2 (c)). Existing co-segmentation methods (Rubinstein et al. 2013; Joulin, Bach, and Ponce 2010) typically assume that the bounding box of the object is not available, while we have gotten the bounding box through the saliency extraction process, which is helpful even though it is not accurate enough. Then we modify the co-segmentation algorithm in (Krause et al. 2015), which has been proved effective and efficient to get the more accurate bounding box. Different from (Krause et al. 2015), we do not use any annotation at both training and testing stage.

Given an image I, let θ_f^I be a foreground model and θ_b^I be a background model, both of which are represented as Gaussian mixture models. y_p^I denotes the pixel p of an image I either foreground or background, its corresponding RGB value is v_p^I, the set of segmentation assignments across all

images is Y, and p_f is a pixelwise foreground prior. The objective is:

$$\max_{Y,\theta} \sum_I (\sum_p E(y_p^I, \theta^I; p_f^I) + \sum_{p,q} E(y_p^I, y_q^I)) \quad (1)$$

where

$$E(y_p^I, \theta^I; p_f^I) = (1 - y_p^I) log(p(v_p^I; \theta_b^I)) + \frac{y_p^I}{2} log(p(v_p^I; \theta_f^I)) + E(y_p^I; p_f), \quad (2)$$

$$E(y_p^I, p_f) = \begin{cases} log(p_f) & y_p^I = 1 \\ log(1 - p_f) & y_p^I = 0 \end{cases} \quad (3)$$

$E(y_p^I, y_q^I)$ is to enforce consistency between neighboring pixels p and q with respect to their assigned binary foreground or background values. The optimization process follows (Krause et al. 2015).

Table 1 shows that we can get more accurate bounding box of the object through co-segmentation than only through saliency extraction. With the accurate bounding box, we can get the semantic parts through the proposed part selection method with spatial constraints.

Part Selection

Since the semantic parts are crucial for fine-grained image classification, the previous works (Xiao et al. 2015; Zhang et al. 2016b) always select semantic parts from the region proposals produced by some objectness methods (e.g. selective search (Uijlings et al. 2013)). These works ignore the spatial relationships between the object and its parts as well as the interaction of the parts. In this section, a new part selection method with two spatial constraints is described, which consists of two stages: generating part proposals and spatial constraints. The first stage is to generate part proposals for selecting distinguishing parts. Selective search is applied to extract the part proposals from each image. Then the second stage selects the parts which denote the key parts distinguished from other classes.

Generating Part Proposals The raw candidate part proposals are produced in a bottom-up process, grouping pixels into regions that highlight the likelihood of parts of some objects. In this stage, we apply selective search to produce the candidate part proposals for each image. These proposals contain some key parts for fine-grained image classification, but with high recall and low precision. It is necessary to filter these proposals to get the real distinguishing parts.

Spatial Constraints With the bounding box and part proposals produced in the previous stage, spatial constraints are proposed for part selection. Two spatial relationships are considered: the relationship between object and its parts, called box constraint, and the relationship of parts, called parts constraint.

Given an image I, its saliency map M is produced at saliency extraction stage and its bounding box b is generated at object localization stage, parts selection process is conducted as follows. Let $L = \{l_1, l_2, ..., l_n\}$ denotes the locations of n parts for each image. In order to select the distinguishing parts from the part proposals, we consider the joint of two spatial constraints by solving the following optimization problem:

$$L^* = arg \max_L \Delta(L) \quad (4)$$

where $\Delta(L)$ is defined as a scoring function over two spatial constraints and its formulation is defined as follows:

$$\Delta(L) = \Delta_{box}(L)\Delta_{parts}(L) \quad (5)$$

in which $\Delta_{box}(L)$ denotes the box constraint and $\Delta_{parts}(L)$ denotes the parts constraint, detailed in the following paragraphs.

Box constraint. As we know, the distinguishing parts must be inside the object region. So a intuitive spatial constraint function is defined as:

$$\Delta_{box}(L) = \prod_{i=1}^{n} f_b(l_i) \quad (6)$$

where

$$f_b(l) = \begin{cases} 1 & IoU(l) > 0.7 \\ 0 & otherwise \end{cases} \quad (7)$$

and $IoU(l)$ defines the proportion of Intersection-over-Union (IoU) overlap of the part region and the object bounding box. It is important to note that the object bounding box is obtained automatically in object localization, not the ground truth bounding box. The object and part annotations are not used in any stage of our method.

Parts constraint. Since the spatial relationship of parts is ignored in previous works, the selected parts may have large overlap with each other. This issue makes some selected parts redundant for classification. Therefore, we consider the parts constraint in our method as follows:

$$\Delta_{parts}(L) = log(A_U - A_I - A_O) + \beta log(Mean(M_{A_U})) \quad (8)$$

where A denotes the area of the object or part region. In detail, A_U is the union area of the n parts, A_I is the intersection area of the n parts, A_O is the area outside the object region and

$$Mean(M_{A_U}) = \sum_{i,j} M_{ij} \quad (9)$$

where pixel (i, j) locates in the union area of the parts. And β is the trade-off configure, in our experiments, $\beta = 1$.

Table 2: Performances of different variants of our method on CUB-200-2011. "TSC" refers to the proposed part selection method with two spatial constraints, "BC" refers to box constraint, "PC" refers to parts constraint , and "without-TSC" refers to part selection without any spatial constraints.

Method	Acc. (%)
VGG-ft-TSC	**84.69**
VGG-ft-BC	81.01
VGG-ft-PC	77.06
VGG-ft-without-TSC	75.94
VGG-ft (Baseline)	74.91

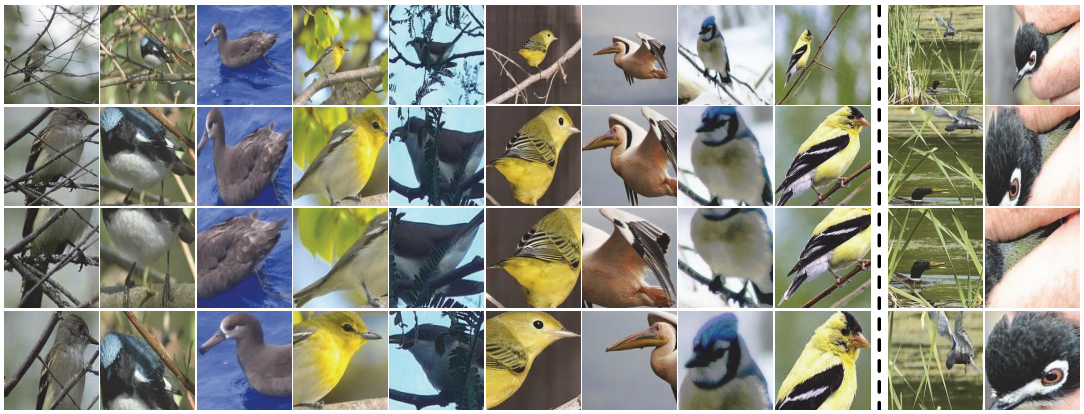

Figure 3: Some results of object localization and parts selection. The first row denotes the original images, the localized objects of the original images are shown in the second row and the selected parts are shown in the third and fourth rows respectively. The last two columns show some failure cases.

Experiments

This section presents the evaluations and analyses of the proposed method on the challenging CUB-200-2011 (Wah et al. 2011) fine-grained image classification benchmark. It contains 11,788 images of 200 types of birds, 5,994 for training and 5,794 for testing. Every image has detailed annotations: 15 part locations, 312 binary attributes and 1 bounding box.

Implementation Details

CNN models In our experiments, we apply the widely used model of VGGNet (Simonyan and Zisserman 2014). It is important to note that the model used in our proposed method can be replaced with any CNN model. The reason of choosing VGGNet is for fair comparison with state-of-the-art methods. The model is pre-trained on ILSVRC 2012, and then fine-tuned on the CUB-200-2011 dataset. In the fine-tuning step, we followed the strategy of (Xiao et al. 2015). First, we apply the selective search to generate patches for each image. Then the pre-trained CNN model on ILSVRC 2012 is used as a filter net for selecting the patches relevant to the object. With the selected patches, we fine-tune the pre-trained model, called ObjectNet.

Classification framework At training stage, we generate the object and its parts of each training image first. Then a part-level CNN model is trained from the ObjectNet with the data of the generated parts, called PartNet. Now we have two models which have the same network structure but are applied for different purposes. At testing stage, we first obtain the object and its parts automatically by our method for each testing image, then use the ObjectNet to get prediction scores of the object and original image, and use the PartNet to get prediction scores of the selected parts. In our experiments, the number of the selected parts is 2, and the final classification is obtained by fusing the above predictions.

Results and Analyses

This part presents the results and detailed analyses of the proposed part selection method with two spatial constraints.

Results of the object localization and its parts selection Fig. 3 shows some results of our method. The first row denotes the original images, the localized objects of the original images are shown in the second row and the selected parts are shown in the third and fourth rows respectively. We can see that the selected parts have the semantic meanings, the third row denotes the body of the object and the third denotes the head. In some cases (e.g. two or more birds are in the same picture, the bird is in heavy occlusion), our method may be out of work. The last two columns show some failure cases.

Detailed analyses of the proposed method We perform detailed analyses by comparing different variants of the proposed method. "VGG-ft" denotes the baseline of our method without any knowledge of the object or its parts. "TSC" refers to the proposed part selection method with two spatial constraints, "BC" refers to box constraint, "PC" refers to parts constraint , and "without-TSC" refers to part selection without any spatial constraints. Each constraint can boost the accuracy respectively, and the combination of two constraints further improve the classification accuracy. From Table 2, we can observe that:

(1) Part selection with two spatial constraints boosts the performance significantly. Comparing with the baseline, TSC brings about a nearly 10% (74.91% $\rightarrow$ 84.69%) improvement. If we select the parts without the two spatial constraints (i.e., only based on the constraint of formula (9)), it only has about 1% improvement over the baseline.

(2) Box constraint plays a more important role than parts constraint. In our experiments, we find that only considering box constraints can achieve a better performance than only considering parts constraint (81.01% vs. 77.06%). And from Table 1 we can see that the performance of object localization is not accurate enough, so if it could be more accurate, the performance of the proposed method would be better.

(3) Combining box constraint and parts constraint can achieve more accurate result than only one constraint is considered (84.69% vs. 81.01% and 77.06%). It proves the

Table 3: Comparisons with state-of-the-art methods on CUB-200-2011.

Method	Train Anno.		Test Anno.		Acc. (%)	Net
	Bbox	Parts	Bbox	Parts		
Our TSC					**84.69**	VGGNet
PD+FC-CNN (Zhang et al. 2016b)					82.60	VGGNet
NAC (Simon and Rodner 2015)					81.01	VGGNet
TL Atten (Xiao et al. 2015)					77.90	VGGNet
VGG-BGLm (Zhou and Lin 2015)					75.90	VGGNet
Spatial Transformer (Jaderberg et al. 2015)					84.10	GoogleNet
Bilinear-CNN (Lin, RoyChowdhury, and Maji 2015)					84.10	VGGNet&VGG-M
PG Alignment (Krause et al. 2015)	√		√		82.80	VGGNet
Triplet-A (64) (Cui et al. 2015)	√		√		80.70	GoogleNet
VGG-BGLm (Zhou and Lin 2015)	√		√		80.40	VGGNet
PN-CNN (Branson et al. 2014)	√	√			75.70	AlexNet
Part RCNN (Zhang et al. 2014)	√	√			73.50	AlexNet
SPDA-CNN (Zhang et al. 2016a)	√	√	√		**85.14**	VGGNet
PN-CNN (Branson et al. 2014)	√	√	√	√	**85.40**	AlexNet
Part RCNN (Zhang et al. 2014)	√	√	√	√	76.37	AlexNet
POOF (Berg and Belhumeur 2013)	√	√	√	√	73.30	
GPP (Xie et al. 2013)	√	√	√	√	66.35	

complementarity of box constraint and parts constraint in fine-grained image classification. The two spatial constraints have the different but complementary focuses: the box constraint focuses on the selected parts must definitely located in the object region and have the largest overlap with the object region; the parts constraint focuses on the selected parts must have little overlap with each other to avoid the information redundancy.

Comparisons with state-of-the-art methods Table 3 shows the comparison results of the proposed method with state-of-the-art methods on CUB-200-2011. Bounding box, part annotations and the CNN model used in the methods are listed for fair comparison. Early works (Berg and Belhumeur 2013; Xie et al. 2013) choose SIFT (Lowe 2004) as features, and the performance is limited. When applying CNN model, our method is the best among methods under the same setting (Zhang et al. 2016b; Simon and Rodner 2015; Xiao et al. 2015; Zhou and Lin 2015), and obtains a 2.09% higher accuracy than the best performing result of PD (Zhang et al. 2016b) (82.60%). PD also has another result of 84.54% which is 0.15% lower than our method. PD has two contributions: part detection and feature encoding, 82.60% is the result of part detection and 84.54% is the result of combining part detection and feature encoding. Since the goal of this paper is to improve the performance of part selection or detection, we do not focus on the influence of feature encoding. Even without considering feature encoding, our proposed method also achieves a better performance than PD. Further more, our method performs better than the methods focusing on the CNN architectures (Jaderberg et al. 2015; Lin, RoyChowdhury, and Maji 2015): the former uses the GoogleNet (Szegedy et al. 2015) with batch normalization (Ioffe and Szegedy 2015) and achieves the accuracy of 82.30% only fine-tuned on the CUB-200-2011 without any other process; the latter uses two different CNN models: VGGNet which is same with our method and VGG-M (Chatfield et al. 2014). Moreover, our method even outperforms methods which use bounding box (Krause et al. 2015) (82.50%) or even part annotations (Zhang et al. 2014) (76.37%), only beaten by (Zhang et al. 2016a) (85.14%) and (Branson et al. 2014) (85.40%) which uses the annotations (e.g. both bounding box and part annotations) at both training and testing stage.

Conclusions

In this paper, we have proposed a weakly supervised part selection method with two spatial constraints which is free of any object or part annotations. The first spatial constraint defines the relationship between the object and its parts, and aims to ensure that the selected parts are definitely located inside the object, and have the largest overlap with the object. The second spatial constraint defines the relationship of parts and reduces the overlap with each other to avoid the information redundancy. We combine two spatial constraints to promote part selection and achieve the best results on CUB-200-2011 dataset under the weakly supervised condition (only class label used). The experiments point out a few future directions. First, since the box constraint plays an import role, we will exploit the method for obtaining the more accurate bounding box to improve the fine-grained image classification performance. Second, inspired by (Zhang et al. 2016b), we will focus on the work of feature encoding for further improvement.

Acknowledgments

This work was supported by National Hi-Tech Research and Development Program of China (863 Program) under Grant 2014AA015102, and National Natural Science Foundation of China under Grants 61371128 and 61532005.

References

Angelova, A., and Zhu, S. 2013. Efficient object detection and segmentation for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 811–818.

Azizpour, H., and Laptev, I. 2012. Object detection using strongly-supervised deformable part models. In *Proceedings of the European Conference on Computer Vision(ECCV)*, 836–849. Springer.

Berg, T., and Belhumeur, P. 2013. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 955–962.

Branson, S.; Van Horn, G.; Belongie, S.; and Perona, P. 2014. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*.

Chai, Y.; Lempitsky, V.; and Zisserman, A. 2013. Symbiotic segmentation and part localization for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 321–328.

Chatfield, K.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*.

Cui, Y.; Zhou, F.; Lin, Y.; and Belongie, S. 2015. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. *arXiv preprint arXiv:1512.05227*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Proceedings of Advances in Neural Information Processing Systems(NIPS)*, 2017–2025.

Joulin, A.; Bach, F.; and Ponce, J. 2010. Discriminative clustering for image co-segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 1943–1950.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Li, F.-F. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2.

Krause, J.; Jin, H.; Yang, J.; and Fei-Fei, L. 2015. Fine-grained recognition without part annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 5546–5555.

Lin, T.-Y.; RoyChowdhury, A.; and Maji, S. 2015. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 1449–1457.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision(IJCV)* 60(2):91–110.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Rubinstein, M.; Joulin, A.; Kopf, J.; and Liu, C. 2013. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 1939–1946.

Simon, M., and Rodner, E. 2015. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 1143–1151.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 1–9.

Uijlings, J. R.; van de Sande, K. E.; Gevers, T.; and Smeulders, A. W. 2013. Selective search for object recognition. *International Journal of Computer Vision(IJCV)* 104(2):154–171.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; and Zhang, Z. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 842–850.

Xie, L.; Tian, Q.; Hong, R.; Yan, S.; and Zhang, B. 2013. Hierarchical part matching for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 1641–1648.

Yang, S.; Bo, L.; Wang, J.; and Shapiro, L. G. 2012. Unsupervised template learning for fine-grained object recognition. In *Proceedings of Advances in Neural Information Processing Systems(NIPS)*, 3122–3130.

Zhang, N.; Farrell, R.; Iandola, F.; and Darrell, T. 2013. Deformable part descriptors for fine-grained recognition and attribute prediction. In *Proceedings of the IEEE International Conference on Computer Vision(ICCV)*, 729–736.

Zhang, N.; Donahue, J.; Girshick, R.; and Darrell, T. 2014. Part-based r-cnns for fine-grained category detection. In *Proceedings of the Proceedings of the International Conference on Machine Learning(ICML)*, 834–849.

Zhang, H.; Xu, T.; Elhoseiny, M.; Huang, X.; Zhang, S.; Elgammal, A.; and Metaxas, D. 2016a. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. 1143–1152.

Zhang, X.; Xiong, H.; Zhou, W.; Lin, W.; and Tian, Q. 2016b. Picking deep filter responses for fine-grained image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 1134–1142.

Zhou, F., and Lin, Y. 2015. Fine-grained image classification by exploring bipartite-graph labels. *arXiv preprint arXiv:1512.02665*.