Fine-grained Discriminative Localization via Saliency-guided Faster R-CNN

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

ABSTRACT

Discriminative localization is essential for fine-grained image classification task, which devotes to recognizing hundreds of subcategories in the same basic-level category. Reflecting on discriminative regions of objects, key differences among different subcategories are subtle and local. Existing methods generally adopt a two-stage learning framework: *The first stage* is to localize the discriminative regions of objects, and *the second* is to encode the discriminative features for training classifiers. However, these methods generally have two limitations: (1) *Separation* of the two-stage learning is *time-consuming*. (2) *Dependence* on object and parts annotations for discriminative localization learning leads to heavily *labor-consuming* labeling. It is highly challenging to address these two important limitations *simultaneously*. Existing methods only focus on one of them. Therefore, this paper proposes *the discriminative localization approach via saliency-guided Faster R-CNN* to address the above two limitations at the same time, and our main novelties and advantages are: (1) *End-to-end network* based on Faster R-CNN is designed to *simultaneously* localize discriminative regions and encode discriminative features, which accelerates classification speed. (2) *Saliency-guided localization learning* is proposed to localize the discriminative region automatically, avoiding labor-consuming labeling. Both are jointly employed to simultaneously accelerate classification speed and eliminate dependence on object and parts annotations. Comparing with the state-of-the-art methods on the widely-used CUB-200-2011 dataset, our approach achieves both the best classification accuracy and efficiency.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Object detection**; **Object recognition**;

KEYWORDS

Discriminative localization, saliency-guided Faster R-CNN, weakly supervised, fine-grained image classification

Figure 1: Examples of CUB-200-2011 dataset [1]. First row shows large variance in the same subcategory, and second row shows small variance among different subcategories.

1 INTRODUCTION

Fine-grained image classification is a highly challenging task due to large variance in the same subcategory and small variance among different subcategories, subcategories, as shown in Figure 1, which is to recognize hundreds of subcategories belonging to the same basic-level category. These subcategories look similar in global appearances, but have subtle differences at discriminative regions of objects, which are crucial for classification. Therefore, most researchers focus on localizing discriminative regions of objects to promote the performance of fine-grained image classification.

Most existing methods [2–8] generally follow a two-stage learning framework: The first learning stage is to localize discriminative regions of objects, and the second is to encode the discriminative features for training classifiers. Girshick et al. [9] propose a simple and scalable detection algorithm, called R-CNN. It generates thousands of region proposals for each image via bottom-up process [10] first. And then extracts features of objects via convolutional neural network (CNN) to train an object detector for each class, which is used to discriminate the probabilities of the region proposals being objects. This framework is widely used in fine-grained classification. Zhang et al. [2] utilize R-CNN with geometric constraints to detect object and its parts first, and then extract features for the object and its parts, finally train a one-versus-all linear SVM for classification. It needs both object and parts annotations. Krause et al. [4] adopt the box constraint of Part-based R-CNN [2] to train part detectors with only object annotation. These methods generally have two limitations: (1) Separation of the two-stage learning is time-consuming. (2) Dependence on object and parts annotations for discriminative localization learning leads to heavily labor-consuming labeling. It is

*Corresponding author.

MM'17, October 23–27, 2017, Mountain View, CA, USA.

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123319

highly challenging to address these two limitations simultaneously. Existing works only focus on one of them.

For addressing the first limitation, researchers focus on the end-to-end network. Zhang et al. [11] propose a Part-Stacked CNN architecture, which first utilizes a fully convolutional network to localize parts of object, and then adopts a two-stream classification network to encode object-level and part-level features simultaneously. It is over two order of magnitude faster than Part-based R-CNN [2], but relies heavily on object and parts annotations that are *labor consuming*.

For addressing the second limitation, researchers focus on how to localize the discriminative regions under the weakly supervised setting, which means neither object nor parts annotations are used in training or testing phase. Xiao et al. [5] propose a two-level attention model: object-level attention is to select relevant region proposals to a certain object, and part-level attention is to localize discriminative parts of object. It is the first work to classify fine-grained images without using object or parts annotations in both training and testing phase, but still achieves promising results [12]. Simon and Rodner [7] propose a constellation model to localize parts of object, leveraging CNN to find the constellations of neural activation patterns. A part model is estimated by selecting part detectors via constellation model. And then the part model is used to extract features for classification. Zhang et al. [6] incorporate deep convolutional filters for both parts selection and description. He and Peng [8] integrate two spatial constraints for improving the performance of parts selection. These methods rarely depend on object or parts annotations, but their classification speeds are *time consuming* due to the separation of localization and encoding.

Different from them, this paper proposes a discriminative localization approach via saliency-guided Faster R-CNN, which is the first attempt based on discriminative localization to simultaneously accelerate classification speed and eliminate dependence on object and parts annotations. Its main novelties and contributions can be summarized as follows:

- ***End-to-end network.*** Most existing discriminative localization based methods [5–7] generally localize discriminative regions first, and then encode discriminative features. The separated processes cause highly *time-consuming* classification. For addressing this important problem, we propose an *end-to-end network* based on Faster R-CNN to *accelerate* the classification speed by simultaneously localizing discriminative regions and encoding discriminative features. Localization exploits discriminative regions with subtle but distinguishing features from other subcategories, and encoding generates representative descriptions. They have synergistic effect with each other, which further improves the classification performance.
- ***Saliency-guided localization learning.*** Existing methods as [13] combine localization and encoding to accelerate classification speed. However, localization learning relies heavily on object or parts annotations, which is *labor consuming*. For addressing this important problem, we propose a *saliency-guided localization learning* approach, which *eliminates the heavy dependence on object and parts annotations* by localizing the discriminative regions automatically. We adopt a neural network with global average pooling (GAP) layer, which is called saliency extraction network (SEN), to extract the saliency information for each image. And then share convolutional weights between SEN and Faster R-CNN to transfer knowledge of discriminative features. This takes the advantages of both SEN and Faster R-CNN to boost the discriminative localization and avoid the labor-consuming labeling simultaneously.

The rest of this paper is organized as follows: Section 2 presents our approach in detail, and Section 3 introduces the experiments as well as the results analyses. Finally Section 4 concludes this paper.

2 SALIENCY-GUIDED FASTER R-CNN

We propose a discriminative localization approach via saliency-guided Faster R-CNN without using object or parts annotations. Saliency-guided Faster R-CNN is an end-to-end network to localize discriminative regions and encode discriminative features simultaneously, which not only achieves a notable classification performance but also accelerates classification speed. It consists of two components: saliency extraction network (SEN) and Faster R-CNN. SEN extracts saliency information of each image for generating the bounding box which is used to guide the discriminative localization learning of Faster R-CNN. They are two localization learning stages, and their jointly learning further achieves better performance. An overview of our approach is shown as Figure 2.

2.1 Weakly supervised Faster R-CNN

We propose a weakly supervised Faster R-CNN to accelerate classification speed and achieve promising results simultaneously without using object or parts annotations. A saliency extraction network (SEN) is proposed to generate bounding box information for Faster R-CNN first. It takes a resized image as an input and outputs a saliency map for generating the bounding box of discriminative region. We follow the work of Zhou et al. [14] to model this process by utilizing global average pooling (GAP) to produce the saliency map. We sum the feature maps of last convolutional layer with weights to generate the saliency map for each image. Figure 3 shows some examples of saliency maps obtained by our approach. Finally we perform binarization operation on the saliency map with a adaptive threshold, which is obtained via OTSU algorithm [15], and take the bounding box that covers the largest connected area as the discriminative region of object. For a given image I, the value of spatial location (x, y) in saliency map for subcategory c is defined as follows:

$$M_c(x, y) = \sum_u w_u^c f_u(x, y) \tag{1}$$

where $M_c(x, y)$ directly indicates the importance of activation at spatial location (x, y) leading to the classification of an image to subcategory c, $f_u(x, y)$ denotes the activation of neuron u in the last convolutional layer at spatial location (x, y), and w_u^c denotes the weight that corresponding to subcategory c for neuron u. Instead of using the image-level subcategory labels, we use the predicted result as the subcategory c.

Faster R-CNN [16] is proposed to accelerate the process of detection as well as achieve promising detection performance. However, the training phase needs ground truth bounding boxes of objects for supervised learning, which is heavily labor consuming. In this paper, we propose weakly supervised Faster R-CNN to localize

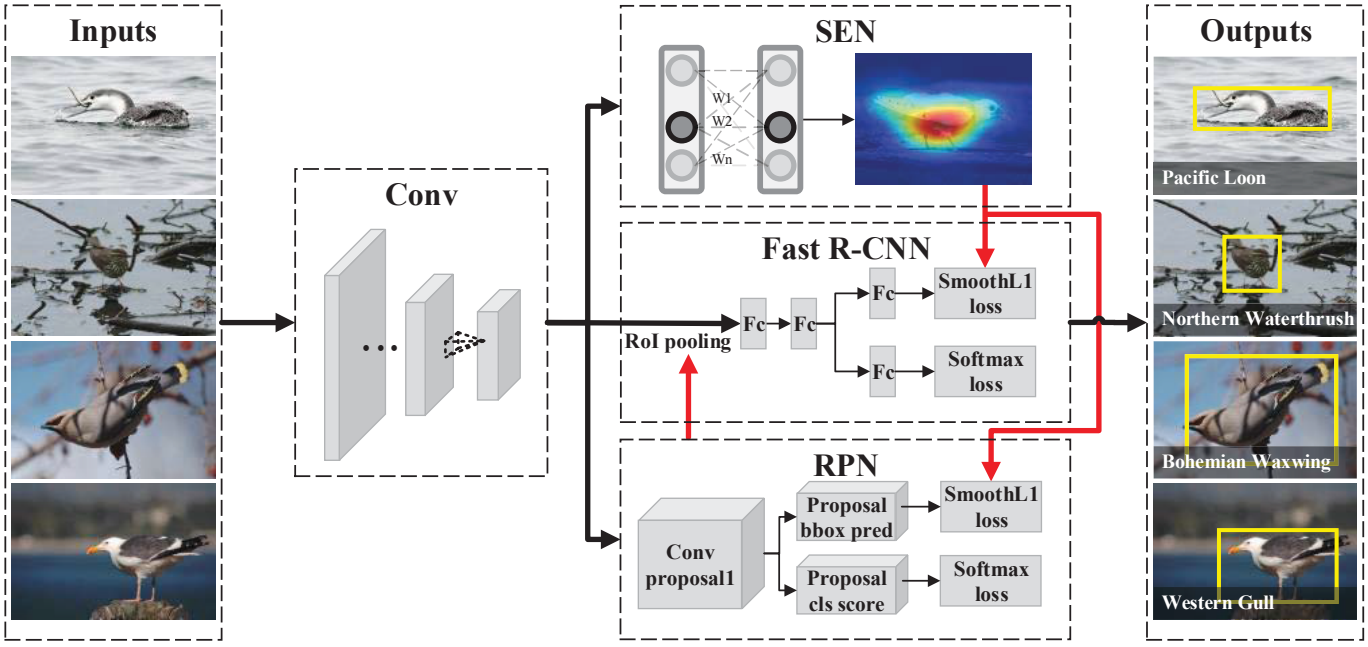

Figure 2: An overview of our Saliency-guided Faster R-CNN approach. Saliency extraction network (SEN) extracts the saliency information to provide the bounding box for training region proposal network (RPN) and Fast R-CNN, RPN produces the region proposal to accelerate the process of proposal generation, and Fast R-CNN learns to localize the discriminative region. The outputs show the predicted discriminative regions and subcategories.

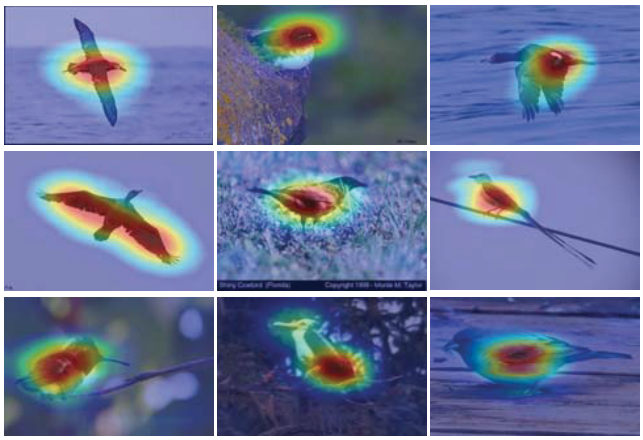

Figure 3: Some examples of saliency maps extracted by SEN in our Saliency-guided Faster R-CNN approach.

the discriminative region, which is guided by the saliency information extracted by SEN. Faster R-CNN is composed by region proposal network (PRN) and Fast R-CNN [17], both of them share convolutional layers for better performance.

Instead of using time-consuming bottom-up process such as selective search [10], RPN is adopted to quickly generate region proposals of images by sliding a small network over the feature map of last shared conolutional layer. At each sliding-window location, k region proposals are simultaneously predicted, and they are parameterized relative to k anchors. We apply 9 anchors with 3 scales and 3 aspect ratios as Faster R-CNN. For training RPN, a binary class label of being an object or not is assigned to each anchor, which depends on the Intersection-over-Union (IoU) [18] overlap with a ground truth bounding box of object. But in our approach, we compute the IoU overlap with the bounding box of discriminative region generated by SEN rather than the ground truth bounding box of object, which avoids using the labor-consuming object and parts annotations. And the loss function for an image is defined as:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (2)$$

where i denotes the index of an anchor in a mini-batch, p_i denotes the predicted probability of anchor i being a discriminative region, p_i^* denotes the label of being a discriminative region of object or not depending on the bounding box t_i^* generated by SEN , t_i is the predicted bounding box of discriminative region, L_{cls} is the classification loss defined by log loss, and L_{reg} is the regression loss defined by the robust loss function (smooth L_1) [17].

For the localization network, Fast R-CNN [17] is adopted. In Fast R-CNN, a region of interest (RoI) pooling layer is employed to extract a fixed-length feature vector from feature map for each region proposal generated by RPN. And each feature vector passes forward for two outputs: one is predicted subcategory and the other is predicted bounding box of discriminative region. Through Faster

R-CNN, we obtain the discriminative region and subcategory of each image simultaneously, accelerating classification speed.

2.2 Saliency-guided localization learning

The saliency-guided localization learning schedules the training process of SEN and Faster R-CNN to make full use of their advantages: (1) SEN learns the saliency information of image to tell which region is important and discriminative for classification, and saliency information guides the training of Faster R-CNN, and (2) RPN in Faster R-CNN generates region proposals that relevant to the discriminative regions of images, which accelerates the process of region proposal rather than using bottom-up process as selective search [10]. Considering that training RPN needs bounding boxes of discriminative regions provided by SEN, and Fast R-CNN utilizes the proposals generated by RPN, we adopt the strategy of sharing convolutional weights between SEN and Faster R-CNN to promote the localization learning.

First, we train the SEN. This network is first pre-trained on the ImageNet 1K dataset [19], and then fine-tuned on the fine-grained image classification dataset, such as CUB-200-2011 [1] in our experiment. And then, we train the PRN. Its initial weights of convolutional layers are cloned from SEN. Instead of fixing the shared convolutional layers, all layers are fine-tuned in the training phase. Besides, we train RPN and Fast R-CNN follows the strategy in Ren et al. [16].

3 EXPERIMENTS

3.1 Dataset and evaluation metrics

We conduct experiments on the widely-used CUB-200-2011 [1] dataset in fine-grained image classification. Our proposed Saliency-guided Faster R-CNN approach is compared with 18 state-of-the-art methods to verify its effectiveness.

CUB-200-2011 [1] is the most widely-used dataset in fine-grained image classification task, which contains 11788 images of 200 subcategories belonging to bird, 5994 images for training and 5794 images for testing. And each image has detailed annotations: a image-level subcategory label, a bounding box of object, and 15 part locations. In our experiments, only image-level subcategory label is used to train the networks.

Accuracy is adopted to comprehensively evaluate the classification performances of our Saliency-guided Faster R-CNN approach and compared methods, which is widely used in fine-grained image classification [2, 6, 12], and its definition is as follow:

$$Accuracy = \frac{R_a}{R} \tag{3}$$

where R denotes the number of images in testing set, and R_a denotes the number of images that are correctly classified.

Intersection-over-Union (IoU) [18] is adopted to evaluate whether the predicted bounding box of discriminative region is a correct localization, and its formula is defined as:

$$IoU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{4}$$

where B_p denotes the predicted bounding box of discriminative region, B_{gt} denotes the ground truth bounding box of object, $B_P \cap B_{gt}$ denotes the intersection of the predicted and ground truth bounding boxes, and $B_p \cup B_{gt}$ denotes their union. We consider the predicted bounding box of discriminative region is correctly localized, if the IoU exceeds 0.5.

3.2 Details of the networks

Our Saliency-guided Faster R-CNN approach consists of three networks: saliency extraction network (SEN), region proposal network (RPN) and Fast R-CNN. They are all based on 16-layer VGGNet [20], which is widely used in image classification task. The basic CNN can be replaced with the other CNN. SEN extracts the saliency information of images to provide bounding boxes needed by Faster R-CNN. For VGGNet in SEN, we remove the layers after conv5_3 and add a convolutional layer of size 3×3, stride 1, pad 1 with 1024 neurons, which is followed by a global average pooling layer and a softmax layer [14]. We adopt the object-level attention of Xiao et al. [5] to select relevant image patches for data extension. And then we utilize the extended data to fine-tune SEN for learning discriminative features. The number of neurons in softmax layer is set as the number of subcategories in the dataset. Faster R-CNN shares the weights of layers before conv5_3 with SEN for better discriminative localization as well as classification performance. The architecture of Fast R-CNN is the same with VGGNet except that pool5 layer is replaced by a RoI pooling layer, and has two outputs: one is predicted subcategory and the other is predicted bounding box of discriminative region.

At training phase, for SEN, we initialize the weights with the network pre-trained on the ImageNet 1K dataset, and then use SGD with a minibatch size of 20. We use a weight decay of 0.0005 with a momentum of 0.9 and set the initial learning rate to 0.001. The learning rate is divided by 10 every 5K iterations. We terminate training at 35K iterations. For Faster RCNN, we initialize the weights with the SEN, and then start SGD with a minibatch size of 128. We use a weight decay of 0.0005 with a momentum of 0.9 and set the initial learning rate to 0.001. We divide the learning rate by 10 at 30K iterations, and terminate training at 50K iterations.

3.3 Comparisons with state-of-the-art methods

This subsection presents the experimental results and analyses of our Saliency-guided Faster R-CNN approach as well as the state-of-the-art methods on the widely-used CUB-200-2011 [1] dataset. We verify the effectiveness of our approach from accuracy and efficiency of classification.

3.3.1 Accuracy of classification. Table 1 shows the comparison results on CUB-200-2011 dataset at the aspect of classification accuracy. Object, parts annotations and CNN used in these methods are listed for fair comparison. Traditional methods as [29] choose SIFT [30] as features, even using both object and parts annotations its performance is limited and much lower than our approach. Our approach achieves the highest classification accuracy among all methods under the same weakly supervised setting that neither object nor parts annotations are used in training or testing phase, and obtains 0.45% higher accuracy than the best result of TSC [8] (85.14% vs. 84.69%), which jointly considers two spatial constraints in parts selection. Despite achieving better classification accuracy, our approach is over two order of magnitude faster than TSC, due

Table 1: Comparisons with State-of-the-art Methods on CUB-200-2011 dataset.

Method	Train Annotation		Test Annotation		Accuracy (%)	Net
	Object	Parts	Object	Parts		
Our Saliency-guided Faster R-CNN Approach					**85.14**	VGGNet
TSC [8]					84.69	VGGNet
FOAF [21]					84.63	VGGNet
PD [6]					84.54	VGGNet
Bilinear-CNN [22]					84.10	VGGNet&VGG-M
NAC [7]					81.01	VGGNet
PIR [12]					79.34	VGGNet
TL Atten [5]					77.90	VGGNet
MIL [23]					77.40	VGGNet
Coarse-to-Fine [24]	√		√		82.90	VGGNet
PG Alignment [4]	√		√		82.80	VGGNet
VGG-BGLm [25]	√		√		80.40	VGGNet
Webly-supervised [26]	√	√			78.60	AlexNet
PN-CNN [27]	√	√			75.70	AlexNet
Part-based R-CNN [2]	√	√			73.50	AlexNet
SPDA-CNN [11]	√	√	√		85.14	VGGNet
Deep LAC [28]	√	√	√		84.10	AlexNet
PS-CNN [13]	√	√	√		76.20	AlexNet
PN-CNN [27]	√	√	√	√	85.40	AlexNet
POOF [29]	√	√	√	√	73.30	

to the end-to-end network, as shown in Table 2. The efficiency analysis will be described in Section 3.3.2. And our approach performs better than the method of Bilinear-CNN [22], which combines two different CNNs: VGGNet [20] and VGG-M [31]. Its classification accuracy is 84.10%, which is lower than our approach by 1.04%. Furthermore, our approach even outperforms these methods using object annotation in both training and testing phase by at least 2.24%, such as Coarse-to-Fine [24], PG Alignment [4] and VGG-BGLm [25]. Moreover, our approach outperforms these methods that use both object and parts annotations [2, 26]. Neither object nor parts annotations are used in our Saliency-guided Faster R-CNN approach, which leads fine-grained image classification to practical application. Besides, end-to-end network in our approach simultaneously localizes discriminative region and encodes discriminative feature for each image, and discriminative localization promotes the classification performance.

3.3.2 Efficiency of classification. Experimental results at the aspect of efficiency on CUB-200-2011 dataset is presented in Table 2. We get the testing speed on the computer with NVIDIA TITAN X @1417MHZ and Intel Core i7-6900K @3.2GHZ, and use frames per second (fps) to evaluate the efficiency. Comparing with state-of-the-art methods, our Saliency-guided Faster R-CNN approach achieves the best performance on not only the classification accuracy but also the efficiency. We split state-of-the-art methods into two groups by the basic CNNs used in their methods: VGGNet [20] and AlexNet [32]. Results of these methods in first group are obtained by their authors' source codes. Comparing with these methods, our approach improves about 123% than Bilinear-CNN at the aspect of classification speed (10.07 fps vs. 4.52 fps). Besides, our classification accuracy is also 1.04% higher than Bilinear-CNN. Even more, our approach is over two orders of magnitude faster than these methods with two separated stages of localization and encoding. When utilizing AlexNet as the basic network, our approach is still faster than PS-CNN [13] and improves about 19.51%, which also utilizes AlexNet. And when applying AlexNet as basic CNN in our approach, the classification accuracy is 73.58%. It is noted that neither object nor parts annotations are used in our approach, while all used in PS-CNN. The classification speed of PS-CNN [13] is reported as 20 fps in their paper. They provide a reference that a single CaffeNet [33] runs at 50 fps under their experimental setting (NVIDIA Tesla K80). In our experiments, a single CaffeNet runs at 35.75 fps, so we calculate the speed of PS-CNN in the same experimental setting with ours as 20*35.75/50=14.30 fps. Our approach avoids the time-consuming classification process by the design of end-to-end network, and achieves the best classification performance by the mutual promotion between localization and classification. This leads the fine-grained image classification to practical application.

3.4 Effectiveness of discriminative localization

Saliency-guided localization learning is proposed to train SEN and Faster R-CNN for improving the localization and classification performance simultaneously. Since we devote to localizing the discriminative region which is generally located at the object, we adopt the IoU overlap between discriminative region and ground truth bounding box of object to evaluate the correctness of localization. We consider a bounding box of discriminative region to be correctly

Table 2: Comparison of average classification speed (frames per second) with state-of-the-art methods on CUB-200-2011 dataset. The results are obtained on the computer with NVIDIA TITAN X @1417MHZ and Intel Core i7-6900K @3.2GHZ.

Methods	Testing Speed (fps)	Net
Our Saliency-guided Faster R-CNN Approach	**10.07**	VGGNet
Bilinear-CNN [22]	4.52	VGGNet&VGG-M
TSC [8]	0.34	VGGNet
TL Atten [5]	0.25	VGGNet
NAC [7]	0.10	VGGNet
Our Saliency-guided Faster R-CNN Approach	**17.09**	AlexNet
PS-CNN [13]	14.30	AlexNet

Table 3: Classification and localization Accuracies.

Methods	Classification Accuracy(%)	Localization Accuracy(%)
Our Saliency-guided Faster R-CNN Approach	**85.14**	**46.05**
SEN	77.50	44.93

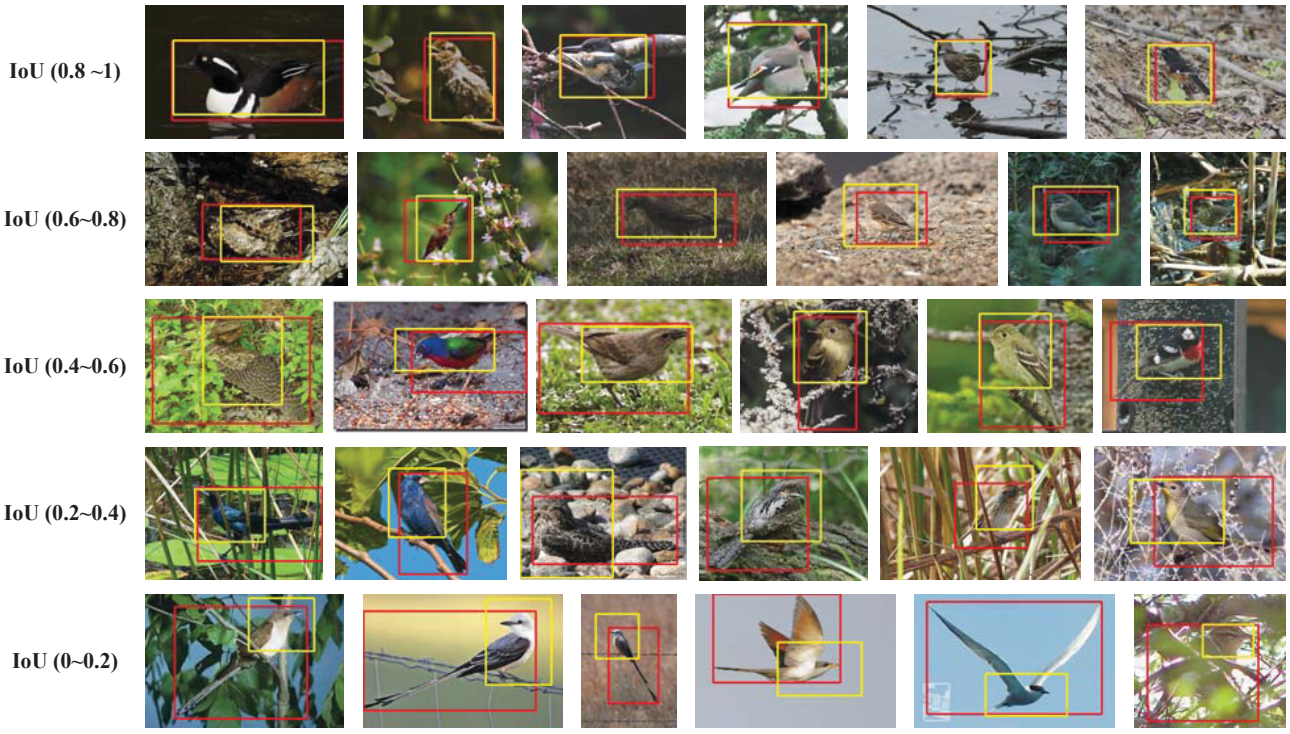

Figure 4: Samples of predicted bounding boxes of discriminative regions (yellow rectangles) and ground truth bounding boxes of objects (red rectangles) at different ranges of IoU on CUB-200-2011 dataset.

predicted if IoU with ground truth bounding box of object is larger than 0.5. The accuracy of localization is shown in Table 3.

Our Saliency-guided Faster R-CNN approach achieves the accuracy of 46.05%. Considering that neither object nor parts annotations are used, it is a promising result. And comparing with "SEN" which means directly using SEN to generate bounding box, our approach achieves improvements both in classification and localization, which verifies the effectiveness of our saliency-guided localization learning approach. We show some samples of predicted bounding boxes of discriminative regions and ground truth bounding boxes of objects at different ranges of IoU (e.g. 0~0.2, 0.2~0.4, 0.4~0.6, 0.6~0.8, 0.8~1) on CUB-200-2011 dataset, as Figure 4. We have some predicted bounding boxes whose IoUs with ground truth bounding boxes of objects are lower than 0.5. But these predicted bounding boxes contain discriminative regions of objects, such as heads and bodies. It verifies the effectiveness of our approach in

Table 4: *PCL* for each part of object in the CUB-200-2011 testing set.

Parts	back	beak	belly	breast	crown	forehead	left eye	left leg
PCL (%)	96.33	96.49	94.00	95.29	97.38	97.07	97.49	89.92
Parts	left wing	nape	right eye	right leg	right wing	tail	throat	**average**
PCL (%)	92.60	96.60	96.79	91.85	97.00	85.03	96.38	**94.68**

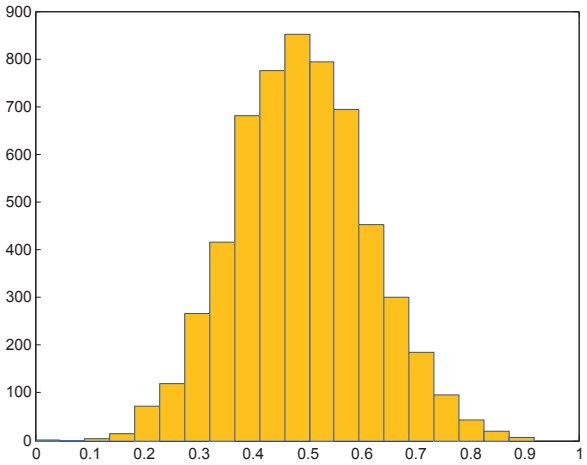

Figure 5: IoU histogram. Abscissa denotes the IoU overlap between predicted bounding box of discriminative region and ground truth bounding box of object. And ordinate means the number of testing images that have the IoU overlap at the range.

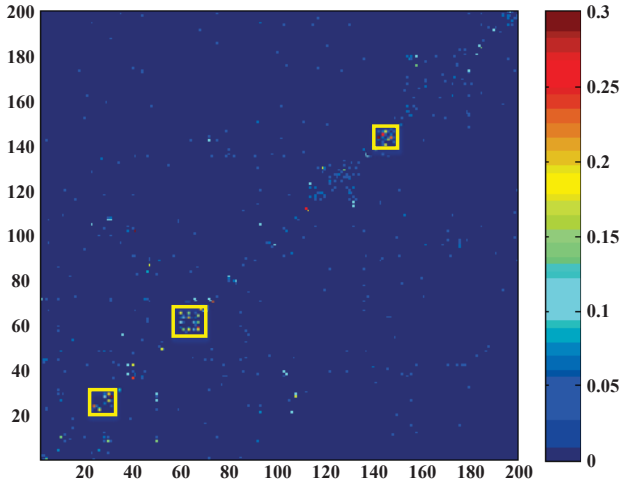

Figure 6: Classification confusion matrix on CUB-200-2011 dataset with 200 subcategories. The yellow rectangles show the sets of subcategories with the higher probability of misclassification.

localizing discriminative region of object for achieving better classification performance. Figure 5 shows the histogram of IoU. We can observe that most testing images lie in the range of 0.4~1. To further verify the effectiveness of discriminative localization in our approach, results are given in terms of the Percentage of Correctly Localization (PCL) in Table 4, estimating whether the predicted bounding box contains the parts of object or not. CUB-200-2011 dataset provides 15 part locations, which denote the pixel locations of centers of parts. We consider our predicted bounding box contain a part if the part location lies in the area of the predicted bounding box. Table 4 shows that about average 94.68% of the parts located in our predicted bounding boxes. It shows that our discriminative localization can detect the distinguishing information of objects to promote classification performance.

3.5 Analysis of misclassification

Figure 6 shows the classification confusion matrix for our approach, where coordinate axes denote subcategories and different colors denote different probabilities of misclassification. The yellow rectangles show the sets of subcategories with the higher probability of misclassification. We can observe that these sets of subcategories locate near the diagonal of the confusion matrix, which means that these misclassification subcategories generally belong to the same genus with small variance. The small variance is not easy to measure from the image, which leads the high challenge of fine-grained image classification. Figure 7 shows some examples of the most probably confused subcategory pairs. One subcategory is most confidently classified as the other in the same row. The subcategories in the same row look almost the same, and belong to the same genus. For example, "Brandt Cormorant" and "Pelagic Cormorant" look the same in the appearance, both of them have the same attributes of black feather and long neck, and belong to the genus of "Phalacrocorax". It is extremely difficult for us to distinguish between them.

3.6 Comparison with baselines

Our Saliency-guided Faster R-CNN approach is based on Faster-RCNN [16], and adopts VGGNet [20] as the basic model. To verify the effectiveness of our approach, we present the results of our approach as well as the baselines in Table 5. "VGGNet" denotes the result of directly using fine-tuned VGGNet, and "Faster R-CNN (gt)" denotes the result of directly adopting Faster R-CNN with ground truth bounding box to guide training phase. Our approach achieves the best performance even without using object or parts annotations. We adopt VGGNet as the basic model in our approach, but its classification accuracy is only 70.42%, which is much lower than ours. It shows that the discriminative localization has promoting effect to classification. With discriminative localization, we find the most important regions of images for classification, which contains the key variance from other subcategories. Comparing with "Faster R-CNN (gt)", our approach also achieves better performance.

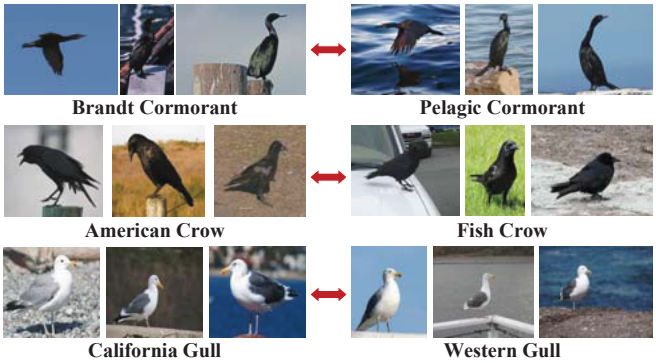

Figure 7: Examples of the most confused subcategory pairs. One subcategory is mostly confidently classified as the other in the same row when in the testing phase.

Table 5: Comparison with baselines.

Methods	Accuracy (%)
Our Saliency-guided Faster R-CNN Approach	**85.14**
Ours (without shared conv layers)	83.95
Faster R-CNN (gt)	82.46
VGGNet	70.42

It is an interesting phenomenon that worth thinking about. From the last row in Figure 4, we observe that not all the areas in the ground truth bounding boxes are helpful for classification. Some ground truth bounding boxes contain large area of background noise that has less useful information and even leads to misclassification. So discriminative localization is significantly helpful for achieving better classification performance. And comparison with "Ours (without shared conv layers)" verifies the effectiveness of our saliency-guided localization learning represented in Section 2.2, which promotes not only discriminative localization but also classification.

4 CONCLUSION

In this paper, discriminative localization approach via saliency-guided Faster R-CNN has been proposed for weakly supervised fine-grained image classification. We first propose saliency-guided localization learning approach to localize discriminative region automatically for each image, which uses neither object nor parts annotations to avoid using labor-consuming annotations. And then an end-to-end network based on Faster R-CNN with guide of saliency information is proposed to simultaneously localize discriminative region and encode discriminative features, which not only achieves a notable classification performance but also accelerates classification speed. And combining them, we simultaneously accelerate classification speed and eliminate dependence on object and parts annotations. Comprehensive experimental results show our Saliency-guided Faster R-CNN approach is more effective and efficient compared with state-of-the-art methods on the widely-used CUB-200-2011 dataset.

The future works lie in two aspects: First, we will focus on learning better discriminative localization via exploiting the effectiveness of fully convolutional networks. Second, we will also attempt to localize several discriminative regions with different semantic meanings simultaneously, such as the head or body of bird, to improve fine-grained image classification performance.

5 ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China under Grants 61371128 and 61532005.

REFERENCES

[1] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.
[2] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. *International Conference on Machine Learning (ICML)*, pages 834–849, 2014.
[3] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, and Qi Tian. Fused one-vs-all mid-level features for fine-grained visual categorization. *ACM International Conference on Multimedia (ACM MM)*, pages 287–296, 2014.
[4] Jonathan Krause, Hailin Jin, Jianchao Yang, and Li Fei-Fei. Fine-grained recognition without part annotations. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5546–5555, 2015.
[5] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 842–850, 2015.
[6] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin, and Qi Tian. Picking deep filter responses for fine-grained image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1134–1142, 2016.
[7] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. *International Conference of Computer Vision (ICCV)*, pages 1143–1151, 2015.
[8] Xiangteng He and Yuxin Peng. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. *AAAI Conference on Artificial Intelligence (AAAI)*, pages 4075–4081, 2017.
[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.
[10] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision (IJCV)*, 104(2):154–171, 2013.
[11] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1143–1152, 2016.
[12] Yu Zhang, Xiu-Shen Wei, Jianxin Wu, Jianfei Cai, Jiangbo Lu, Viet-Anh Nguyen, and Minh N Do. Weakly supervised fine-grained categorization with part-based image representation. *IEEE Transactions on Image Processing (TIP)*, 25(4):1713–1725, 2016.
[13] Shaoli Huang, Zhe Xu, Dacheng Tao, and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1173–1182, 2016.
[14] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[15] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.
[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015.
[17] Ross Girshick. Fast r-cnn. *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
[18] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, 2015.
[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.
[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arxiv:1409.1556*, 2014.

[21] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, and Qi Tian. Fused one-vs-all features with semantic alignments for fine-grained visual categorization. *IEEE Transactions on Image Processing (TIP)*, 25(2):878–892, 2016.

[22] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. *International Conference of Computer Vision (ICCV)*, pages 1449–1457, 2015.

[23] Zhe Xu, Dacheng Tao, Shaoli Huang, and Ya Zhang. Friend or foe: Fine-grained categorization with weak supervision. *IEEE Transactions on Image Processing (TIP)*, 26(1):135–146, 2017.

[24] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li, and Qi Tian. Coarse-to-fine description for fine-grained visual categorization. *IEEE Transactions on Image Processing (TIP)*, 25(10):4858–4872, 2016.

[25] Feng Zhou and Yuanqing Lin. Fine-grained image classification by exploring bipartite-graph labels. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1124–1133, 2016.

[26] Zhe Xu, Shaoli Huang, Ya Zhang, and Dacheng Tao. Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

[27] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arxiv:1406.2952*, 2014.

[28] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1666–1674, 2015.

[29] Thomas Berg and Peter Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 955–962, 2013.

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

[31] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arxiv:1405.3531*, 2014.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[33] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *ACM International Conference on Multimedia (ACM MM)*, pages 675–678, 2014.