An Improved Deep Neural Network Model Based on Searching Space Limitation for Object Localization

Renchao Wu
International College, Beijing University of Posts and Telecommunications, Beijing, China

Abstract. Image object localization is the primary research direction in computer vision. A lots of algorithms had been proposed based on the ResNet structure and good results have been obtained on mainstream datasets. However, these methods still have the problem of too large ROI search space on weakly supervised data. This paper proposes a method that further limits the search space by directly extracting the high-level parameters of ResNet. The better search results and search efficiency are achieved and the candidate ROI search space is better reduced. The space information of the original images can be retained and training/testing time can be saved. The experimental results on the benchmark datasets show that this network structure has good object localization accuracy under weakly supervised data and the search efficiency is improved compared to other methods.

Keywords: Object localization; ResNet; ROI search space; Space information.

1. Introduction
Human’s vision has a strong spatio-temporal positioning ability. Whether it is a high-speed vehicle or a pedestrian crossing a zebra crossing, it can be quickly noticed, identified, and reacted accordingly. Correspondingly, advanced human-computer interaction and environmental intelligent cognitive computer systems also need to provide this ability that extracting from the scene multiple types of objects, which include hearing, vision, and their detailed locations. This cognitive ability includes classification or recognition of objects, spatial/temporal location, and extraction and detection of these objects.

Artificial intelligence processing solutions for computer vision have been studied for many years, and many algorithms and structures had been continuously made. Most of these algorithms need to be based on a large amount of visual scene data. The problem is that the labeling information that matches these massive data is often scarce. If manually labeling the unlabeled data manually, it is very expensive and consumes a lot of time with inevitably errors. For example, pedestrians are identified as the background, or different people have different views on the location of the object. Therefore, it is necessary to propose the better performance system on the weakly labeled dataset of visual scenes.

The SPM method [1] was an improved version of the BOF feature pack. BOF is to calculate the distribution characteristics of feature points and generate vertical histograms in the entire image. SPM can overcome the inherent shortcomings of BOF and obtain local information of the image by counting the distribution of image feature points at different resolutions.

Deep Residual Learning [2] is residual learning framework. Assuming that multiple non-linear layer combinations can approximate a complex function, the residuals of the hidden layer can also approximate a complex function. The residual network makes it easier to optimize the network by adding shortcut connections. The output of the residual unit is activated by summing ReLU between
the concatenated output and input elements of multiple convolutional layers. These structures are cascaded to obtain the entire residual network.

Krizhevsky applied AlexNet [3] with an 8-layer convolutional neural network and won the ImageNet 2012 image recognition challenge with a great advantage. It proved for the first time that the learned features can surpass the features of manual design. Compared with the relatively small LeNet, AlexNet contains 8 layers of transformations, including 5 layers of convolution and 2 layers of fully connected hidden layers, and 1 fully connected output layer.

Chatfield analysis delving deep into convolutional nets [4]. The results of several deep network models and traditional feature extraction methods in image classification are compared. The experiments show the effects of data expansion, feature normalization, and dimensional changes on the performance of the final model. It also proves that the depth model trained on ILSVRC is still very effective in the classification of other data. If targeted fine-tune is added, it will be more effective.

Weakly Supervised Deep Detection Networks [5] was proposed as one end-to-end object detection method. Weakly supervised object detection means that the data has only image-level annotations and there were no bounding boxes and only class information. The end-to-end weakly supervised deep detection network uses the CNN network to pre-trained on the i-th universal image data set as the initialization feature, performs region proposals on the feature map obtained from the last convolution layer, and adds an SPP layer to output recognition and detection, and finally merged to get the predicted labels.

Compared with supervised algorithms, weakly supervised methods that lack fine labeling still have obvious problems of low accuracy in object detection and recognition. Literature [6] proposed a pipeline detection method for object detection, recognition and segmentation. First, a temporary result is constructed from weakly labeled data, and then supervised algorithms continue to be used for a given task. Intermediate fusion uses multiple resolution mechanisms to implement specific tasks such as positioning and detection. This method has achieved good performance results on various weakly supervised competition datasets.

Image datasets usually have ground truth image-level tags. Images crawled from the web have no reliable labeling information. Unlike image recognition tasks, fully supervised semantic segmentation requires that we have pixel-level sample annotation. On the other hand, the network training requires a large increase in the sample size, so the lack of sample size and the unity of sample types have become the primary factors of full supervision methods. Literature [7] proposed to solve the problem of semantic segmentation in the case of weak supervision through two steps. First, a bootstrapping process is used to generate accurate pixel-level segmentation masks, and the results are then used as a training set to transform the problem into a fully supervised problem.

In weakly supervised learning, label data is usually used to make coarse-to-fine judgments. For example, two cascaded networks can be used, the first to generate bounding boxes and the second to classify them. Similar to weakly supervised positioning, attention-based models also select relevant regions to support decision-making [8]. However, the existing methods still have the problem that the search space of ROI is too large, which will cause operational-time cost. This paper proposes a method that preserves the original spatial information while reducing the computational cost. By directly extracting the high-level parameters of the neural network, it can obtain better search results and search efficiency.

The paper’s organization is as follows: The second part gives the general weakly supervised learning methods for image classification and localization, then the searching space limitation principle is proposed, finally the experimental results on Pascal VOC datasets are showed, which followed by a conclusion.

2. General Weakly Supervised Learning

Weakly supervised problems often encountered in image detection and recognition. In literature [9], the method of searching and segmenting target regions and merging them is used to deal with the difficult situation of small label data sets, and an unsupervised method is adopted. The network is pre-trained, then the sparsely pre-trained CNN is used to further complete the target region segmentation and recognition tasks, and finally, the fully connected layer is used.
Andreas Opelt studied the existing methods of weak supervision of object locations [10]. According to this literature, these methods are usually tested on datasets with common and obvious objects. The objects of these datasets are obvious and well-labeled, which cannot verify the effectiveness of weakly supervised algorithms. They proposed a new data set that can verify the ability of weakly supervised algorithms. Unlike the existing data set, this new data set is more suitable for implementing weakly supervised algorithms, and uses improved object recognition methods for object localization and gives results.

Matthew et al. considered that the general object detection and localization is to fully supervise learning the region of interest to adjust the model continuously reaching the minimum loss [11]. If the regions of interest are not provided with sufficient labels, such as the type of object or its location, weakly supervised learning methods can be used only. They proposed that treating good labels as latent variables and use output formulas to perform weak supervision, and improve the detection level by ranking methods. It is verified on the classic data set that this method is close to the level of supervised learning.

A further question was raised that how to find those discriminative regions that are decisive for classification in the picture, when there is only an image-level label and no labels like detection and segmentation [12]. The traditional weakly-supervised object positioning literature is usually not in an end-to-end form, requires a lot of preprocessing, and does not evaluate the positioning capability, so it is difficult to use in real scenarios. Network in GoogLeNet try to avoid using a fully connecting layer. Its characteristic is to use global average pooling to avoid overfitting, in fact, it can also improve object positioning capabilities. Global average pooling encourages the network to identify complete object areas, because average pooling has less loss than discriminative areas where the entire object is identified by max pooling. On the other hand, by introducing a class activation map to represent the weight of the activation map in each graph, it is possible to further accurately locate the distinguishing regions.

The latest network architecture WILDCAT (shown in Fig.1) is divided into three parts [13]: FCN (ResNet-101) to extract feature maps; Multi-map transfer layer decomposes feature maps into multi-channel features, each channel corresponding to a significant local feature; Pooling The generated multi-channel feature map is aggregated; finally, image level ground truth information is learned.

### 3. Searching Space Limitation Method

The selection of the fully convolutional architecture network in Fig.1 is very important for the performance of the entire model. The correct network selection should retain a higher spatial resolution. During training, the model parameters pre-trained on the ImageNet dataset firstly. The ResNet-101 was selected and its last layer was removed and replaced with the transfer layer and pooling layer. The Multi-map transfer layer uses 1x1 convolution, which is the channels number in the past layer. The converted according to the sum of categories through a convolutional layer, then the same number of channels is fixedly assigned to each category, which help the convolution kernel corresponding to the channel learning some features of its corresponding category.

Since the ground truth is a label, it is necessary to aggregate the transfer layer's information into the image level. The pooling layer has these operations: first progressive pooling, then pooling in space. The Class-wise is a pooling in the corresponding channel in each category to get a heatmap; spatial pooling uses the equation.1 for pooling of each class's map to calculate a score for the image with objects of that category:
\[ s_c^e = \max_{k^+, k^-} \frac{1}{k^+} \sum_{i,j} h_{i,j} \bar{z}_{i,j}^c + \alpha \left( \min_{k^-, k^+} \frac{1}{k^-} \sum_{i,j} h_{i,j} \bar{z}_{i,j}^c \right) \]  

(1)

Where \( c \) is the class, \( k^+, k^- \) and \( \alpha \) are hyper-parameters, \( \bar{z}_{i,j}^c \) is the i-th and j-th row/column of the output from class c in class-wise pooling.

In the object localization part, we propose an improved searching space limitation method, which is considered to select a small region in the image and training it through the entire network to obtain the score of the small region corresponding to each category. Therefore, as long as we iteratively divide this step to obtain the score of each area corresponding to each category, and then take the area with a high score, it can be approximated to the location of an object. Each time, the high result was choose to obtain the final object positioning. By this searching process, the amount of calculation is relatively large. The calculation cost also needs to be considered, so it is directly found from the output layer of ResNet-101.

Fig.2 shows the search space limitation principle. The search is performed on a 5*5 convolutional layer. The first subgraph represents the search results of the previous convolutional layer, and several other subgraphs represent the results. Continue to calculate the score of the pre-selection box until the best score is obtained, and then map back to the original image.

4. Experimental Results

The experimental data was selected from Pascal VOC. Specifically, the pre-trained ResNet101 models on the ImageNet dataset and the 2009-2012 Pascal VOC dataset were integrated into the training part. The monitoring signal given in the data set contains only one image tag. The content annotation of the image is part of the dataset, and it also contains bounding boxes corresponding to various objects. The ultimate goal of the experiment is to perform multi-label classification on the image and locate it under the weak supervision of the object.

| Training | Testing |
|----------|---------|
| mAP      | 96.1    | 90.0    |
| macc     | 98.6    | 98.2    |
| wacc     | 98.9    | 97.5    |

The results of multi-label classification are shown in Table 1. The details are as follows: For models that have been pre-trained with ImageNet data, the learning rate is set to 0.02; for other models, the learning rate is set to 0.01, and \( k^+ \) and \( k^- \) are set to 20% of the entire matrix size. After entering the pictures, adjust all images to 448 * 448.
In the image localization experiment, the optimal model trained by the multi-label classification problem is directly used. In order to limit the search space, the aspect ratio of the candidate frames to be searched is additionally limited; the maximum value of the log of all prediction frames is selected as the score. In addition, only one box with the highest score was kept for each class in the test. Fig. 3 shows an example of the prediction positioning block diagram of our proposed algorithm. Red represents the ground truth, while green represents the predicted results.

It can be observed from the experimental result list and the result display diagram in the fourth part that the search space restriction scheme proposed in this paper still has the correct target positioning accuracy under the condition of weak labels. One reason is that this scheme can exist in the raw data to maintain the spatial state; another reason is that this search space limitation can directly extract the advanced parameters of the Resnet output layer to improve search efficiency.

5. Conclusions
As the effective image classification network, the FCN-based method shows good results on the classic datasets. However, on weakly supervised dataset, the FCN-based method still has the problem that the search speed of ROI is large with the slow operation speed. Aiming at these problems, this paper proposes a search space limitation scheme that can improve the computational processing efficiency. Its ability is to further reduce the search space of candidate ROIs and achieve correct target positioning accuracy under weak label conditions. This method can exist in the spatial state of the raw data to maintain, and by directly extracting the advanced parameters of the neural network can achieve better search results and search efficiency. The results of object localization tests on standard data sets show that the method has satisfactory object localization accuracy and search efficiency on weakly supervised data sets.

References
[1] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories[C]. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006, 2: 2169-2178
[2] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]. Advances in neural information processing systems. 2012: 1097-1105.
[4] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: Delving deep into convolutional nets[J]. arXiv preprint arXiv:1405.3531, 2014.
[5] Bilen H, Vedaldi A. Weakly supervised deep detection networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2846-2854.
[6] Ge W, Yang S, Yu Y. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1277-1286.
[7] Shen T, Lin G, Shen C, et al. Bootstrapping the performance of weakly supervised semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1363-1371.
[8] Opelt A, Pinz A. Object localization with boosting and weak supervision for generic object recognition[C]. Scandinavian Conference on Image Analysis. Springer, Berlin, Heidelberg, 2005: 862-871.

[9] Raj A, Gandhi K, Nalla B T, et al. Object detection and recognition using small labeled datasets[M]. Computational Intelligence: Theories, Applications and Future Directions-Volume II. Springer, Singapore, 2019: 407-419.

[10] Blaschko M, Vedaldi A, Zisserman A. Simultaneous object detection and ranking with weak supervision[C]. Advances in neural information processing systems. 2010: 235-243.

[11] Zhang J, Bargal S A, Lin Z, et al. Top-down neural attention by excitation backprop[J]. International Journal of Computer Vision, 2018, 126(10): 1084-1102.

[12] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.

[13] Durand T, Mordan T, Thome N, et al. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 642-651.