Object Localization with weakly supervised learning

Jun He
Beijing Normal University
N0.19 XinJieKouWai Street, Beijing
hejun@bnu.edu.cn

Huanqing Yan
Beijing Normal University
N0.19 XinJieKouWai Street, Beijing
yanhq@mail.bnu.edu.cn

Abstract

Object localization is a task to find the location of the objects in the image, which is easy when given the labels like pixels or boxes. However, due to the high cost of data annotation, researchers are committed to find a way to localize objects under weak supervision, which is only image category label. In this paper, we present our solutions for the Weakly-supervised Object Localization of Learning from Imperfect Data (LID)2020.

1. Introduction

In recent years, Weakly-supervised Object Localization (WSOL) is a research hotspot of computer vision because of its little dependence on labels. The key problem is how to accurately locate the target in the image without the location label. Existing approaches are mainly utilize the discriminative features in the Convolutional Neural Networks (CNN) classifier. The key idea is to find out the most confident region that make classifier to decide the final class result. And we also use these methods as visualization of CNN. In this way, [1-3] visualize the location of the object by generating a class activation map.

However, because the goal of classifier is to find the most distinguishable region, the results are often concentrated in the local area and can not completely cover the whole object. For example, when locating a cat, only the head of the cat is found and the body part is missed. This situation results in inaccurate results and is quite common in WSOL. Therefore, some methods [4-5] remove the most discriminative regions which are found first, so as to find other regions. [6] erases regions randomly on the image to effectively reduce the classifier's attention to the most distinguishable regions.

However, the above methods either need to modify the model structure, such as adding a classifier or adding a loss function which make it inefficient. [7] proposes an Attention-based Dropout Layer (ADL), an efficient and effective method which utilizes self-attention mechanism to remove the most discriminative region.

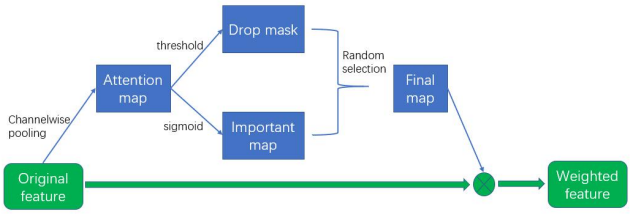

Figure 1: Block diagram of ADL.

2. Method

We use the solution in [7] to localize object. ADL is applied on each feature map to help model to learn the entire region of the object. It generates a drop mask and an importance map from input features. The procedure of the method is shown in Figure 1.

Based on the CNN features of the whole image obtained from the front convolution layer, we first use channelwise pooling to get one-channel features. Then, the importance image and drop mask are obtained by normalizing and thresholding the self-attention map. Then randomly select one to multiply into the input feature image. See the original text for specific model details.

Now, we get a weighted feature that can focus on areas beyond the most discriminative areas. Then the feature is sent to the subsequent classifier to get a classification result. After that, the process of generating mask is the same as that in [1]. After removing the full connection layer, a class activation graph is generated as the target location result. The whole method is easy, and it can be attached to any model without modifying the model structure, and is effective. Therefore we use it as our solution.

3. Experiments

The data set used in the challenge [8] is ImageNet Large Scale Visual Recognition Competition (ILSVRC) [9], which totally includes 1.2 million training images from 1000 categories. It provides pixel-level annotations of 44271 images (validation/testing: 23151/21120) for

evaluation.

We use VGG [10] as backbone networks. We replace the last pooling layer and fully connected layers with GAP (Global Average Pooling) layer. During training, put the image into the CNN model and get feature of image, and generates the weighted feature after the module of ADL. It doesn't need ADL when in testing. And we extract the heatmap from classification model using CAM as final mask.

The result of mask on test set are shown in Figure 2. As shown in the figure, the results of the model can better cover the whole object, rather than gather in some areas. And there are two metrics (Peak IoU, Peak Threshold) for weakly supervised object localization based on intersection over union. In the ideal curve, the highest IoU score is expected to close to 1.0. The threshold value corresponding to the highest IoU score is expected to be 255 since the higher threshold values can reflect a higher contrast between the target object and the background.

In our experiments, we get 0.48 on Peak IoU with Peak Threshold of 42.00.

Figure 2: Results on test set.

References

[1] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In CVPR, 2016.

[2] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, et al. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. In ICCV, 2017.

[3] Aditya Chattopadhyay, Anirban Sarkar, et al. Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks. In WACV, 2018.

[4] Dahun Kim, Donghyeon Cho, Donggeun Yoo, et al. Two-Phase Learning for Weakly Supervised Object Localization. In ICCV, 2017.

[5] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, et al. Adversarial Complementary Learning for Weakly Supervised Object Localization. In CVPR, 2018.

[6] Krishma Kumar Singh and Yong Jae Lee. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization. In ICCV, 2017.

[7] Junsuk Choe, Hyunjung Shim. Attention-based Dropout Layer for Weakly Supervised Object Localization. In CVPR, 2019.

[8] Yunchao Wei, Shuai Zheng, Ming-Ming Cheng, Hang Zhao. LID 2020: The Learning from Imperfect Data Challenge Results, 2020.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: a large-scale hierarchical image database. In CVPR, 2009.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.