Weakly Supervised Object Detection Using Class Activation Map

1st Anh Tuan Dao Tran
University of Information Technology
Ho Chi Minh City, Vietnam
20522107@gm.uit.edu.vn

2nd Phu Vinh Tran
University of Information Technology
Ho Chi Minh City, Vietnam
20522161@gm.uit.edu.vn

3rd Dung Mai Tien
University of Information Technology
Ho Chi Minh City, Vietnam
dungmt@uit.edu.vn

***Abstract*—Class activation map helps visualize the region of a given object category in an image that an accurate classifier uses for prediction. The original work also extended this method to solve weakly supervised object localization. However, results have shown that class activation map only shows the discriminative part of an object instead of the whole object. Recent proposed solutions help alleviate this problem and allow class activation map to achieve state-of-the-art localizing performance. In this paper, we take advantage of those solutions and propose modifications to use CAM for detecting multiple objects instead. The proposed modification achieves promising results on PASCAL VOC 2007 and 2012.**

***Index Terms*—class activation map, weakly supervised object detection**

I. Introduction

Class Activation Map (CAM) introduced by Zhou et al. [1] to visualize the region of a given object category in an image. This method utilizes the weights from the fully connected layer and the feature maps before the global average pooling layer to generate CAM. CAM allows us to know which region of the image is used by an accurate classifier to make a prediction. Zhou et al. extended their work to solve weakly supervised object localization (WSOL). WSOL aims to find a region of a targeted object in an image using image-level labels as training data. Weakly supervised approaches are great alternatives to fully supervised approaches [2] [3] as they require less effort and cost to collect, label, and train data.

However, research has shown that CAM only highlights the most discriminative part of an image, thus affecting the localization performance. Kim et al. [4] solved this problem by addressing issues when training a classifier to generate CAM. The proposed method expands the discriminative region to the whole object, thus achieving state-of-the-art localization performance on ImageNet-1k [5] and CUB-200-2011 [6]. Another solution from Yang et al. [7] is to use a function to combine CAM from different categories.

Although the discriminative part has been solved, CAM only focuses on finding the region of a single object. Utilizing Kim et al. [4] work, we propose modifications to use CAM for object detection instead of object localization.

In summary, the contributions of this paper are as follows:

- We propose a way to combine [4] and [7] to use CAM for object detection instead of object localization.
- We show how using Otsu as a post-processing step can affect the result of the proposed method.
- We also build a simple demo to show the result of the proposed method.

II. Related Works

MIL-based
Pseudo-labeling

III. Proposed Method

A. Addressing CAM limitations

It is widely observed that CAM only highlights the most discriminative part of an object.

1) Bridging: Kim et al. [4] proposed a method to address this issue.

Briefly describe the method.

We propose to use this method to improve CAM performance.

2) CCAM: Yang et al. [7] proposed a method to combine CAM from different categories.

Briefly describe the method.

We propose to use one of the solutions proposed by this method to improve CAM performance.

B. From multi-class to multi-label

In the original work [1], CAM is inferred from a multi-class classifier.

We propose to use a multi-label classifier.

C. Post processing

In the original work [1], CAM is segmented by keeping region of which value is above 20%.

We propose to use Otsu as a post-processing step.

IV. Experiments

A. Experimental Setup

Dataset.
Metric.
Implementation Details.

B. Results

Result on PASCAL VOC 2007.
Result on PASCAL VOC 2012.
Comparison with the referenced work.
Result on semantic segmentation.

V. Conclusions

In this paper, we propose modifications to Kim et al. [4] work to use CAM for object detection instead object localization in the original work. In addition, we experiment with Yang et al. [7] method to combine CAM from different object categories. We conduct experiments on PASCAL VOC 2007 and 2012 datasets and achieve promising results. We also show that our modifications enable CAM to detect multiple objects in a single image. In the future, we plan to focus on working on occluded objects and tight objects to improve the result.

Acknowledgment

We want to thank our advisors ... Thanh Duc Ngo and ... Dung Tien Mai for their invaluable support and guidance. We would also like to thank every lab member for their helpful comments and suggestions.

References

[1] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization." *CVPR*, 2016.

[2] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[4] E. Kim, S. Kim, J. Lee, H. Kim, and S. Yoon, "Bridging the gap between classification and localization for weakly supervised object localization," 2022.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[7] S. Yang, Y. Kim, Y. Kim, and C. Kim, "Combinational class activation maps for weakly supervised object localization," 2019.