Dual-attention Guided Dropblock Module for Weakly Supervised Object

Junhui Yin, Siqing Zhang, Dongliang Chang, Zhanyu Ma*, and Jun Guo

Beijing University of Posts and Telecommunications

Abstract: Attention mechanisms is frequently used to learn the discriminative features for better feature representations. Attention mechanisms is frequently used to learn the discriminative features for better feature representations. In this paper, we extend the attention mechanism to the task of weakly supervised object localization (WSOL) and propose the dual-attention guided dropblock module (DGDM), which aims at learning the informative and complementary visual patterns for WSOL. This module contains two key components, the channel attention guided dropout (CAGD) and the spatial attention guided dropblock (SAGD). To model channel interdependencies, the CAGD ranks the channel attentions and treats the top-k attentions with the largest magnitudes as the important ones. It also keeps some low-valued elements to increase their value if they become important during training. The SAGD can efficiently remove the most discriminative information by erasing the contiguous regions of feature maps rather than individual pixels. This guides the model to capture the less discriminative parts for classification. Furthermore, it can also distinguish the foreground objects from the background regions to alleviate the attention misdirection. Experimental results demonstrate that the proposed method achieves new state-of-the-art localization performance.

1. Introduction

Weakly supervised object localization (WSOL) requires less detailed annotations to identify the object location in a given image [1] compared to the fully-supervised learning. WSOL is a challenging task since neural networks have access to only image-level labels (``cat" or ``no cat") that confirms the existence of the target object, but not the guidance of the expensive bounding box annotations in an image.

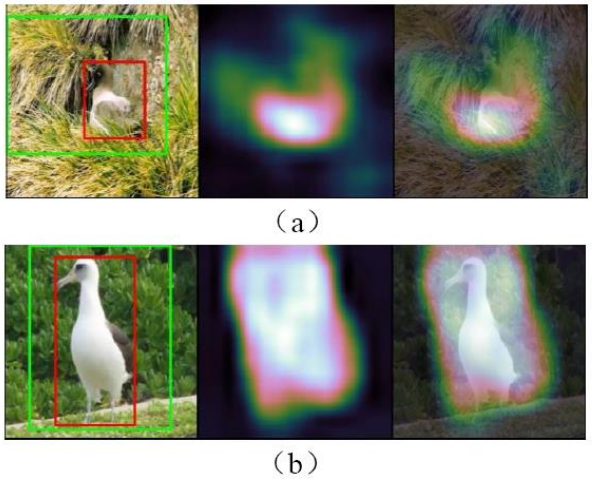


Figure. 1. Example images obtained by ResNet50-ADL. From left to right in each sub figure: the input image, the heatmap, and the overlap between the heatmap and the input image. In input image, the ground-truth bounding boxes are marked in red and the predicted are in green. The erasing operation sometimes leads to the attention spreading into the background. Meanwhile, the bounding box is too large to precisely locate the object.

Erasing the most discriminative parts is a simple yet powerful method for WSOL. For example, ADL [2] uses the self-attention mechanism as supervision to encourage the model to learn the more useful information of the object. However, the erasing methods abandon a lot of information on the most discriminative regions. This forces the model to highlight the less discriminative parts and sometimes captures useless information of the background, which leads to the attention misdirection and the biased localization. As shown in Figure, the bounding box is too large to precisely locate the object, and the classification performance is not as good as before since the focused attention has been changed to other objects.

In this paper, we propose a dual-attention guided dropblock module (DGDM), a lightweight yet powerful method, for WSOL. It contains two key components, the channel attention guided dropout (CAGD) and the spatial attention guided dropblock (SAGD), to learn the discriminative and complementary features by using the spatial and the channel attentions, respectively.

2. Method

Deep networks implemented with DGDM incorporate the image classification and WSOL. In an end-to-end learning manner, the proposed method captures the complementary and discriminative visual features for precise object localization and achieves good result of image classification.

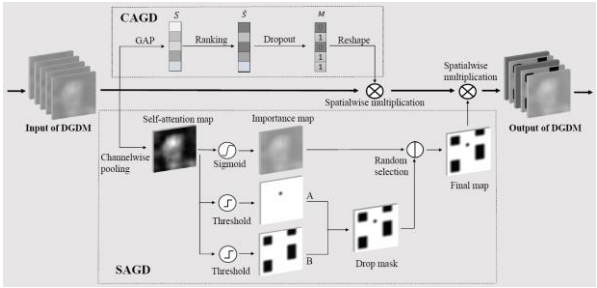


Figure. 2. Overall structure of the DGDM. It contains two key components, CAGD and SAGD. In CAGD, we rank channel attention and consider the attentions with the top-k largest magnitudes as important ones. Some low-valued elements are kept to increase their value if they become important during training. For SAGD, the drop mask can not only efficiently erase the information by removing contiguous regions of feature maps rather than individual pixels, but also sense the foreground objects and background regions to alleviate the attention misdirection. The importance map is used to highlight the most discriminative regions of target object and suppress less useful ones. Finally, we randomly select one of these two maps at each iteration and then multiply it to the input feature map. It is worth noting that this figure shows the case when the drop mask is chose.

2.1 Channel attention guided dropout

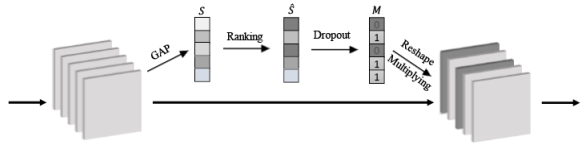
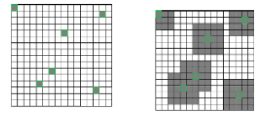


Fig. 5. Diagram of CAGD . As illustrated, we first compress spatial information of a feature map by GAP to generate channel attention. We also rank channel attention according to a fast measure of importance (magnitude), and then discard some elements with low importance. Regarding to the channel selection, the binary mask is generated to indicate whether each channel is selected or not.

2.2 Spatial attention guided dropblock

For SAGD, the drop mask can not only efficiently erase the information by removing contiguous regions of feature maps rather than individual pixels, but also sense the foreground objects and background regions to alleviate the attention misdirection.



(a) Original drop mask (b) Our drop mask

The importance map is used to highlight the most discriminative regions of target object and suppress less useful ones. Finally, we randomly select one of these two maps at each iteration and then multiply it to the input feature map.

3. Results

Method	Backbone	FLOPs (Gb)	# of Params (Mb)	Overheads Computation(%)	Overheads Parameter(%)	CUB-200-2011 Top-1 Loc (%)	CUB-200-2011 Top-1 Clas (%)
CAM	VGG-GAP	18.20	29.08	0	0	34.41	67.55
ACoL	VGG-GAP	31.98	37.63	71.51	75.71	45.92	71.90
ADL	VGG-GAP	18.20	29.08	0	0	52.36	65.27
DANet	VGG-GAP	24.12	48.56	32.53	66.99	52.52	75.40
Ours	VGG-GAP	18.20	29.08	0	0	**54.34**	69.85
ADL	ResNet50	62.32	23.92	0	0	46.29	79.72
DANet	ResNet50	74.33	32.63	19.27	36.41	51.10	81.60
Ours	ResNet50	62.32	23.92	0	0	**59.40**	76.20
CAM	InceptionV3	4.84	25.69	0	0	43.67	-
SPG	InceptionV3	31.98	37.63	560.74	46.48	46.64	-
ADL	InceptionV3	4.84	25.69	0	0	**53.04**	74.55
DANet	InceptionV3	7.23	30.62	49.38	18.47	49.45	71.20
Ours	InceptionV3	4.84	25.69	0	0	52.62	72.23

Table. 2. Quantitative evaluation results on CUB-200-2011 test set with the state-of-the-art results.

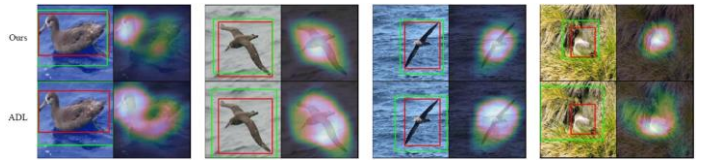


Figure. 2. Visualization results of ResNet50 on CUB-200-2011.

4. Conlusions

In this paper, we proposed a simple yet effective dual-attention guided dropblock module (DGDM) for weakly supervised object localization (WSOL). We designed two key components of DGDM, namely the channel attention guided dropout (CAGD) and the spatial attention guided dropblock (SAGD), and integrated them with the deep learning framework. The proposed method hides the most discriminative part and then encourages the CNNs model to discover the less discriminative part. We defined a pruning strategy so that CAGD can be adapted to model the interdependencies across the channels. In addition, SAGD can not only efficiently remove the information by erasing the contiguous regions of feature maps rather than the independent individual pixels, but also sense the target objects and background regions to alleviate the attention misdirection. Compared to some existing WSOL techniques, the proposed method is lightweight, and can be easily employed to different CNNs classifiers. We also have achieved new SOTA localization accuracy on CUB-200-2011, Stanford Cars, and ILSVRC.

References

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, Learning deep features for discriminative localization, in CVPR, 2016, pp. 2921–2929.
[2] J. Choe and H. Shim, Attention-based dropout layer for weakly supervised object localization, in CVPR, 2019, pp. 2219–2228.

Acknowledgements: This work was supported in part by the National Key R&D Program of China under Grant 2019YFF0303300 and under Subject II No. 2019YFF0303302.

* Please contact *mazhanyu@bupt.edu.cn* for further information.