Localizing Common Objects Using Common Component Activation Map

Weihao Li, Omid Hosseini Jafari, Carsten Rother
HCI/IWR, Heidelberg University, Germany
(weihao.li,omid.hosseini_jafari,carsten.rother)@iwr.uni-heidelberg.de

Abstract

In this work, we propose an approach to localize common objects from novel object categories in a set of images. We solve this problem using a new common component activation map (CCAM) in which we treat the class-specific activation maps (CAM) as components to discover the common components in the image set. We show that our approach can generalize on novel object categories in our experiments.

1. Introduction

Learning to classify and localize visual objects is a fundamental problem in visual recognition. The task of object localization aims to recognize the category of the main object presents in the image and locate it with an axis-aligned bounding box [11]. Recently, most of the state-of-the-art object detection or localization methods [13, 10] are trained with a strong supervised manner, which requires a large amount of human labeled bounding box annotations. However, these annotations are expensive, particularly for the large-scale datasets, such as ImageNet [11].

Currently, there have been a lot of works solving object localization task using weakly supervised setting [9, 17, 15, 16], which learn object locations in a given image only using image-level category labels. Weakly supervised object localization is getting more attention since it does not need massive bonding box annotations for training. Zhou *et al*. [17] proposed Class Activation Maps (CAM) to generate *class-specific* localization maps using classification-trained Convolutional Neural Networks (CNNs) with global average pooling. For a particular category, the class activation map shows the discriminative important image regions used by a CNN to recognize that category. However, these CAM related *class-specific* object localization methods [17, 12, 5, 15, 16] can only generate the localization maps of predefined object categories, which are not suitable for localizing the image regions for the unseen or unknown object classes.

Common object localization, also known as object co-localization, is the problem of localizing common objects of the same class across a set of distinct images [14, 3, 1, 8, 6, 7]. In contrast to weakly supervised object localization methods, the co-localization problem is not limited to predefined object categories.

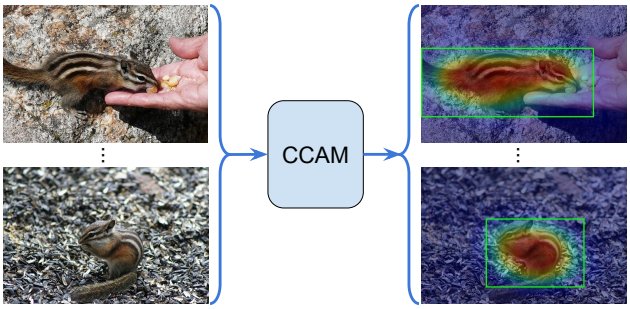

Figure 1. Common Object Localization. Given a set of images containing objects from novel classes (unseen during training), CCAM localizes the common objects in these images. Best viewed in colour

In this work, we consider both weakly supervised object localization and object co-localization to propose a simple yet effective common object localization method for unseen object categories. Unlike previous works [14, 3, 1, 8], our approach is proposal-free, which does not need any object proposals to perform object localization and only requires a CNN model with similar architecture as [17], pre-trained on a classification task. We regard the output of the last fully-connected layer as a component vector for an input object, instead of the categorical output for probability map. For a group of images, we first compute the average of the component vectors to find the group common vector. Then, we pick k largest entries from the group common vector. Finally, for each image, we compute a weighted sum of feature maps of the last convolutional layer to get the common component activation map according to the top k components. We test our method on six unseen ImageNet classes [8], which are not included in the 1000 categories used for training the CNN classification model. We show the effectiveness of our method in the result section.

	chipmunk	rhino	stoat	raccoon	rake	wheelchair	mean
[8]	44.0	81.8	**67.3**	41.8	14.5	39.3	48.1
[6]	44.9	**86.4**	56.7	**66.0**	10.3	32.4	49.5
ours	**48.2**	77.9	55.7	57.3	**46.4**	**48.6**	**55.7**

Table 1. Object co-localization on subset of ImaegeNet.

2. Method

For making the paper self-contained, we first briefly review the class activation map (CAM) for the class-specific heatmap generation, then we show how to generalize the CAM to common component activation map (CCAM) to localize the common objects.

2.1. Class Activation Map

For a specific object category, the CAM indicates the discriminative image regions used by a CNN to identify the importance of that category. Given an input image I, we first pass it through a classification network [17], which uses global average pooling on the last convolutional layer and use those as features for a fully-connected layer to produce the object categorical output. Let F represent the feature maps of the last fully convolutional layer. The size of F is $H \times W \times C$, where $H \times W$ is the spatial size and C is the number of feature channels. We denote the weight matrix of the fully-connected layer as W, in which W_c^s is the weight corresponding to class s for the channel c and indicates the importance of the channel c for the specific class s. Then, the class activation map for the class s is defined as

$$M_s(h, w) = \sum_c W_c^s F_c(h, w). \tag{1}$$

For the specific class s, $M_s(x, y)$ can directly show up the importance of the activation at the spatial grid (h, w).

2.2. Common Component Activation Map

For a given image with known categories, the CAM can identify the image regions which are most relevant to these particular categories. However, this method is incapable to find the important regions for the unseen object classes, which are not included in the training dataset. In order to generate the activation map for the unseen object, we treat the output of the fully connected layer as a component vector for the input image, instead of categorical probability maps. For a given group of N images $\mathcal{I} = \{I_1, ..., I_N\}$ containing objects from an unseen category, let the vectors $\mathcal{V} = \{V_1, ..., V_N\}$ be the outputs of the softmax function. Then we obtain the common component of the group by computing the average of output vectors $\mathcal{V}$ as

$$G = \frac{1}{N} \sum_i V_i.$$

Given the vector G, we represent $K(G)$ as a set of indices of the K largest entries. For each image I_i, we compute a weighted sum of feature maps of the last convolutional layer to get the CCAM according to the top K components.

$$M^i(h, w) = \sum_{k \in K(G)} G_k \sum_c W_c^k F_c(h, w). \tag{2}$$

To perform localization, we can generate a bounding box (see section 3.1) given the CCAM for the image I_i.

Using CCAM, we can decompose the neural activations of the common novel object into semantically interpretable components which are pre-trained with known object categories. In Fig. 2, the percentage of the contribution of each component and its corresponding known object class-specific CAM is shown.

3. Experiments

For a fair compassion with other approaches [8, 6], we evaluate the effect of our method using AlexNet [4], which is pre-trained by [17] using ILSVRC with 1000 image categories [11]. In the AlexNet, the penultimate fully-connected layer is replaced with a global average pooling layer.

3.1. Generating Boxes

To produce a bounding box from CCAM, we use a similar threshold method as [17] to segment the heatmap. In particular, we segment the regions of which the value is above a fixed threshold. In contrast to [17], we only take a single box which covers the largest connected component in the segmentation map and includes the max value of the CCAM. In our experiment, we set the threshold to 25% of the max value of the CCAM. We take top $K = 200$ components for computing the CCAM.

3.2. Evaluation Metric

Following [2, 14], we use CorLoc as the evaluation metric, which is defined as the percentage of images in which a method correctly localizes the common objects. If there is one ground-truth box of the common object having more than 0.5 intersection-over-union with the predicted box, then we count this image as a correctly localized one.

3.3. Dataset

In order to evaluate our method for unseen object categories, we follow [8] and test our method on the six subsets

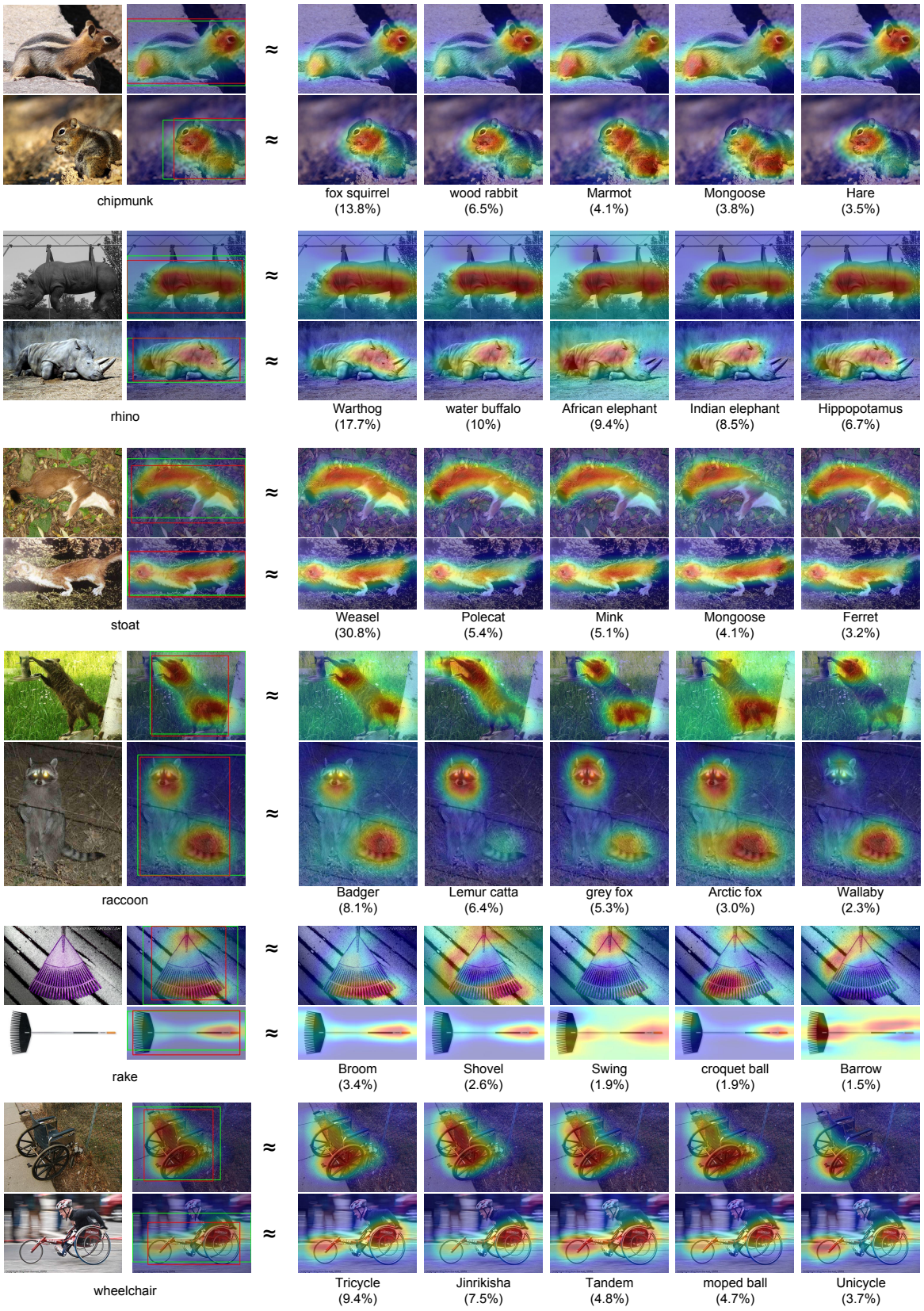

Figure 2. Viusal examples of common object localization on the subset of the ImageNet. Red boxes are ground-truth and green boxes are our predictions. Best viewed in colour.

of the ImageNet, which are not included in the ILSVRC. These unseen objects are *chipmunk*, *rhino*, *stoat*, *raccoon*, *rake*, and *wheelchair*.

3.4. Results

In Table. 1, we show the quantitative results of our method as well as the state-of-the-art approaches [8] and [6]. Clearly, our approach outperforms [8, 6] by a large margin. It is important to note that [8] use object proposal method and [6] use the over-segmentation method. Our method is proposal-free and superpixel-free. Visual examples of common object localization on the subset of the ImageNet can be found in Fig. 2. The ground-truth boxes are in red and the predicted boxes are in green.

4. Conclusion

In this work, we propose an approach to localize common objects from novel object categories in a set of images. We solve this problem by using CAMs as components instead of class-specific activation maps. As we show in the experiment section, our approach can localize novel object categories.

References

[1] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[2] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *European conference on computer vision*, pages 452–466. Springer, 2010.

[3] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision*, pages 253–268. Springer, 2014.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[5] K. Kumar Singh and Y. Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[6] H. Le, C.-P. Yu, G. Zelinsky, and D. Samaras. Co-localization with category-consistent features and geodesic distance propagation. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[7] W. Li, O. Hosseini Jafari, and C. Rother. Deep object co-segmentation. In *ACCV*, 2018.

[8] Y. Li, L. Liu, C. Shen, and A. van den Hengel. Image co-localization by mimicking a good detectors confidence score distribution. In *European Conference on Computer Vision*, pages 19–34. Springer, 2016.

[9] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[10] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations*, 2013.

[14] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[15] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang. Adversarial complementary learning for weakly supervised object localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang. Self-produced guidance for weakly-supervised object localization. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.