Weakly-Supervised Object Localization by Cutting Background with Deep Reinforcement Learning

Wu Zheng[1,2,4] and Zhaoxiang Zhang[1,2,3,4]

[1] Center for Research on Intelligent Perception and Computing, CASIA, China
[2] National Laboratory of Pattern Recognition, CASIA, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, China
[4] University of Chinese Academy of Sciences, China
{zhengwu2016, zhaoxiang.zhang}@ia.ac.cn

Abstract. Weakly-supervised object localization only depends on image-level labels to obtain object locations and attracts more attention recently. Taking inspiration from the human visual mechanism that human searches and localizes the region of interest by shrinking the view from a wide range and ignoring the unrelated background gradually, we propose a novel weakly-supervised localization method of cutting background of an object iteratively to achieve object localization with deep reinforcement learning. This approach can train an agent as a detector, which searches through the image and tries to cut off all regions unrelated to classification performance. An effective refinement approach is also proposed, which generates a heat-map by sum-pooling all feature maps to refine the location cropped by the agent. As a result, by combining the top-down cutting process and the bottom-up evidence for refinement, we can achieve a good performance on object localization in only several steps. To the best of our knowledge, this may be the first attempt to apply deep reinforcement learning to weakly-supervised object localization. We perform our experiments on PASCAL VOC dataset and the results show our method is effective.

Keywords: Weakly-Supervised Object Localization · Deep Reinforcement Learning · Convolutional Neural Network.

1 Introduction

The current state-of-art localization results come from approaches of fully supervision, such as [1–5]. Fully supervision means providing both bounding boxes and labels of objects in the image during training. However, labelling the samples is time-consuming and expensive, which limits the usability of localization task significantly. In contrast, weakly-supervised object localization [7–10,12,21] does not require annotated bounding boxes but only the image-level labels. Though such methods are usually less accurate than fully-supervised methods, it is often considered as an acceptable sacrifice to reduce dependency for annotated datasets.

Fig. 1. An simplified demonstration of our localization process.

Exciting recent weakly-supervised object localization methods [6, 13–15] has shown that the discriminative object part can be localized using class activation map (CAM) [6], which is a kind of heat-map generated by grouping class-specific convolutional feature maps. However, these methods usually generate large amounts of candidate proposals based on CAM and select the candidate with highest confidence as target location, which is time-consuming and inconsistent with human visual mechanism.

By contrast, our proposed method tries to seek a different approach to achieve weakly-supervised object localization. Taking inspiration from the human visual mechanism that human searches and localizes the region of interest by shrinking the view from a wide range and ignoring unrelated background gradually, we propose a novel weakly-supervised localization method of cutting background of an object iteratively to obtain object locations with deep reinforcement learning. We train an agent as a detector, which searches through the image and tries to cut off all regions unrelated to classification performance, as depicted in Figure 1. To achieve better localization performances, we propose a bottom-up approach that sum-pooling all feature maps to generate a heat-map for refining the cropped result. Compared with previous CAM-based methods, our approach conforms more to the human visual mechanism and can localize the object intelligently. To the best of our knowledge, this may be the first attempt to apply deep reinforcement learning to weakly-supervised object localization. We believe that it may provide a brand new perspective to address this problem. Overall, our contributions can be summarized as follows:

- We propose a novel deep reinforcement learning approach to achieve weakly-supervised object localization. Compared with previous methods, we can crop the object location quickly in several steps without generating large amount of region candidates and selecting the best one.
- We propose a new approach to refine the predictions of weakly-supervised object localization and improve the localization performance.

2 Methodology

2.1 Overview

We aim to achieve weakly-supervised object localization by training an agent with deep reinforcement learning for cutting background intelligently. Firstly,

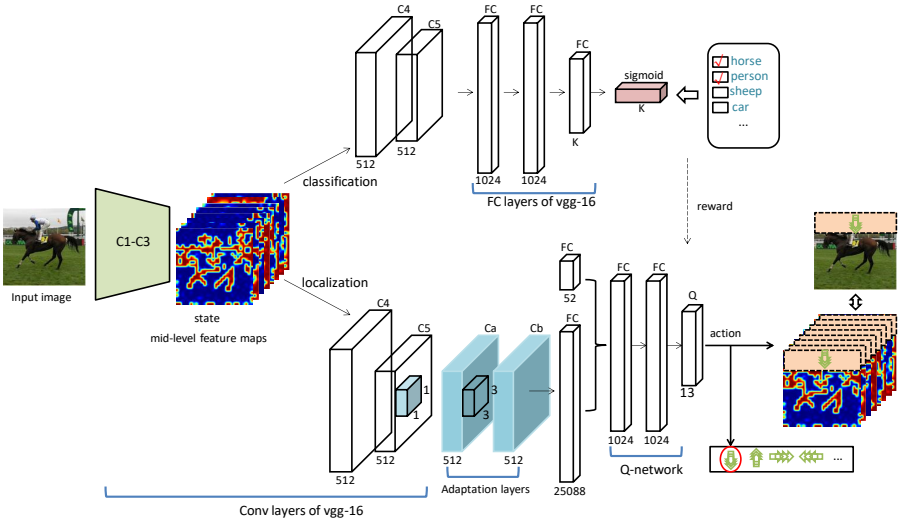

Fig. 2. Overview of the proposed weakly-supervised object localization framework.

we replace the last softmax layer of the vgg-16 pretrained on [20] with a sigmoid layer and fine-tune it for performing multi-label classification. Secondly, we train a deep Q-network (DQN) [16] for cutting unrelated background to achieve object localization. Finally, the location cropped by the agent will be refined by the heat-map generated by sum-pooling the feature maps.

2.2 Deep Reinforcement Learning for Localization

The proposed framework is depicted in Figure 2. We perform the cutting actions on the mid-level feature maps due to a trade-off between the speed and fineness of cutting. In the training process, there are two streams in our framework, the upper stream is the modified vgg-16 for multi-label classification and the lower stream is a DQN for localization, which takes advantage of the last two group convolutional layers of vgg-16 for extracting senior features. To adapt to the task of DQN, we add two convolutional adaptation layers following the final convolutional layers of vgg-16. In the testing process, we just use the lower stream to achieve weakly-supervised object localization. Next we will detail on the reinforcement learning settings.

State The state representation consists of two elements. One is a vector of senior feature information extracted from the cropped mid-level feature maps, we pad the cut-off region of mid-level feature maps with zeros to keep the aspect ratio unchanged and so avoid the dramatic decrease of classification performance. We cut feature maps on the dimensions of width and height and apply the cutting action to all channels. The other element is a vector of 4 past executed actions, which are encoded as an one-hot vector.

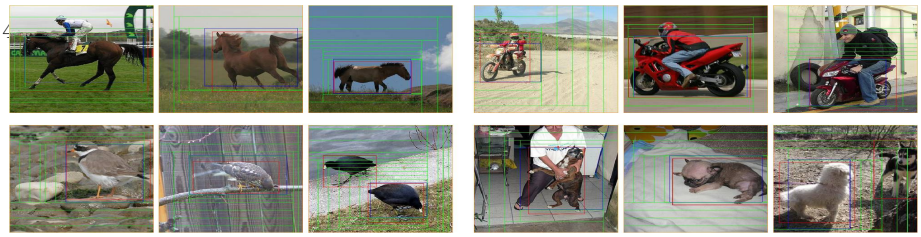

Fig. 3. Examples of object localized by the agent. The green, blue, and red lines show the cutting process, final cropped results and ground truth bounding boxes separately.

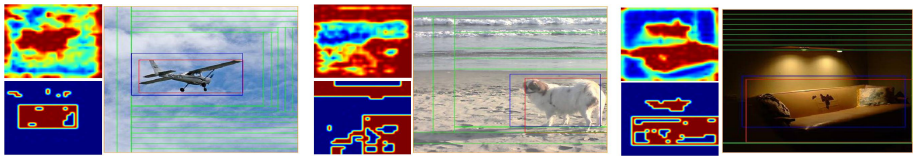

Fig. 4. Examples of refinement with heat-maps. In each group pictures, the left-top and left-bottom pictures show the original heat-map and the heat-map with the drawn maximal region. In the right picture, the smallest green rectangle and the blue rectangle represent the prediction of agent and final result refined by the heat-map, respectively.

Action There are two types of actions: cutting actions for cutting feature maps and terminal action for terminating the cutting process. The cutting actions involve different directions and scales, which means we can crop the square feature map from the four sides $\{up, down, left, right\}$ in three scales $\{\frac{1}{28}, \frac{2}{28}, \frac{3}{28}\}$. Combining multiple scales of actions helps us balance the speed and accuracy of cutting. There are two cases of terminating the cutting process, one is that the terminal action is selected by the agent, the other is that the classification score of the cropped feature maps is less than the pre-defined threshold. Thus we have totally 13 actions for agent to select. Figure 3 gives some examples of object localization process based on the action space.

Reward In weakly-supervised setting, we can roughly judge whether the object is retained or cut off only from the classification score. However, sometimes cutting off the specific background will result in a significant decrease of classification score, while cutting off parts of an object will have no effect. Therefore, we define a descent threshold of classification score to balance when to cut or stop. And we set different scales of rewards corresponding to different scales of actions, e.g. rewards=1, 2, 3 for actions=1, 2, 3, which aims to encourage large scales of cutting actions in the early stage to finish the cutting process as soon as possible, and encourage small scales of actions in the late stage to get a accurate result. Let denote the classification score of the original and the cropped feature maps as $\bar{p}(\mathrm{c}|x)$ and $\hat{p}(\mathrm{c}|x)$, respectively. Denote the threshold as δ. Then the reward function for non-termination case is

$$R(s, s') = \begin{cases} +scale, & \bar{p} - \hat{p} \leq \delta, \\ -scale, & \text{otherwise.} \end{cases} \tag{1}$$

For termination case, the reward function is

$$R(s) = \begin{cases} +\eta, & \bar{p} - \hat{p} \leq \delta, \\ -\eta, & \text{otherwise.} \end{cases} \tag{2}$$

The scalar η represents the absolute value size of reward for termination action, it is a hyperparameter.

Q-learning Q-learning is a reinforcement learning algorithm, which can train the agent to select the optimal action according to a specific state. We build a DQN to estimate the action-state value function $Q(s, a)$, which is an approximation of the expected rewards after executing the action a on current state s. The agent will choose the action corresponding to the maximal Q value. γ is a discount factor, then the $Q(s, a)$ and its update rule are denoted as

$$Q(s, a) = r + \gamma \max_{a'} Q(s', a') \tag{3}$$

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)] \tag{4}$$

2.3 Refinement

Motivated by [17], we think that the heat-map generated by sum-pooling all feature maps may provide extra information beyond classification score, which can be used to refine the result cropped by the agent. As depicted in Figure 4, we find the maximal response region of heat-map can cover the object itself very well in most cases. Thus, we firstly convert the heat-map to a binary image according and draw a rectangle to surround the area of maximal response, the thresholds used to generate binary image are evaluated on a small validation subset. Then we take the intersection between the cropped location and the rectangle of the maximal response region in heat-map as our final result.

3 Experiment

In our experiments, we firstly fine-tune the modified vgg-16 for multi-label classification on the VOC2007 trainval and VOC2012 training set, and test it on the VOC2007 [18] test set and VOC2012 [19] validation set, with results summarized in Table 1. Secondly, we train the DQN based on the trained vgg-16 and test it for localization on VOC2007 test set and VOC2012 validation set, with results summarized in Table 2 and 3. Finally, we explore the effects of our proposed refinement method, and the results are summarized in Table 4.

3.1 Network Training

We replace the last softmax layer of pre-trained vgg-16 with sigmoid layer and fine-tune it on VOC2007 trainval set+VOC2012 training set for multi-label classification. We train each agent for 50 epoches. In the training process, we use RMSProp optimizer with a decaying learning rate ranging from 1e-5 to 1e-6 in

the early 10 epoches. We use the ϵ-*greedy* policy to explore more state-action pairs and the ϵ is decayed in steps of 0.1 from 1 to 0.1 over the first 10 epoches. Discount factor γ is set to 0.95. The descent threshold δ of classification score for each class is evaluated on a small validation set.

Table 1. Average precisions of multi-label classification tested on VOC2007 test set and VOC2012 validation set.

Category	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	mAP
Accuracy-voc07	98.7	85.6	94.3	89.7	66.3	89.4	83.7	96.4	73.7	84.4	79.3	95.1	94.3	92.6	95.9	58.7	87.5	69.8	95.7	82.3	85.7
Accuracy-voc12	98.6	85.5	94.4	89.6	66.4	89.5	83.5	96.3	73.6	84.2	79.4	94.9	94.5	92.5	95.8	58.5	87.3	70.0	95.5	82.2	85.6

Table 2. Horizontal comparison of average precisions for object localization on VOC2007 test set.

Method	mAP
Wang [10]	**30.9**
Bency [17]	25.7
Gudi [15]	30.2
Ours	**30.5**

Table 3. Horizontal comparison of average precisions for object localization on VOC2012 validation set.

Method	mAP
Qquab [13]	11.7
Bency [17]	26.5
Gudi [15]	25.4
Ours	**28.8**

3.2 Localization Prediction Metric

We use the standard object detection bounding box overlap metric Intersection-Over-Union(IOU) to determine correctness of the predicted location. If the IOU between the predicted location and the ground truth bounding box exceeds 0.5, the predicted location will be labeled as correct. Otherwise, we count the prediction as a false positive and increment the false negative count. We define the confidence of the predicted location as its classification score. The average precision is calculated according to the standard algorithm.

3.3 Performance and Analysis

Classification performance Table 1 concludes the results of multi-label classification on VOC2007 test set and VOC2012 validation set. We can see that the average precisions of most categories exceed 80%, which makes a good basis for our localization experiment.

Localization performance Localization results are summarized in Table 2 and 3. The average precision of our method is computed by localizing one object of same category per image while baselines are for multiple objects detection, thus we make a horizonal comparison. We find the proposed approach achieves a competitive results with baselines, which indicates the validity of our method.

Localization refinement We show the refinement results on VOC2007 test set and VOC2012 validation set in Table 4. Our refinement method is based on the location information in heat-map. We take the intersection between this

Table 4. mAP of refinement on VOC2007 test set and VOC2012 validation set.

Dataset	no refinement	refinement
VOC2007	25.2	**30.5**
VOC2012	24.4	**28.8**

maximal region in heat-map and the location cropped by the agent as our final result, which aims to remove some background outside the rectangle of maximal response region. It bring a $4 \sim 5\%$ improvement to the performance of agent.

Table 5. Comparison of number of proposals.

Model	Proposals
SS-Based	2000
Ours	9

Proposals analysis Many recent methods are based on CAM, like [15], and use the unsupervised method like Selective Search [11] to generate large amount of candidate proposals, with each proposal to be classified separately to determine as the positive or negative sample. As shown in Table 5, compared to the methods that generate about 2000 proposals for each image, our proposed method generate average 9 proposals and do 9 times classification per image. Therefore, our approach are much faster than those based on the unsupervised method, which can save large amount of time in detection process. Besides, our search process guided by reinforcement learning is more human-like than the unsupervised exhaustive search methods.

4 Conclusion

This paper presents a deep reinforcement learning solution to weakly-supervised object localization by cutting background iteratively, which is more human-like and consistent with human visual mechanism compared with those methods depending on unsupervised region proposals. Also, by combining the top-down cutting process and bottom-up refinement, we can cut object background out intelligently only in several steps and achieve a good localization performance, which is much faster than those methods based on unsupervised generation of proposals. We believe it may provide a brand new perspective to address weakly-supervised object localization.

Acknowledgement

This work was supported in part by the National Key R&D Program of China(No. 2018YFB1004600), the National Natural Science Foundation of China (No. 61773375, No. 61375036, No. 61602481, No. 61702510), and in part by the Microsoft Collaborative Research Project.

References

1. R. Girshick. Fast r-cnn. *Computer Science*, 2015.
2. R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.
3. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.
4. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.
5. S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2015.
6. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.
7. D. Li, J. B. Huang, Y. Li, S. Wang, and M. H. Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, 2016.
8. R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):189, 2015.
9. H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. *arXiv preprint arXiv:1403.1024*, 2014.
10. C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*, 2014.
11. J. R. Uijlings, K. E. Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.
12. H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.
13. M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015.
14. T. Durand, T. Mordan, N. Thome, and M. Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *CVPR*, 2017.
15. A. Gudi, N. van Rosmalen, M. Loog, and J. van Gemert. Object-extent pooling for weakly supervised single-shot localization. *arXiv preprint arXiv:1707.06180*, 2017.
16. V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
17. A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, and B. Manjunath. Weakly supervised localization using deep feature maps. In *ECCV*, 2016.
18. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.
19. M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2012 (voc2012) results (2012). In *URL http://www. pascal-network. org/challenges/VOC/voc2012/workshop/index. html*, 2012.
20. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and F. F. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.
21. R. Gokberk Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014.