Multiple Instance Detection Network with Online Instance Classifier Refinement

Zhang, Liqiang

June 2, 2018

Abstract

Of late, weakly supervised object detection is with great importance in object recognition. Based on deep learning, weakly supervised detectors have achieved many promising results. However, compared with fully supervised detection, it is more challenging to train deep network based detectors in a weakly supervised manner. Here we formulate weakly supervised detection as a Multiple Instance Learning (MIL) problem, where instance classifiers (object detectors) are put into the network as hidden nodes. We propose a novel online instance classifier refinement algorithm to integrate MIL and the instance classifier refinement procedure into a single deep network, and train the network end-to-end with only image-level supervision, i.e., without object location information.

1. Introduction

With the development of Convolutional Neural Network (CNN), great improvements have been achieved on object detection, due to the availability of large scale datasets with accurate boundingbox-level annotations [1]. However, collecting such accurate annotations can be very labor-intensive and time-consuming, whereas achieving only image-level annotations is much easier, as these annotations are often available at the Interne.In this paper, the author aim at the Weakly Supervised Object Detection (WSOD) problem, only image tags are available during training to indicate whether an object exists in an image.

Though many promising results have been achieved in WSOD, they are still far from comparable to fully supervised ones. Weakly supervised object detection only requires supervision at image category level. Bilen and Vedaldi presents an end-to-end deep network for WSOD, in which final image classification score is the weighted sum of proposal scores, that is, each proposal contributes a percentage to the final image classification. As shown in Fig. 1 (left), the top-ranking proposal A is too small. Meanwhile, proposals B, C, and D have similar detection scores [3].

Nevertheless, forcing spatially overlapped proposals to have the same features seems too rigorous. Rather than taking the rigorous constraint, the aurhor think the features of spatially overlapped proposals are in the same manifold [4]. Then these overlapped proposals could share similar label information. As shown in Fig. 1 (right), They except the label information of A can propagate to B and C which has large overlap with A, and then the label information of B and C can propagate to D to correctly localize object.

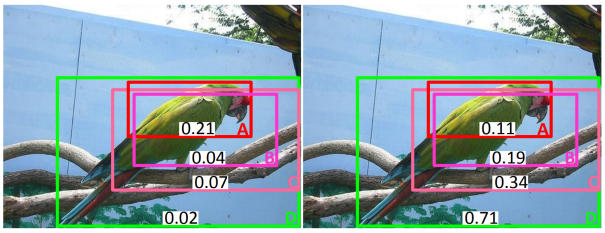

Figure 1. Detection results without/with classifier refinement (left/right). Detection scores are plotted in the bottom of the sampled proposals A, B, C, and D. In the left, the top ranking proposal A does not correctly localize the object. After instance classifier refinement, in the right, the correct proposal D is detected and more discriminative performance of instance classifier is shown.

2. Method

It is necessary to achieve instance-level supervision to train refined classifier, yet such supervision is unavailable. The top-scoring proposal by instance classifiers and its adjacent proposals can be labelled to its image label as supervision. So the author first introduce our MIDN to generate the basic instance classifier. Notice that this network is independent of special MIL methods, so any method that can be trained end-to-end could be embedded into our network.

As shown in the Multiple instance detection network block of Fig. 2, proposal features are branched into two streams to produce two matrices $x^c, x^d \in \mathbb{R}^{C\times|R|}$ of image by two fc layers, where C denotes the number of image classes and $|R|$ denotes the number of proposals. Then the two matrices are passing through two softmax layer along d-

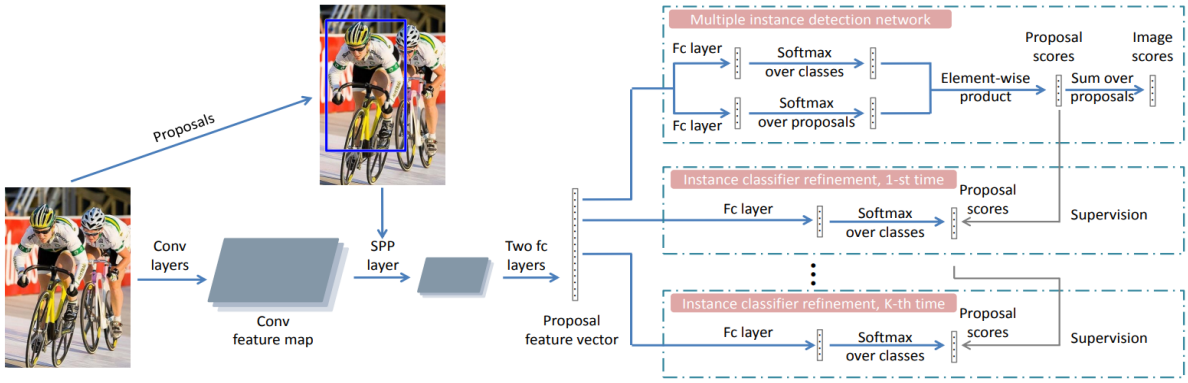

Figure 2. The architecture of MIDN with OICR. Proposal/instance feature is generated by the spatial pyramid pooling layer on the convolutional feature map of image and two fully connected layers. These proposal feature vectors are branched into many streams for different stages: the first one for the basic multiple instance detection network and others for instance classifier refinement. Supervision for classifier refinement is decided by outputs from their preceding stages. All these stages share the same proposal representations.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep
WSDDN-VGG F [2]	42.9	56.0	32.0	17.6	10.2	61.8	50.2	29.0	3.8	36.2	18.5	31.1	45.8	54.5	10.2	15.4	13.2
WSDDN-VGG M [2]	43.6	50.4	32.2	26.0	9.8	58.5	50.4	30.9	7.9	36.1	18.2	31.7	41.4	52.6	8.8	14.0	26.3
WSDDN-VGG16 [2]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	35.3
WSDDN+context	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	49.2	42.0	47.3	56.6	15.3	12.8	14.6
OICR-VGG M	53.1	57.1	32.4	12.3	15.8	58.2	56.7	39.6	0.9	44.8	39.9	31.0	54.0	62.4	4.5	20.6	34.2
OICR-VGG16	58.0	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	30.6	25.3	37.8	65.5	15.7	24.1	52.3

Table 1. Average precision (in %) for different methods on VOC 2007 test set. The upper part shows results using a single model. The lower part shows results of combing multiple models.

ifferent directions: $[\sigma(x^c)]_{ij} = \frac{e^{x^c_{ij}}}{\sum_{k=1}^{c} e^{x^c_{kj}}}$. the author train the basic instance classifier by standard multi-class cross entropy loss, as shown in Eq. 1, then the instance classifier can be obtained according to the proposal score x^R.

$$L_b = -\sum_{c=1}^{c}\{y_c log\phi_c + (1-y_c)log(1-\phi_c)\}, \quad (1)$$

3. Experiments

The author report their results for each class on VOC 2007 and 2012 in Table 1. Compared with other methods, their method achieves the state-of-the-art performance using single model, and even outperforms the results by combining multiple different models. Specially, their methods achieves much better performance than the method by Bilen and Vedaldi using the same CNN model.

References

[1] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, pages 1081–1089, 2017. 1

[2] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, pages 2846–2854, 2015. 2

[3] J. Deng, W. Dong, R. Socher, and K. Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE TPAMI*, 38(1):142–158, 2016. 1