WeBox: locating small objects from weak edges*

CHAN Sixian (产思贤)**, **LIU Peng** (刘鹏), **and ZHANG Zhuo** (张卓)
College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 31000, China

(Received 20 May 2020; Revised 28 June 2020)

In the object detection task, how to better deal with small objects is a great challenge. The detection accuracy of small objects greatly affects the final detection performance. Our propose a detection framework WeBox based on weak edges for small object detection in dense scenes, and proposes to train the richer convolutional features (RCF) edges detection network in a weakly supervised way to generate multi-instance proposals. Then through the region proposal network (RPN) network to locate each object in the multi-instance proposals, in order to ensure the effectiveness of the multi-instance proposals, we correspondingly proposed a multi-instance proposals evaluation criterion. Finally, we use faster region-based convolutional neural network (R-CNN) to process WeBox single-instance proposals and fine-tune the final results at the pixel level. The experiments have been carried out on BDCI and TT100K proves that our method maintains high computational efficiency while effectively improving the accuracy of small objects detection.

Document code: A **Article ID:** 1673-1905(2021)06-0349-5

DOI https://doi.org/10.1007/s11801-021-0085-7

Small objects are a direction that must be faced in the actual scene, and are the future direction of object detection. The existing object detection methods are successful for large and medium-sized objects, but do not perform well for small objects. For example, in the ranking of COCO detection tasks, the average accuracy of small objects (APS) is usually 2 times lower than the average accuracy of medium objects (APM) or the average accuracy of large objects (APL). Considering that there are many small objects (40%) in COCO, this problem must be solved.

Most current target detection algorithms use window scoring methods[1] to generate proposals or directly generate detection results[2,3]. These methods cannot be well adapted to multi-scale object detection. In scenes involving a large number of small objects, such as traffic signs[4-6] and pedestrian detection[7,8], connected edges can be used to locate the position of the object, but this is expensive and it is difficult to accurately mark the edge on each pixel. Therefore, we intend to use the bounding box as a weak edge to train the edge detection network to generate multi-instance proposals and summarize the proposal-by-proposal detection results.

In this paper, we propose a novel detection framework for small object detection in dense scenes to solve the above problems. We perceive the edge of the object through weak supervision, and then use the connected area analysis algorithm to generate a multi-instance proposal. The area contains multiple objects to be detected. The existing average accuracy precision (AP) is for single-instance proposals. Therefore, we used F1-score to weigh the pros and cons of the proposal, and revised the determination of the true positive (TP) for the multi-instance proposal, and then proposed a new evaluation standard for multi-instance proposals. After obtaining the multi-instance proposal, we use region proposal network (RPN) to locate a single instance in the multi-instance proposal and summarize the detection results of each proposal. Finally, we use Faster Convolutional Neural Network (Faster-CNN) to further fine-tune the final positioning results of the WeBox to adapt to the detection of small objects and general objects.

As we all know, the group proposal method shines in the early application of deep learning research, but with the development of deep learning, this method is gradually replaced. The WeBox small object detection method belongs to the updated application of the group proposal method. The specific contributions mainly include the following aspects:

We proposed a grouping proposal method based on deep learning, and proposed a new multi-instance proposal evaluation method based on bottom-up thinking;

We have designed a new network structure RCF-RPN. It can enhance the generalization ability of the network, and effectively detect adhered or adjacent objects, and has a wide range of practical application scenarios;

The verification of the authoritative BDCI and TT100K data sets fully proves the effectiveness and robustness of our proposal method.

In the existing object detection literature, most of them are for general objects, such as the classic single-stage methods YOLO[2] and Single Shot MultiBox Detector

* This work has been supported by the National Natural Science Foundation of China (No.61906168).
** E-mail: liu114850@163.com

(SSD)[3], the two-stage method Faster Region-Based Convolutional Neural Network (R-CNN)[1], etc. The solution is mainly designed for the general object data set, so for small objects in the image, the detection effect is ordinary. In recent years, more and more people have paid attention to small object detection, but most of the methods are to improve and optimize the existing object detection method.

Lin et al[9] believed that the key to small object detection is how to deal with multi-scale features, and proposed a multi-scale feature fusion method, which uses the results of different feature layer fusion to make predictions. Cai et al[10] proposed a multi-stage object detection architecture for the setting of the Intersection over Union (IoU) threshold in object detection. The detector is composed of detectors trained by the continuously improved IoU threshold, so that it can select close false positive sequences. Li et al[11] improved the detection of small objects by reducing the representation difference between small objects and large objects, and proposed a new model of perceptual generative adversarial networks (GAN). Li et al[12] used the backbone network as an entry point. In view of the shortcomings of large objects returning to weak and small objects difficult to find in the existing network, a backbone network dedicated to detection tasks was proposed. Good results have been obtained in the detection of small objects. At present, most mainstream methods focus on and how to narrow the difference between small objects and general objects. For the weak edge information that is easy to obtain for small objects, we propose to connect them into multi-instance proposals, and then locate small objects.

For small object detection, we propose a WeBox detection framework, which aims to effectively process small objects that are currently difficult to detect by learning the weak edge feature information of the object. The flow of this method is shown in Fig.1. It mainly includes three parts: the extraction of weak edge information, the generation of multi-instance proposals, and the finer detection of each proposal.

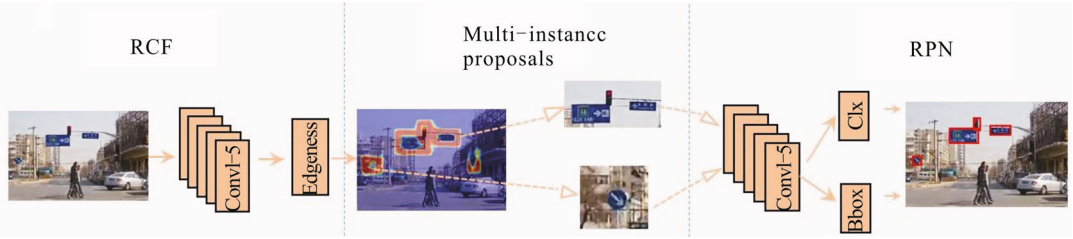

Fig.1 The pipeline of the WeBox approach

The edges of objects are concise, and the edges of connected objects form contours, which provide good semantic information at the individual level. Therefore, the gathered edges can be used for object detection. Object detection tasks often only have rough bounding boxes as supervision information. We use them as weak edges to perform object detection, and let the entire deep convolutional network learn to predict rough object edges. Because the RCF[13] network has richer feature expressions, it can obtain multi-scale information better, so that a weak edge detection model with better characterization ability can be obtained. We regard the labeled bounding box of the object as a weakened edge to train the RCF network. After the obtained edge map is visualized in the form of a heatmap, a significant target object area is observed, and the background area hardly produces a response. According to this feature, we directly obtain the rough target area on the edge map generated by the RCF network.

Most of the existing network structures predict the classification results in the last layer. The fusion of the network is easily affected by the gradient divergence phenomenon, and the last layer is also susceptible to excessive update, which affects the network performance. Therefore, we propose prediction classification and calculate the loss function after different depths of the network layer, which can generate regulatory information at different depths. The loss function of the RCF network in the training phase is the weighted sum of multiple auxiliary loss functions and the loss function on the fusion feature, which is expressed as

$$L = L_{\text{fuse}} + \sum_{m=1}^{M} \alpha_m L_{\text{side}}^{(m)}(W, w^{(m)}), \tag{1}$$

where L_{fuse} is the loss value of the fusion feature, L_{side} is the loss value of each branch, W is the sum of the output weights, M is the order of the network, α_m is the coefficient set in each stage, and each branch occupies a different proportion. Each branch output performs pixel classification. Because of the imbalance between categories, the cost-sensitive cross-entropy loss function is used here:

$$L_{\text{side}}^{(m)}(W, w^{(m)}) = -\beta \sum_{j \in Y_+} \log(y_j = 1 \mid X; W, w^{(m)}) - (1-\beta) \sum_{j \in Y_-} \log(y_j = 0 \mid X; W, w^{(m)}), \tag{2}$$

where β is the balance factor, and Y_+ and Y_- represent the edge and non-edge truth value set.

By cascading a single object proposal method to a weakly supervised edge network, it will generate a single proposal with better positioning accuracy. In addition, because labeling the bounding box is much easier and

lower cost than labeling the edge by pixel, our WeBox method fully utilizes the advantages of the edge detection network and the object proposal network, and achieves better performance for small object detection tasks.

The RCF weak edge detection network processes the input image and outputs a weak edge map of the same size. Since we use the bounding box as the edge of the detected object, it is not accurate from the pixel level, but relatively accurate from the object perspective. According to the obtained weak edge graph, we use the connected region analysis algorithm to separate the candidate object detection region. It includes the following steps:

(1) Binarization of weak edge map: used to filter out single or very small (less than 10 pixel×10 pixel) areas and low confidence areas;

(2) Determine the connected area: scan the entire binarized weak edge map, and use the union data structure to mark the connected pixels as the same area number. The number starts from 1 and the maximum number indicates the number of connected areas determined;

(3) Determine the minimum enclosing frame: Use the connected area number information in (2) to calculate the maximum and minimum coordinates of the horizontal and vertical directions in each connected area respectively as the descriptor of the minimum enclosing frame of the area.

We have noticed that two or more detection objects that are close to each other or have occlusion will cause adhesion on the weak edge map, then the rough candidate detection area determined by the above (1)-(3) contain one or more target instances. This kind of sticking is unavoidable, because the weak edge network will generate corresponding values at the edges of the object to be detected and around it. In order to ensure that the connected edges form a closed curve as far as possible to contain the complete object, the distance Weak edges between nearby objects will be retained. Through this method, a large amount of background area is removed, which can be regarded as a rough proposed area. Considering that in the work related to object detection, the default proposal area contains only a single instance, we let go of this limitation and generate a multi-instance proposal, and then provide a new evaluation criterion for multi-instance proposals.

After getting the multi-instance proposal containing objects, we use it as the input of the RPN, and use the real bounding box in the region as the positive sample label of the network. The RPN can predict single-instance proposals with higher APs for the multi-instance proposals generated by RCF weak edges, and the detection results on each multi-instance proposal are summarized to obtain the single-instance proposal of RCF + RPN. We consider that the performance of RPN can be further improved after being equipped with Faster R-CNN, so we use Faster R-CNN to refine the location and category of target instances in the region on the multi-instance proposal generated by the RCF network.

The experiment consists of three parts, the first is the introduction of data sets and evaluation methods. The second is the ablation experiment, which evaluates the quality of the multi-instance suggestion region generated by the weak edge model and the region-by-region test results. Finally, there is a comparison with the existing objects detection algorithm.

We used two datasets, BDCI and TT100K. BDCI's rematch dataset, a total of 10 000 images, the image size is 720×1 280. Each image contains an average of 5 objects; the number of small objects is 70% of all object instances. We randomly used 8 000 training sessions and 2 000 tests. The Tsinghua-Tencent100K dataset also contains small-sized traffic signs. Contains about 9 000 images, each with a size of 2 048×2 048. The average image contains 7 small objects, and the small object number is 35%. Of these, 6 088 images were for training and the remaining 3 055 images for testing.

The proposals were divided into two categories: multi-instance proposals and single-instance proposals. The existing proposals algorithm was for single instance, and its evaluation index AP was not suitable for the multi-instance proposals. In view of this situation, we proposed an evaluation protocol suitable for the multi-instance proposals. In the case where the multi-instance proposal contained multiple truth-valued mark areas, the IoU of each truth-valued area and the recommended area was likely to be less than the threshold. According to the original standard, it could be regarded as false detection, but in fact the proposal detected the object area in a looser way. Therefore, as long as one or more true value regions were included, it was regarded as True Positive (TP), which was a criterion applicable to the multi-instance suggestion region, and was referred to herein as "including rules".

The multi-instance proposals were determined by the RCF weak edge without a confidence score. Hence, the P-R curve and the AP corresponding to the P-R curve are not suitable. Using F1-score to weigh the Precision and Recall on the unordered result was a better approach, the higher the score corresponding to the better the performance of the algorithm. We used F1-score to evaluate the quality of a multi-instance proposals, where the TP decision rules used the newly proposed inclusion rules.

We separately evaluate the quality of the multi-instance recommendation area generated by the RCF weak edge network and the RPN network. The evaluation criteria are the evaluation rules introduced in 4.1. In the experiment, two RCF networks were compared, and the basic learning rates were 1e-5 and 1e-6, respectively, denoted RCF_B5 and RCF_B6. Considering that the RPN performance in the prediction phase is affected by the TopN parameters, the top 20 and 50 proposals are selected for evaluation. Tab.1 shows the Recall, Precision, and F1 scores obtained after the evaluation of each of the four schemes. Among them, the RCF_B5 and RCF_B6 have the highest F1 scores, respectively 65.52% and 73.92%. According to the

evaluation criteria proposed in this paper, RCF_B6 has the best performance; Under different TopN values of the RPN network, the F1 scores are all below 20%, and the performance is poor. In addition, note that the F1 score is the harmonic average of precision and recall. Only when precision and recall scores are close and both are high, can you get a high F1 score.

Tab.1 Evaluation results for different scenarios of RCF and RPN

	TopN	Base_lr	Recall	Precision	F1
RCF_B5	-(<20)	1×10^{-6}	91.34	51.09	65.52
RCF_B6	-(<20)	1×10^{-5}	86.22	64.7	73.92
RPN_T20	20	1×10^{-3}	25.80	9.37	13.74
RPN_T50	50	1×10^{-3}	72.01	10.46	18.27

In this part, we mainly explore the performance of the branch loss function on the RCF weak edge network under the multi-instance proposal evaluation standard. In the experiment, the two schemes are compared. RCF-DSN and RCF-L5. There is also a loss function on the features gathered at each stage. RCF-L5 removes the branch loss function from stage 1 to stage 4, and only retains the loss function of stage 5 and converging features. RCF-DSN and RCF-L5 have the same training conditions except for different loss functions in the network structure, including using 1e-6 as the initial learning rate, batch size of 12, and performing 60 000 iterations. Tab.2 shows the experimental results. As the number of iterations increases, Recall decreases while Precision and F1-score increase, and is basically stable after 30 000 iterations. In terms of F1-score, RCF-DSN is always higher than RCF-L5, so the multi-loss network structure of DSN has a higher score on the F1-score of the multi-instance recommendation area.

Tab.2 Effect of branch loss function on RCF

	RCF-DSN			RCF-L5		
Iter	Recall	Precision	F1	Recall	Precision	F1
5k	96.09	32.46	48.52	97.98	27.21	42.59
10k	95.43	33.54	49.64	96.42	28.35	43.82
15k	93.76	39.55	55.63	94.74	35.04	51.16
20k	90.53	57.74	70.51	92.94	42.97	58.77
25k	86.73	63.36	73.22	89.58	57.34	69.92
30k	89.29	61.99	73.17	88.60	59.92	71.49
35k	85.69	65.20	74.06	88.19	60.80	71.19
40k	89.25	62.05	73.20	87.57	61.87	72.51
45k	85.01	66.24	74.46	87.35	62.24	72.69
50k	89.22	62.06	73.20	87.28	62.32	72.72
55k	85.03	66.20	74.40	87.20	62.31	72.68
60k	89.22	62.04	73.19	87.11	62.37	72.69

We compare our proposed method with the current mainstream methods on the T100K datasets. The results in Tabs.3 and 4 show that our method greatly improves the detection accuracy of small target objects while ensuring robustness.

The results in Tab.4 indicate that our proposed method has achieved excellent performance in both AP and AP^S. The Webox achieving 20.6 of AP, and has significantly better precision than other algorithms under Recall conditions greater than 0.45, mainly because the algorithm is based on RCF weak edge, which can achieve a close to 50% F1 score with a small number (less than 20, average of 6) of multi-instance suggestion regions, so the entire algorithm can achieve a higher AP.

Tab.3 Comparisons of final results on TT100K dataset

Method	Backbone	AP^S	AP	AP^{50}
Faster R-CNN[1]	VGG16	9.7	34.6	76.6
LOCO[14]	VGG16	31.7	51.5	91.7
WeBox	VGG16-Resnet50	43.8	38.2	76.3

Tab.4 Comparisons of final results on BCDI dataset

Method	Backbone	AP^S	AP	AP^{50}
Faster R-CNN[1]	VGG-16	7.0	12.1	37.8
Faster R-CNN[1]	RF_VGG16_atrous	14.1	17.2	51.2
LOCO[14]	VGG-16	9.3	12.9	32.9
WeBox	VGG16-Resnet50	17.7	20.6	52.4

In this paper, a new object detection framework WeBox was proposed. By learning and predicting weak edges, it could effectively cope with small and general size targets. The multi-instance proposals were further accurately detected and classified to obtain the result of the whole image. The whole algorithm could be used as a proposal algorithm or as a multi-class detection algorithm, and only needed to specify the second network separately. WeBox better solved the weakness of the proposed regional algorithm based on regional scoring, and naturally detected the existence of small objects.

References

[1] Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun, IEEE Transactions on Pattern Analysis and Machine Intelligence **6**, 1137 (2017).

[2] Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi, IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu and Alexander C. Berg, SSD: Single Shot MultiBox Detector, European Conference on Computer Vision, 21 (2016).

[4] Junqi Jin, Kun Fu and Changshui Zhang, IEEE Transactions on Intelligent Transportation Systems **15**, 1991 (2014).

[5] Tam T. Le, Son T. Tran, Seichii Mita and Thuc D.

Nguyen, Real Time Traffic Sign Detection Using Color and Shape-Based Features, Asian Conference on Intelligent Information and Database Systems, 268 (2010).

[6] Yingying Zhu, Chengquan Zhang, Duoyou Zhou, Xinggang Wang, Xiang Bai and Wenyu Liu, Neurocomputing **214**, 758 (2016).

[7] Yonglong Tian, Ping Luo, Xiaogang Wang and Xiaoou Tang, Pedestrian Detection Aided by Deep Learning Semantic Tasks, IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[8] Liliang Zhang, Liang Lin, Xiaodan Liang and Kaiming He, Is Faster R-CNN Doing Well for Pedestrian Detection? European Conference on Computer Vision, (443) 2016.

[9] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan and Serge Belongie, Feature Pyramid Networks for Object Detection, IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[10] Zhaowei Cai, Nuno Vasconcelos, Cascade R-CNN: Delving into High Quality Object Detection, Computer Vision and Pattern Recognition, 2017.

[11] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng and Shuicheng Yan, Perceptual Generative Adversarial Networks for Small Object Detection, Computer Vision and Pattern Recognition, 2017.

[12] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Den and Jian Sun, DetNet: A Backbone Network for Object Detection, Computer Vision and Pattern Recognition, 2018.

[13] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Jia-Wang Bian, Le Zhang, Xiang Bai and Jinhui Tang, IEEE Transactions on Pattern Analysis and Machine Intelligence **41**, 1939 (2019).

[14] Peng Cheng, Wu Liu, Yifan Zhang and Huadong Ma, LOCO: Local Context Based Faster R-CNN for Small Traffic Sign Detection, International Conference on Multimedia Modeling, (329) 2018.