Small Object Detection Using Context Information Fusion in Faster R-CNN

Pengcheng Fang
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: pengchengfang@bupt.edu.cn

Yijie Shi
State Key Laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: yijieshi2000@bupt.edu.cn

***Abstract*—Currently, most of the object detection research focuses on detecting a big object covering large part of the image. The problems of detecting the small object covering small part of the image are largely ignored. The difficulty of small object detection is that small objects have large quantity and less pixel (less information) and cover small part of the images. In this paper, we aim at improving the accuracy of small object detection. Firstly, we use a subset of the COCO [1] dataset to build a benchmark database. This benchmark database is specifically designed to evaluate the performance of object detection algorithms on small object detection. Secondly, we improve Faster R-CNN [2]. The improvements include a more flexible context information integration method. The experiments show that the improved Faster R-CNN algorithm has a good performance on the accuracy and recall rate of small object detection. Our small object detection algorithm is able to strike the balance between detection speed and detection accuracy.**

Keywords-small object detection; context information; neural network

I. INTRODUCTION

In the past decade, with the development of deep learning and convolutional neural networks, the field of object detection has made great progress, but there are still many challenges. Small object which covers a small part of an image is difficult to be detected due to its' lower resolution and more noise. All in all, the detection accuracy of small object is low.

The mainstream object detection algorithms include single-stage detector like SSD [3] and YOLO [4] and two-stage detector like Faster R-CNN. Both of these two types of object detection algorithms have limitations in detecting small objects.

SSD relies on the anchor box to train the network and each pixel on each feature map corresponds to several anchor boxes. For small object, there are few effective corresponding anchor boxes and ground truth box and anchor box have less than 0.5 overlap. As result, it's difficult for small object to carry on sufficient training. The large regions of interest may cover many anchor boxes, and these anchor boxes have the opportunity to be trained, while the small object does not cover a lot of anchor boxes and can not be fully trained. YOLO gets the prediction through the global feature characteristics, which completely relies on the data accumulation and has bad performance in small object detection.

The essential reason for the low accuracy of small object detection is that the small objects have less effective information and weakly features extracted from neural network. To improve the accuracy of small object detection, we must enhance the features extracted from small objects. In this paper, we proposed a more flexible contextual information fusing method that only fuses the classification information with the region proposal windows. Instead of using both classification information and bounding box information. This method can enhance the efficiency of region proposal's classification information without introducing the bounding box regression error. Faster R-CNN algorithm is improved by adding more flexible context information fusion method to improve the accuracy of small object detection.

The rest of this paper is organized as follows. Section 2 briefly reviews the related work and state-of-the-art approaches. Section 3 explains the details of the proposed methods for improving the small object detection performance. Section4 describes the utilized dataset, experiments, and results. Finally, Section 5 concludes the paper with the summary and discussion.

II. RELATED WORK

Benefited from the rapid development of deep convolutional networks [5, 6, 7], the great progress has been made for the object detection problem. R-CNN [8] is the first to utilize deep neural network features into detection system. Hand-engineered methods, such as Selective Search [9] and Edge Boxes [10], are involved to generate proposals for R-CNN. Then Fast R-CNN [11] is proposed to join train object classification and bounding box regression, which improves the performance by multi-task training. Following Fast R-CNN, Faster R-CNN introduces Region Proposal Network (RPN) to generate proposals by using network features. Because of richer proposals, it marginally performs the higher accuracy. Faster R-CNN is regarded as a milestone of R-CNN serials detectors. Most of the following works strengthen Faster R-CNN.

Generally, context is useful for improving the object detection performance in natural scenes [14, 15]. The approach which is proposed in [17] uses both segmentation and context to improve object detection accuracy. The approach which is proposed in [18] studies the role of context in existing object detection approaches and further proposed a model that exploits both the local and global context. In our work, we also leverage the context

978-1-5386-8339-2/18/$31.00 ©2018 IEEE

information to get better performance for small object detection. We propose a more flexible context information fusion method that does well in improving the accuracy of object detection and does better in small object detection.

R-CNN for small object detection [16] proposes a context information fusion method which is shown in Figure 1. The candidate window is up-sampled 2 times, 3 times or 5 times directly, and then this up-sampling window is put through the convolutional neural layers and the fully connected layers to get the 1024-dimensional feature. Finally, the 1024-dimensioanl feature extracted from the candidate window and the 1024-dimensional feature extracted from the up-sampling window are concatenated into 2048-dimensional feature. This 2048-dimensional feature is used for classification and position regression. The current context information fusion method needs to extract the features of the region proposals and the features of the context information windows respectively. Since the context information window contains the region proposals, there are amount of repeated calculations in the process of extracting features. This 2048-dimensional feature not only includes the classification information and bounding-box regression information of region proposal, but also includes the classification information and bounding-box regression information of up-sampling window. Even if the classification information of the region proposal is enhanced to some extent, the bounding box regression information has been greatly weakened. A large amount of regression errors will lead to more inaccurate detection and positioning of objects. Thus the accuracy of small object detection does not increase but decreases. Even worse, the features which are fed into the convolutional neural network head have very high dimension (2048), the computational cost is increased, and the model becomes more complicated and difficult to train.

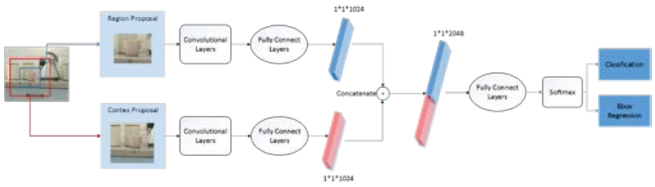

Figure 1. Common context information fusion method.Using two sub-networks to extract region proposal feature and context window feature respectively, the back-end sub-network takes in the concatenation of the two feature vectors and computes the final classification and regression.

III. Flexible and Efficient Integration of Context Information

First, in real life, an object can not exist alone and it must has more or less correlation with other objects or surrounding environment. That is commonly referred to as context information. Small objects are relatively simple in shape and usually occupy a small image region and the features extracted from proposal regions through convolutional neural network are usually low resolution, less discriminative and large noise. Context information plays an important role in improving the performance of small object detection.

Based on Faster R-CNN algorithm, we propose a more flexible context information fusion approach, which does not take the image to frame the region proposal windows and context information windows respectively. We put the image into the convolution neural network body for feature extraction only once to obtain the complete feature maps in which the dimension of the complete feature maps is w*h*1024. Region Proposal Network [2] (RPN) is used to generate region proposals. Before feeding the region proposals into neural network head, non-maximum suppression (NMS) is used to reduce the number of proposals.

Then the context information fusion method is performed on the selected region proposals. As shown in Figure 2. We crop eight corresponding context region enclosing the proposal region. The eight corresponding context regions are set to be the same size as the region proposal, and named as Top-Left, Top-Middle, Top-Right, Middle-Left, Middle-Right, Bottom-Left, Bottom-Middle, Bottom-Right (in Figure 3). These eight contextual child-windows need to be discriminated whether they cross the boundary of the feature map. For example, the Top-Left window contains information that marks its window position (x, y, width, height), where (x, y) is the coordinate representation of the point in the center of Top-Left window and (width, height) is the width and height of the Top-Left window.

Figure 2. Flexible Context Information Fusion Method.

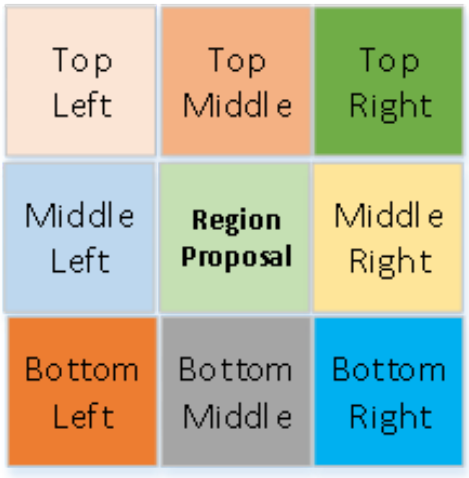

Figure 3. The eight context information child windows.

In Figure 4, we give the spatial position relation and coordinate description among the Top-Left context window, the region proposal and the feature map. Then it can be calculated whether the Top-Left child window crosses the boundary of the feature map, and the calculation process is

shown below. Definitely, the region proposal must be located in the feature map, the point B must be in the feature map, and we only need to determine whether the point A (x – height / 2, y – width / 2) is in the feature map. In this way, it can be determined whether the Top-Left context information child window is out of boundary. The other seven child windows can also be calculated in the same way. If a child-window crosses out of the boundary of the feature map, we will discard the whole window directly and do not add its' context information to the region proposal.

We add the valid context windows to the region proposals and only add the classification information instead of adding full information which includes classification information and regression information. In this way, the classification information of the regional proposals is enhanced without introducing the bounding box regression error.

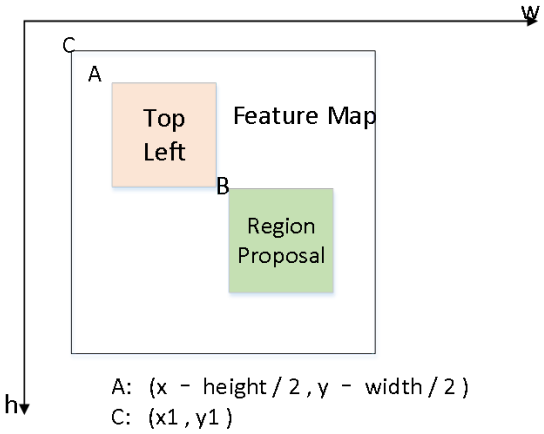

Figure 4. Coordiante relationship among context information child-window, region proposal and feature map.

IV. Performance Analysis

In this section, we review details of the utilized dataset, evaluation metrics, and the conducted experiments. The experiment results are discussed in detail to verify the validation.

A. The Dataset

TABLE I. Data Analysis of Small Object Dataset.

Category	Number of Instances	Average relative area
Knife	5536	0.024765
Bottle	16983	0.013355
Wine Glass	5618	0.025976
Cup	14513	0.021304
Spoon	4287	0.021943
Fork	3918	0.031746
Bowl	10064	0.080524
Sports Ball	4392	0.004020
Orange	4597	0.037154
Vase	4623	0.046501

COCO is used as the base dataset which is famous for rich small objects. We pick out the smaller object categories (Knife, Fork , etc.) that are really small and easy to cluster from the complex scene, such as kitchen, to compose the subset. The dataset contains the training set, the test set, and the validation set. The relative areas of small objects and big objects in COCO are compared in Table I. and Table II. The detail of the dataset is showed in Table I.

TABLE II. Data Analysis of Big Objects in COCO Dataset.

Category	Average relative area
Bed	0.024765
Train	0.013355
Dog	0.025976
Refrigerator	0.212214
Horse	0.127856
Oven	0.170634
Elephant	0.162298
Table	0.356729

Object detection is evaluated by the COCO-style Average Precision (AP). We evaluate the COCO-style mean Average Precision (mAP) and AP on small objects (APs). The COCO Average Recall (AR) and AR on small objects (ARs) are evaluated following the definitions in [1].

B. The Comparative Experiment of Context Information Fusion

Faster R-CNN with FPN is used as the basic object detection algorithm to obtain the baselines of object detection accuracy and recall rate.

The context region of the 1x is one time larger than the proposal region in both height and width dimension. The 2x model is defined in a similar way and it uses much larger context region. The impact of using different context fusion methods are shown in Table 3 and Table 4. It can be found that the models with context integration method achieve better performance than the baseline model (without leveraging context information). The relative mAP improvement over the baseline is 1% and 2% for the 1x and 2x context information directly fusion methods, respectively. Integrating two times of context information introduces more noise and errors. Therefore, the detection accuracy of object detection after introducing 2 times of context information does not rise but decreases. At the same time, neither 1x and 2x context information directly fusion method have almost no influence on the recall rate of object detection.

TABLE III. Object Detection Accuracy Using Faster R-CNN Evaluated on the COCO Minival Set. Models Are Iterated 6k Times and Use ResNet-50.

Method	*mAP*	*APs*
Faster R-CNN	16.24	3.25
Faster R-CNN + Context(1x)	17.13	5.49
Faster R-CNN + Context(2x)	18.03	3.16
Faster R-CNN + Context our	21.24	11.16
Faster R-CNN + Context our + FPN	**23.21**	**12.3**

Additionally, the more flexible context information fusion approach which is proposed in Section III. is compared with the 1x and 2x context information directly fusion methods. In our experiments, under the same basic setting, the flexible contextual information fusion approach gets higher accuracy and recall rate in small object detection (Fig. 5.). The baseline model with flexible context information fusion method improves mAP to **23.21** (Table III.), which increases 6.97 points over the single-scale RPN baseline (Table III.). In addition, the performance on small objects (APs) is boosted by **9.05** points. The detection recall rate is compared in Table IV. and the proposed context fusion approach improves AR100 to **39.1** and ARs to **28.9**.

TABLE IV. OBJECT DETECTION RECALL RATE USING FASTER R-CNN EVALUATED ON THE COCO MINIVAL SET. ALL MODELS ARE TRAINED ON THE TRAIN SET.FPN USE RESNET-50.

Method	***AR100***	***ARs***
Faster R-CNN	29.1	13.0
Faster R-CNN + Context(1x)	36.2	23.7
Faster R-CNN + Context(2x)	35.7	24.6
Faster R-CNN + Context our	38.4	25.4
Faster R-CNN + Context our + FPN	**39.1**	**28.9**

V. CONCLUSIONS

In this paper, we confirm that flexible context information fusion method is essential to improve the accuracy of small object detection.

Moreover, the method proposed in this paper does not exploit many popular improvements, such as iterative regression online hard negative mining [19] and soft NMS [20], etc. These improvements may be complementary to our approach and should improve accuracy further.

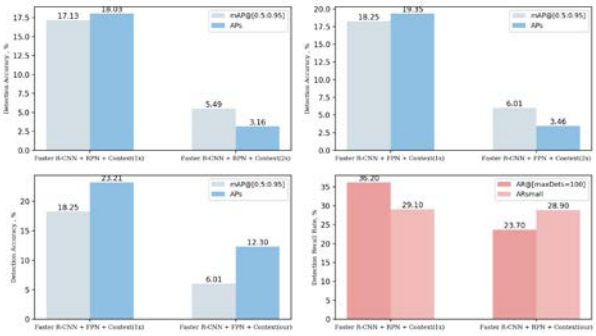

Figure 5. The horizontal axis represents the method, and the vertical axis represents the detection accuracy or recall rate.

ACKNOWLEDGMENT

This work is supported by Detectorn (Detectron is Facebook AI Research's software system that implements state-of-the-art object detection algorithms, including Mask R-CNN. It is written in Python and powered by the Caffe2 deep learning framework.).

REFERENCES

[1] Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollar, P.,Zitnick, C.L.: Microsoft coco: Common objects in context. european conference on computer vision (2014) 740–755

[2] Ren, S., He, K., Girshick, R.B., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 1137–1149

[3] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: Ssd:Single shot multibox detector. european conference on computer vision (2016) 21– 555 37

[4] Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. computer vision and pattern recognition (2016) 779–788

[5] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. (2012) 1097–1105

[6] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. international conference on learning representations (2015)

[7] He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. european conference on computer vision (2016) 630–645

[8] Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. computer vision and pattern recognition (2014) 580–587

[9] Uijlings, J.R.R., De Sande, K.E.A.V., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. International Journal of Computer Vision 104 (2013) 571 154–171

[10] Zitnick, C.L., Dollar, P.: Edge boxes: Locating object proposals from edges. (2014) 391–405

[11] Girshick, R.B.: Fast r-cnn. international conference on computer vision (2015) 1440–1448

[12] Lin, T., Dollar, P., Girshick, R.B., He, K., Hariharan, B., Belongie, S.J.: Feature pyramid networks for object detection. computer vision and pattern recognition 577 (2017) 936–944

[13] Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. computer vision and pattern recognition (2017) 6517–6525

[14] Divvala, S.K., Hoiem, D., Hays, J., Efros, A.A., Hebert, M.: An empirical study of context in object detection. (2009) 1271–1278

[15] Torralba, Murphy, Freeman, Rubin: Context-based vision system for place and object recognition. (2003) 273–280

[16] Chen, C., Liu, M., Tuzel, O., Xiao, J.: R-cnn for small object detection. (2016) 214–230

[17] Zhu, Y., Urtasun, R., Salakhutdinov, R., Fidler, S.: segdeepm: Exploiting segmen-and pattern recognition (2015) 4703–4711

[18] Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R.,Yuille, A.L.: The role of context for object detection and semantic segmentation in the wild. (2014) 891–898

[19] Shrivastava, A., Gupta, A., Girshick, R.B.: Training region-based object detectors with online hard example mining. computer vision and pattern recognition (2016) 761–769

[20] Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms — improving object detection with one line of code. international conference on computer vision (2017) 5562–5570