A Solution for Product Detection in Densely Packed Scenes

Jun Yu, Haonian Xie, Guochen Xie, Mengyan Li, and Qiang Ling
Department of Automation, University of Science and Technology of China
{xie233, xiegc, limmy}@mail.ustc.edu.cn,{harryjun, qling}@ustc.edu.cn

1. Task Introduction

The task of the challenge is to detect products in crowded store displays. This challenge is based on the SKU-110K dataset [5]. The SKU-110K dataset collects 11,762 densely packed shelf images from thousands of supermarkets around the world, including locations in the United States, Europe, and East Asia.

SKU-110K images are partitioned into train, test, and validate splits. Training split consists of 70% of the images (8,233 images) and their associated 1,210,431 bounding boxes; 5% of the images (588), are used for validation (with their 90,968 bounding boxes). The rest, 2,941 images (432,312 bounding boxes) are used for testing. Images were selected at random, ensuring that the same shelf display from the same shop does not appear in more than one of these subsets.

Finally, all methods will be evaluated on a new test set, which is published without annotations. The evaluation metrics is similar to those used by COCO [7], reporting the average precision (AP) at IoU=0.50:0.05:0.95.

2. Method

The overall pipeline of our solution is shown in Figure 1. Our solution is consist of two models. They both are based on Adaptive Training Sample Selection (ATSS) [14] , which can automatically select positive and negative samples according to statistical characteristics of object. The difference between them is the backbone and neck, which is the part that connects the backbone and detection head. The mdoel 1 adopts HRNet-W32 [11] as backbone, HRFPN [11] as neck and ATSS as detection head. The mdoel 2 employs Res2Net-101 [4] as backbone, Balanced FPN [8] as neck and ATSS as detection head. The ATSS head including a centerness branch [12], a regression branch, a classification branch, GroupNorm [13] and a trainable scalar for each level feature pyramid. We adopt Soft-NMS [1] as post-processing to obtain bounding boxes. To improve the performance of detector, we employ the Weighted Boxes Fusion (WBF) [10] to ensemble these bounding boxes detected by Model 1 and Model 2.

During training, we adopt the DIoU loss [15] as regression loss function and Focal loss [6] as classification loss function.

3. Experiments

We compare the detection accuracy of our proposed solution and recent state-of-the-art on the SKU-110K benchmark. All experiments are implemented on PyTorch [9] and mmdetection [3].

Default configuration: For fair comparisons, We train both model 1 and model 2 on the trainset of the SKU-110K, and evaluate them by validation subset and testing subset. we resize the input images to keep their shorter side being 800 and their longer side less or equal to 1,333. The whole network is trained using the Stochastic Gradient Descent (SGD) algorithm with 0.9 momentum and 0.0001 weight decay. We train detectors with 4 GPUs (1 images per GPU) for 24 epochs with an initial learning rate of 0.0025, and decrease it by 0.1 after 16 and 22 epochs respectively. All other hyper-parameters follow the settings in ATSS.

We compare different models on the SKU-110K validation subset and testing subset in Table 1. The symbol $^+$ means that the result is from the original paper. The symbol * represents testing with both flip and multi-scale. Ensemble means the fused results predicted by model 1 and model 2. In the WBF, the weight assigned for model 1 and model 2 is 2:1 and average value is employed for calculating confidence in weighted boxes.

As shown in Table 1, our model 1 and model 2 achieve 56.1% AP and 56.0% AP respectively, which are better than other methods, including Cascade R-CNN [2], FCOS [12] and baseline [5]. The best result 58.4% is achieved with model ensemble and multi-scale testing, outperforming all the previous detectors by a large margin.

Finally, we can further improve the AP accuracy of the proposed method by following methods. Firstly, input image resolution is increased from 1333 $\times$ 800 to 1800 $\times$ 1120. What's more, we train model 1 and model 2 on the train+test split of SKU-110K, and validation split is used as validation. Moreover, we adpot Synchronized BN instead of BN. Under the above improvements, the experimental

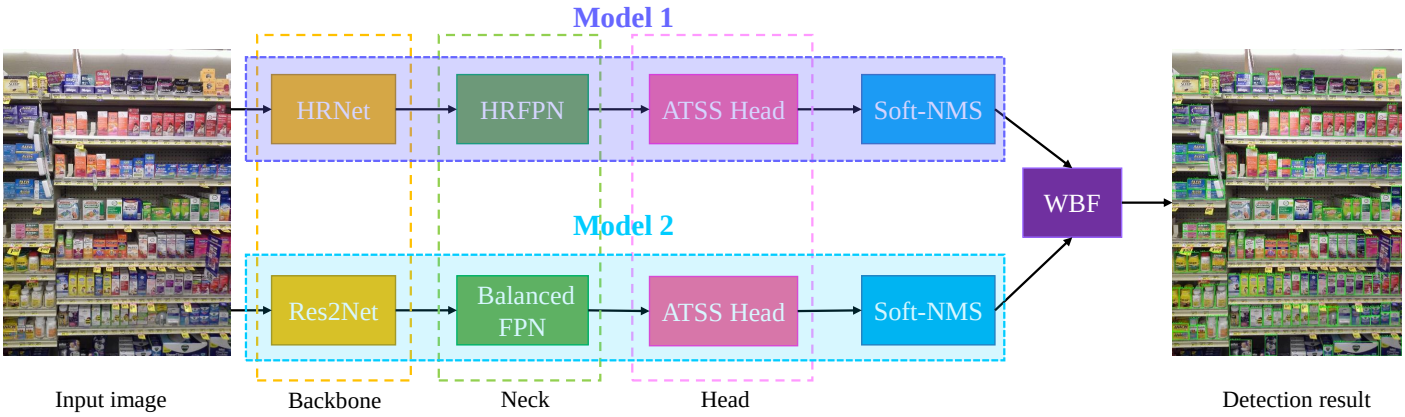

Figure 1. An overview of our solution.

Table 1. Comparisons with different methods trained on train subset of SKU-110K. The symbol $^+$ means that the result is from the original paper. The symbol * means test with the multi-scale testing strategy.

method	backbone	Input	AP (Testset)	AP (Valset)
RetinaNet$^+$[6]	ResNet-50	(1333,800)	45.5	-
Baseline$^+$ [5]	ResNet-50	(1333,800)	49.2	-
FCOS [12]	HRNet-W32	(1333,800)	54.3	53.0
Cascade R-CNN [2]	ResNet-50	(1333,800)	53.3	51.6
Cascade R-CNN	HRNet-W32	(1333,800)	54.2	52.4
ATSS [14]	Resnet-50	(1333,800)	53.1	51.5
ATSS	ResNet-101	(1333,800)	54.2	52.6
ATSS	Res2Net-101	(1333,800)	56.0	54.4
ATSS	HRNet-W32	(1333,800)	56.1	54.7
ATSS*	Res2Net-101	(1333,800)	57.5	56.3
ATSS*	HRNet-W32	(1333,800)	57.6	56.4
Ensemble	HRNet-W32+Res2Net-101	(1333,800)	57.4	56.1
Ensemble*	HRNet-W32+Res2Net-101	(1333,800)	58.4	57.3

results are shown in Table 2. Some qualitative results are shown in Figure 2. As shown in the figure, our method can detect a wide range of objects including crowded, occluded and extremely small objects.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms – improving object detection with one line of code. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[4] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[5] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[8] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2019.

[9] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Al-

Table 2. Object detection results with model trained on train+test subset of SKU-110K. The symbol * means test with the multi-scale testing strategy.

method	backbone	Input	AP (Valset)
ATSS	Res2Net-101	(1800, 1120)	55.9
ATSS	HRNet-W32	(1800, 1120)	55.8
ATSS*	Res2Net-101	(1800, 1120)	57.9
ATSS*	HRNet-W32	(1800, 1120)	57.6
Ensemble*	Res2Net-101+ HRNet-W32	(1800, 1120)	58.7

Figure 2. Qualitative detection results on SKU-110K.

ban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[10] Roman Solovyev and Weimin Wang. Weighted boxes fusion: ensembling boxes for object detection models. *arXiv preprint arXiv:1910.13302*, 2019.

[11] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[12] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[13] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[14] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. *arXiv preprint arXiv:1911.08287*, 2019.