A SOLUTION TO PRODUCT DETECTION IN DENSELY PACKED SCENES

Tianze Rong, Yanjia Zhu, Hongxiang Cai, Yichao Xiong
Media Intelligence Technology Co.,Ltd
{tianze.rong, yanjia.zhu, hongxiang.cai ,yichao.xiong}@media-smart.cn

October 28, 2020

1 Introduction

Densely packed scene detection is a extension of the object detection, but with a larger number, denser arrangement. We adopted a modified random crop strategy and a optimized Cascade R-CNN to solve the problems. Finally, we achieved a mmAP as 58.7% on SKU-110k [1].

2 Method

2.1 Data Description and Analysis

Statistics of Image After eliminating the invalid data, the size of the data set is shown below:

	Image Number	Annotation Number
Train Set	8219	1208482
Validation Set	588	90968
Test Set1	2936	431546
Total	11743	1730996

Table 1: Quantitative Statistics of SKU-110k

Dense object scene is mostly with a huge number of objects in a single image. The table below states the statistics of number of annotations in a single image:

	Mean	Max	Min	99.5% percentile	0.5% percentile
Train Set	147	576	1	356	61
Validation Set	154	759	40	582	59

Table 2: Statistics of Number of Annotations in Single Image

Statistics of Annotations Assume the input size is (1333,800), and the scale of a single bounding box represents as $\sqrt{wh}$, where w, h stand for the width and height of bounding box. Here, the figure 1, is the histogram and cumulative ratio curve of the scales of the data set.

As we can see, a third of bounding box is small object by the definition of MS COCO data set [2] after rescaling, the scale of the object is smaller than 32 pixels.

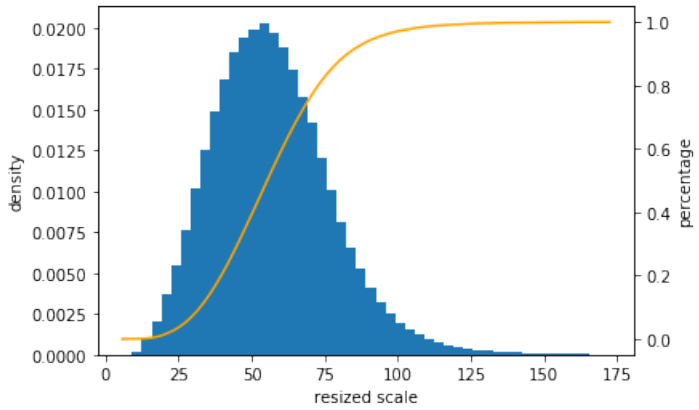

Figure 1: the Histogram and Cumulative Ratio Curve of the Scales

2.2 Solution to Densely Packed Scene

2.2.1 Promotion via Data

Rescale As the analysis before, it is apparent that small object problem is a vital factor should be solved to keep a better performance of model. The naivest solution to small objects is to enlarge the input size of images, which is able to make the small objects into larger ones. Regarding that SKU-110K data set possesses a high resolution level, a large input size would not take the side-effect significantly.

Random Crop Theoretically, we should input the images as large as possible. However, practically, the GPU memory capacity usually limits the input size in most of the cases. We adopt random crop to combines these two urges. But something need to be paid more attention here:

- When bounding boxes right on the cropping border, it would be clipped and remain the reserved region.
- Random sampling is probable to make some of the objects never be sampled.

Random Seven Crop We designed a strategy to relieve these two disadvantage. Clipping the bounding box may cause some fake box whose entity in box has been clipped out but background still remains. These fake box can lead to confusion to model in training. Hence we only remain a clipped box whose IoU to origin box higher than a threshold. Regular random crop sample the position of crop region from a uniform distribution. Random Seven Crop is designed to sample the region from only seven certain position: Four corners of image, center point and two end points of short axis.

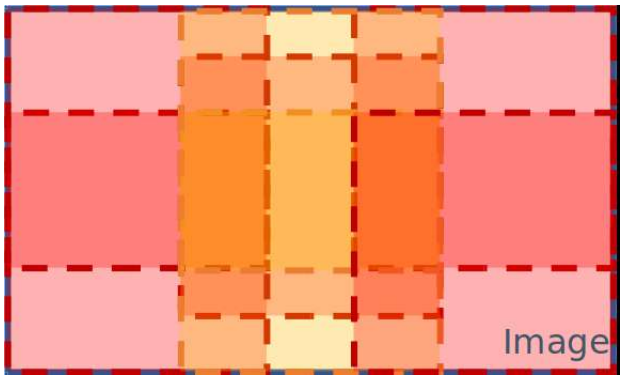

Figure 2: The seven sample areas of Random Seven Crop

2.2.2 Adjustment of Detector

Sampler Hyper-parameters Besides small object problem, another problem is about counts of the bounding boxes in a single image. As the statistics, a single images has a average as about 150 boxes, a number far more than MS COCO. But most of the default hyper-parameters are based on MS COCO data set which has a lower density of objects than SKU-110K. Hence we adjust the max positive sample number of both RPN and R-CNN sampler to release the limits.

Cascade R-CNN The model can achieve a relatively high AR@50 after sorts of optimization, while mAP(0.5:0.95) declined rather rapidly from AP@50. Cascade R-CNN [3] can refine those bounding box whose location is not that accurate by cascading bounding box heads. In this case, more accurate localization not only makes tight bound, but suppresses duplicated boxes so that reduces false positive boxes.

2.3 Other Modification

Inference Hyper-parameters Inference hyper-parameters is occasionally neglected by developers during improvement, in this case a.k.a NMS hyper-parameters optimizing. We adopt grid search eliminate by orthogonal design to search the optimal combination of the hyper-parameter.

Backbone and Neck Empirically, larger networks with more parameters bring better performance. To make an overview on the architecture of two-staged detector with neck. Backbone is the one of the easiest module to be adjust. So we replace the ResNet to ResNeXt [4]. Neck is also available now to be replaced as easy as backbone. BFP [**?**] is a neck structure with a balanced integration between octaves, meanwhile it can broadly improve the performance of the model.

3 Experiment

All experiment are conducted on MMDetection Platform with single GPU and run evaluation by library pycocotools on validation set. Since that, all mmAPs presented in this work are COCO-style and the IoU threshold of mmAP is [0.5:0.95:0.05]. Because of the differences between SKU-110K and MS COCO, we set the maxDet parameter of pycocotools on 400 instead of the default setting as 100.

3.1 Baseline

We built a baseline via Faster R-CNN [6] with FPN [7] and ResNet-50 [8] as backbone, also all configures are inherited from the Faster R-CNN with FPN benchmark on MS COCO from MMDetection [9]. The result mmAP is 50.6%, which is shown in the first row of the result table.

3.2 Performance Improving of Inference

We designed a orthogonal experiment to optimize NMS parameters using $L_9(3^4)$. As for the result of orthogonal table, we analyzed with ANOR. And the optimal parameters combination is following:

- Input and Output boxes number = 3000
- Minimal confidence threshold = 0.05
- IoU threshold of positive sample = 0.7
- Max output boxes number = 400

3.3 Adjustment on Sampler of Detector

According to the statistics, we already knew that number of ground truth in single image impacts the model vitally. We aim on the samplers among the all modules in detector by ruling the others out one by one, since sampler is one of the strongest modules coupled with the quantitative characteristics of a data set. The default RPN sampler number is 256 on COCO and there is a fraction of 0.5 on positive sample, which is not enough for SKU-110K apparently. And we counted positive sample number in all image. The result is illustrated in figure 3. So we enlarge the sample number to 512. Furthermore, we set the R-CNN sampler number to 3072 in accordance with RPN, since we got the called for conclusion on R-CNN sampler.

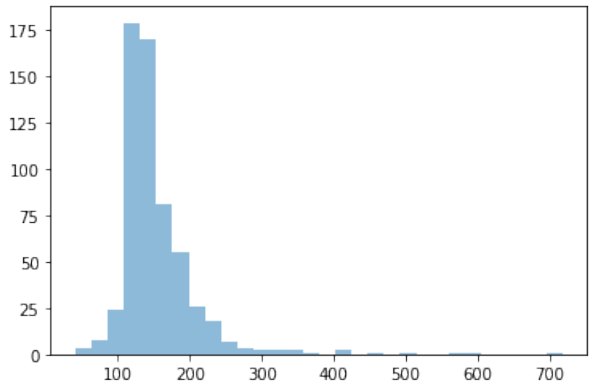

Figure 3: The histogram of positive sample number in a single image(x axis is positive sample number, y axis is count of this bin)

3.4 Random Crop and its Modification

The regular random crop has a problem that it can hardly sample the most of the ground truth sufficiently. We ran a Monte-Carlo simulation to estimate the coverage of random crop on 12 epoch and characterized the probability of the coverage of the IoU of sampled region on the whole image. When we prolong the training epochs to 18, the coverage got a ascendance. Although the prolonging of the epochs can avoid the insufficient sampling somewhat, the distribution of coverage still keeps a fairly wide peak.

We ran a same experiment with random seven crop. The following histogram is the result:

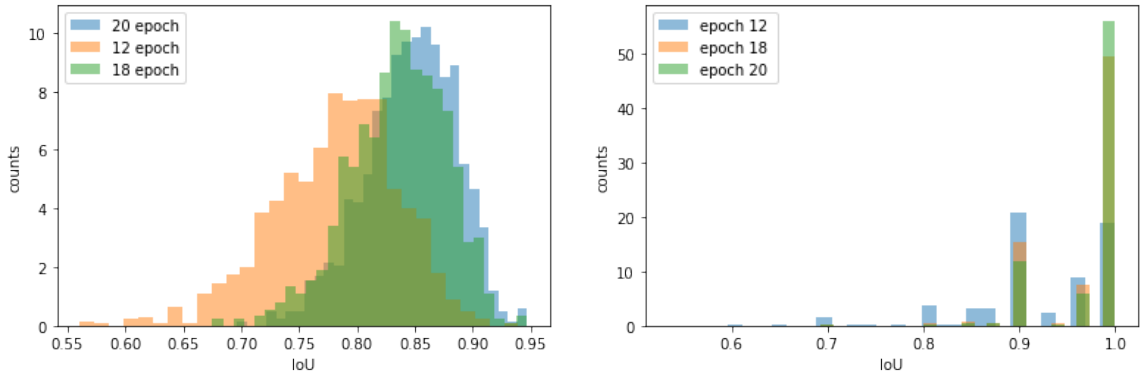

Figure 4: Histogram of Crop Coverage: RandomCrop(left), RandomSevenCrop(right)

Then we ran the random seven crop with the average input size and the crop size as big as our GPU memory can afford.

3.5 Cascade R-CNN

The model can perform a considerable recall by the former operations of us. Ideally, precision is more important to raise the mmAP. The score of detection mmAP is calculated based on IoU, so we analyzed iou of all positive samples and plot a histogram as figure 5.

Numbers and parameters of cascade head is determined by the experiment as table3.

3.6 Results

Here is the list of adjustment used in the model with abbreviation:

- ON: Optimal NMS - NMS with the parameters shown in 3.2
- OS: Optimal Sampler - The random sampler of RPN and R-CNN with optimal parameters

Head Number	IoU Threshold	mmAP
1	0.5	55.3
2	[0.5, 0.6]	**56.9**
3	[0.5, 0.6, 0.7]	56.7

Table 3: Parameter and Results of Cascade R-CNN

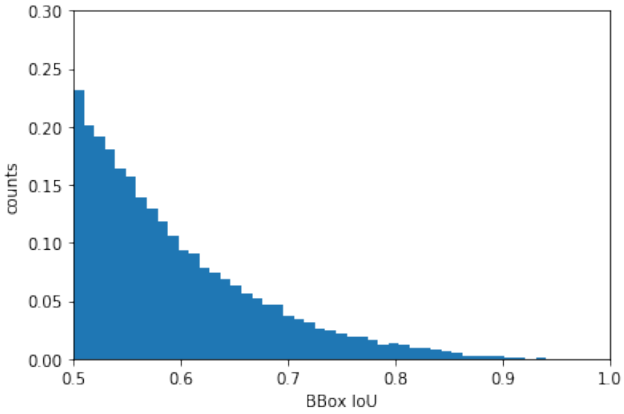

Figure 5: Positive Sample IoU of R-CNN

- OHEM - RPN sampler is the same with OS. R-CNN sampler is replaced by OHEM sampler [10] instead.
- C: Cascade Head - The RCNN head cascaded with multi-head.
- ResNeXt - ResNeXt with 32 groups and base width of 4.
- BFP - BFP with 2 refine levels.

ON	OS	C	Random Crop	Crop Size	Backbone	Neck	Input Size	Epoch	mmAP(%)	mmAP(test)
					ResNet-50	FPN	(1333,800)	12	50.6	
✓					ResNet-50	FPN	(1333,800)	12	**52.8**	
✓	✓				ResNet-50	FPN	(1333,800)	12	**54.0**	
✓	OHEM				ResNet-50	FPN	(1333,800)	12	52.1	
✓	✓		Uniform	(800,800)	ResNet-50	FPN	(1333,800)	12	53.0	
✓	✓				ResNet-50	FPN	(1800,1080)	12	**55.3**	
✓	✓		Uniform	(1200,1200)	ResNet-50	FPN	No Rescale	12	54.6	
✓	✓		Uniform	(1200,1200)	ResNet-50	FPN	(3000,1800)	12	55.3	
✓	✓		Uniform	(1200,1200)	ResNet-50	FPN	(3000,1800)	18	**56.1**	
✓	✓		Seven	(1200,1200)	ResNet-50	FPN	(3000,1800)	18	55.1	
✓	✓		Seven	(1200,1200)	ResNet-50	FPN	(3000,1800)	24	**56.6**	
✓	✓	✓	Uniform	(1200,1200)	ResNet-50	FPN	(3000,1800)	18	**56.9**	
✓	✓	✓	Seven	(1200,1200)	ResNet-50	FPN	(3000,1800)	24	**57.4**	
✓	✓	✓	Seven	(1200,1200)	ResNet-50	BFP	(3000,1800)	24	**57.7**	
✓	✓	✓	Seven	(1200,1200)	ResNeXt-101	BFP	(3000,1800)	24	**58.0**	**58.7**

Table 4: Results and Conditions of All Experiment

References

[1] 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.

[2] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015.

[7] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[9] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Mmdetection: Open mmlab detection toolbox and benchmark. *CoRR*, abs/1906.07155, 2019.

[10] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.