Joint COCO and Mapillary Workshop at ICCV 2019: COCO Instance Segmentation Challenge Track
Technical Report: HSSLab

Baoyu Fan, Runze Zhang, Liang Jin
Inspur State Key Laboratory of High-end Server & Storage Technology
1036 Langchao Rd., Jinan, Shandong, China
valuefish@gmail.com

Abstract

In this report, we do experiments for the COCO Instance Segmentation Challenge Track. The final single-model results are 45.5 AP for segmentation. Through the experiments, we find an interesting issue, that is the recently proposed algorithms and strategies for uncascaded models are hard to obtain the better results when they are used in cascaded models, especially cascade RCNN series models.

1. Baseline

The baseline model in our experiments are the HTC [1] models, the baseline algorithm for the 2018 COCO instance segmentation Challenge Champion. The backbone network we use is Resnet-50 which adopts the Deformable Convolution Network(DCN) [13].To boost the performance, we add recently proposed methods or strategies. Libra-RCNN [7], which is focused on the imbalanced issue during the object detection pipeline. The instance segmentation greatly relies on the performance of object detection, so we think this method can also solve the imbalance issue for the instance segmentation. Guided-anchoring [10], aimed at solving the imbalance between the receptive field and semantic scope of the same Roi. Double Head RCNN [11] decouples the classification and the regression and finds the best combination for solving the tasks. Mask Scoring RCNN [5] presents an extra head to predict MaskIoU score for each segmentation mask, whose goal is to solve the inconsistency of classification probability and bbox localization score. Grid RCNN [6] replaces linear bounding box regressor with the principle of locating corner keypoints corner-based mechanism. All of the methods or strategies above improve the mAP for the Faster-RCNN [8] series ,which adopt uncascaded network.

2. Experiments

For fair comparisons, all experiments are implemented on mmdetection [2].The baseline results on COCO val-2017 are listed in Table 1. It is demonstrated that the new methods or strategies can not boost the performance. We think maybe the cascaded structure incur the strategies such as the sampling strategies in Libra RCNN,the mask scoring hypothesis in Mask scoring RCNN. To verify the conclusion, we add all these methods above to the Mask RCNN [3] and the performance has a 5.8 point gain. Averagely, each method can raise by more than one point.

During the test evaluation process, especially in the mask AP stage, we also found an interesting issue. There exists some wrong detections in some certain cases, which incurs the mAP performance. Firstly, the ground truth may not cover all of the categories about one single image but our model successfully detects these bounding bboxes. We could see relevant cases in Figure 1. The giraffe ,which our model detects should have emerged in ground truth, on the third row, second column. Secondly, there exists the circumstance that some small objects are grouped under the same target ground truth bounding box but our model detects them one by one. The orange should have been detected one by one but the ground truth only has one bounding box. All of these issues may get low IOU but get high classcification scores. However, the evalluation metrics give priority to the classification scores. So the circumstances above could degenerate the mAP performance. To verify this conclusion, we do such ablation experiments. During test, we lowers the classification scores artificially according to the IOU between our detected results and the ground truth on val-2019 COCO datasets. The results are listed in Table 2.

For the final results, we compare different deeper backbone networks. We use the pretrained models from ImageNet to finetune the model. And the results on COCO val-2017 dataset are listes in Table 3. We select three backbones, Resnext101-64*4d [12], SENet-154 [4], HRNet-w48 [9] and choose the best model Resnext101-64*4d. It is abnormal that the result finetuned from SENet can not exceed the result from Resnext-101. Perhaps it is due to the SENet we used is not provided by the mmdetection. Finally,the results on COCO val-2017 are 45.5 mAP. Baseline means the result from "HTC + DCN" in Table 1. The score/2(IOU¡0.1) means the strategy when Iou of the detected bounding bbox is lower than 0.1, the relevant classification score is divided by 2. And so on , for each of the remaining strategies. From the results we could see the upper limit when removing the cases of the lower IOU but higher score. It is about 3 percents. So solving this issue could lead a better performance.

Table 1: Results(mask AP) with different methods or strategies on COCO val-2017 dataset (%).

Methods	**AP**
HTC + DCN	39.7
HTC + DCN + GA-RPN	39.5
HTC + DCN + Libra-RCNN	37.9
HTC + DCN + Mask Scoring RCNN	35.2
HTC + DCN + Grid RCNN	39.0
HTC + DCN + Double Head RCNN	38.1
Mask-RCNN	34.2
Mask-RCNN + ALL	40.0

Table 2: Results(mask AP) with different test strategies on COCO val-2017 dataset (%).

Methods	**AP**
Baseline	39.7
Baseline + score/2(IOU <0.1)	41.7
Baseline + score/5(IOU <0.1)	42.3
Baseline + score=0(IOU <0.1)	42.3
Baseline + score/5(IOU <0.2)	42.4
Baseline + score/5(IOU <0.3)	42.6

Table 3: Results(mask AP) with different methods or strategies on COCO val-2017 dataset (%).

Methods	**AP**
RexNext101(64*4d)	42.2
SENet154	41.5
HRNet-w48	41.3
RexNext101(64*4d) + MS Training	43.9
RexNext101(64*4d) + MS Training + MS testing	44.9

References

[1] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019. 1

[2] 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. 1

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[5] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6409–6418, 2019. 1

[6] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7363–7372, 2019. 1

[7] 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. 1

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[9] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 2

[10] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2974, 2019. 1

[11] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization in r-cnn. *arXiv preprint arXiv:1904.06493*, 2019. 1

[12] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 2

[13] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 1

Figure 1: Sample images on val-2019 COCO datasets for the special cases which cound hurt the mAP performance .