Joint COCO and Mapillary Workshop at ICCV 2019: COCO instance segmentation Challenge Track

Technical Report: Improved-HTC

Yuqing Wang
Meituan-Dianping Group
wangyuqing06@meituan.com

Baoshan Cheng
Meituan-Dianping Group
chengbaoshan02@meituan.com

Haibo Su
Northwestern Polytechnical University
2018262230@mail.nwpu.edu.cn

Abstract

Instance segmentation is a challenging computer vision task, which has both the characters of object detection and semantic segmentation. Different from semantic segmentation, instance segmentation needs to separate instances from the same class. Therefore it utilizes the ROIs extracted from the object detection model to separate them respectively. In this COCO Instance Segmentation challenge, our method is built on the Hybrid Task Cascade network. We achieved better performance with deformable convolution, attention mechanism and GCNet module. After multi-scale testing and model ensemble, our final model achieve 0.462 mAP scores on the test-challenge dataset. These models are trained without any extra data.

1. Our Baseline

We employ the Hybrid Task Cascade network (HTC)[3] as our baseline, the HTC network interleaves bounding box regression and mask prediction instead of executing them in parallel. Besides instance segmentation, the network also performs semantic segmentation, this multi-task operation can improve the overall segmentation quality. The ResNeXt-101[7] with the Imagenet-pretrained[5] weights are used as the backbone. The semantic branch is supervised by COCO-stuff[1] annotations. The network is trained on the MS COCO[6] train-2017 set.

2. Methods

We add several modules to improve the performance, including Deformable Convolution(Deconv)[4, 9], Spatial Attention Mechanism[8] and GCNet[2]. The performance of these modules are shown in Table 1.

Deformable Convolution is a new kind of convolution which can learn the offsets of convolution based on the shape of the object. We added the deformable convolution in ResNeXt[7] stage 3 to 5.

We also add spatial attention and GCNet module to the network, which can helps to focus on the object area and model global context respectively.

The Spatial Attention Mechanism[8] can help the network learn where to focus, it can help to preserve more details for segmentation task. The Generalized Attention module is inserted after 1x1 conv of the backbone.

The global context block in GCNet is also a kind of self-attention module, which can help to model global contexts. The Global Context(GC) block is also inserted after 1x1 conv of backbone. We used ratio 4 in all GC blocks.

3. Implementation Details

We trained the model on 8 Tesla-v100 GPUs (1 image per GPU) for 20 epochs with an initial learning rate of 0.01, and decrease it by 0.1 after 16 and 19 epochs, the training process costs about 10 days. During the training process, The longer edge and short edge of images are resized to 1333 and 800 respectively without changing the aspect ratio.

4. Ablation Study

As shown in Table 1, the deformable convolution module can improve about 3 points on Mask AP, the attention and GCNet modules can improve 1.2 points on Mask AP.

In addition to adding modules on the network structure, we also tried common tricks to improve the final result, including multi-scale testing and model ensemble. The eval-

Table 1: The performance of all useful modules on HTC. All results are evaluated on MS COCO val-2017 set (single-scale, no-flip test).

Method	box AP	mask AP
HTC baseline	46.9	40.8
+Deconv	50.6	43.8
+Deconv+Attention+GCNet	51.0	45.0

Table 2: The performance of multi-scale testing on MS COCO val-2017 dataset using HTC+deconv model.

Test scale	mask AP
(1333,800)	44.2
(1600,1600)	45.2
(2000,2000)	45.0
(1000,1600), (1200,1900), (1400, 2200)	45.3

uation results of testing in different scales on val-2017 are shown in Table 2. We choose the scales of [(1000,1600), (1200,1900), (1400, 2200)] for final prediction. After that, the performance on the test-challenge2019 is 0.459. In the end, we ensemble the final two models(epoch-19 and epoch-20) to reach the final 0.462 AP.

5. Discussion

The MS COCO dataset is a complex dataset for computer vision research. As the annotation is in polygen format, the evaluation server will not reward methods that can segment objects more precisely. Besides, when there are two objects overlaps, the pixels in the boundary area might belonging to two objects, which makes the network hard to decision.

State-of-the art instance segmentation methods relies on object detection results. We hope there will be methods that can directly output masks in the future.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018. 1

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *arXiv preprint arXiv:1904.11492*, 2019. 1

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 2017. 1

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[6] 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 *ECCV*, 2014. 1

[7] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016. 1

[8] Xizhou Zhu, Dazhi Cheng, Zheng Zhang, Stephen Lin, and Jifeng Dai. An empirical study of spatial attention mechanisms in deep networks. *arXiv preprint arXiv:1904.05873*, 2019. 1

[9] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. *arXiv preprint arXiv:1811.11168*, 2018. 1