Instance segmentation model CP-Condinst

Jianye Yuan¹, Xinwang ding²*, Lele Sun³

¹College of Electrical Engineering, Xinjiang University, Urumqi 830047
²Technology and Information Center, Wuhan Marine Machinery Plant Co., LTD, Wuhan 430084
³College of Electrical Engineering, Shandong Huayu Institute of Technology, Dezhou 253071

*Corresponding author: aku_cintapadamu@163.com

Abstract—With the rapid development of deep learning image recognition technology, people have higher and higher requirements on the accuracy and speed of segmentation recognition. So instance segmentation becomes more and more important. The accuracy and speed of segmentation recognition are increasingly demanded. Recently, Shen Chunhua et al. proposed a very quantized instance segmentation model CondInst. In the paper, which used CSPResNet instead of ResNet, and added a Prediction module method to the head to improve model accuracy. Finally, on the MS COCO2017 data set, the simulation results show that CondInst’s AP, AP50, AP75, APS, APM and APL were increased by 0.2%, 0.4%, 0.3%, 0.1% and 0.4% respectively in the 50 layer residual model of our model, compared with CondInst's AP, AP50, AP75, APS, APM and APL. The experimental results show that the method has certain practical value.

1. INTRODUCTION

With the rapid development of deep learning, the segmentation of images by Convolutional Neural Network (CNN) has become more and more popular. However, the traditional semantic segmentation and scene segmentation no longer meet people's needs. Convolutional Neural Network based on target detection (R-CNN) indirectly solves the difficult problem of inaccurate segmentation effect. Instance Segmentation is the result of the pixel-by-pixel Segmentation of a target based on the target detection problem. At the same time, the emergence of instance segmentation not only solves the problem of inaccurate image classification, but also solves the problem of inaccurate target positioning. At present, instance segmentation is mainly divided into two types: single stage and double stage.

Among the single stage and double stage instance segmentation models, double stage has become the mainstream segmentation model at present, most of which adopt the Mask R-CNN series. The basic idea detected the rectangular frame of the object first, and after the detection, the rectangular frame is segmented pixel by pixel. Currently, there are many double stage instance segmentation models derived from Mask R-CNN. For example, Chen et al. designed Multitask and Multistage Hybrid Task Cascade (HTC) by combining with Mask R-CNN, and integrated semantic segmentation to enhance spatial information. Liu et al. proposed an instance segmentation model through improved Feature Pyramid Networks (FPN) and mask branch. Huang et al. designed the...
model of Mask R-CNN, and the segmentation effect was better than that of Mask R-CNN. All the above results were optimized based on the double stage Mask R-CNN. Therefore, Mask R-CNN achieves a good segmentation reference effect, but it also has the following shortcomings, which need to be further improved. For example: 1) The segmentation effect of Mask R-CNN is completely limited by the accuracy of the rectangular box; 2) A large amount of Parameter(Params) and number of floating point of operations(FLOPs) will be generated by using per-pixel segmentation, and the use of down-sampling to reduce the size of the rectangular box will reduce the accuracy; 3) per-pixel segmentation is not suitable for some tasks, such as: text type, etc.

In order to solve the problem of Mask R-CNN, Shen[12] et al. proposed a new instance segmentation model, which was CondInst in 2020 CVPR. For the first time, the conditional convolution is applied to the instance segmentation task to solve the instance segmentation task from the perspective. At the same time, the Mask R-CNN series avoids using ROI Pool or ROI Align to get the instance Mask, but adopts the dynamic perceptive network model[13] of instance segmentation to get the instance Mask. Therefore, the segmentation effect is greatly improved. In addition, this model proposed that instance segmentation is based on a single stage model, which directly avoids a series of problems above mentioned in the double stage model Mask R-CNN.

2. RELATED WORK

According to the re-calibration method for Mask R-CNN by YOLACT[14] and BlendMask[15], the single stage instance segmentation model was continuously optimized, such as the FCOS single stage detection model proposed by TIAN[16] recently. The paper is optimized in the latest 2020 CVPR literature Conditional Convolutions for Instance Segmentation[12]. It is optimized by reference to the three classic modules of Instance Segmentation, backbone, neck and head, which respectively proposed to use CSPResNet as the backbone model. Prediction model was increased at CondInst. Therefore, the network model in the paper is named CP-Condinst.

In view of this, the contribution of this paper is summarized as follows:

(1) Using CSPResNet instead of ResNet It introduced a Prediction module in the head and improved the Prediction accuracy of the model by adding concat.

(2) Propose a new architecture that can be used for instance segmentation model, and the network structure diagram is shown in Figure 1.

Fig.1 CP-Condinst network structure diagram, where C3, C4 and C5 represent the feature maps of the backbone network, the network model we use is CSPResNet; the feature level of the final predicted image from P3 to P7, we use FPN; all feature maps We set the input size to 800×800, and the other parts are set to be the same as the original text.
3. SEGMENTATION MODEL

The instance segmentation model can be divided into three parts: neck, backbone and head. The paper, by replicating the latest instance segmentation paper of CVPR 2020, further optimize the model of neck, backbone and head, and redesign a better instance segmentation model CP-Condinst.

The backbone network functions as the characteristic extraction of pictures, which the backbone network used by CondInst is the characteristic extraction function of ResNet-50 and ResNet-101 models\textsuperscript{[18]}. The paper used on the idea of yolov4 which had a high realtime capability, and introduces the Cross Stage Partial Network(CSPNet)\textsuperscript{[19]} model as the backbone network in CondInst. Therefore, the backbone network of this article combines CSPNet to be CSPResNet. When the optimization is complete, the network structure of the backbone network used in this article and that of CondInst is shown in Figure 2.

![Fig.2](image1)

The role of head is to predict the instance category and mask position by obtaining the output result of the neck module. At present, most of the methods used for instance segmentation are based on the segmentation of the Fully Convolutional Networks(FCN), and lags behind Region of Intersect(ROI)\textsuperscript{[20]}. CondInst performs an analysis of each feature graph instance based on Conditional convolution(CondConv)\textsuperscript{[21]}. Mask FCN model parameters are generated dynamically, and the mask of the feature map is predicted based on the central region. The CondInst head is optimized in FCOS, a new shared head is proposed as a mask generator with a model structure of convolution(ConV) and ReLU. The paper further optimized the shared head output and was followed by a feature extraction module. The model referred to the Prediction part of snake instance segmentation model from Peng\textsuperscript{[22]}, which was placed directly in the CondInst shared head module and then further integrated the predicted instance information. The Prediction network structure is shown in Figure 3.

![Fig.3](image2)
4. EXPERIMENT

The paper evaluates the performance of the model CP-Condinst on a Microsoft (MS) open COCO2017\(^{[23]}\) dataset containing 80 classes.

The indicators used in this study are whether they can apply Mosaic Data augg (MAug), Average Precision (AP)\(^{[24]}\), APx, and the level of accuracy in dividing small (AP\(_S\)), medium (AP\(_M\)), and large (AP\(_L\)) objects.

The experiment uses Ubuntu16.04 operating system, and the graphics card is a computer device with 32G memory of NVIDIA RTX 2080i dual-card graphics card.

The network model of the paper, CP-Condinst, is optimized based on the latest CondInst model, as the code used is modified by the author on the official CondInst open source code on GitHub. The official code uses an 8-card V100 GPU, but this experiment uses a dual-card GPU due to the influence of experimental conditions.

As mentioned above, the network model of this article uses CSPResNet. Meanwhile, a Prediction module added to the head feature extraction module. Because CondInst uses ResNet50. Therefore, the paper selects ResNet50 for experimental comparison, and other unstated uniform Settings are the same as the original text.

It can be seen from Table 1. Among them, 1X means 90K iterations, 3X means 180K iterations, and so on. Using R to represent ResNet; Using CSPR to represent CSPResNet; In this experiment, 3X, 6X and 9X were selected for comparison. It can be found that in iteration 3X and 6X, CondInst model effect is better than Mask R-CNN in accuracy. When the number of iterations increased to 6X, the accuracy of the model was further improved. Compared with the CP-Condinst model in the paper, the accuracy of Mask R-CNN and CondInst was slightly lower. Therefore, it can be concluded that with the number of iterations from 3X to 6X, the accuracy rate keeps increasing, and the model has not converged when 3X is trained.

When the number of iterations reaches 9X, the network model in the paper converges. In the paper, resnet-50-FPN of Mask R-CNN and CondInst were compared with CSPR-50-FPN of CP-Condinst respectively. According to the experimental data in Table 3, the network model CP-Condinst presented performs better in AP, AP50, AP75, APS, APM and APL than that in Mask R-CNN and CondInst in the paper. Among them, the overall effect of CondInst is better than that of Mask R-CNN. Using the 50-layer residual network model AP, CP-Condinst is 0.2% more accurate than CondInst. Therefore, the paper comes to the conclusion that CP-Condinst is superior to CondInst network model in instance segmentation and has practical application value.
Table 1   Comparison of data between our model and the most advanced model

| method     | backbone   | Sched | AP  | AP50 | AP75 | APs | APm | APc |
|------------|------------|-------|-----|------|------|-----|-----|-----|
| Mask R-CNN | R-50-FPN   | 3X    | 29.8| 48.5 | 30.7 | 14.1| 32.1| 40.0|
| CondInst   | R-50-FPN   | 3X    | 31.0| 49.8 | 31.5 | 15.6| 33.1| 42.2|
| CP-CondInst| CSPR-50-FPN| 3X    | 31.0| 50.3 | 32.1 | 16.2| 33.2| 43.8|
| Mask R-CNN | R-50-FPN   | 6X    | 34.2| 55.3 | 35.1 | 18.3| 36.5| 47.1|
| CondInst   | R-50-FPN   | 6X    | 36.8| 57.9 | 39.4 | 20.2| 39.9| 48.1|
| CP-CondInst| CSPR-50-FPN| 6X    | 37.9| 58.2 | 40.4 | 20.9| 40.2| 48.5|
| Mask R-CNN | R-50-FPN   | 9X    | 37.5| 58.0 | 37.9 | 19.3| 38.4| 48.8|
| CondInst   | R-50-FPN   | 9X    | 39.9| 59.8 | 41.3 | 22.8| 43.9| 49.5|
| CP-CondInst| CSPR-50-FPN| 9X    | 40.1| 60.2 | 41.5 | 23.1| 44.0| 49.9|

5. CONCLUSION

In the paper, optimization and modification are carried out in the latest 2020 CVPR paper, and a new instance segmentation model CP-Condinst is proposed. Through experiments on MS COCO2017, it is found that the CP-Condinst instance segmentation network model proposed in this paper has a better effect than the classic Mask R-CNN and the original CondInst instance segmentation model. Therefore, on the premise of increasing accuracy, a small amount of model complexity and computation can be accepted. Finally, experiments show that the CP-Condinst model is a powerful instance segmentation model.

REFERENCES

[1] Gu, Jiuxiang & Wang, Zhenhua & Kuen, Jason.(2015). Recent Advances in Convolutional Neural Networks. Pattern Recognition. 10.1016/j.patcog.2017.10.013.
[2] LONG, Jonathan, SHELHAMER, Evan, DARRELL, Trevor. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 3431-3440.
[3] FU, Jun, et al. Dual attention network for scene segmentation[J]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. p. 3146-3154.
[4] GIRSHICK, Ross, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[J]. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 580-587.
[5] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey, et al. ImageNet Classification with Deep Convolutional Neural Networks[J]. Neural Information Processing Systems. (2012), 25. 10.1145/3065386.

[6] HAFIZ, Abdul Mueed, BHAT, Ghulam Mohiuddin. A survey on instance segmentation: state of the art[J]. International Journal of Multimedia Information Retrieval, 2020, 1-19.

[7] He, Kaiming, et al. "Mask R-CNN[C]." international conference on computer vision (2017): 2980-2988.

[8] K. Chen, J. Pang, J. Wang, et al. “Hybrid task cascade for instance segmentation[J],” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 4974–4983, 2019.

[9] S. Liu, L. Qi, H. Qin, et al. “Path aggregation network for instance segmentation[J],” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 8759–8768, 2018.

[10] Lin, Tsungyi, et al. "Feature Pyramid Networks for Object Detection[J]." computer vision and pattern recognition (2017): 936-944.

[11] Z. Huang, L. Huang, Y. Gong, et al. “Mask scoring r-cnn[J],” in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pp. 6409–6418, 2019.

[12] TIAN, Zhi, SHEN, Chunhua. CHEN, Hao. Conditional Convolutions for Instance Segmentation[J]. arXiv preprint arXiv:2003.05664, 2020.

[13] X. Jia, B. De Brabandere, T. Tuytelaars, et al. “Dynamic filter networks[J],” in Proc. Advances in Neural Inf. Process. Syst., pp. 667–675, 2016.

[14] D. Bolya, C. Zhou, F. Xiao, et al. “YOLACT: realtime instance segmentation[J],” in Proc. IEEE Int. Conf. Comp. Vis., pp. 9157–9166, 2019.

[15] CHEN, Hao, et al. BlendMask: Top-down meets bottom-up for instance segmentation[J]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. p. 8573–8581.

[16] TIAN, Zhi, et al. FCOS: Fully convolutional one-stage object detection[C]. In: Proceedings of the IEEE international conference on computer vision. 2019. p. 9627–9636.

[17] BOCHKOVSKIY, Alexey; WANG, Chien-Yao; LIAO, Hong-Yuan Mark. YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[18] He, Kaiming, et al. "Deep Residual Learning for Image Recognition[J]." computer vision and pattern recognition (2016): 770-778.

[19] WANG, Chien-Yao, et al. CSPNet: A new backbone that can enhance learning capability of cnn[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020. p. 390-391.

[20] GIRSHICK, Ross. “Fast r-cnn[C],” In: Proceedings of the IEEE international conference on computer vision. 2015. p. 1440-1448.

[21] B. Yang, G. Bender, Q. V, et al. “Condconv: Conditionally parameterized convolutions for efficient inference[J],” in Proc. Advances in Neural Inf. Process. Syst. pp. 1305–1316, 2019.

[22] PENG, Sida, et al. “Deep Snake for Real-Time Instance Segmentation[C],” In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. p. 8533-8542.

[23] T. Lin, M. Maire, S. Belongie, et al. “Microsoft coco: Common objects in context[J],” in Proc. Eur. Conf. Comp. Vis., pp. 740–755, Springer, 2014.

[24] GUPTA, Saurabh, et al. “Learning rich features from RGB-D images for object detection and segmentation[J],” In: European conference on computer vision. Springer, Cham, 2014. p. 345-360.