CarfRCNN: A Two-stage Effective Model For Instance Segmentation

Yujie Chen¹, Pengbiao Zhao² and Julin Chen¹*

Network and Data Security Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China

*Corresponding author’s e-mail: 202021090126@std.uestc.edu.cn

Abstract. With the rapid development of deep learning, many instance segmentation models have achieved good results in accuracy and time. But here are still many problems. In this paper, we proposed a two-stage model CarfRCNN. We proposed CAResNet to change the structure of the backbone, making the feature extraction of the input image more refined. At the same time, we also added the CRF module to add smooth constraints to the pixels, so that the segmentation mask fits the target contour more closely. We train and test on the COCO datasets, and compare with the Mask R-CNN. The experimental results prove that the model we proposed can greatly improve the accuracy.

1. Introduction

Instance segmentation provides different labels for individual instances of objects belonging to the same class. Therefore, instance segmentation can be defined as a technology that simultaneously solves the problem of object detection and semantic segmentation. Instance segmentation[1] has become one of the relatively important, complex and challenging fields in computer vision research. Instance segmentation aims to predict object class labels[2] and pixel-specific object instance masks[22]. It can locate object instances in various images with different classes. It does a great favour for robotics, autonomous driving, surveillance[10], etc. With the rapid development of deep learning[4][14], specifically Convolutional Neural Networks (CNN)[5], many instance segmentation frameworks have been proposed, and the accuracy of instance segmentation has rapidly improved.

Under the lead of Fast/Faster R-CNN, Fully Convolutional Networks (FCN) have been used to predict segmentation masks[18], alongside box regression and box classification[20]. In order to obtain high performance, Feature Pyramid Network (FPN) has been used to extract phased network features[24], in which a top-down network path with horizontal connections has been used to obtain semantically strong features.

Although the effects of many models have been greatly improved, there are still many problems, which need to be further improved. For example, 1) the background of the input image is complex, making it difficult for the network to find the segmentation target, which will cause the performance of the model to decrease, 2) traditional CNN has a large receptive field, which leads to rough results when outputting pixel-level labels and it lacks smooth constraint.

We proposed a two-stage model CarfRCNN to solve the problems. The contributions can be summarized as follows:

- Using CAResNet instead of ResNet. It introduced an Attention module and improved the prediction accuracy of the model.
- We add a conditional random field on the branch of the mask.
Our model has a good advantage in performance, which achieved better results compared with Mask R-CNN.

2. Related work

Instance segmentation. Instance segmentation is a highly active research area. At present, there are mainly two frameworks for instance segmentation, mainly two-stage frameworks and single-stage frameworks. Two-stage instance segmentation often formulates this task in two ways, namely the bottom-up semantic segmentation-based method and the top-down detection-based method. The main representative model is Mask R-CNN. One-stage instance segmentation is inspired by one-stage, anchor-based detection models such as YOLO and RetinaNet, whose representative works include YOLACT and SOLO. For platforms with rich computing resources, the two-stage framework has better accuracy than the single-stage framework. Because the two-stage framework is flexible and more suitable for area-based detection. Due to the lack of preprocessing in single-stage, a lightweight backbone network, fewer candidate areas and the use of fully convolutional detection subnets, it is usually faster than the two-stage.

Attention mechanism. The attention mechanism can be regarded as a resource allocation mechanism, which reallocates resources according to the importance of the attention object for the resources that were originally evenly allocated. The core idea is to find the correlation between them based on the original data, and then highlight some of its important features. There are several types of visual attention, such as channel attention, pixel attention, multi-level attention, etc. In our CAResNet, we exploit channel-wise attention based on an efficient architecture and add it to each residual block of the residual network.

CRF. CRF refers to conditional random field, random refers to random variables, and conditional refers to conditional probability. This means that CRF is a discriminant model. The original labels from CNN are generally patchy images. In the image, the labels of some small areas may be incorrect, so they cannot match the surrounding pixel labels. In order to solve this discontinuity, we add a conditional random field on the branch of the mask, which takes advantage of the similarity of pixels in the original image to re-refine the CNN label.

3. Methodology

Our method is built on Mask R-CNN, because it is a simple instance segmentation model that also achieves state-of-the-art results. The network diagram is shown in Figure 1.

![Figure 1. The structure of our model CarfRCNN](image)

3.1. CAResNet

We proposed CAResNet, which adds a channel attention mechanism to the original resnet and embeds it in each residual block. It will be used as a new backbone to perform feature extraction on the input image. Channel attention can be seen as what the neural network is looking at. Each layer of the CNN has many convolution kernels, and each convolution kernel corresponds to a feature channel. Channel attention is to allocate resources between each convolution channel.

The structure of CAResNet is shown in Figure 2. We use max pooling and average pooling to aggregate the channel information of the feature map, so that we can get the spatial descriptors.
representing the average pooling feature $F_{avg\_pool}$ and the max pooling feature $F_{max\_pool}$ at the same time, and then put these two descriptors into two FC layers, the channel attention map $M_{ca}$ is finally generated. Then we add the channel attention module to each residual block of Resnet.

**Figure 2.** The structure of the block CAResnet

### 3.2. CrfFCN

We added a conditional random field on the branch of the mask, which called CrfFCN. We first use FCN for feature extraction, then use CRF to optimize the output of FCN, and finally get the segmentation mask map. We assume that the pixel $i$ has a class label $x_i$ and a corresponding output of FCN $f_i$, the relationship between $x_i$ and $f_i$ can be characterized as follows:

$$\Psi_p(x_i, x_j) = u(x_i, x_j) \sum_{m=1}^{M} \omega(m) k_G(m)(f_i, f_j)$$

It describes the relationship between pixel and pixel, and encourages similar pixels to be assigned the same class, while pixels with large difference are assigned different class. The definition of this "distance" is related to the color value and the actual relative distance. So CrfFCN can make the picture as far as possible segmentation at the boundary.

### 4. Experiments

We test the model performance on the COCO2014\[13\] datasets. Following common practice, we train using the union of 80K train images and train images in coco_2014_valminusminival, then we test using coco_2014_minival datasets. ResNet-50-FPN and ResNet-101-FPN are used as our backbone networks. Specifically, our network is trained with stochastic gradient descent (SGD) for 90K iterations with the initial learning rate being 0.02 and a mini-batch of 8 images.

Table 1 shows the comparison of the results of our experiments with others, we can see that the results of our experiment have achieved better performance on all evaluation indicators.

**Figure 3.** The visualization of our experiments and Mask R-CNN, our method can better focus on the object, distinguish the foreground from the background, and better fit the object’s contour.

| backbone | AP | AP50 | AP75 | APs | APm | APi |
|----------|----|------|------|-----|-----|-----|
| Mask R-CNN | 25.68 | 41.9 | 27.3 | 7.3 | 25.4 | 43.9 |
| CarRCCNN(ours) | **25.92** | **42.7** | **27.5** | **7.5** | **25.4** | **44.8** |
| Mask R-CNN | 26.1 | 42.3 | 27.8 | **10.5** | 25.9 | 45.5 |
| CarRCCNN(ours) | **26.6** | **43.1** | **28.4** | 6.9 | **26.7** | **46.3** |
5. Conclusion
In this paper, we propose a two-stage instance segmentation model CarfRCNN, including two blocks CAResNet and CrfFCN, in order to solve the exiting problems in instance segmentation. CAResNet as a backbone can extract more refined feature map, CrfFCN adds a smooth constraint to make the segmentation mask more closely fit the contour of the object itself. The extensive experiments have proved that the proposed method has achieved better results on various evaluation indicators.

Acknowledgments
This work was supported in part by the National Natural Science Foundation of China (No.62072074, No.62076054,No.62072827, No.61902054), the Frontier Science and Technology Innovation Projects of National Key R&D Program (No.2019QY1405), the Sichuan Science and Technology Innovation Platform and Talent Plan (No.2020JDJQ0020), the Sichuan Science and Technology Support Plan (No. 2020YFSY0010), and the Natural Science Foundation of Guangdong Province (No. 2018A030313354).

References
[1] Abdul Mueed Hafiz, Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. 2020, 9(3):171-189.
[2] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.
[3] Z. Qin et al., "Demographic information prediction based on smartphone application usage," 2014 International Conference on Smart Computing, Hong Kong, 2014, pp. 183-190, doi: 10.1109/SMARTCOMP.2014.7043857.
[4] Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT press.
[5] Krizhevsky A, Sutskever I, Hinton GE Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, 2012. pp 1097-1105.
[6] He K, Gkioxari G, Dollar P, Girshick R (2018) Mask R-CNN. IEEE transactions on pattern analysis and machine intelligence. doi:10.1109/tpami.2018.2844175.
[7] Redmon J , Farhadi A. YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.
[8] Lin, T.Y., Goyal, P., Girshick, R., He, K., Doll’ar, P.: Focal loss for dense object detection. In: Proc. IEEE Int. Conf. Comp. Vis. (2017).
[9] Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: YOLACT: Real-time instance segmentation. In: Proc. IEEE Int. Conf. Comp. Vis. (2019).
[10] Wang, Yilei, Tang, Y., Ma, J., & Qin, Z. (2015). Gender Prediction Based on Data Streams of Smartphone Applications. In Yu Wang, H. Xiong, S. Argamon, X. Li, & J. Li (Eds.), Big Data Computing and Communications (pp. 115–125). Cham: Springer International Publishing.
[11] Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: ICLR, (2017).
[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In ICLR, 2015.
[13] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollar P, Zitnick L (2014) Microsoft COCO: Common Objects in Context. 65.
[14] O. Kwabena, Z. Qin, T. Zhuang and Z. Qin, "MSCryptoNet: Multi-Scheme Privacy-Preserving Deep Learning in Cloud Computing," in IEEE Access, vol. 7, pp. 29344-29354, 2019, doi: 10.1109/ACCESS.2019.2901219.
[15] Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., Luo, P.: Polar- Mask: Single shot instance segmentation with polar representation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recog. (2020).
[16] Z. Qin et al., "Learning-Aided User Identification Using Smartphone Sensors for Smart Homes," in IEEE Internet of Things Journal, vol. 6, no. 5, pp. 7760-7772, Oct. 2019, doi: 10.1109/JIOT.2019.2900862.
[17] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In CVPR, 2018.
[18] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), pages 325–341, 2018.
[19] Redmon J , Farhadi A . YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.
[20] Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikäinen M (2020) Deep Learning for Generic Object Detection: A Survey. International Journal of Computer Vision 128 (2):261-318. doi:10.1007/s11263-019-01247-4 28.
[21] Wang X, Kong T, Shen C, Jiang Y, Li L (2019) SOLO: Segmenting Objects by Locations. arXiv preprint arXiv:191204488.
[22] Nekrasov V , Shen C , Reid I . Diagnostics in Semantic Segmentation[J]. 2018.
[23] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018.
[24] Zhen Qin, Gu Huang, Hu Xiong, Zhiguang Qin, and Kim-Kwang Raymond Choo. A Fuzzy Authentication System Based on Neural Network Learning and Extreme Value Statistics. IEEE Transactions on Fuzzy Systems. 2019. DOI: 10.1109/TFUZZ.2019.2956896.