Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

An overview of important models of semantic segmentation

Tao Wang*, Kun Wang, Chen Yang

The College of Nuclear Technology and Automation Engineering, Chengdu University of Technology, Chengdu, Sichuan, China

Abstract: With the continuous development of deep learning and computer vision technology, image segmentation technology has been fully developed and applied, and more than a hundred related models and methods have been produced. This article starts with the FCN model of the landmark in the field of image segmentation, and analyzes the seven most representative methods in the field of image segmentation. It mainly covers mainstream processing technologies such as fully connected network, encoder-decoder structure and its variants, specialized pooling structure, multi-scale processing technology, hole convolution structure and conditional random field structure. Explains the basic ideas of these methods, main innovations and existing problems, as well as subsequent models to improve the previous work.
Keywords: Image segmentation, codec, hole convolution, multi-scale.

1. INTRODUCTION

With the development of deep learning, computer vision technology has been continuously developed. Image segmentation technology is an extremely important branch of computer vision. Image segmentation can be roughly understood as a matting operation, which is to select the region of interest. Image segmentation technology is widely used, including semantic segmentation, target positioning, medical image processing, satellite image analysis and so on. Image segmentation has a long history of development, but the real rise of image segmentation technology based on deep learning began with the proposal of the Fully Convolutional Network (FCN) model[1], the pioneering work in the field of image segmentation. After the FCN model was proposed, many researchers improved and upgraded it, making the image segmentation technology basically mature.

Due to the rapid development of image segmentation technology, various models and ideas emerge in an endless stream, so it is not easy for beginners to quickly enter this field. Existing review articles in the field of image segmentation have elaborated on all representative methods in this field, but for beginners, it is extremely important to quickly grasp the most classic centralized model. Therefore, this article mainly analyzes the most several important models, including their innovations and main existing problems, as well as subsequent improvements to them.

2. MODEL STRUCTURE

2.1 FCN

Fully Convolution Networks (FCN)[1] is a landmark structure in the field of image semantic segmentation, and it first achieved pixel-level classification. FCN replaces the fully connected layer of the traditional CNN network with a fully convolutional layer, and then outputs a feature map of the same size as the input image, and then obtains the classification information of each pixel through softmax to complete the image segmentation finally. The original image extracts feature map through five convolution and maximum pooling operations, then connects two full convolution layers, and finally restores the feature map to the original image size through multiple upsampling. In order to better restore image details, FCN combines the maximum pooling feature map and up-sampling feature map to improve the accuracy of image segmentation.

2.2 SegNet

FCN uses deconvolution upsampling to restore the size of the original image when upsampling, but this method will lose some of the details in the image. The SegNet[2] structure uses a location index method to solve this problem. When SegNet performs maximum pooling downsampling, it first stores the maximum position index of each pooling area in memory, and then when upsampling, it completes the upsampling assignment in the depooling process according to the stored index position. This method guarantees the preservation of image details during the image restoration process.

2.3 U-Net

In the image restoration process, the FCN structure directly adds the down-sampling feature and the restored up-sampling feature point by point. U-Net network[3] uses another more effective method to complete the fusion of feature images. U-Net directly stitches the down-sampling feature map and the up-sampling restored feature map into a thicker feature map. It is guaranteed that the down-sampling feature information can be directly output to the restored image feature map without being lost, and the accuracy loss problem caused by the up-sampling process is solved.

2.4 PSPNet

The main problems in the field of segmentation are manifested as mismatched relationships, types of confusion, and insignificant but ignored important categories. The author analyzes that the main reason for the error of semantic segmentation is that the existing model does not introduce enough context information. The PSPNet[4] model uses the pyramid pooling module to design a global scene-level a priori scene analysis network. The convolutional network of the PSPNet network uses the ResNet model. After the feature map extracted by the ResNet convolutional network is pooled in four different sizes, the feature information of the four different sizes is obtained, and then the channel thickness of the feature information is compressed by 1*1 convolution. Then the image size is restored through upsampling and spliced together with the original image feature information, and finally the image segmentation is completed. Another important feature of the PSPNet model is the result of assisting Loss to optimize image segmentation in the ResNet convolutional network.

2.5 R-CNN

The RCNN series of algorithms propose a novel image segmentation idea, that is, a deep learning target detection algorithm based on candidate regions. R-CNN[5] draws on the idea of sliding window,

extracts 2000 independent candidate regions by processing the input image, and then extracts a fixed-length feature vector for each region, and finally uses SVM for each region to complete the target classification.

Since R-CNN needs to extract 2000 regions of interest and consumes huge computer resources, Fast R-CNN[6] is proposed. Fast R-CNN sends the entire picture to the cnn network for feature extraction during training. It does not need to input each candidate region into the CNN for processing like the RCNN model, thereby improving the efficiency of the use of cnn. At the same time, the SVM classifier is no longer used, but the extracted feature information and candidate regions are directly input into the optimizer.

Fast-RCNN[7] is still not fast enough to generate candidate regions, so the Faster-RCNN model is proposed. The biggest innovation of the Faster-RCNN model is the use of RPN networks to generate candidate regions. When generating the candidate area, anchors are generated, and then it is judged whether the anchors belong to the foreground or the background, and finally the accurate candidate area is obtained through border regression.

2.6 DeepLab

DeepLab is a new semantic segmentation network proposed by the Google research team, and a total of 4 versions have been updated so far. DeepLab v1[8] mainly solves two main problems faced by CNN when processing image segmentation. Downsampling leads to loss of image detail information. The spatial invariance of the CNN model makes the features extracted by CNN not fine enough. DeepLab v1 mainly uses hole convolution and conditional random fields to solve the above problems. DeepLab v1 uses hole convolution instead of maximum pooling convolution, which can obtain more context information without reducing the resolution of the feature map. At the same time, a conditional random field is introduced to better recover the boundary information of the image.

Compared with the first version of DeepLab, DeepLab v2 replaces the CNN network from the VGG model with the ResNet model. At the same time, the spatial pyramid pooling is proposed to solve the multi-scale problem of images. The image is transformed into different scales by spatial pyramid pooling, and the multi-scale image features are obtained by using different void ratios under different branches.

The main contribution of DeepLab v3[10] is to improve the spatial pyramid pooling model. Serial hole convolution is used in the residual block of the ResNet network. In the same residual block, multiple holes of different sizes are used in parallel to extract the multi-scale features of the image. In order to solve the gridding problem of hollow convolution, multi-grid is proposed. Finally, 1*1 convolution is introduced, and image-level features are directly introduced to ensure the accuracy of segmentation.

2.7 RefineNet

In the image segmentation model described above, although FCN and its derivative networks can retain the low-dimensional and high-dimensional features of the image, deconvolution cannot restore the low-level feature information, and the spatial information of the image is lost. The hollow convolution of the DeepLab model, for large-size images, will not only increase the consumption of

computing power, but also require more memory. Not only that, but the feature of the hole convolution also loses some detailed information.

RefineNet[11] proposes a multi-path enhanced network to complete image segmentation. Use multi-level abstract feature information to achieve high-resolution semantic segmentation. RefineNet mainly includes three core modules: residual convolution module, multi-scale fusion module and chain residual pooling module. The residual convolution module contains multiple ReLU activations and 3*3 convolution operations, and then uses addition to fuse the feature information before and after. In the multi-scale fusion module, the input feature information of multiple resolutions is restored to the same size through an up-sampling operation, and all feature maps are fused. In the chain residual pooling module, a series of different pooling operations are used to obtain more contextual information from the large background, and the effective features obtained from multiple pooling windows are combined and merged.

3. CONCLUSION

This article mainly analyzes the classic models and frameworks in the field of image segmentation, analyzes the main contributions of these models, and the main innovations relative to the previous works. Compared with other review articles, this article selects the 7 most representative models in the field of image segmentation, and introduces them in a more concise language and short space. We started with the FCN model, a pioneering work in the field of image segmentation, and discussed the improved models SegNet model and U-Net model proposed for its main problems. Then discussed the classic model PSPNet that started to consider global information and context information, then discussed the RCNN model based on candidate regions to complete image segmentation, then discussed a new model DeepLab series model proposed by Google Labs, and finally discussed an efficient The image segmentation model RefineNet.

Image segmentation technology is a field full of challenges and opportunities. At present, many important problems in the field of image segmentation have been solved, but there are still many detailed problems that significantly affect the results. We look forward to the new technologies and new ideas that will be brought forward with the resolution of these detailed problems.

REFERENCES

[1] Long J , Shelhamer E , Darrell T . Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Tra nsactions on Pattern Analysis and Machine Intelligence, 2015, 39(4):640-651.

[2] Badrinarayanan V, Handa A, Cipolla R (2015) Segnet: A deep convolutional encoderdecoder architecture for image segmentation. arXiv:151100561

[3] Ronneberger O , Fischer P , Brox T . U-Net: Convolutional Networks for Biomedical Image Segmentatio n[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Spri nger International Publishing, 2015.

[4] Zhao H , Shi J , Qi X , et al. Pyramid Scene Parsing Network[J]. 2016.

[5] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object dete ction and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[6] Girshick, R. (2015). Fast r-cnn. arXiv preprint arXiv:1504.08083.

[7] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with regi on proposal networks. In Advances in neural information processing systems (pp. 91-99).

[8] Liangchieh C , Papandreou G , Kokkinos I , et al. Semantic Image Segmentation with Deep Convolutiona l Nets and Fully Connected CRFs[J]. Computer ence, 2014(4):357-361.
[9] “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully conn ected crfs,” arXiv preprint arXiv:1606.00915, 2016.
[10] Chen L C , Papandreou G , Schroff F , et al. Rethinking Atrous Convolution for Semantic Image Segmen tation[J]. 2017.
[11] Lin, G. , Milan, A. , Shen, C. , & Reid, I. . (2017). RefineNet: Multi-path Refinement Networks for High -Resolution Semantic Segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognitio n (CVPR). IEEE.