High-Resolution Remote Sensing Image Semantic Segmentation Method Based on Improved Encoder-Decoder Convolutional Neural Network

Xinyu Zhang, Ying Zhang, Jianfei Chen, Huijun Du

High-resolution Remote Sensing Image Semantic Segmentation Method Based on Improved Encoder-Decoder Convolutional Neural Network

Zhang Xinyu[1][*], Zhang Ying[1], Chen Jianfei[1] and Du Huijun[1]

[1] State Grid Shandong electric power company Tai'an power supply company, 8 Dongyue street, Taishan district, Tai'an, Shandong Province, China
Weike.liu@163.com.com

Abstract. In recent years, Convolutional neural network with encoder-decoder structure is a kind of image semantic segmentation method with high accuracy. However, the characteristics of large amount of parameters and high requirements for computing power restrict its application in the fields of limited computing power and high real-time requirement, such as unmanned driving, road monitoring, remote sensing classification and mobile object detection. To solve the above problems, this thesis firstly designs the dilated convolution combination module, which solves the gridding problems while ensuring large receptive field; then, a double-channel encoder-decoder convolutional neural network is built by using the dilated convolution module combined with the depth separable convolution. Using this network, the parameters and computation of semantic segmentation convolution model of high resolution remote sensing image are greatly reduced while maintaining high segmentation accuracy. Through experiments on GID data sets, and compared with a variety of semantic segmentation methods, this thesis verifies the effectiveness and light-weight of this method.

Keywords: Semantic Segmentation, Dilated Convolution Combination, Receptive Field, Gridding Problems, Depthwise Separable Convolution.

1 Introduction

With the development of remote sensing technology, high-resolution remote sensing images have become the key research object in the field of remote sensing because they have more abundant and accurate ground object information. High resolution remote sensing images have the characteristics of changeable spectral characteristics, detailed texture characteristics, obvious geometric structure and clear context information, which provides more convenience for automatic classification of high resolution remote sensing images. However, because it contains more semantic and detailed information, some traditional methods, such as SVM [1], Watershed, Random-Forest and so on, do not perform well in the classification of high-resolution remote sensing images. The interference factors of the task are mostly the same spectrum foreign matter and the

same object different spectrum. Therefore, we need to find a more accurate classification algorithm.

In 2014, using the deep learning, LONG[2] proposed the first image semantic segmentation network: Full convolutional networks (FCN)[3], and then advanced semantic segmentation networks such as SegNet[3], U-Net[4][5], DeepLab[6] series and ESPNet[7] series were proposed. The encoder-decoder structure[8][9] network represented by U-Net has attracted the attention of scholars because of its high segmentation accuracy, such as U-Net++ [10], Refine-net, SegNet and Deeplab-v3+[11]. The convolutional neural network of encoder decoder structure is divided into two parts. The encoding part obtains the feature expression from detail to whole through convolution and pooling, and the decoding part obtains the semantic label from whole to detail through deconvolution. However, this kind of network has large volume, many parameters and high requirements for computing power. In order to reduce the computational power requirements, a variety of lightweight semantic segmentation networks have been proposed, such as Refine Net-LW[12], LiteSeg[13], ESPNet-v2[14], which greatly improves the network training speed, but this performance is obtained at the cost of reducing the accuracy of semantic segmentation. To solve this problem, on the premise of ensuring the accuracy, in order to improve the convolution efficiency and reduce the amount of network parameters, this thesis designs the dilated convolution combination based on the encoder-decoder structure convolution neural network structure and aiming at expanding the receptive field, and solves the gridding problem in the dilated convolution by standardizing the dilation rate in the dilated convolution combination, The deep separable convolution is used to reduce the parameters of the model, and the learning of feature details is improved by establishing the double connection structure of long connection and pooled index of the corresponding codec layer. And finally, the encoder-decoder structure network is built, and an efficient light-weight semantic segmentation method is obtained.

2 Network Structure

In view of the excellent performance of convolutional neural network of the encoder-decoder structure in the field of semantic segmentation, it can be used for ground object classification and extraction with high-resolution remote sensing images. In order to better deal with the problem of large parameters and large amount of calculation of convolution neural network caused by large spatial size and rich detailed features of high-resolution remote sensing images, this thesis establishes a double connection between encoder-decoder structure by using deep separable convolution[15], and combines the dilated convolution non-gridding problem, and has large receptive field, It enhances the learning ability of encoder-decoder network to detail features and large-scale features, and reduces the amount of network parameters and computation.

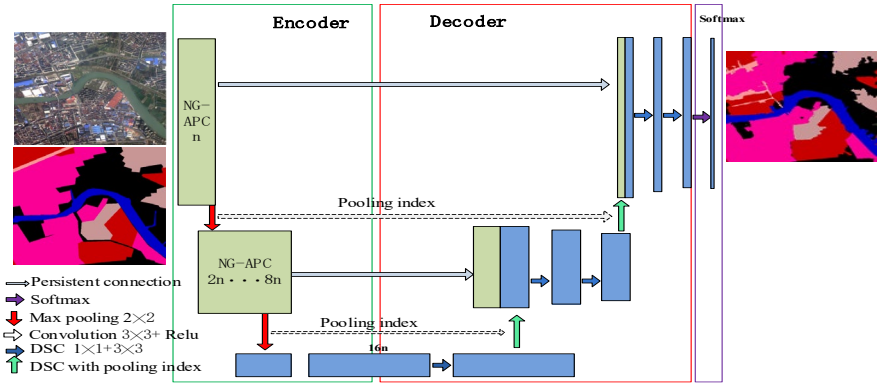


Fig. 1. Structure the encoder decoder network

2.1 Light-weight

In order to realize light-weight convolution neural network, it is an effective method to transform convolution into deep separable convolution. Its principle is to decouple the mapping of correlation and spatial correlation between channels in convolution layer, and map them separately, which can achieve the same convolution effect, and the amount of parameters can be greatly reduced. Deep separable convolution is used in lightweight semantic segmentation models such as Xception [17], Mobile-Net [18], ShuffleNet [19], Squeeze-Net [20]. Therefore, a deep separable convolution encoder-decoder network structure is proposed, which not only improves the network training speed, but also reduces the amount of parameters and calculation of the network.

Specifically, the standard convolution (except the input layer) in the encoder-decoder network is improved to a separable two-steps convolution structure. Firstly, the characteristic diagram of each input channel is spatially convoluted, and the characteristic combination of an input channel is obtained through the training of the parameter value of the convolution kernel. Secondly, the small convolution kernel is used to confuse the channel dimension to obtain the combined characteristics of the channel dimension. And the parameters fell from $M \times N \times n^2$ to $M \times n^2 + M \times N \times 1^2$, which the number of input channels is M, the number of output channels is N, and the kernel size is n^2.

2.2 Double Connection

According to the convolutional neural network of various codec structures, it can be seen that the key of codec structure is how to accurately complete the sampling on the image and the feature extraction after sampling. To solve this problem, two channel connection is adopted in the network:

i. Establish the feature splicing of the corresponding coding layer and decoding layer. The long connection structure is a larger scale cross layer connection, which can

effectively solve the problem of gradient disappearance and improve the learning of detailed features

ii. Establish the deconvolution connection with the maximum pool index, establish the maximum pool index in the coding layer, and pass it to the deconvolution operation of the corresponding decoding layer by establishing pooled index, we can reduce the loss of detail features and improve the learning of detail features

The cross layer connection of two channels can effectively solve the problem of gradient disappearance and improve the learning of detail features.

2.3 Receptive fields

In order to meet the needs of high resolution remote sensing images for large receptive fields, on the premise of solving the gridding problem, in order to maximize the receptive fields and keep the number of parameters unchanged, we designed a three-layer cascade and multi branch parallel dilated convolution combination[21]. The first part is the cascade part, which is the cascade dilated convolution with a dilation rate of 3, after the standard convolution. If the dilated convolution is selected for the first layer, the characteristics around the sampling center point will be missed. After the cascade, this phenomenon of missing sampling will be exacerbated and cannot be completely no-gridding. So the standard convolution [22] is selected for the first layer, the dilated convolution with a dilation rate of 3 is selected for the second layer. The third layer is two branch Parallel Grouping convolution. The first branch is the dilated convolution with a dilation rate of 9, which can achieve the dilated convolution of the non-gridding problem. The receptive field is 27 pixels. In order to further expand the receptive field, the second branch is the dilated convolution with a dilation rate of 18, and the characteristics of the missing sampling part are supplemented by the first branch. Thus, the gridding problem can be maximized, and the dilation convolution combination of receptive field is 45 pixels. After the activation function and feature splicing, a convolution module is formed, and the module structure is shown in Figure 2.

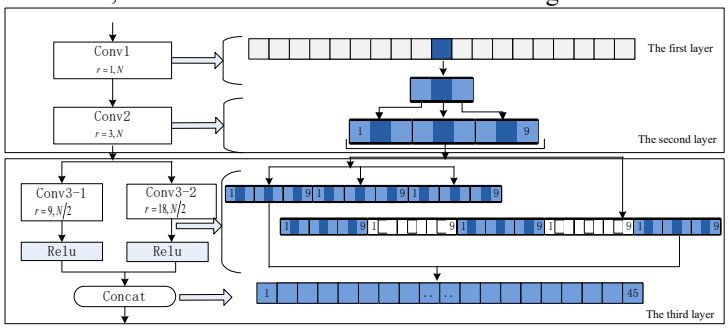


Fig. 2. Three layers and two branches of NG-APC module diagram

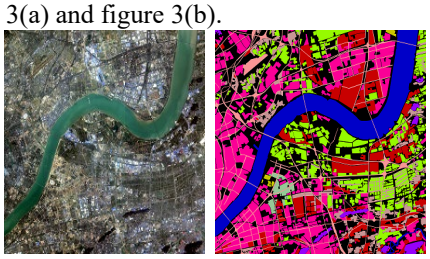
3 Experiment verification and analysis

In order to verify the effectiveness of the method proposed in this thesis in high resolution remote sensing image classification, this paper uses the GID dataset [23] with GF-2 images as the data source for training and testing.

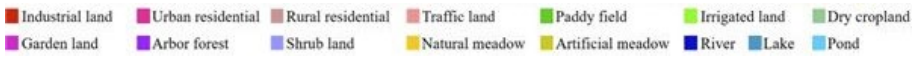
3.1 Dataset and Training process

The feature classification dataset released by Wuhan University, 2018, using GF-2 images, is a large dataset for land use and land cover classification. It contains over of the 150 images, whose images including more than 50,000 KM^2 of 60 different cities in China, the high-precision land cover includes 10 images of 15 types with the size of 7200*6800*3 pixels.

In the experiment, an image from southern China is selected, as shown in Figure. 3(No.gf2_pms2_l1a0001471436-mss2). Because the image is too large, direct training using the original image will lead to computer overload, and the positive sample ratio is low (the effective data ratio is very low), so a series of processing such as atmospheric correction, orthographic correction, image registration and image fusion should be carried out first, Then, the fused high resolution image is preprocessed with a series of data, such as annotation, cutting and data amplification. After the preprocessing operation of the above process, the learning data set that can be used to identify the model is obtained, including training set, verification set and test set, with a total of about 10000 samples. Among them, there are 10000 training sets, 1000 verification sets and tests, the sample size is 600*600, and 12 pixels are overlapped between adjacent samples to avoid damaging small target objects. The sample includes image data and annotation data, as shown in Figure 3(a) and figure 3(b).



(a)Panchromatic image (b) Image label



(c)Labels to different features

Fig. 3. Experimental image of GID datasets

3.2 Experiment results and analysis

The parameters and computation quantity are compared with the classical encoder-decoder network DeconvNet, U-Net, SegNet, and lightweight encoder-decoder network ESP-Netv2, LiteSeg and RefineNet-LW. The result of comparison is shown in Table1.

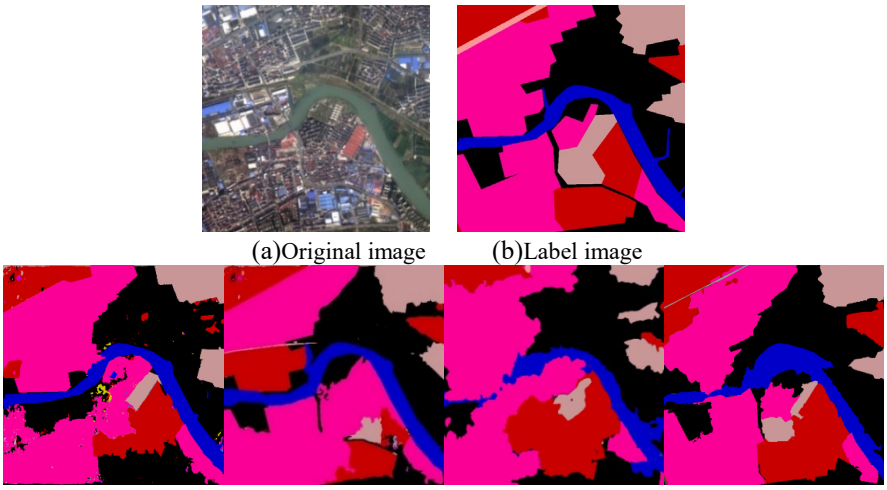


Fig. 4. Comparison of classification results of first GID images by different methods

Table 1. Comparison of operation results of different image semantic segmentation algorithms

CNN	Encoder Parameters /MB	Decoder Parameters /MB	Softmax Parameters /KB	Training Speed Frames/S	Total Parameters /MB	Calculation Quantity GFLOPs
DeconvNet[16]	131.3	131.3	0.13	0.1	262.7	1814
U-Net	18.85	12.18	0.13	0.3	31.03	217.
SegNet	64.1	52.9	0.13	0.2	117	218
ESPNetv2	3.49	3.28	0.13	1.3	6.79	22.6
LightSeg					20.55	57.4
RefineNet-LW					46	52
This thesis method	4.96	2.42	0.13	1.0	7.39	43.27

Table 2. Accuracy list of the first GID image classification results by different methods

	U-Net	MobileNet	ESPNetv2	This thesis method
Accuracy /%	68.68	67.51	69.29	78.84
mIoU /%	52.9	50.8	54.5	60.0

It can be seen from the Table1 that among a variety of encoder decoder convolutional neural networks, the parameters and computation quantity of this thesis method belong to lightweight networks, which are much smaller than DeconvNet, U-Net and Segnet, and are equivalent to lightweight ESPNetv2. At the same time, it is not far from the current mainstream lightweight convolutional neural networks LightSeg and RefineNet-LW networks. Therefore, in general, this thesis method is a lightweight encoder decoder convolutional neural network.

The accuracy and mIou value are compared with the classical encoder-decoder network U-Net, lightweight encoder-decoder network ESP-Netv2, LiteSeg and RefineNet-LW. The result of comparison is shown in Table2.It can be seen from the segmentation results in Table2 that this method can distinguish high resolution images well. Compared with U-Net, MobileNet and ESPNetv2, the accuracy of this method is improved by 10.16%, 11.33% and 9.55% respectively, and the mIou value is improved by 7.1%, 9.2% and 5.5% respectively, which verifies the advantages of the new method in algorithm accuracy.

4 Conclusion

Based on the U-Net model, using dilated convolution combination and depthwise separable convolution, this thesis method constructs a lightweight semantic segmentation model based on convolution neural network. At the encoder part, the NG-APC dilated convolution combination module is used to solve the gridding problem in dilated convolution combination. At the decoder part, depthwise separable convolution is used to reduce the amount of parameters and calculation of this model, and reduce the dependence of the model on computational power. Through experimental verification on GF-2 image dataset, accuracy reaches 78.84%, and the parameter quantity is only 7.39M, which is smaller than U-Net model and many lightweight semantic segmentation models based on U-Net. The results show that this thesis method is a lightweight and efficient image semantic segmentation method with encoder-decoder structure.

References

1. Madan, S., Pranjali, C.: A Review of Machine Learning Techniques Using Decision Tree and Support Vector Machine. In: International conference on computing Communication Control& Automation, pp.1-7. IEEE Piscataway, Pune, India (2017).
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(4), 640-651(2014).
3. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12), 2481-2495(2017).
4. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation.In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp.1-8. Springer International Publishing, London, UK (2015).
5. Nanjun, H., Leyuan, F., Plaza, A.: Hybrid first and second order attention U-Net for building Segmentation in remote sensing images. Information Sciences 63(140305), 69-80 (2020).
6. Liang, C., George, P., Iasonas, K., et al.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(4), 834-848 (2018).
7. Mehta, S., Rastegri, M., Caspi, A., et al.: ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In: European Conference on Computer Vision,pp.561-580.Springer, Cham (2018).

8. Huihui, H., Weitao, L., Jianping, W., etc.: Semantic segmentation of encoder-decoder structure.Journal of Image and Graphics 25(02), 255-266(2020).
9. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentaion. IEEE Trans on pattern analysis and machine intelligence 39(12), 2481-2495(2017).
10. Norman, B., Pedoia, V., Majumdar, S.: Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry. Radiology 288(1), 1109-1122(2018).
11. Wang, Y., Sun, S., Yu, J., et al.: Skin lesion segmentation using atrous convolution via DeepLab v3. ArXiv vol(1),1-4(2018).
12. Nekrasov, V., Shen, C., Reid, I.: Light-weight refine-net for real-time semantic segmentation. In: BMVC, Newcastle upon Tyne, pp.1-15, England(2018).
13. Emara, T., Hossam, E., Abd, E.: LiteSeg: a novel lightweight convnet for semantic segmentation. In: 2019 Digital Image Computing: Techniques and Applications (DICTA), pp.1-7. Perth, Australia (2019).
14. Mehta, S., Rastegari, M., Shapiro, L., etc. ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network. In: CVPR, pp.1-10.IEEE, CA, USA (2019).
15. Xiaoqing, Z., Yongguo, Z., Weike, L., et al.: An improved architecture for urban building extraction based on depthwise separable convolution. Journal of Intelligent and Fuzzy Systems 38(11), 1-9(2020).
16. Noh, H., Hong, S., Han, B.: Learning Deconvolution Network for Semantic Segmentation. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1520-1528.IEEE, Santiago (2015).
17. Francois, C.: Xception: Deep Learning with Depthwise Separable Convolutions. Maryland：CVPR, pp. 1800-1807. IEEE, Honolulu, HI, USA (2017).
18. Akay, M.， Du, Y.， Sershen, CL., et al.: Deep Learning Classification of Systemic Sclerosis Skin using the MobileNetV2 Model [J]. IEEE Open Journal of Engineering in Medicine and Biology (99), 104-110(2021).
19. Zhang, X., Zhou, X., Lin, M., et al.: ShuffleNet： an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 6848-6856.IEEE, Salt Lake City, US(2018).
20. Qi, Z., Nauman, R., Shuchang, L., et al.: RSNet: A Compact Relative Squeezing Net for Image Recognition. In: VCIP, pp.1-4. IEEE, NSW, Australia (2019).
21. Xiaoqing, Z., Yongguo, Z., Weike, L., et al.: A hyperspectral image classification algorithm based on atrous convolution. EURASIP Journal on Wireless Communications and Networking 1(1), 1-12(2019).
22. Chen, LC., Papandreou, G., Kokkinos, I., et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. Computer Science 22(4), 357-361(2014).
23. Zhang, Y., Chi, M.: Mask-R-FCN: A Deep Fusion Network for Semantic Segmentation. IEEE Access (8), 155753-155765(2020).