Semantic Segmentation of Remote Sensing Images Based on Multi-Model Fusion

Zuohua Miao¹,², Likun Liu¹, Lei Ren¹, Yang Tang¹, Yong Chen¹,² and Jun Li¹,²

¹School of Resources and Environmental Engineering, Wuhan University of Science and Technology, Wuhan 430081, China
²Hubei Key Laboratory for Efficient Utilization and Agglomeration of Metallurgic Mineral Resources, Wuhan 430081, China
Email: liulikun0313@163.com

Abstract. Convolutional neural networks have created a new field in the research of semantic segmentation of remote sensing images. However, different network structures have different effects on the semantic segmentation of different land types. In this paper, the original data set is expanded, and an improved U-Net model is used to train a model for each type of feature target. Then combined with conditional random field (CRF) and image overlapping strategy for optimization processing; Finally, the two binary classification models obtained by training are fused to obtain multi-classified semantic segmentation images. Solve the obvious problem of large-scale remote sensing image edge stitching. The experimental results show that this method has higher accuracy in solving the segmentation of large-scale remote sensing images.

1. Introduction

Semantic segmentation of an image refers to adding a semantic label to each pixel in the image, and segmenting the image into multiple regions with different semantic labels based on the semantic unit, enabling the computer to recognize the image content. The semantic segmentation of remote sensing images is widely used in urban planning, disaster prevention, and GIS information construction.

When computer hardware is not enough to support deep learning neural networks, researchers have come up with various methods to implement image semantic segmentation technology. For example: the use of pixel-level threshold segmentation [1], pixel-based clustering segmentation [2], "graph cut" image segmentation [3], pixel-based decision tree classification [4], etc. Among them, the most commonly used is the graph segmentation method to achieve semantic segmentation of images.

Until 2015, the IEEE International Conference on Computer Vision and Pattern Recognition Fully Convolutional Networks for segmentation (FCN) was proposed by Long [5] et al. Segmentation technology has entered the period of FCN. FCN have shown great potential in deep learning, and deep learning methods have become the mainstream of solving semantic segmentation problems today. Among them, deep convolutional neural network (DCNN) has achieved great success in image classification, target detection, scene understanding, etc. Compared with traditional image segmentation algorithms, the accuracy of image segmentation has been greatly improved.

The U-Net network model proposed by Olaf Ronneberger [6] et al has achieved very good results in the field of medical image segmentation. The most basic U-Net network structure is used for medical image segmentation. The segmentation accuracy is very high when processing images with single background, low complexity, and binary classification [7]. Since then, various improved versions based on U-Net network models have been widely used each field.
Due to the complexity of the image information contained in the remote sensing image [8], the conventional machine learning-based feature difference detection method has low detection efficiency, and the deep neural network-based semantic segmentation has rarely been studied in the field of remote sensing image difference recognition. In this paper, through the improved U-Net network structure, each type of feature is divided into two categories for semantic segmentation, and finally the predicted subgraphs are fused to obtain the final semantic segmentation image.

2. Basic Theory

2.1. Improved U-Net Semantic Segmentation Network

Many semantic segmentation network structures are improved based on FCN, including U-Net. Generally speaking, U-Net includes two parts: the first part is down-sampling and feature extraction; the second part is up-sampling. It is because of the overall U shape in the network architecture, which looks very elegant, so it is called U-Net. The network structure is shown in Figure 1:

![Figure 1. Structure diagram of the improved U-Net network](image)

As shown in the figure: the input and output image sizes of the network structure are 256 x 256, and there are 19 convolutional layers, including 4 downsampling and 4 upsampling.

Introduce a residual module in the U-Net network. This residual module contains skip connections, while retaining the same skip connections between the corresponding feature maps of the encoder and decoder as U-Net. Introducing the residual module can speed up the network convergence speed, and it is more adaptable to deep network structures. At the same time, the residual link not only carries the low-level features of the pre-order layer, but also contains the high-level features obtained in the post-order layer. Thus, more efficient feature reuse can be achieved.

The left half of the network structure is down-sampling, which is the most typical structure in convolutional neural networks [9]. It includes 4 sets of convolutions and maximum pooling. Every two convolutions are pooled. When the convolution depth is sufficient, the deep features are extracted to the maximum extent. Moreover, each layer of the network is normalized to make the distribution of features in each layer more uniform, so as to achieve the purpose of speeding up the model convergence speed and improving the model's fault tolerance.

The right half of the network is axisymmetric with the left half, and consists of a series of upsampling slaves, including 4 groups of Upsampling and conv. The convolution of each group is in addition to the deep abstract features obtained by upsampling the upper layer. It will also be merged with the shallow features in the down-sampling layer corresponding to the left, so that the deep abstract features can be better extracted while the shallow detailed features are retained to the maximum extent, and the corresponding spatial information dimension is unchanged.

The special structure of U-Net network structure makes it have the advantages that other network structures do not have:
(1) Local perception; the convolutional neural network only connects the local areas of the image, which greatly reduces the amount of redundant parameters and effectively improves the speed and accuracy of training.

(2) Weight sharing; the convolutional neural network draws on the relevant concepts of the visual neural receptive field, so that one of the local neurons will perform weighted calculations on all regional pixels in the image, so as to achieve weight sharing, which can make the network model have the invariability of image translation [10].

2.2. Conditional Random Field (CRF)
Conditional Random Field (CRF) is a discriminant probability model. It was first proposed by Lafferty [11] and so on. It combines the characteristics of maximum entropy model and hidden Markov model, and is widely used in the fields of part-of-speech tagging and named entity recognition. The conditional random field can be expressed as an undirected graph model \( G = \{ V, E \} \), where \( V \) represents a set of nodes on the way and \( E \) represents a set of undirected edges between nodes. Each node \( V \) corresponds to a random variable \( Y_v \), so \( Y = Y_v \mid v \in V \). Under the condition of observing the sequence \( X \), each random variable \( Y_v \) satisfies Markov characteristics, that is:

\[
P = (Y_v \mid X, Y_{\omega}) = p(Y_v \mid X, Y_{\omega}, \omega \sim v)
\]

(1)

Where \( \omega \sim v \) indicates that \( w \) and \( v \) are two adjacent nodes in the graph.

2.3. Image Overlay Strategy
Faced with the prediction of large-scale remote sensing images, we cannot directly predict large-scale images, we need to cut the large-scale remote sensing images into several small parts, and then predict them one by one, and finally stitch them together. It is precisely because the thumbnails used for prediction are artificially cut, so there may be incomplete buildings at the edges of each thumbnail, so that the trained model cannot accurately identify it, and the stitching traces will appear obviously when stitching again.

To solve this problem, overlap is performed when cutting large-scale remote sensing images, as shown in Figure 2:

![Figure 2. Overlapping cuts](image)

The large yellow frame and the red frame are the edge tracks of the large-scale remote sensing image, and then the large yellow and red maps are used for prediction. The predicted maps are cut according to the small yellow and red frames in the above figure, and then stitched together. Solve the crack feeling after splicing.
3. Experimental Analysis

3.1. Experimental Parameter Settings

This experiment is based on the deep learning framework of Keras. The experimental environment is the operating system: window10 Professional Edition; CPU: Intel Xeon (R) Gold 5120; memory: 128G; GPU: NVIDIA GeForce RTX 2080Ti.

The input size of the remote sensing image is 256 × 256, the total training round is 200 rounds, and the initial benchmark learning rate is 0.0025. Each round of training includes a process of forward propagation and back propagation. Value comparison generates cross-entropy loss, and then backpropagation updates the model weight parameters.

3.2. Data Enhancement

The remote sensing image selected in the experiment comes from the "AI Classification and Recognition Competition of CCF Satellite Images". The data is a high-resolution remote sensing image. The spatial resolution of the image is sub-meter level. The spectral visible band (R, G, B) has been removed Coordinate information. Use labelme to manually label the buildings, roads and water bodies of the map. In order to expand the experimental sample, the original image and the marked image are randomly cut into 256 × 256 small images, and the small images are flipped, color adjusted, and noise added. Through the above operations, the training set is expanded to 20,000 thumbnails.

3.3. Evaluation Index

In the experiment, pixel accuracy (PA), mean pixel accuracy (MPA) and mean intersection over Union (MIoU) were selected as the evaluation indicators of semantic segmentation.

\[
PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}}
\]

Among them, \( p_{ii} \) represents the pixels predicted correctly, and \( p_{ij} \) and \( p_{ji} \) represent false positives and false negatives.

On this basis, MPA is an improvement of PA, calculating the proportion of each pixel correctly classified, and then averaging all classes.

\[
MPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}}
\]

MIoU represents the ratio of the intersection and union of the true and predicted values:

\[
MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}
\]

3.4. Analysis of Experimental Results

After about 40 hours of training, 98.80% training accuracy and 98.12% test accuracy were obtained on the data set.

As shown in Fig. 3: from left to right are the prediction images of three types of features, building prediction maps, road prediction maps, and water prediction maps. From top to bottom are the images under the traditional U-Net, improved U-Net, and optimized U-Net model.

It can be seen from the figure that the edges of the traditional U-Net prediction map are blurred, and the stitching trace of the small image is obvious. It can be clearly seen from the figure: Compared with the traditional U-Net network, the improved U-Net has greatly improved the segmentation effect. On this basis, conditional random field post-processing and image overlapping strategy are used. Finally, the building model, road model and water model are fused. Since the three prediction maps are fused together, there may be a problem that one pixel belongs to two or three categories at the same time. Therefore, the integrated learning voting strategy can be used to remove some pixels with obvious classification errors when the model is fused. Point, greatly improving the predictive ability of the model. Because the image is a remote sensing image of a city, the largest proportion of buildings in the figure is followed by the road and finally the water body, so the priority set when the model is
fused is building > road > water body, on this basis, for each prediction map the pixels are voted to predict the category of the pixel at the corresponding position of each picture. The category with the most votes is the category of the pixel. As shown in Figure 3:

![Figure 3. Semantic segmentation results](image)

Compared with the Segnet multi-classification model, this paper trains multiple binary classification models, and then merges the multiple binary classification models. This method has the advantages of more precise segmentation effect and smoother edges of each type of feature. The segmentation effect is shown in Figure 4:

![Figure 4. Comparison between Segnet's multi-class semantic segmentation and the method in this paper](image)

|                | Traditional U-Net | Improved U-Net | Optimized U-Net |
|----------------|-------------------|----------------|-----------------|
|                | Building | Water | Road | Building | Water | Road | Building | Water | Road |
| **PA**         | 0.8500   | 0.8893 | 0.8763 | 0.8830  | 0.9765 | 0.9128 | 0.9011  | 0.9965 | 0.9528 |
| **MPA**        | 0.8123   | 0.8511 | 0.8435 | 0.8492  | 0.9310 | 0.9014 | 0.8502  | 0.9418 | 0.9092 |
| **MIoU**       | 0.6673   | 0.7034 | 0.6995 | 0.7030  | 0.7489 | 0.7229 | 0.7552  | 0.8589 | 0.7728 |
As can be seen from Table 1, compared with the traditional U-Net network, the improved U-Net deepens the number of convolution layers, introduces a residual module, and expands the training set before training, so whether it is in pixel accuracy (PA), mean pixel accuracy (MPA) or mean intersection over Union (MIoU) have greatly improved compared to traditional U-Net networks. Based on the improved U-Net network model, combined with the conditional random field and image overlapping strategy, the pixel accuracy for building segmentation was increased from 0.8500 to 0.9011, and the pixel accuracy for water segmentation was increased from 0.8893 to 0.9965, but The proportion of water in the whole image is too small, so it seems to be particularly good for the segmentation of water. In addition to pixel accuracy (PA), mean pixel accuracy (MPA) and mean intersection over Union (MIoU) have been further improved.

4. Conclusion
Facing the semantic segmentation of remote sensing images with large-scale and complex image information, this paper proposes an improved U-Net network model, which extracts the underlying features by deepening the number of convolution layers, recovers the shallow detail features by upsampling, and introduces a residual module So as to achieve a more accurate two-category segmentation effect. The trained multiple binary classification models are combined with conditional random field and influence overlap strategy and integrated learning voting strategy to optimize the segmentation effect. Finally, the trained multiple binary classification models are fused to obtain the final semantic segmentation image. From the experimental data: the improved U-Net network model achieved 98.80% training accuracy and 98.12% test accuracy on the data set. Compared with the basic U-Net network, the method in this paper has greatly improved pixel accuracy (PA) and mean pixel accuracy (MPA) mean intersection over Union (MIoU). The experimental results show that: by improving the U-Net network, and then through the conditional random field, the optimized multi-model fusion of the image overlapping strategy can significantly improve the accuracy of the semantic segmentation of remote sensing images.

5. Acknowledgments
This research was funded by the National Natural Science Foundation of China (41071242 and 41701624 and 41271449) and Educational Commission of Hubei Province of China (D20131104).

6. References
[1] Haiyang W, Dehua P and Deshen X. Fast implementation of two- dimensional Otsu adaptive threshold selection algorithm[J]. Acta Automatica Sinica, 2007 (9): 968-971.
[2] Coates A, Ng A Y. Learning feature representations with k-means [M] Neural networks:Tricks of the trade. 2012:561-580.
[3] Songtao L and Fuliang Y. Image segmentation method based on graph cut and its new development [J].Acta Automatica Sinica, 2012, 38 (6): 911-922.
[4] Shotton J, Johnson M and Cipolla R. Semantic texton forests for image categorization and segmentation [C] Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008: 1-8.
[5] Long J, Shelhamer E and Darrell T. Fully convolutional networks for semantic segmentation [C] Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015:3431-3440.
[6] Ronneberger O, Fischer P and Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation [C] International Conference on Medical Image Computing and Computer-Assisted Intervention. 2015.
[7] Han K, Rongguo Z and Kuan C. Medical image segmentation technology based on deep learning [J]. Artificial Intelligence, 2018, (4): 30-37.
[8] Wenjie Z. Research on aircraft detection algorithm of remote sensing image based on visual saliency [D].Hubei: Huazhong University of Science and Technology, 2016.
[9] Kaijun G, Shao yuan S and Guangshun Y. Semantic Segmentation of Unmanned Vehicle Night Vision Images Based on Deep Learning [J]. Applied Optics, 2017, 38 (3): 421-428.
[10] Risheng Z, Guibin Z and Yanqin Z. A satellite image region segmentation and recognition method based on CNN and gradient watershed algorithm [J]. Infrared Technology, 2017, 39 (12): 1114-1119.

[11] Lafferty J, McCallam A and Pereir A. Conditional Random Fields: Probabilistic Models for segmenting and Labeling sequence Data [C] Eighteenth International Conference on Machine Learning, 2001: 282—289.