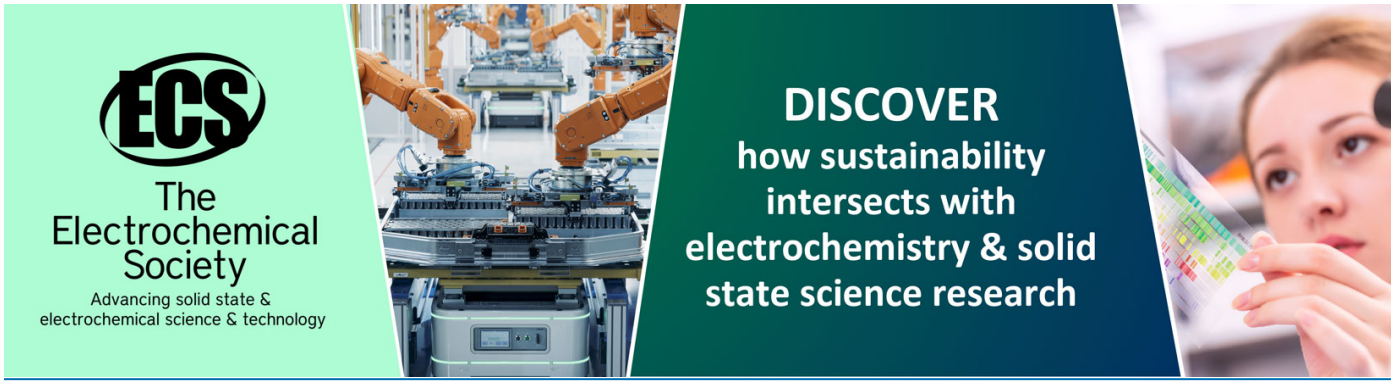
High Resolution Remote Sensing Image Classification Algorithm Based on Improved FCN



High Resolution Remote Sensing Image Classification Algorithm Based on Improved FCN

Kong Yunbo*, Fu Haojun, Yangfan, Zhouhai, Wen Na, Zhang Fan

The Mapping Terminal of Xi'an, Xi'an 710053,China

*Corresponding author's e-mail: kongyunbo123@163.com

Abstract: Aiming at the problems of less bands of high resolution remote sensing image data and limited learning richness of model features, this paper proposes a high resolution remote sensing image classification algorithm based on improved full convolution neural network. Firstly, a standardization layer is added to batch process the image, and then a pooling index is added to the image to realize the up-sampling. Finally, the pooling index, the transposed convolution and the convolution eigenvalue are combined into a feature group to restore the class pixels of the image to a great extent. It can improve the prediction ability of the model. A simulation experiment is carried out to verify the effectiveness and feasibility of the proposed algorithm.

1. Introduction

Classification of remote sensing images is a basic problem in the field of remote sensing, and it is the basis of the application of various remote sensing technologies. It refers to the process of recognizing different objects according to the spectral, shape and other characteristic information of the objects, so that each pixel in the image can be endowed with the real object category. Therefore, the study of remote sensing image classification has certain practical significance for subsequent remote sensing applications. The initial classification of remote sensing images was done by manual visual interpretation. It is a process in which the interpreter obtains the target information by observing the image directly based on experience and knowledge. This method is labor intensive, subject to subjective images, requires high interpreters and is difficult to update. With the rapid development of computer technology, computer-aided classification of remote sensing images is a widely used classification method. It can make up for the inherent shortcomings of visual interpretation by simulating human brain to distinguish remote sensing images, and classify remote sensing images quickly and efficiently. However, the emergence of high resolution images also requires higher classification and counting of remote sensing images. Therefore, it is necessary to study a more efficient and accurate classification method of high resolution remote sensing images.

In 2017, Maggiori et al. constructed a classification framework of remote sensing images using FCN to realize pixel-by-pixel classification of high-resolution remote sensing images, and on this basis combined with multi-scale feature maps. But the trade-off between localization and recognition accuracy has always been an important issue in semantic annotation of FCN. In order to solve this problem, many scholars have made improvements on the basis of FCN. Chen et al. introduced multi-scale and conditional random fields on the basis of FCN, which can get better results. RefineNet adds the design of residual connection, Nicolas uses SegNet and multi-core convolution to propose features, and uses residual correction method to combine the characteristics of multi-source data.

Although U-Net has a strong application potential in high-resolution remote sensing image classification, there are some shortcomings in practical application. On the one hand, the bands of high-resolution remote sensing image are limited and the spectral information is not rich, which limits the richness of model feature learning to a certain extent. It is necessary to add multi-source data. On the other hand, the influence of classification usually has small staggered areas, and the boundary of terrain is slightly smooth, so it needs to take certain methods to optimize. In order to solve these problems, this paper proposes a high-resolution remote sensing image classification algorithm based on SegNet and U-Net model fusion. The algorithm combines multi-source image data with the original image, synthesizes spatial and spectral resolution, adopts the idea of fusion of Seg-Net and U-Net models, and adopts the idea of voting on each pixel of each result map. The classification of pixels in the corresponding position of each image is predicted. The category with the largest number of votes is the pixel. Categories, which can remove some pixels with obvious classification errors, greatly improve the prediction ability of the model. Simulation results verify the effectiveness and feasibility of the proposed algorithm.

2. Classification algorithm based on full convolutional neural network

Convolution neural network is a neurocognitive machine based on the concept of receptive field, consisting of multilayer structure. Generally, convolution neural network is composed of several convolution layers and pool layers nested, and the following operations are usually performed:

2.1 volume stacking and pooling layer

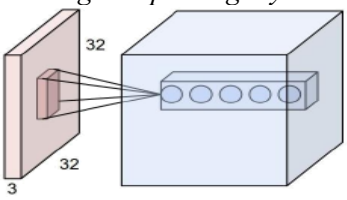


Fig. 1 Schematic diagram of convolution kernel

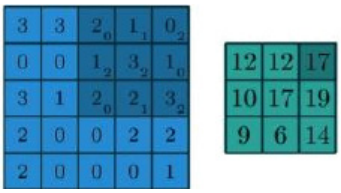


Fig. 2 forward propagation process of coiling layer

Convolutional neural network is called convolution network because the most important part of it is convolution layer, convolution layer is also often called filter or convolution kernel. For different size image data, convolution kernel size is also different, its shape Ewing rectangle, length and width are generally specified manually. The commonly used convolution kernel size is 3*3 or 5*5. The single convolution kernel extracts local image specific features by analyzing local images. In the actual operation, the number of convolution kernels need to be specified manually. Different convolution kernels can extract different features of graphics, which can enrich the extracted features. Figure 1 shows a schematic diagram of convolution kernels. The small matrix on the left is the size of convolution kernels. The depth of the small rectangle on the right represents the number of convolution kernels. Figure 2 shows the convolution layer forward propagation process of the 3*3 convolution kernel with a step size of 1 on the 5*5 matrix. The left image is shown as an image matrix or a characteristic matrix, and the right picture is the global feature after convolution. In this process, the convolution kernel is applied to the upper left corner of the matrix, and the result is obtained by weighting the corresponding position of the convolution kernel and saving it as the upper corner of the new matrix. Then the filter is moved to the right with a step size of 1. After moving two steps to the left, it moves one step downward, then does the product sum from left to right again, and moves one time until the filter is moved to the right. The lower right corner matrix. Once the convolution kernel moves each time, a value can be obtained by calculating the weighted sum of the nodes. These values are spelled into a new matrix. As shown in the right-hand diagram of Figure 2, the propagating process ahead of the convolution layer is completed, and it can be seen intuitively that the dimension is reduced from 5 * 5 to 3 * 3 through the convolution process.

The convolution calculation method from L-1 layer to L layer is:

$$X_j^l = f\left(\sum_{i \in N_j} X_i^{l-1} * K_{ij}^l + \varepsilon_j^l\right) \quad (1)$$

Among them, K is a convolution kernel, ε is a bias, and f is an activation function.

Although convolution layer can reduce the image dimension, but for large images, the amount of data is still too large, so we need to add a pool layer behind the convolution layer. The most important role of the pooling layer is to reduce the dimension of the matrix, thereby reducing the number of parameters of the final fully connected layer and speeding up the calculation. At the same time, by reducing the influence of some unnecessary eigenvectors, the pool layer can also effectively prevent over-fitting. The operation process of the pool layer is similar to the sliding process of the convolution kernel of the volume layer on the image. The only difference is that the window in the pooling layer does not do weighted sums at the corresponding locations, but uses the method of maximum or average values. At present, the most commonly used method is to take the maximum value of the image. This method can not only retain the most effective features of the image, but also some less significant changes in the image do not affect the extracted features. The structure of this method is also called the maximum pooling layer.

The pooling process from L-1 layer to L layer is calculated as follows:

$$X_j^l = f(\alpha_j^l Max(X_j^{l-1}) + \beta_j^l) \quad (2)$$

Where *Max* (x) represents the maximum value for dimension reduction.

2.2 full convolution neural network FCN

The main change of FCN of full convolution neural network is to replace the last several fully connected layers of original convolution neural network with convolution layer. the original convolutional neural network layers 6 and 7 are all fully connected layers, which are one-dimensional vectors of length 4096, and the eighth layer is a classification layer, which is a one-dimensional vector of length 1000. The FCN replaces the three fully connected layers with convolution layers. The convolution kernels are all 1*1 in size and the output channels are 4096, 4096 and 1000, respectively.

It can be seen that after many times of convolution kernel pooling, the size of the image is gradually shrinking, obviously it is impossible to classify each pixel of the original image. In order to restore the resolution of the original image, FCN uses transposition convolution to increase the size of the image. Convolution connects pixels of multiple cells to the active region of a single cell through a filter, while transposed convolution connects the active region of a single cell to the output of multiple cells. A 4 * 4 input image is convoluted by 3 * 3 convolution kernels with a step size of 1. Finally, a 2 * 2 characteristic image is obtained. Firstly, the edge of the 2*2 feature image is complemented by 0, and then the 3*3 transposed convolution kernel is used to perform the transposed convolution with step size 1. Finally, the 4*4 restored image is obtained. Convolution operation outputs 4 * 4 images to 2 * 2 images, which leads to the loss of image information, while the transposed convolution operation makes up 0 for the edge of the feature map. All the restored images have the loss of edge information.

2.3 U-Net

U-Net model is the improvement and extension of FCN. It follows the idea of image semantic segmentation based on FCN, which uses convolution layer and pool layer to extract features, and then uses deconvolution layer to restore image size. However, U-Net combines the characteristics of coding-decoding structure and hopping network, and is more elegant and ingenious in model structure, which is mainly reflected in the following two points:

(1) U-Net model is an encoding-decoding structure, compression channel is an encoder for extracting image features layer by layer, extended channel is a decoder for restoring image location information, and each hidden layer of U-Net model has more feature dimensions, which can be used to learn more diverse and complete models. The characteristics of the surface.

(2) The "U-shaped" structure of U-Net model makes the process of cutting and stitching more intuitive and reasonable. The stitching of high-level feature map and low-level feature map, as well as the repeated and continuous operation of convolution, enable the model to combine context information

and detail information to get more accurate output feature map. Experiments show that U-Net model can get more accurate classification results with fewer training samples.

3. Improved full convolutional neural network

3.1 batch of standardization

Deep learning requires training a deep-level network model in which the input of each layer is the output of the previous layer, i.e. the parameter distribution received by each layer is also the parameter distribution of the output of the previous layer. Number is ReLUctant, it can only activate the eigenvalue greater than 0, that is to say, the output parameters are greater than 0 after the activation function, which leads to two problems: First, each layer should learn the parameter distribution again according to the parameter distribution changes of the previous layer, which does not meet our requirements for the stability of the distribution; second, it is not consistent with our requirements for the stability of the distribution. The parameter distribution of the former layer changes, and this change will be magnified after the deep network, which will result in the backward propagation algorithm can not effectively update the value of the former layer.

$$\hat{x}^k = \frac{x^{(k)} - E[x^{(k)}]}{\sqrt{Var[x^{(k)}]}} \tag{3}$$

This paper proposes to use batch standardization to solve this problem. Assuming that x is the input sample of the current layer, the method subtracts its own mean value from each dimension and divides it by its own standard deviation, so as to realize the data distribution with the mean value of 0 and the variance of 1. The calculation formula is as follows:

However, the data processing will lead to a decline in the ability to express data, so two parameters scale and shift are added. These two parameters are trained in the network to enhance the network's ability to express. The calculation method is as follows:

$$y^{(k)} = \kappa^{(k)} \hat{x}^{(k)} + \varepsilon^{(k)} \tag{4}$$

In particular, when $\kappa^{(k)}$ equals the standard deviation of the previous data sample and $\varepsilon^{(k)}$ equals the mean of the previous data sample, the transformation restores the data to the original distribution. This has an advantage. First, we can use a larger learning rate. If the data distribution of each layer is inconsistent, a smaller learning rate must be used in order to reduce the loss function better. After using batch standardization, the distribution of data is uniform, and the learning rate can be higher, which can effectively improve the convergence speed of the algorithm.

Secondly, Dropout prevents the final result from over-fitting by ignoring a part of the feature information between each layer. The reason of over-fitting is generally at the boundary of the data distribution, and the initialization weights of the data are left inside the data by over-fitting, which effectively alleviates the over-fitting phenomenon. At the same time, the L2 weight attenuation system is also proposed to prevent model overfitting.

3.2 pooled index

In the process of coding, the coordinate information of the location of the maximum value in the pooling process will be preserved every time the pooling occurs. In the decoding process, the value will be restored according to the coordinate information. if maximum pooling is adopted in the downsampling coding process, the coordinate position of the maximum pooling value in the 2*2 network is recorded and saved as Pooling Index. In the subsequent up-sampling decoding process, the maximum value is put into the restored position according to the position information of the value recorded by Pooling Index. The corresponding position in the image, and the other three positions are all zero, so that the details of the part of the information is restored.

3.3 algorithm flow

The activation functions used in the model are all ReLUctant activation functions, the loss functions are

cross entropy cost functions, and the optimization method is stochastic gradient descent method. BN is Batch Normalization, Conv is convolution layer, Maxpool is maximum pooling layer, UNConv is transposed convolution layer, and UNPool is upper sampling layer. Each convolution layer and transposed convolution layer denotes the size of the convolution kernel, and @ denotes the number of output layers. The total pooling and up-sampling sizes of the model are both 2*2. The overall structure of the model is still coded and decoded. In the process of encoding, batch normalization is added to the back of each convolution layer, and the position and record of each maximum pooling pool are recorded. In the process of decoding, the original convolution features, the features of transposed convolution restoration and the features of up-sampling restoration are still used. The method of up-sampling and restoring is improved. The image is restored by using the maximum position recorded in the previous pool layer. Finally, the result image of pixel-by-pixel classification with the same size as the original image is obtained.

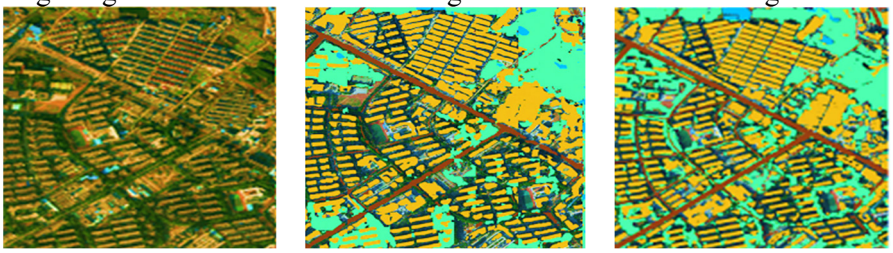
4. Simulation experiment and analysis

4.1 experimental sample data

The data used in this experiment are from the AI classification and recognition contest of satellite images in the 2017 Grand Data and Computer Intelligence Competition. The data provided by the competition (a high-definition remote sensing image of a city in southern China in 2015) is a small data set containing five large-scale RGB remote sensing images (ranging in size from 3000 *3000 to 6000 *6000) marked with four types of objects, vegetation (Mark 1), architecture (Mark 2), and water (Mark 2). 3), road (Mark 4) and others (Mark 0). Among them, cultivated land, woodland and grassland are classified as vegetation. In order to better observe the annotation, we visualized three training pictures as follows: blue-water, yellow-house, green-vegetation, brown-road. Now let's talk about the steps of data processing. We now have five large-scale remote sensing images, and we can't put them directly into the network for training because memory can't stand them and their sizes vary. Therefore, we first cut them randomly, that is, generate x, y coordinates randomly, then cut out the coordinates 256 * 256 small map, after the above data enhancement operation, we get a larger training set: 100000 256 * 256 pictures.

4.2 experimental results and analysis

The algorithm proposed in this paper is compared with SegNet and U-Net algorithms, and is evaluated by visual evaluation method and data evaluation method respectively. Figure 3 (b) and figure3 (c) are the graphs after the real target classification of test set3 (a). Figure3(b) is a visual map after Uet network is used to classify objects. Figure 3 (c) is the rendering of classification visualization of the algorithm. From the experimental results, we can see that the algorithm proposed in this paper can classify the road, water and other targets very well. Compared with the U-Net algorithm, this paper deals with the details of remote sensing images better and restores the original information of the target better.



（a） （b） （c）

Fig. 3 target classification results of remote sensing images

Figs 4 and 5 show the change curves of accuracy and loss function on the training set respectively. From the change of the curve, we can see that the accuracy of the two algorithms increases with the increase of training times, while the value of loss function decreases. And from the experimental results, the proposed method is due to U-Net, and has a high accuracy. From the 15th epoch to the 25th epoch,

it can be clearly seen that the loss function of the proposed method has a lower value. With the increase of training times, the training appears to be saturated. Both methods obtain the final

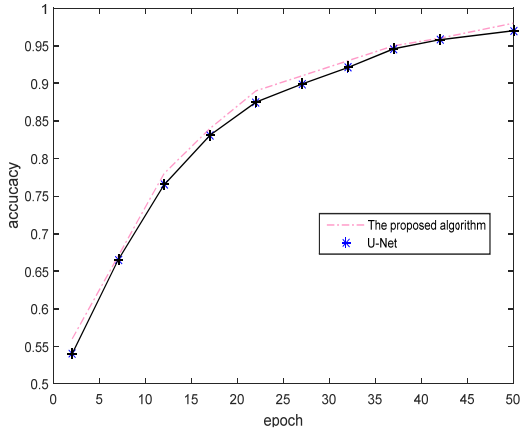


Fig.4 The figure of accuracy rate variation curve

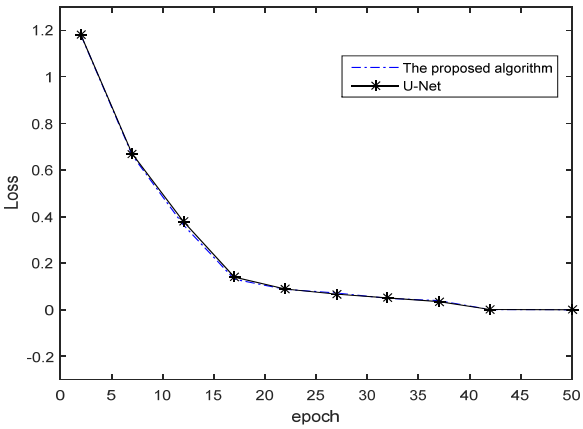


Fig. 5 The figure of loss function variation curve

loss function. Table 1 shows the accuracy of the two methods on the test set. The results show that the accuracy of the proposed algorithm is higher than that of U-Net. It shows that the proposed method has a good effect on remote sensing image classification.

Table 1 Comparison of accuracy of two methods

The algorithm	Test set accuracy rate
U-Net	0.9467
The proposed algorithm	0.9568

5. Conclusion

This paper proposes a high resolution remote sensing image classification algorithm based on improved full convolution neural network. Firstly, a standardization layer is added to batch process the image, and then a pooling index is added to the image to realize the up-sampling. Finally, the pooling index, the transposed convolution and the convolution eigenvalue are combined into a feature group to restore the class pixels of the image to a great extent. It can improve the prediction ability of the model. Finally, a simulation experiment is carried out to verify the effectiveness and feasibility of the proposed algorithm.

References

[1] Xiaoming Ye, Wei Gao, Yi Wang. Research on SAR Images Recognition Based on ART2Neural Network [C]. 7th IEEE Conference on Industrial Electronics and Applications(ICIEA), 2012, 1888-1891.

[2] Zongjie Cao, Zongyong Cui, Yong Fan. SAR Automatic Target Recognition Using aHierarchical Multi feature Fusion Strategy [C]. GC'12 Workshop: Radar and Sonar Networks,2012, 1450-1454.

[3] Ring C, Ma L, Yang X. Stacked denoise autoencoder based feature extraction and classification for hyperspectral images[J].Journal of Sensors, 2015, 2016.

[4] Midhun M E, Nair S R, Prabhakar V T, et al. Deep model for classification of hyperspectral image using restricted boltzmann machine[C].Proceedings of the 2014 International Conference on Interdisciplinary Advances in Applied Computing. ACM, 2014: 35.

[5] Li T, Zhang J, Zhang Y. Classification of hyperspectral image based on deep belief networks[C].Image Processing (ICIP), 2014 IEEE International Conference on.IEEE, 2014: 5132-5136.

[6] Chen Y, Zhao X, Jia X. Spectral-spatial classification of hyperspectral data based on deep belief network[J].IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015, 8(6): 2381 — 2392.