UAV Remote Sensing Image Rural Building Detection Based on Convolutional Neural Network

Xing CHEN, Jing-wen XU*, Jun-cheng WU and Meng-jiao ZHANG
College of Resources, Sichuan Agricultural University, Chengdu, China
*Corresponding author

Keywords: Convolutional neural network, UAV remote sensing image, Rural building, Detection.

Abstract. In order to improve the accuracy of intelligent detection of rural buildings, this paper uses massive remote sensing data of UAV to train convolutional neural network and improves CaffeNet model. Experimental results are as the followings: With reference to the traditional CaffeNet, the improved a-CaffeNet has an accuracy rate of 97.30%. The test time and the training time are shortened by 30.4% and 11.84% respectively. According to the experiments, this method is more suitable for remote sensing image of rural buildings detection.

Introduction

With the implementation of rural revitalization strategy, rural buildings need to be counted, especially the regional location, quantity, size and distribution of rural construction land. Traditional detection methods mainly rely on the expression of artificial design features, which is subjective to a certain extent. Moreover, this method is difficult to be used in massive data, and the detection effect is poor. In order to detect rural buildings in UAV remote sensing images, it is necessary to seek a method that can automatically learn image features from a large number of samples, obtain the most effective feature information in the image, construct a complex network structure, fully mine the correlation between feature information and establish an effective classifier.

In recent years, the popular convolutional neural network provides an end-to-end automatic learning method. The parameters in the network are trained by gradient descent method. The trained convolutional neural network can learn the features in the image and complete the extraction and classification of image features. This method is also very effective in UAV remote sensing image processing [1]. In this paper, several common convolutional neural networks are trained by using massive remote sensing data of UAV, and the network is improved according to the shortcomings of CaffeNet. Finally, by comparing the improved network with the traditional network, it is concluded that the improved network has better detection effect on rural buildings in UAV remote sensing images under the same conditions.

Dataset

In the experiment, 1:2000 remote sensing image of UAV in danling county was used, which covers all types of construction land in the residential area. By manual pre-cutting, the images containing the construction land of rural residential areas are cut down randomly, and the size is limited within the range of 500*500 pixels to ensure that the computer can successfully extract the area of interest and identify the construction land. During the preparation of the experiment, a total of 1500 samples were selected for study, among which 1000 were randomly selected for training and the remaining 500 were used as test pictures.
Convolutional Neural Network Model

The Basic Structure of Convolutional Neural Networks

The hierarchical structure of the convolutional neural network consists of input layer, CONV layer, incentive layer, Pooling layer and FC layer[2].

![Basic structure of a convolutional neural network](image)

The input of the convolutional neural network is the original image $X$. In this paper, $H_i$ is used to represent the feature graph of the $i$th layer of convolutional neural network ($H_i = X$). Assuming $H_i$ convolution layer, the generation process of $H_i$ can be described as follows:

$$H_i = f(H_{i-1} \otimes W_i + b_i).$$

(1)

Where, $W_i$ represents the weight vector of the convolution kernel at the $i$th layer. The $\otimes$ operator represents the convolution operation of the convolution kernel with the $i - 1$th layer image or the feature image. The output of the convolution with the first layer of the offset vector addition, ultimately through the nonlinear excitation function $f(x)$ is the first $i$ figure $H_i$ layer features.

AlexNet Model

AlexNet has eight layers of network architecture, including five convolutional layers and three fully connected layers. The first and second layers of convolution are followed by local corresponding normalization. The first, second and fifth layers have overlapping pooling. The output of each layer of convolution layer and full connection layer has ReLU[3]. At the same time, dropout function was added after the full connection layer, which could temporarily discard neurons from the network according to a certain probability to prevent overfitting[4].

CaffeNet Model

CaffeNet model is a model with fine-tuning on the basis of AlexNet model [5]. Its protobuf definition is different from AlexNet in the local normalization layer and the pooling layer. CaffeNet puts the pooling layer in front of the local normalization layer. After the second convolution, the pooling layer is still placed in front of the local normalization layer, and the difference of deviation size[6].

a-CaffeNet Model

The a-caffenet model is an improvement of CaffeNet. Its improvement method is as follows: (1) in order to obtain more features, the size of the convolution kernel of the first layer is reduced from $11*11$ to $8*8$. (2) By reducing the step size of the first layer from 4 to 2, the first convolutional
layer can obtain more features. (3) Since the function of the third convolutional layer is similar to that of the fourth convolutional layer, the fourth convolutional layer needs to be removed under the premise of saving computer budget resources.

Selective Search Algorithm

The selective search algorithm adopts a variety of strategies such as color, texture and size to merge regions, so it can be applicable to a variety of targets [7]. As shown as Fig.1, the merging process from left to right is from small image block to large image block.

![Selective Search Algorithm](image)

Figure 1. Selective Search Algorithm.

Results and Analysis

In order to evaluate the ability of convolutional neural network to detect rural buildings in UAV remote sensing images, AlexNet model, CaffeNet model and a-caffenet were adopted in this paper. All the three algorithms adopt the 64*64 scale standardized region of interest for training. In the training process, all the parameters of the convolutional neural network adopt the same optimal parameter setting, which is the optimal performance parameter of the model. 1000 training samples and 500 test samples. Model performance is reflected by accuracy. Accuracy reflects the judgment of the model on all data, whether positive samples or negative samples, as long as they are correctly classified as the numerator of the fraction, and all test samples as the denominator.

As can be seen from table 1, The convolutional neural network with the highest accuracy is an improved and optimized a-caffenet model, with an accuracy of 97.30%. From the perspective of training time and test time, the optimized and improved a-caffenet model ranks top in the training model in terms of speed, and only takes 27.6 minutes to train 1000 pictures. From the perspective of test time, the shortest model is a-caffenet model, whose test time for 500 remote sensing images is 94.5s. Therefore, the a-caffenet model is suitable for the remote sensing image recognition of existing UAV mounted on a high-performance computer.

![Experimental data](image)

Figure 2. Experimental data.

Table 1. Comparison of different model recognition performance.

| methods      | Accuracy(%) | Training time(min) | test time(s) |
|--------------|-------------|--------------------|--------------|
| AlexNet      | 91.60       | 36.1               | 117.6        |
| CaffeNet     | 96.70       | 39.7               | 114.0        |
| a-CaffeNet   | 97.30       | 27.6               | 94.5         |

After region of interest extraction and a-caffenet network identification, the research area was returned to the original image for frame drawing. The final results are shown in Fig. 3.
Summary

On the whole, most buildings can be calibrated correctly, but some buildings cannot be calibrated. The main reason is that the building is not extracted in the process of region of interest (ROI) extraction, which needs to be further improved and optimized. Secondly, the training samples are identified as non-construction objects in the process of insufficient recognition. For continuous identification of construction land, some parts are still unmarked, and the identification and block method of continuous buildings still need further research and discussion.

References

[1] Vaduva C., Gavat I., Datcu M., Deep learning in very high resolution remote sensing image information mining communication concept, IEEE signal Processing Conference. 20(12)2506-2510.

[2] Lawrence S., Giles C.L., Tsoi A.C., et al, Face recognition a convolutional neural-network approach, IEEE Transactions on Neural Networks. 8(1)98-113.

[3] Ballester P., Araujo R.M., On the performance of GoogLeNet and AlexNet applied to sketches, Thirtieth AAAI Conference on Artificial Intelligence. 20(16)1124-1128.

[4] Cheng P.M., Malhi H.S., Transfer Learning with Convolutional Neural Networks for Classification of Abdominal Ultrasound Images, Journal of Digital Imaging. 30(2)234-243.

[5] Faria S.H., The symmetry group of the CAFFE model, Journal of Glaciology. 54(54)643-645.

[6] Umemura M., Hotta K., Nonaka H., et al, Image Labeling for LIDAR Intensity Image Using K-Nn of Feature Obtained by Convolutional Neural Network, International Archives of the Photogrammetry Remote Sensing & S. 20(16)931-935.

[7] Uijlings J.R., Gevers T., et al, Selective Search for Object Recognition, International Journal of Computer Vision. 104(2)154-171.