An Improved Full Convolution Neural Network for Image Semantic Segmentation

Xuejing Ding
Computer engineering college,
Anhui Sanlian University
Anhui Hefei, China
330545497@qq.com

***Abstract*--In this paper, the current mainstream image semantics segmentation methods based on full convolution network are analyzed. Through analysis, it is found that FCN weakens the function of classifiable features while acquiring semantic information of the target, resulting in insufficient expression ability of the feature, resulting in insufficient accuracy of target recognition, incomplete segmentation and even loss of the target, etc. To solve this problem, an image semantics segmentation method based on multi-scale feature extraction is proposed. By constructing a full convolution network and using different scale images as input of the network, the features of different scale images are extracted. Finally, the segmented image is obtained by feature fusion. Experiments show that better image semantics segmentation can be achieved by extracting and fusing multi-scale features.**

Keywords--Image semantic segmentation, Multi-scale features, Full convolution network

I. INTRODUCTION

Image semantics segmentation is a basic problem in the field of computer vision. As the first part of the whole algorithm, image is preprocessed. The segmented results will be used for subsequent analysis. Therefore, the quality of semantics segmentation has a great impact on the accuracy of the final results. The goal of image semantics segmentation is to classify each pixel in a color image, so that RGB image can be transformed into annotated image, and the same color pixels in the annotated image represent the same kind of objects. Traditional image segmentation methods include threshold segmentation, edge-based segmentation and region-based segmentation. Most of these methods are based on artificial feature extraction, such as color and texture information. However, the characteristics of artificial design are often shallow, which makes the traditional methods have great limitations, and the space for performance improvement is very limited. In this paper, the current mainstream image semantics segmentation methods based on Full Convolution Network (FCN) are analyzed, and put forward a kind of image semantic segmentation method based on multi-scale feature extraction[1-2]. By constructing a full convolution network, and by using different scale images as the input of Network to extract the features of different scale images. Finally, the segmentation image is obtained by feature fusion. Experiments show that the effect of semantic segmentation can be improved by extracting and fusing multi-scale features.

II. FULL CONVOLUTION NETWORK

Full convolution network FCN can extract image features from a large number of sample data, which is better than manual labeling features, making it a great success in high-level computer vision such as image semantics segmentation and object detection[3]. The application of FCN image semantics segmentation model can break the limitation of manual labeling feature through a large number of sample self-learning features, so as to achieve a very good segmentation effect. Although, FCN successfully extracted the semantic features of the image by changing the structure of convolutional neural network, and completed the recognition and segmentation of the target. However, the algorithm still has the following shortcomings[4]:

1) Insufficient target recognition ability: Although FCN uses multi-layer feature fusion to improve the accuracy of semantic segmentation, it ignores the important feature of full-connection layer to distinguish object categories when using full-convolution instead of full-connection, which results in insufficient recognition ability of the network to target.

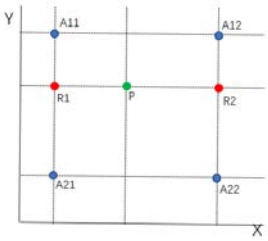

Fig.1. The bilinear interpolation

2) It is insensitive to the details of the image: FCN only uses a simple bilinear interpolation method to sample the feature map. As shown in Fig.1, assuming the pixel values of known points A11, A12, A21 and A22, the pixel values of R1 and R2 are first obtained by linear interpolation in the Y direction, and then the pixel values of point P are obtained by linear interpolation in the X direction. Through this interpolation method, FCN fills in the original image size with the feature map pooled many times. However, the semantic features obtained by this rough up-sampling method are not detailed enough.

  DOI: 10.25236/iccpb.2018.018

Although the effect of 8-fold up-sampling is much better than that of 32-fold up-sampling, the regional semantics is still blurred, insensitive to the details of the image and easy to lose the target.

This paper mainly deals with how to make full use of the context information of images and extract rich features through FCN for semantic segmentation of images. Recently, the algorithm of multi-scale input training CNNs for image semantic segmentation has achieved good results. In this paper, a multi-scale image is used as input to extract features and complete semantic segmentation of images by FCN.

III. Semantic Segmentation Algorithms For Multi-Scale Feature Extraction

In the training stage, each scale image will produce a subgraph through the network, which will be scaled to the same size through the up-sampling. The output graph is generated through the multi-scale fusion layer, and then the loss is calculated with the segmentation label, and then the back-propagation is completed. In the testing stage, each scale image will generate a subgraph through the network, and the output graph is generated by the multi-scale fusion algorithm. The basic flow chart of the algorithm is shown in Fig.2.

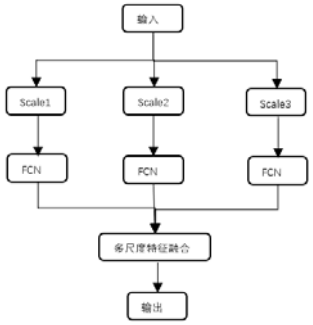

Fig.2. The flow chart of the algorithm

A. Network Model Structure

The network in this paper is an extension of full convolution network, the network in this paper is an extension of full convolution network. The hopping network structure is replaced by the sharing network structure. Different scale feature maps are generated by different scale image input to ensure the effect of the output graph. The network structure is shown in Figure 3. First, the input image is scaled to three different sizes and put into the convolution network to generate three sub-graphs, which are fused by multi-scale fusion layer. Fusion into a segmented image, so that using different scales of images, we can better perceive the rich spatial information in the image.

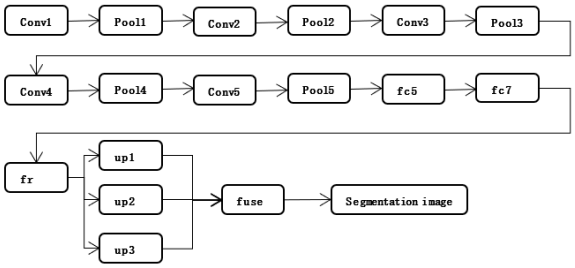

Fig.3. Network structure of this paper

For each scale image, put it into the network of Fig.3 for training and learning. The network consists of five convolution layers (Conv1, Conv2, Conv3, Conv4, Conv5). Behind each convolution layer, a maximum pool layer (pool 1, pool 2, pool 3, pool 4, pool 5) is used to reduce the amount of calculation. The full connection layer (fc6, fc7) in convolution network is transformed into convolution layer, and several features images are output, fr layer. The size of the feature image is 1/32 times that of the original image. The deconvolution operation of (2) is needed to change the feature image into the original image size.

$$w_1=(w_0+2\times padding-kernelsize)/stride+1 \quad (1)$$

$$w_0=(w_1-1)\times stride+kernelsize-2\times padding \quad (2)$$

Among them, $w_0\times h_0$ is the input image size, $w_1\times h_1$ is the image size after convolution calculation, *padding* is the padding value, *kernelsize* is the convolution kernel size, *stride* is the sliding window size.

Each layer of convolution layer is activated by ReLU (Rectified-Linear Units) function. Different scales correspond to different output feature maps. The final segmentation map can be obtained by fuse function.

Given the image I_s with input scale s, the parameters of the training network are $\theta_s=(w^l, b^l)$, where $l=1, 2, \ldots, n$ is the

network hierarchy .*w* denotes the weight of the network, *b* denotes the bias, assuming that the output after L-level convolution and pooling operation is O_s^l, then there is (3):

$$O_s^l=pool(ReLU(w^l\times O_s^{l-1}+b^l)) \tag{3}$$

From the input image I_s, the output values of each neuron in the network are computed in turn until the final subgraph is computed. This process is called Forward Propagation. The output O_s^l of the last layer requires a softmax activation function for normalization, such as (4):

$$O_s^l=softmax(w^l\times O_s^l+b^l) \tag{4}$$

In the training process, the errors between the actual output and the correct output of each neuron in the network need to be calculated from the direction of the output to the input image, and the parameters of the network need to be updated by the random gradient descent algorithm (SGD). This process is called Backward Propagation.

segmentation graph, the probability of each pixel being divided into each class can be calculated. Let the pixel be a_i, $i=1,2,\ldots,N$ denotes the index of the pixel, p_{ij} denotes the probability that the pixel a_i is assigned to the label c_j, $j=1,2,\ldots,C$ denotes the category of the object. Then the loss function can be expressed as (5)

$$e = -\sum_{i=1}^{N}\sum_{j=1}^{C} 1\{y_i = j\} \ln p_{ij} \tag{5}$$

Among them, y_i is the label that the pixel a_i outputs through the network. The whole training process is to optimize the loss function.

B. Multi-scale feature extraction and fusion

In this paper, we use the multi-scale method of scaling the input image to sense the image. The image is scaled to 1, 0.6 and 1.2 scales as the input of the network. Objects that are too large or too small can be better perceived at three scales. It can be divided into three scales to better perceive objects that are too large or too small. Three sub-graphs of different scales are generated from the network, and the final output graph is obtained by fusion. (3) ~ (5) give the calculation and error function of a single network. The network structure of Fig.3 is used to extract feature maps of different scales. Fuse is used to calculate the fusion of multiple scale feature graphs, and the specific process is described as follows.

Suppose the input image I is scaled to S input images, and $f_s(I_s, \theta_s)$ is used to represent the output subgraph of the image Is on the s network, where s=1,2,…,S. The S sub-graphs perceive different scale information, and then use (6) to fuse them into the final output result graph G (I, C), where C represents the semantic category label.

$$G(I,C) = \sum_{s=1}^{S}(w_s \times f_s(I_s, \theta_s)) \tag{6}$$

The weight W_S represents the importance of the sub-graph generated by the S-Scale image. In this paper, it is set to 1/3 to represent the equivalence between sub-graphs of different scales. Fusion only requires linear superposition to produce the final output graph.

IV. Experiments And Results Analysis

Using Stanford background dataset 8 data sets, 500 images were randomly selected for training and 120 images for verification and testing. With the increase of training iterations, the accuracy of image semantics segmentation is also improving. After 12,000 iterations, the accuracy changes smoothly. As shown in TABLE 1, the accuracy of each semantics class after 12,000 iterations can be well distinguished from different semantics classes.

TABLE I. PER-CLASS ACCURACY

Class	Accuracy of this paper	Accuracy of FCN
sky	91.80	90.75
tree	86.43	84.1
road	92.23	90.7
grass	89.57	88.90
water	85.36	83.81
building	82.95	80.95
mountain	78.26	78.11
foreground	80.45	80.30

From the above table, it can be seen that the pixels in the image can be roughly classified as correct semantic labels, which is better than FCN.

V. CONCLUSION

In this paper, a learning model of FCN using different scales of images as input is proposed. Through repeated iteration training, the feature information of different scales of images can be extracted and fused to complete the task of image semantics segmentation. The experimental results show that image features of different scales have a vital impact on the accuracy of image semantic segmentation. Combining multi-scale feature extraction, pixels can be classified correctly and the accuracy of image semantic segmentation can be improved.

Acknowledgements

Fund Project: Key Natural Science Projects in Anhui Province.(NO: SK2018A0668).

References

[1] Jiang Yingfeng, Zhang Hua, Xue Yanbing. A new method of multi-scale depth learning for image semantic understanding [J]. Photoelectron Laser,2016,(02):224-230.

[2] Liu Dan, Liu Xuejun, Wang Meizhen. A Multi-scale CNN Image Semantic Segmentation Algorithm [J]. Remote Sensing Information,2017,(01):57-64.

[3] Xu Xinzheng, Ding Shifei, Shi Zhongzhi. New Theory and New Method of Image Segmentation [J].Journal of Electronic Science,2010,38(b02):76-82.

[4] Ma xiao. Image Semantic Segmentation Based on Deep Convolution Neural Network [D]. University of Chinese Academy of Sciences