Semantic Segmentation Based on Deep Convolution Neural Network

IOP Conf. Series: Journal of Physics: Conf. Series **1069** (2018) 012169 doi:10.1088/1742-6596/1069/1/012169

Semantic Segmentation Based on Deep Convolution Neural Network

Jichao Shan[1, 2], Xiuzhi Li[1, 2], Songmin Jia[1, 2] and Xiangyin Zhang[1, 2]

[1]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
[2]Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124, China.
Email: 18810133216@163.com

Abstract. Semantic segmentation using full convolutional neural network (FCN) avoids the problems of repeated calculation and storage due to using of pixel blocks. However, the results obtained by FCN are still not precise enough, and the results of upsampling are still relatively fuzzy and smooth. It is not sensitive to the details such as small object in the image. Therefore, this paper proposes an image segmentation method based on simplified deep residual network. A simplified Deep Residual Network (DRN) is proposed to replace the VGG-16 network in the original FCN framework. The net combines deep residual network architecture with 30% lesser parameters than the VGG model and the method of skip connection that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentation results. The results showed that the effect of segmentation is more refined and the recognition and segmentation of small objects are significantly improved by the improved network model.

1. Introduction

The semantic segmentation is to group the pixels in the image according to the prediction of semantic meaning that is to assign a label for each pixel, and to complete the tasks of simultaneous image segmentation and multi-object classification. Semantic segmentation has important research significance in the field of machine vision. It is widely used in robot navigation, automatic driving, intelligent security and other technical fields. Semantic image segmentation algorithms have been developed for decades, and various types of theories and methods have been proposed in recent years. With the emergence of large-scale training data and the rapid development of computer hardware, deep learning has becoming an efficient method in semantic segmentation of images. Convolution Neural Network (CNN) is an important deep learning architecture. It can extract the image features automatically and has a high classify accuracy. [1] uses DCNN (Deep CNN) for the first time about semantic segmentation called SDS. SDS proposed a simultaneous detection and segmentation method. The disadvantage of SDS is that it depends on a large number of proposals results in high computational load. Long J and Shelhamer E [2] proposed a full convolutional network for semantic segmentation and transformed the traditional image classification network directly into a pixel classification network. After this, a large number of FCN-based semantic segmentation algorithms [3-5] were proposed, which further promoted the development of image semantic segmentation.

In order to obtain more accurate results and increase the sensitivity to the details of image. This paper proposes an image segmentation method based on deep residual network. Firstly, we propose a simplified DRN network which with 30% lesser parameters than the VGG model for higher

recognition accuracy and real-time performance. Then, we fused the method of skip connection that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentations. The experiment results validate our proposed method.

2. Overview of FCN and Residual Network

Traditional methods for segmentation based on CNN uses the pixel blocks as input for training and prediction, resulted in repeated calculation and storage. To deal with this problem, FCN employed deconvolution [6] for upsampling and define a skip architecture that can enrich the partial loss information of the upsampling to produce accurate and detailed segmentation results.

In theory, the depth of the network is directly proportional to the learning ability of the network. However, as the number of network layers deeper, the network becomes more difficult to train, and the recognition rate begins to decrease after reaching a certain depth. To deal with this problem, [7] proposed deep residual networks which explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. Deep residual nets won the 1st places on the tasks of ImageNet detection, COCO detection, and COCO segmentation.

3. CNN Architecture: Residual Network with Skip Connection and Deconvolution

The convolution neural network structure designed in this paper is shown in figure 1, which is a 22-layer structure. It consists of 3 deconvolutions layers which are used to upsample and 5 types of convolution layer. Every type convolution layer which has special colour includes several convolution layers which is an operation required of the residual network.

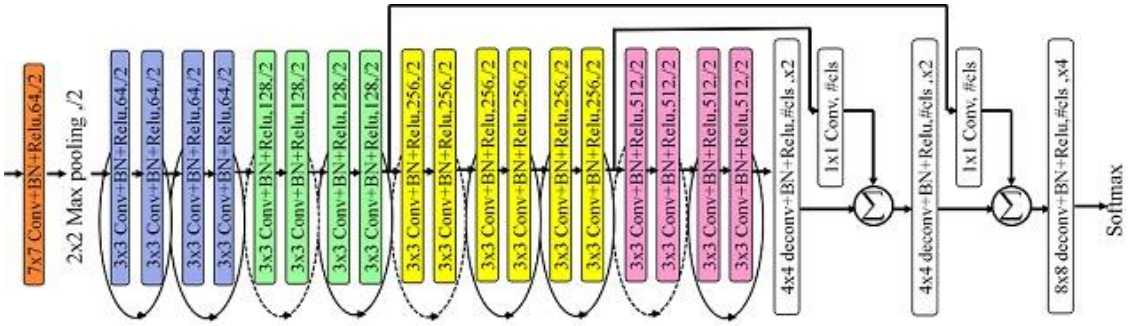


Figure 1. CNN Architecture: Residual network with skip connection and deconvolution.

We rescale incoming images to 224*224 resolution for our CNN, using bilinear interpolation for RGB. As is shown in figure 1, the first type convolution layer uses 64 convolution kernels of 7*7 to obtain 64 feature maps of 112*112 by convolution operation, and then take the 64 feature maps through a pooling layer with a window whose size is 3*3 (moving step is 2 pixels) obtaining 64 feature maps of 56*56(M1). The second type of convolution layer takes M1 as input and uses 4 same structures with 64 convolution kernels of 3*3 to obtain 64 feature maps of 56*56(M2). The third type of convolution layer takes M2 as input and uses 4 same structures with 128 convolution kernels of 3*3 to obtain 128 feature maps of 28*28(M3). The fourth type of convolution layer takes the 128 slices feature map as input and uses 6 same structures with 256 convolution kernels of 3*3 to obtain 256 feature maps of 14*14(M4). The fifth type of convolution layer takes the 256 slices feature map as input and uses 4 same structures with 512 convolution kernels of 3*3 to obtain 512 feature maps of 7*7(M5). After multiple convolutions and pooling, the feature maps become smaller and smaller and the resolution is getting lower. In order to recover from the low resolution image to the resolution of the original image, we adopt the skip connection method to reduce the step size of the upsampling at the shallow layer, combining the output feature maps of the 3rd and 4th layer with the feature map after the deconvolution of the next layer, thus enrich the partial loss information of the upsampling, And finally we use 8 times upsampling to realize the pixel-level segmentation. Meanwhile, residual

learning is used in every type of convolution layer (except the first type), each adjacent convolution layer is treated as a residual structure [7].

The convolution layers use BN [8] and ReLU [9] as the activation function to solve gradient disappearance and gradient explosion of the deep network. Also, the role of the BN layer is to speed up the network learning rate. The ReLU employs the method of unilateral inhibition that makes neurons in the neural network have sparse activation. Their function expressions are shown in figure 2.

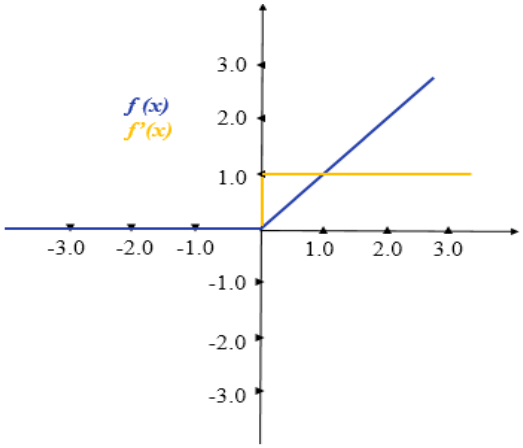


Figure 2. Function expression of ReLU.

Therefore, given the input image $I_t(u), u=(x,y)\subset Z^2, 0\le x<W,\ 0\le y<H$. CNN output a semantic segmentation map as a set of semantic class probabilities, i.e.

$$\tilde{S}(v)=P(c\mid I_t) \tag{1}$$

where $v=(s,t\subset Z^2), 0\le s<W/8, 0\le t<H/8$ and $P(c)$ denotes a class probability, $P(c)\subset R, 0\le P(c)\le 1, c\subset Z, 0\le c<N$ with N being the number of categories. The symbol $\sim$ denotes instead hereinafter a map of size $W/8\times H/8$.

4. Experimental Results and Analysis

4.1. Dataset

PASCAL VOC [10] provides a complete set of standard data sets for image recognition and segmentation, which is used to evaluate the performance of computer vision technology. Images were captured from real scenes. Figure 3 shows some images in dataset.

Figure 3. Some images from the PASCAL VOC dataset.

In this experiment, subset from PASCAL VOC, including 9963 training images and corresponding label objects are used for model training. There are 21 categories of datasets, such as the car, person, truck, and bicycle and so on. The label consists of two parts: class segmentation which marked category with each pixel and object segmentation which marked out which object each pixel belongs to. In this paper, we choose class segmentation label that is used to evaluate.

4.2. Network Training

A laptop with an Intel i7-6700 CPU@3.40GHz, a GTX1070 GPU with 16G memory, and operating system Ubuntu14.04 LTS is used for training and testing. The development language is python. We initialise our CNN with weights from Noh et. al. [11] trained for segmentation on the PASCAL VOC 2012 segmentation dataset. For optimisation we used standard stochastic gradient descent, with a learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. After 10k iterations we reduced the learning rate to 0.001. We use a mini-batch size of 64, and trained the networks for a total of 30k iterations over the course of 2 days on an Nvidia GTX 1070.

4.3. Accuracy Evaluation

We test our method on semantic segmentation and scene parsing, exploring PASCAL VOC. Semantic segmentation and scene analysis usually use the following four metrics: Pixel Accuracy, Mean Accuracy, Mean IoU(Intersection over Union),Weighted IoU. This paper uses Mean IoU as an evaluation metric compared with other methods. Mean IoU formula is as follows:

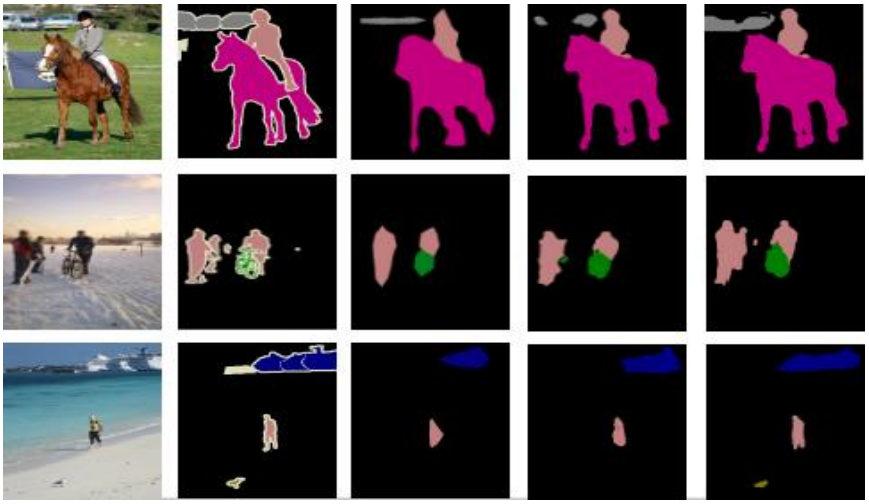
$$(1/n_{cl})\sum_i n_{ii} / (t_i + \sum_j n_{ji} - n_{ii}) \tag{2}$$

Let n_{ij} be the number of pixels of class i predicted to belong to class j, where there are n_{cl} different classes, and let $t_i = \sum_j n_{ij}$ be the total number of pixels of class i. The segmentation results on the PASCAL VOC-2012 test set are shown in Table 1. The method in this paper has been improved on the index of Mean IoU and is superior to FCN-8s.

Table 1. PASCAL VOC test set .

Method	Accuracy(Mean IoU)	Forward time
FCN-32s	59.2	~160ms
FCN-8s	62.2	~175ms
Our method	67.6	~ 98ms

The visual effects produced by this method are shown in figure 4: the results showed that the effect of segmentation is more refined and the recognition and segmentation of small objects are significantly improved by the improved network model.



(a) Image (b) Ground Truth (c) FCN-32s (d) FCN-8s (e) Our method

Figure 4. Semantic segmentation results.

5. Conclusions

In this paper, we propose an image segmentation method based on deep residual network. Firstly, a simplified DRN network is proposed which has higher recognition accuracy. Then, skip connection method is employed which combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer to produce accurate and detailed segmentation results. It can be seen from the results of the experiment, the effect of segmentation is more refined and the recognition and segmentation of small objects are significantly improved compared with FCN-8s by the modified network model.

6. Acknowledgments

This work is supported by the Ri-Xin Talents Project of Beijing University of Technology (2015-RX-L03) and the 2017 BJUT United Grand Scientific Research Program on Intelligent Manufacturing (040000546317552).

7. References

[1] Hariharan B, Arbelaez P, Girshick R 2014 Simultaneous detection and segmentation *European Conference on Computer Vision*, 297 - 312
[2] Long J, Shelhamer E, Darrell T 2015 Fully Convolutional Networks for Semantic Segmentation Conference *on Computer Vision and Pattern Recognition. IEEE*, 3431 - 3440
[3] Liu Z, Li X, Luo P 2015 Semantic Image Segmentation via Deep Parsing Network *International Conference on Computer Vision. IEEE*, 1377 - 1385
[4] Liang-Chieh C, Jonathan T, George P 2016 Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform *Conference on Computer Vision and Pattern Recognition. IEEE*, 4545 - 4554
[5] Guosheng L, Chunhua S, Anton van den H, Ian R 2016 Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation *Conference on Computer Vision and Pattern Recognition. IEEE, 3194 - 3203*
[6] Hyeonwoo N, Seunghoon H, Bohyung H 2015 Learning Deconvolution Network for Semantic Segmentation *International Conference on Computer Vision. IEEE*, 1520 - 1528
[7] Kaiming H, Xiangyu Z, Shaoqing R, Jian S 2016 Deep Residual Learning for Image Recognition *Conference on Computer Vision and Pattern Recognition. IEEE*, 770 - 778
[8] Sergey L, Christian S 2015 Batch normalization: Accelerating deep network training by reducing internal covariate shift *International Conference on Machine Learning*
[9] Vinod N, Geoffrey H 2015 Rectified linear units improve restricted boltzmann machines *International Conference on Machine Learning*

[10] Mark E, John W, Andrew Z 2010 The pascal visual object classes (VOC) challenge *International Journal of Computer Vision.* 303 – 338
[11] Hyeonwoo N, Seunghoon H, Bohyung H 2015 Learning deconvolution network for semantic segmentation *International Conference on Computer Vision. IEEE,* 1520 - 1528