Journal of KIIT. Vol. 16, No. 7, pp. 1-8, Jul. 31, 2018. pISSN 1598-8619, eISSN 2093-7571 1
http://dx.doi.org/10.14801/jkiit.2018.16.7.1

Research on Image Semantic Segmentation Based on FCN-VGG and Pyramid Pooling Module

Biao Hao*, Dae-Seong Kang**

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (NO. 2017R1D1A1B04030870).

Abstract

This paper proposed an algorithm to extract new feature maps using pyramid pooling module based on FCN-VGG(Fully Convolutional Network-Visual Geometry Group) for image semantic segmentation. First, the VGG was changed to FCN-VGG in order to consider the variability of the input images. The FCN-8 feature is extracted using the changed FCN-VGG. Then the extracted 4-parts features using pyramid pooling module and the previously extracted FCN-8 features are fused to obtain the advanced features. Classification is performed using obtained advanced features. By the comparison experiment between the feature maps of the proposed algorithm and the existing feature maps, it is proved that the method improves the accuracy of image semantic segmentation by about 2%.

요 약

본 논문은 이미지 시맨틱 분할을 위해 FCN-VGG(Fully Convolutional Network-Visual Geometry Group)기반으로 피라미드 플링 모듈을 사용하여 새로운 특징맵을 추출하는 알고리즘을 제안하였다. 먼저 입력 이미지의 가변성을 고려하기 위해 VGG를 FCN-VGG로 변경하여 실험하였다. 변경한 FCN-VGG를 이용하여 FCN-8 특징을 추출한다. 그리고 나서 파라미드 플링 모듈을 거쳐서 추출된 4-parts 특징들과 앞서 추출된 FCN-8 특징들을 융합하여 개선된 특징들을 얻는다. 새롭게 만들어진 특징들을 바탕으로 이미지의 픽셀들을 분류할 수 있다. 제안하는 알고리즘의 특징 맵과 기존의 특징 맵 간의 비교실험을 통해 제안하는 방법이 정확도 측면에서 평균적으로 2% 정도 개선되었다는 것을 증명했다.

Keywords
image semantic segmentation, FCN, VGG, CNN, deconvolution, pyramid pooling module

* Dong-A University Electronic Engineering
- ORCID: https://orcid.org/0000-0001-8127-4908
** Dong-A University Electronic Engineering Professor
- ORCID: https://orcid.org/0000-0003-0186-2430

· Received: Jan. 26, 2018, Revised: Jul. 11, 2018, Accepted: Jul. 14, 2018

· Corresponding Author: Dae-Seong Kang
Dept. of Dong-A University,37 NaKdong-Daero 550, beon-gil
saha-gu, Busan, Korea,
Tel.: +82-51-200-7710, Email: dskang@dau.ac.kr

Ⅰ. Introduce

Visual information is the main source of human information obtained from the objective world and also an important means of human cognition of the world. Image semantic segmentation is the most basic problem in computer vision. Its goal is to classify each pixel of an image. Image classification[1], object detection[2] and image semantic segmentation are the three core research issues in computer vision. Among them, the task of image semantic segmentation is the most challenging. Image semantic segmentation combines the traditional tasks of image segmentation and object recognition. The purpose of the image segmentation is to segment the image into several groups of pixels with some specific semantic meaning and identify the class of each region. Finally an image with pixel semantic annotation can be got. In the field of automatic driving, through semantic segmentation of the scene in front of the car body, it can accurately locate road, cars body, pedestrians and object information to improve the safety. Image semantic segmentation will greatly promote the development of related industries such as automatic driving or assisted driving.

The whole development of society is moving towards the dream of artificial intelligence step by step. With the emergence of deep learning, the birth of convolution neural network, humans take a big step in the field of artificial intelligence. Now deep learning has become a research hotspot, it is widely used in image recognition, semantic understanding and so on. Deep learning is essentially a very complex deep neural network. With the development of technology, deep learning shows great performance on the problem of imagenet. Increasing convolution networks's layer can extract advanced features that are hard to reach by artificially defining. So it is very effective for feature extraction by neural networks. In image processing, more and more people use convolutional neural networks[3]. Convolutional neural network in the image classification have achieved very good results, its recognition rate is even higher than the recognition rate of the human eye. The AlexNet and VGG network are most used in image classification. Table 1 is comparison of Alexnet and VGG convolutional neural network. Here M is MByte.

Table 1. Comparison of alexnet and VGG

	Alexnet	VGG
Layer	8 layers	16 layers
Parameters	57M	134M
Accuracy	High	Higher than Alexnet
Kernel	11*11, 5*5	Change into 3*3
	Calculation parameters reduced	
Advantage	Learning feature ability of VGG is higher.	

So in this study VGG convolutional neural network[4] is chosen. The problem of image semantic segmentation is transformed into pixels classification though transforming fully connection layers into convolutional layers. This paper adds pyramid pooling model based on VGG full convolutional neural network for pixel classification to achieve image semantic segmentation.

Ⅱ. Correlational Research

2.1 Previously used algorithms

The method based on the edge detection is mainly used to realize image segmentation through the edge detection of the region. By using the inconsistency of features between regions, first it detects the edge points in the image and then connects them into a closed curve according to a certain strategy to form the segmentation region. The edges in an image are usually places that are not continuous in grayscale, color or texture. It often needs to use edge detection operators to carry out the edge detection. They include canny operator[5], laplace operator[6] and so on. The edge detection algorithm is more suitable for the segmentation of simple images whose edge gray value is more obvious and noise is smaller. Due to the more complex edges and the strong noise images, it faces the contradiction between noise immunity and detection accuracy. Therefore, this method is often used as a preprocessing algorithm for image segmentation. It is that first preprocessing is performed on image, and then subsequent method is used to deal with the image.

Conditional Random Field(CRF)[7] and Markov Random Field(MRF)[8] are semantic segmentation methods based on machine learning[9]. Markov Random Field is a theoretical model of undirected probability graphs, which has a wide range of applications in computer vision. The idea is to assign a random value to the pixel and then calculate and classify it in a probabilistic way.

The conditional random field is a discriminant probability undirected graph learning model based on the maximum entropy model and hidden markov model. It is a conditional probability model for annotating and segmenting ordered data. The conditional random field model uses the serialization form to optimize and decode the global parameters, solving the paranoid mark problem that other discriminant models can't avoid. It has a good application effect in the fields of natural language processing and image semantic segmentation.

2.2 Convolutional neural network

The method based on the machine learning through training to build the probabilistic model between image features and semantic categories, build a energy function and use the handwritten feature library to optimize the function through iterative computation, finally the image semantic segmentation model is obtained. This machine learning method rely on hand-marked feature library very much, so it is difficult to express the image features widely. Therefore, it has lot of limitations in practical application. The traditional image segmentation is to extract the features such as texture, spatial structure information, color and other features from different regions. Semantic information of the same area is suppressed and there are certain differences between different regions. Machine learning has restrictive in dealing with the original data. It require senior engineer or experts that for different target design different feature extractor and extract the appropriate feature vector.

Convolutional Neural Network (CNN) can automatically learn the distribution of image features and avoid the need of artificial feature extraction. Compared to traditional methods, CNN can learn more abstract image features from a large number of sample images. The traditional convolutional neural network is divided into two parts. The first part consists of multiple convolutional layers and the pooling layers, they are alternately responsible for the feature extraction. The second part is fully connection layers, its task is classification. For ordinary CNN, the input images need to be fixed due to the existence of the fully connection layer. For the convolutional layer and the pooling layer, the size of each layer's feature map is variable. If fully convolutional network is used to extract features of input image, the input image scale is arbitrary. Deep convolutional network is widely used to solve many image processing problems due to its excellent performance in image processing. The core of deep learning is feature learning. Its purpose is to obtain the hierarchical information through the hierarchical network so as to solve the problem that artificially design feature extraction problem. It is simple to say that make the machine to enter and own output. Learning process is to learn the most suitable network parameters weight and bias, under the conditions of the minimum loss to optimize their own learning parameters, finally it is able to accurately predict input data. Convolutional neural network is a neural network which consist of neurons with learnable weights and bias. The basic principles of convolutional neural networks are local connection, weight sharing, downsample. The proposed local connection is to reduce the training weight parameter. The nodes in the hidden layer are only connected with some nodes of the upper layer, it greatly reduces the amount of calculation. Weight sharing can reduce the computational complexity by convolution kernel applications, and downsample is a further reduction of network computation parameters. Fig. 1 is a simple convolutional neural network. It has input layer, convolutional layer, pooling layer, fully connection layer, output layer, but the last layer of convolutional neural network is commonly softmax layer.

Here only the ReLU function is introduced, because most convolutional neural networks use ReLU activation function. Other activation functions are sigmoid and tanh functions and so on. Fig. 2 is the image of activation function.

Fig. 3 is the process of convolutional calculation. The calculation of convolutional neural network is that image data and convolution kernel are calculated to get a value.

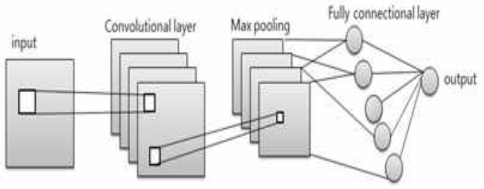


Fig. 1. Convolutional neural network

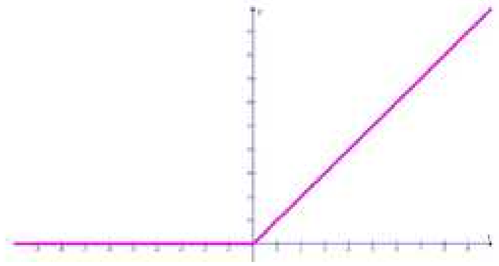
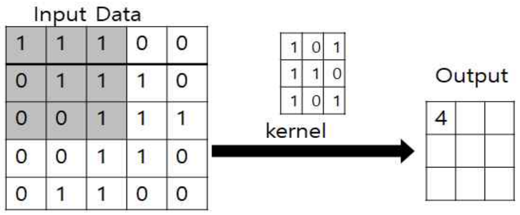


Fig. 2. ReLU function

Fig. 4 is pooling calculation. It contains max pooling and average pooling. The main role is to reduce the amount of computation.

Fig. 5 is a VGG convolutional neural network, this neural network has 16 layers, including 13 convolutional layers, 3 fully connection layers and 5 layers of pooling layer. The specific structure of the neural network in turn are 2 convolutional layers, 1 pooling layer, then 2 convolutional layers, 1 pooling layer, 3 convolutional layers, 1 pooling layer, 3 convolutional layers, 1 pooling layer, 3 convolutional layers, 1 pooling layer, and 3 fully connection layers.

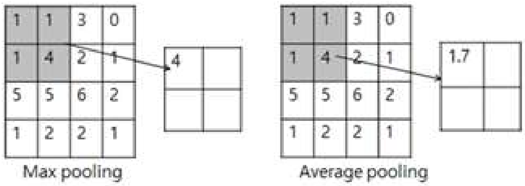


1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

1	0	1
1	1	0
1	0	1

4		

Fig. 3. Convolutional operation



1	1	3	0
1	4	2	1
5	5	6	2
1	2	2	1

4	

1	1	3	0
1	4	2	1
5	5	6	2
1	2	2	1

1.7	

Fig. 4. Pooling calculation

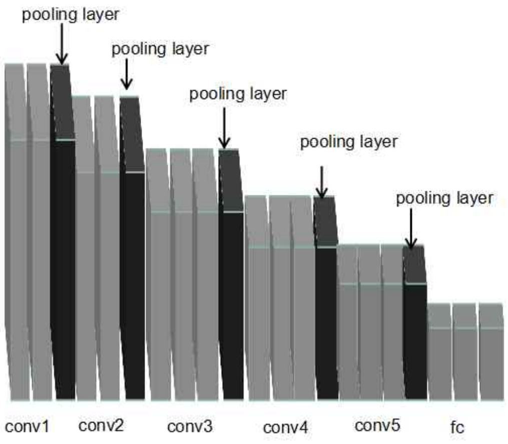


Fig. 5. VGG convolution network

2.3 The selection of the framework

The libraries for machine learning and deep learning are TensorFlow[10] Caffe, Theano, MXNet and so on. Here tensorflow is chosen. TensorFlow is an artificial intelligence platform open source software from Google in 2015. TensorFlow is a system that transmits complex data structures to an artificial neural network for analysis and processing. TensorFlow can be used in variety of deep learning areas such as speech recognition or image recognition, and it can be run on a wide range of devices, from as little as a smart phone to as many as thousands of data center servers. Tensorflow can be divided into tensor and flow, tensor is an N-dimensional array, flow is the meaning of graph, and tensorflow is calculated based on graph. TensorFlow running process is divided into two steps, one is construct the model and the other is training. In the construction phase, we need to build a graph that describes our model and then launch it in the session. The map can be understood as a flow chart that is the data input and output process, but this time it is not an actual operation, because TensorFlow is deferred execution model. It must know what you are calculating, and then your execution graph start sending computational tasks to computers. So first we should create a calculation graph in internal storage and start an execution session. Finally actual training tasks is performed by session.run. Fig. 6 is chart of a calculation that use tensorflow.

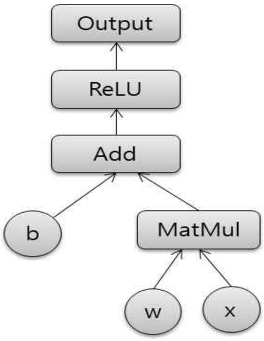


Fig. 6. The use of tensorflow

2.4 Fully convolutional networks

Now most people use convolutional neural networks in image processing. The most used frequently convolutional neural networks are Alexnet, VGG, Googlenet[11] and so on. In the image classification task, VGG convolutional neural network can well achieve the task of classification. FCN-VGG is the change of VGG convolution neural network. By converting the normal VGG network's fully connection layers into convolutional layers, FCN-VGG can be got. FCN-VGG is used to realize image semantic segmentation. Image semantic segmentation is that divides the pixels into different groups according to semantic meaning in the image.

III. Proposed Algorithm

In this paper, proposed algorithm use FCN-VGG network to extract feature maps of input image. Extracted feature map FCN-8 is inputted into pyramid pooling module, through pyramid pooling module, it can extract 4 parts feature maps. Then the extracted 4-parts feature maps and FCN-8 feature map is fused into advanced feature maps. Finally the new feature maps is used to process classification. Fig. 7 is framework of proposed algorithm.

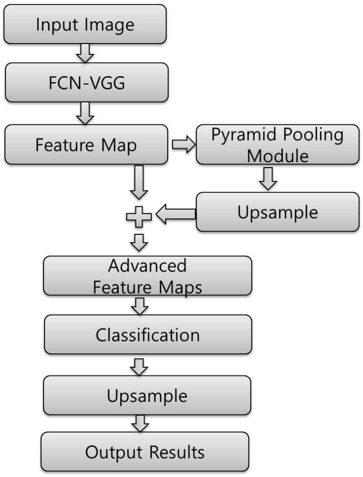


Fig. 7. The structure of proposed algorithm

The training steps of image semantic segmentation using proposed algorithm as follow.

Step 1: First select an appropriate convolutional neural network, in this study VGG network is chosen.

Step 2: Carry out convolutional neural network transformation, transform the VGG network into full convolutional neural network.

Step 3: Training the network, and extract feature map FCN-8 of input image using FCN-VGG.

Step 4: Use extracted FCN-8 features to perform pyramid pooling module, and extract 4-parts features. Then carry out upsample operation, fuse the features and FCN-8 features into advanced features.

Step 5: Here use softmax classifier to classify the pixels of image.

Step 6: Carry out upsample operation, recovery the size of image to the original image.

Step 7: Output semantic segmentation images.

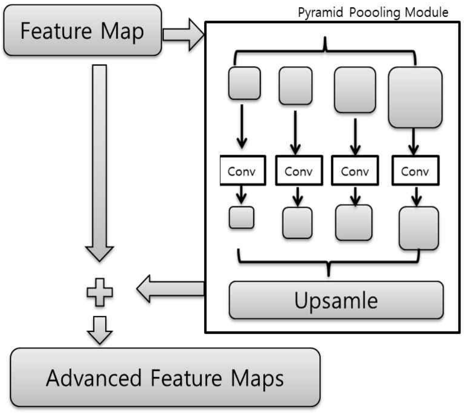


Fig. 8. The structure of pyramid pooling module

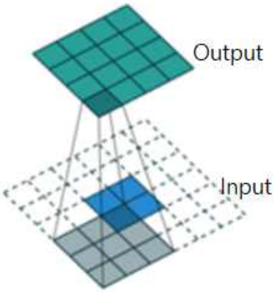


Fig. 9. Upsample operation

3.1 Pyramid pooling module

The pyramid pooling module can learn different features of multi-scale and then combine, it have stronger expression ability. Fig. 8 is the structure of pyramid pooling module.

Due to the increase of the parameter quantity, the 1*1 convolutional layer is used in the pyramid pooling module to reduce the parameters. Then fuse the feature maps and original feature maps to obtain advanced features, finally the new features is used to perform the pixels classification. Here 4 scales pools are used in the pyramid pooling module, they are respectively 1*1, 2*2, 3*3 and 6*6. Then pooling is performed in feature map respectively. In this module it has used upsample module. Fig. 9 is upsample operation.

Because the ordinary pool will reduce the size of the images, such as VGG's five pooling layers can reduce the input image by 32 times. In order to get a output image that the size is as the same as the original image, it need upsample. The size of extracted 4-parts using pyramid pooling module is decrease. So first upsample is used to revert to the size of the original feature map, then classification is performed.

Ⅳ. Dataset and Experimental Results

PASCAL VOC 2012 is usually used to evaluate semantic segmentation algorithms. In this paper the database is used to do research. Table 2 is database introduction.

Table 2. PASCAL VOC 2012 database

Database	Image Size	Class Number	Training Database	Test Images
PASCAL VOC2012	500*375	20 Classes	2913 Images	Choose 50 Images

Table 3. Accuracy of methods

	FCN-VGG(%)	Proposed Method(%)
Pixel accuracy	90.3	92.1
Mean accuracy	75.9	77.3

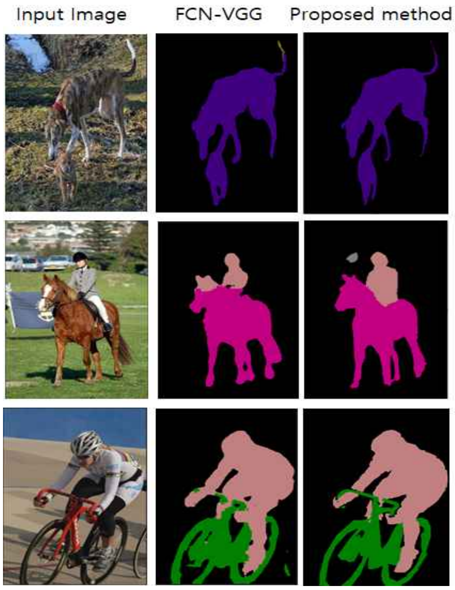


Fig. 10. Results of semantic segmentation

Table 3 is the accuracy of image semantic segmentation based on FCN-VGG and proposed algorithm. Semantic segmentation is a classification based on the pixel level. It can read the classes and location of object in the image. Fig. 10 is the experimental results. From the results we can see the effects of using different algorithm the image semantic segmentation.

Ⅴ. Conclusions

With the development of deep learning, convolution neural network has achieved great success in image processing. For image recognition convolutional neural networks recognition rate is higher than our human recognition rate. At present unmanned restaurant, automatic drive, map search and other technologies become the research hotspot among current scholars. Semantic segmentation is essential for these algorithm. This study uses FCN-VGG to extract FCN-8 feature, then pyramid pooling module is performed based on FCN-8 feature. And then feature fusion is performed. Finally pixels classification is perform. From the experimental results, we can see that the results of image semantic segmentation is better. The accuracy of the proposed method for image semantic segmentation is improved by about 2%.

In the future, semantic segmentation algorithm is valuable. So we will apply different network frameworks to optimize the proposed algorithm for better result of semantic segmentation.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", Communications of the ACM, Vol. 60, No. 6, pp. 84-90, Jun. 2017.

[2] P. B. Felzenszwalb, R. B. Girshick, and D. M. Allester, "Object Detection with Discriminatively Trained Part-Based Models", IEEE, Vol. 32, pp. 1627-1645, Sept. 2010.

[3] H. W. Jang, and S. H. Jung, "Training Artificial Neural Networks and Convolutional Neural Networks using WFSO Algorithm", Journal of Digital Contents Society, Vol. 18, No. 5, pp. 969- 967, Aug. 2017.

[4] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", ICLR, pp. 1-14, Apr. 2015.

[5] Z. Wen, X. Zhang, L. S. Zhou, and L. Zhang, "Modified Edge Detection Algorithm based on Canny", Communications Technology, pp. 2236-2240, Oct. 2017.

[6] Y. Zheng and X. H. Sun, "The Improvement of Laplace Operator in Image Edgedetection", Journal of Shenyang University and Civil Engineering Institute, Vol. 21, No. 3, pp. 268-271, Mar. 2005.

[7] M. Ling and X. Mei, "Image Semantic Segmentation based on Higher-order CRF

Model", Application Research of Computers, Vol. 30, No. 11, pp. 3514-3517, Nov. 2013.

[8] Y. Huang, K. Fu, and Y. R. Wu, "Image Segmentation Method Using K-Means Based on Markov Random Field", ACTA Electronica Sinica, Vol. 37, No. 12, pp. 2700-2704, Dec. 2009.

[9] S. H. Park, S. Y. Ihm, and Y. H. Park, "A study on Application of Machine Learning Algorithm to Visitor Marketing in Sports Stadium", Journal of Digital Contents Society, Vol. 19, No. 1, pp. 27- 33, Dec. 2018.

[10] https://en.wikipedia.org/wiki/TensorFlow. [accessed: Jan. 12, 2018]

[11] C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Rees, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", IEEE Conference on Computer Vision & Pattern Recognition(CVPR), pp. 1-19, Sep. 2014.

Authors

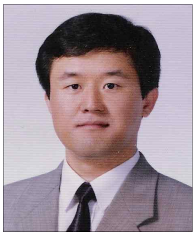
Biao Hao

2015 : B.S., Electronic Information Engineering, Hebei University of Science and Technology.

2016 ~ 2018 : M.S., Electrical Engineering, Dong-A University.

Research interests : signal processing and pattern recognition.

Dae-Seong Kang



1994 : Ph.D., Electric Engineering, Texas A&M University.

1995 ~ present : Professor, Dong-A University.

Research interests : image processing and compression.