A Survey On Image Semantic Segmentation Methods With Convolutional Neural Network

Fude Cao
School of Information & Art
Shandong Institute Of Commerce & Technology
Jinan, China
caofude@sict.edu.cn

Qinghai Bao
Jinan Labour Wu Paragraph
China Railway Jinan Group Co., Ltd.
Jinan, China
357367202@qq.com

***Abstract*—Image semantic segmentation is an important branch in the field of AI. Traditional semantic segmentation algorithms are mostly specific to the problem, and there is no universal segmentation algorithm suitable for all images. But deep learning approaches can solve this problem. Recently convolutional neural network is a widely used model in image segmentation, target recognition and scene classification, and has achieved great success. This survey introduces what image semantic segmentation is, what the semantic segmentation approaches are, and the methods of image segmentation with CNN. Next, we present several data sets that are often used in image segmentation experiments. Finally, these deep learning algorithms of image segmentation are analyzed and compared appropriately.**

Keywords-component; Image semantic segmentation, deep learning, CNN, FCN

I. INTRODUCTION

Image semantic segmentation is a basic computer vision task. There can be no correct identification without proper segmentation. Image semantic segmentation is mainly applied to automatic driving, 3D map reconstruction, image beautification, face modeling. So image semantic segmentation is a key step from image processing to image analysis. Image semantic segmentation methods are divided into traditional semantic segmentation and deep learning semantic segmentation. The basic idea of deep learning semantic segmentation is to obtain a linear decision function by training a multi-layer perceptron, and then use the decision function to classify pixels to achieve the purpose of segmentation.

Convolutional neural network (CNN) has made great achievements and been widely used in image classification and image detection since 2012 [1]. In the early stage CNN-based segmentation method has several disadvantages: First, the storage overhead is large. Second, the calculation efficiency is low. Third, the size of the pixel block limits the size of the perceived area. Usually, the size of a pixel block is much smaller than the size of the entire image, and only some local features can be extracted, resulting in the limited performance of the classification.

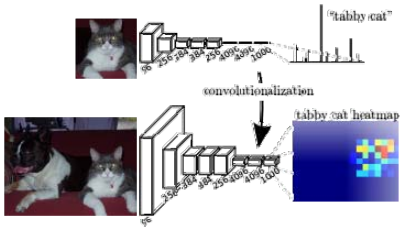
In response to these problems, UC Berkeley's Jonathan Long et al. proposed Fully Convolutional Networks [2] for image segmentation. The network attempts to recover the category to which each pixel belongs from the abstract features. That is, the classification from the image level is further extended to the classification at the pixel level. Compared to the traditional methods of image segmentation with CNN, FCN has two distinct advantages: First, it can accept input images of any size without requiring all training images and test images to have the same size. Second, it is more efficient because it avoids the problems of repeated storage and computational convolution due to the use of pixel blocks.

This paper is organized as follows: in Section 2 we intently analyze the image semantic segmentation approaches with CNN. We focus on the popular techniques in recent years. Then we make a review about several datasets for image semantic segmentation in Section 3. After that, we discussion and conclusion in section 4 and section 5.

II. APPROACHES

A. Fully Convolutional Networks [2]

The first successful neural network in image segmentation is Fully Convolutional Networks (FCN). This paper's key insight is to build "fully convolutional" networks that take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning [2]. FCN is to replace all fully connected layers with the convolution layers as shown in Fig.1. Originally, only one category classification network can output one classification result in each pixel of the feature map. Now we can turn the vector of classification into a character graph of classification.

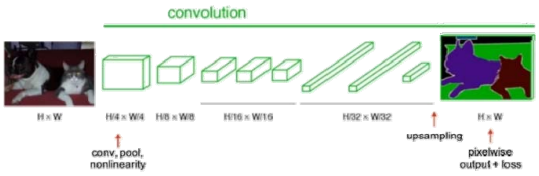


(Figure 1) Replace fully connected layers with convolution layers [2]

The fully connected layer of each neuron becomes a convolution kernel, and then the size of the original image is reduced by convolving the features of the small image through an upsampling layer. The feature map of the last layer is actually softmax of dense, which is equivalent to loss function

978-1-7281-9761-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CISCE50729.2020.00103

with fully connection network at every pixel point as shown in Fig.2.



(Figure 2) FCN Architecture [2]

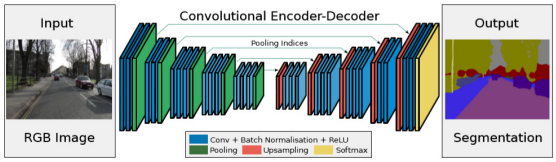
It is equivalent to making a classification network at each point in the original graph size, so FCN is essentially a dense classification network. It simply replaces the middle fully connected layers with the convolution layers, subtly changing the output of only one classification result to the output of any number of classification results.

FCN uses VGG-16 classification neural network and then make it fully convolulational. Moreover, we use skip architectures by concatenating upsampled pool 1 to 4 with the score layer to get finer features. Training was done on two stages, first on PASCAL VOC training dataset, secondly training plus validation datasets.

After that, the core of all image segmentation algorithms using CNN is the improvement of fully convolutional network. The network using FCN includes Dilated Convolutions [3], Feature Pyramid Networks [4], Pyramid Scene Parsing Network [5], Mask R-CNN [6], Conditional Random Fields RNN [7], SegNet [8], U-Net [9], DeepLab v1[10], DeepLab v2[11], RefineNet [12], Large Kernel Matters [13], DeepLab v3[14], DeepLab v3+ [15].

In order to restore the classified feature map to the original size, the upsampling layer is adopted. Upsampling contains two methods. One is the deconvolution. The other is bilinear interpolation. FCN achieved the most advanced segmentation of PASCAL VOC [2].

B. *SegNet [8]*



(Figure 3) SegNet Architecture [8]

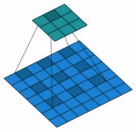
SegNet is designed to solve autonomous driving or intelligent robots. It is based on FCN, modifying the segmentation network obtained by VGG-16. The encoder part of SegNet uses the first 13-layer convolutional network of VGG16. Each encoder layer corresponds to a decoder layer, and the output of the decoder is sent to soft-max classifier to generate class probability for each pixel independently (Fig. 3).

The encoder and the FCN have the same convolutional layer. It removes all fully connected layers. The decoder uses the max-pooling indices received from the corresponding encoders to perform nonlinear upsampling of the input feature map. Indices play an important role in the upsampling process then SegNet can achieve end-to-end training. It makes SegNet more efficient than FCN in memory and training time. On CamVid dataset and the other large dataset SegNet is competitive and scores high for road scene understanding.

C. *Dilated convolutions [3]*

Many neural networks now use dilated convolution, also called atrous convolution. This is because image segmentation requires more global information for segmentation. The dilated convolution can make the visual features larger. Further, the segmentation accuracy is much higher and it does not increase the amount of calculation (Fig. 4).

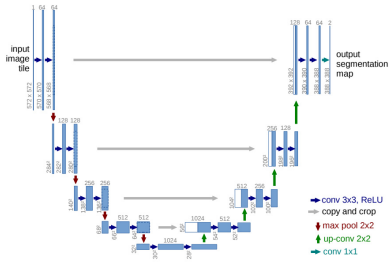
Dilated convolution increases the receptive field of convolution kernel while keeping the number of parameters unchanged. At the same time, it can ensure that the size of feature map output remains unchanged. It is mentioned in this paper that the presented context module increases the accuracy of state-of-the-art semantic segmentation systems.



(Figure 4) The original convolution kernel only covered the 3 by 3 matrix, now it covers the 6 by 6 matrix, and the middle region is set to 0 [21].

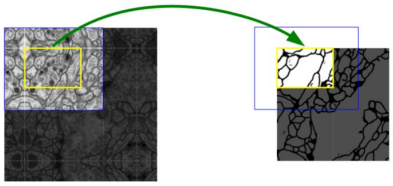
D. *U-Net [9]*

The full name of U-Net is Convolutional Networks for Biomedical Image Segmentation. U-Net consists of two parts, which can be seen in Figure 5. The first part is for feature (context information) extraction that is similar to FCN. The second part is the upsampling part. Because the network structure is U-shaped, it is called U-Net network.



(Figure 5) U-Net Architecture [9]

Its architecture still does not have any fully connected layers, so it can greatly reduce the need for training parameters. Only valid parts of the convolution layer are used. A very important modification to the FCN network structure is the addition of a large number of feature channels in the upsampling part. The U-Net multi-scale feature fusion method is the feature channel dimension splicing, while the FCN is the feature point-by-point addition. And the overlap-tile strategy is adopted. This strategy is for large images such as medical image so that the resolution of large images is not affected (Fig. 6).

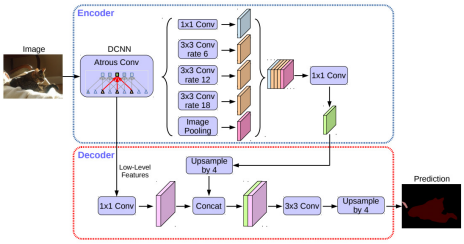


(Figure 6) The use of a seamlessly split overlap-tile strategy for arbitrarily large image is very successful. The image is mirrored around the input image [9].

With U-Net not only medical images, but also natural image segmentation can achieve very good results. U-Net can use data augmentation to train some relatively small samples of data, such as 30 images. The u-net model yielded a final ranking score of 0.9965 on medical image dataset.

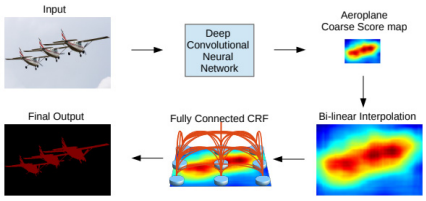
E. DeepLabv3+ [15]

The DeepLabv3+ also is the encoder-decoder architecture. The encoder architecture uses DeepLabv3. The decoder uses a simple but effective module to recover the target boundary details. We can use the dilated convolution to control the resolution of the feature under the specified computing resources (Fig. 7).



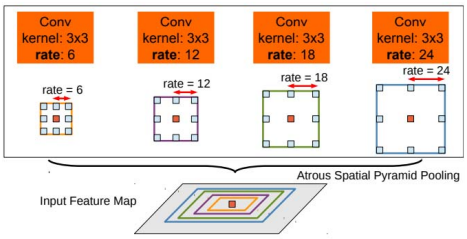
(Figure 7) DeepLabv3+ Architecture [15]

Let me briefly review DeepLab family history. DeepLabv1 [10] uses the atrous/dilated convolutions to extend the receptive field and get more contextual information. The classifier is invariant that requires spatial transformation, which naturally limits the positioning accuracy of the CNN. So DeepLabv1 uses the fully connected CRF to improve the model's ability to capture details (Fig. 8).



(Figure 8) DeepLabv1 Pipeline [10]

DeepLabv2[11] is an optimization based on DeepLabv1. It proposes Atrous Spatial Pyramid Pooling (ASPP) (Fig. 9). DeepLabv1 works hard in three directions, but the problems remain which are reduced feature resolution, multi-scale objects, and translational invariance of DCNN. DeepLabv2 removes subsampling in the last few largest pooling layers and instead uses atrous convolution to calculate feature maps at higher sampling densities. For the third problem, it proposes parallel sampling of atrous convolution at different sampling rates on a given input, which is equivalent to the context of capturing images in multiple scales. DeepLabv2 is the ability to sample fully connected CRFs to enhance the model's ability to capture detail.



(Figure 9) Atrous Spatial Pyramid Pooling module [11]

DeepLabv3+ further uses depth-wise separable convolution in the ASPP and decoder modules. The ASPP (Fig. 9) is designed to capture rich contextual information by pooling operations at different resolutions. The decoder structure is designed to gradually obtain clear object boundaries. The ASPP and decoder modules can improve the speed and robustness of the encoder-decoder network. So DeepLabv3+ is the latest Google semantic image segmentation model with the best performance.

III. DATASETS

Image semantic segmentation most commonly used data sets mainly introduce three: PASCAL VOC, CityScapes and CamVid. The PASCAL VOC2012 is the most vital datasets for semantic segmentation [16] [17].

The PASCAL VOC provides standardized image data sets for object class recognition. It also provides a common set of tools for accessing the data sets and annotations, enables evaluation and comparison of different methods. So it can be evaluated the performance of various image segmentation methods by running a challenge on this dataset [18].

The PASCAL VOC is a relatively old data set, it provides 20 categories, including people, cars and so on. There are 6,929 labeled pictures, which provide class-level labeling and distance-level labeling. In other words, we can do semantic segmentation and only distinguish whether cars are cars or not. You can also do distance segmentation, you can distinguish a few cars, and can mark different cars. Most images in this dataset have a foreground or two surrounded by highly diverse backgrounds. It implicitly leads to bias towards algorithms containing detection techniques.

The Cityscapes Dataset focuses on semantic understanding of urban street scenes [19]. It has 30 detailed categories. Five thousand of the images were finely annotated to the pixel level. There are also 20,000 images with rough markings. It can also provide class-level segmentation and distance-level segmentation.

CamVid (Cambridge-driving Labeled Video Database) is automotive dataset which contains 367 training, 101 validations, and 233 testing images [20]. It is the first video

collection with object class semantic tags. The database provides ground truth labels. There are 11 different classes, such as buildings, trees, sky, cars, roads, etc. The data is taken from the perspective of driving a car. The driver scenario increases the number and heterogeneity of the observed object classes.

IV. Discussion

FCN is the originator of image semantic segmentation. All semantic segmentation networks based on CNNs are based on FCN. Image semantic segmentation is actually a process of encoding and decoding. Encoders use multiple convolutional layers and pooling layers, and decoders often use upsampling layers. In order to improve the accuracy of segmentation, it is necessary to work on the multi-scale feature fusion.

SegNet uses the max-pooling indices technique on the basis of FCN to speed up feature fusion. The Dilated Convolution directly employs atrous convolutional layers to make the view feature larger. Feature channel dimension splicing can be used by U-Net. DeepLabv3+ uses Atrous Spatial Pyramid Pooling and other techniques for making multi-scale features very well integrated.

Specific performance details are shown in table 1, table 2, table 3 and table 4.

(Table 1) Compare on VOC2012 dataset

Approach	Mean Score	Source
FCN-8s	62.2	Leaderboard[21]
SegNet	59.9	Leaderboard[21]
Dilated Convolutions	75.3	Reported in the paper[3]
DeepLabv3+_JFT	89.0	Leaderboard[21]

(Table 2) Compare on Cityscapes dataset

Approach	IoU cla.	Source
FCN	65.3	Reported in the paper[5]
Dilated Convolutions	67.1	Reported in the paper[5]
DeepLabv3+	82.1	Reported in the paper[15]

(IoU = Area of Overlap /Area of Union)

(Table 3) Compare on CamVid dataset

Approach	mIoU	Source
FCN	49.83	Reported in the paper[8]
FCN (learnt deconv)	51.96	Reported in the paper[8]
SegNet	60.10	Reported in the paper[8]

(Table 4) Carvana Image Masking Challenge performance

Approach	Mean Score	Source
U-Net	Winner	Leaderboard[22]

For the above models, it is obvious that DeepLabv3+ is the best segmentation efficiency in the Pascal VOC2012 ranking. However, in practical applications, we must consider factors such as memory usage and training time. The motivation for creating SegNet is to build a smaller, faster and more efficient network. U-Net is also a small network, but its structure can combine high-level and low-level features to restore fine edges. Not only is it far ahead in medical image segmentation, but it also excels in natural image segmentation. In the Carvana image masking challenge of Kaggle, U-Net won the first place, but DeepLabv3+ was behind. When we choose the network model for image semantic segmentation, we should choose the appropriate convolutional neural network architecture based on the actual situation and comprehensive consideration.

V. Conclusion

With the continuous opening of data sets, image semantic segmentation approach based on CNNs has become an efficient and universal segmentation method. This is different from traditional image segmentation. In this paper we investigate the most recent and efficient approaches such as FCN, SegNet U-Net and DeepLabv3+. We also showed the data set which are most commonly used for the image semantic segmentation. Among these approaches DeepLabv3+ presents the best performance but U-Net also has competitive advantages.

Acknowledgment

Thanks to Professor Hyungjeong Yang at Chonnam National University in South Korea for her great help.

References

[1] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton "ImageNet Classification with Deep Convolutional Neural Networks" 2012

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully Convolutional Networks for Semantic Segmentation" pp. 3431–3440, 2015

[3] Fisher Yu, Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions" Published as a conference paper at ICLR 2016

[4] Tsung-Yi Lin, Piotr Doll´ ar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie "Feature Pyramid Networks for Object Detection" pp. 2881–2890, 2016

[5] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang and Jiaya Jia, "Pyramid Scene Parsing Network" 2016

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick, "Mask R-CNN" arXiv:1703.06870v3 [cs.CV] 24 Jan 2018

[7] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr, "Conditional Random Fields as Recurrent Neural Networks" 2016

[8] Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla, Senior Member, IEEE, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation" 2016

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox "U-Net: Convolutional Networks for Biomedical Image Segmentation" 2015

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs", 2015

[11] Liang-Chieh Chen, George Papandreou, Senior Member, IEEE, Iasonas Kokkinos, Member, IEEE, Kevin Murphy, and Alan L. Yuille, Fellow, IEEE "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs" 2017

[12] Guosheng Lin, Anton Milan1, Chunhua Shen, Ian Reid, "RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation" 2016

[13] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, Jian Sun, "Large Kernel Matters — Improve Semantic Segmentation by Global Convolutional Network" 2017

[14] Liang-Chieh Chen, George Papandreou, Florian Schroff, Hartwig Adam, Google Inc, "Rethinking Atrous Convolution for Semantic Image Segmentation" 2017

[15] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation" 2018

[16] Sasank Chilamkurthy, "A 2017 Guide to Semantic Segmentation with Deep Learning" July 5, 2017

[17] A. Garcia-Garcia, S. Orts-Escolano, S.O. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A Review on Deep Learning Techniques Applied to Semantic Segmentation" 2017

[18] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," International Journal of Computer Vision, vol. 111, no. 1, pp. 98–136, Jan. 2015

[19] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016

[20] G. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," PRL, vol. 30(2), pp. 88– 97, 2009

[21] http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?cls=aeroplane&challengeid=11&compid=6&submid=17681

[22] https://www.kaggle.com/c/carvana-image-masking-challenge