Development of Semantic Segmentation Based on Deep Learning

Yang Zhao *

Department of Communication Engineering, Xidian University, Xi'an, China

* Corresponding Author Email: yz2075@hw.ac.uk

Abstract. In recent years, the discipline of computer vision has seen a lot of interest in the study of image semantic segmentation. Deep learning has grown in popularity, and deep learning and image segmentation have combined and improved. These technologies are now widely employed in autonomous vehicles, intelligent robots, and other devices. In the beginning, Fully Convolutional Networks (FCN) or U-net-based semantic segmentation techniques were proposed; FCN realized an end-to-end training network and effectively applied Convolutional Neural Networks (CNN) to the semantic segmentation domain. To improve outcomes in the field of semantic segmentation, the encoder-decoder structure from the FCN approach was later implemented, and the Atrous Convolution approach was also proposed. Transformer-based semantic segmentation techniques are another recent trend, in addition to CNN-based networks. The Transformer model was first proposed in 2017, and subsequent Transformer-based semantic segmentation methods have also achieved good results. In this paper, these various methods will be compared and discussed to provide a guidance for this field.

Keywords: Semantic Segmentation, CNN, FCN, Transformer.

1. Introduction

The goal of image segmentation is to categorize various objects in an image. Additionally, Semantic Image Segmentation will classify objects at the pixel level. With semantic segmentation, a corresponding class of what is being represented is assigned to each pixel of an image. As opposed to instance segmentation, which separates instances of the same class in images, semantic segmentation concentrates on the category of each pixel. Autonomous vehicles and image diagnosis in medicine are just two applications of semantic segmentation models. Semantic segmentation's straightforward objective is to take a grayscale image or an RGB color image and produce a segmentation map where each pixel carries a class label. Up to now, there are many semantic segmentation methods e.g. Fully Convolutional Networks(FCN) [1], U-net [2], other CNN-based model [3], and Transformer-based model [4].

The dominant force in semantic segmentation has been CNN. Network topologies like ResNet arose and performed effectively in the dataset ImageNet on the basis of CNN's prior outstanding accomplishments in image classification initially [5]. To achieve end-to-end training, Jonathan Long et al. presented FCN in a study for image semantic segmentation [1]. Since then, FCN has established itself as the fundamental framework for semantic segmentation, and succeeding algorithms actually build upon it.

There have been various FCN-based algorithms since the invention of FCN, but SegNet was the first to use the encoder-decoder structure for semantic segmentation [6]. To achieve end-to-end pixel-level image segmentation, SegNet is an encoder-decoder symmetric structure built on the semantic segmentation problem of FCN. U-net, which was first created to address the issue of tiny datasets for medical picture segmentation, also makes use of the encoder-decoder structure. In order to provide end-to-end training with better outcomes using fewer photos, u-net employs data enhancement techniques.

In order to control the perceptual field without changing the size of the feature map [7], DeepLab proposes the idea of atrous convolution to address the issue that the encoder-decoder structure needs to produce good results while requiring large computation. This concept makes it easier to extract

multi-scale information. In order to increase the segmentation boundary accuracy, Chen et al. also introduced DeepLab V3+ based on null convolution with encoder-decoder architecture [8]. This version further combines the underlying features with the high-level features.

In addition to the CNN-based semantic segmentation techniques. On the basis of this self-attentive paradigm, the Vision Transformer (ViT) was proposed in 2017, which was based on the Transformer model. The first representative model of ViT-based semantic segmentation, SEgementation TRansformer (SETR) [9], suggests changing the current design of semantic segmentation models by substituting CNN encoders with encoders of pure Transformer structure. Additionally, the semantic segmentation tool swin-Transformer offers a superior answer to the issue of exact semantic segmentation that is more flexible and comprehensive [10].

The purpose of this paper is to sort out the inner logic of the development of some encoder-decoder based algorithms after the emergence of the FCN model. What's more, since many excellent algorithms have emerged in the field of medical image segmentation. This paper is to give a brief introduction based on the recent emergence of excellent Transformer based articles. This paper firstly introduces the FCN and the U-net, both of which are excellent networks for image segmentation in the early days of deep learning. Since many subsequent models are based on FCNs, this paper then analyses FCNs based on the encoder-decoder structure, followed by an analysis of Atruos Convolution applied to FCNs. These two approaches are the mainstream ideas for improving FCNs. Further, due to the recent widespread use of the Transformer model in the field of semantic segmentation, this paper briefly introduces several excellent algorithms applying the Transformer.

2. Method

2.1. Models

With the introduction of network architectures like VGG and Resnet, based on CNNs, have obtained excellent results in ImageNet and have made significant advancements in picture categorization. The perceptual domain is larger and it is possible to learn more abstract information in the deeper convolutional layers. The classification performance is aided by the fact that these abstract properties are less affected by the object's size, location, and orientation. These abstract qualities can be effective at classifying the various item classes that might be seen in an image. At the image level is image classification. However, because CNN loses picture features during convolution and pooling, the feature map size rapidly decreases, making it impossible for it to accurately segment an object or identify the precise object to which each pixel belongs. Fully Convolutional Networks (FCN) were suggested by Jonathan Long et al. as a solution to this issue for image semantic segmentation. Since then, FCN has established itself as the fundamental framework for semantic segmentation, and succeeding algorithms actually build upon it.

Some fully connected layers are added at the network's end for general classification CNNs like VGG and Resnet, and after softmax, the category probability data can be acquired. This fully-connected method is not relevant to image segmentation since this probability information is 1-dimensional, i.e., it can only identify the category of the entire image, not the category of each individual pixel point. In order to produce a 2-dimensional feature map and solve the segmentation problem, FCN suggests replacing the next few full joins with convolution. Softmax will then be used to retrieve the classification information for each pixel.

For encoder and decoder problems applied in segmentation tasks. The job of the encoder is to obtain the feature map of the input image using neural network learning after receiving the input image; the decoder then gradually implements the segmentation process by categorizing each pixel. Most often, the network structures used for classification tasks are used as the basis for the encoder structures in segmentation tasks, which are typically quite comparable. This has the benefit of being able to use the weight parameters of the classification network that were developed by training on a large database to improve performance through migration learning. Therefore, the effectiveness of a segmentation network based on codec structures is substantially determined by the difference in

decoders. SegNet's encoder and decoder structures match up exactly, with a decoder having the same spatial dimension and number of channels as its corresponding encoder. Both have 13 convolutional layers for the basic SegNet structure, with the encoder's convolutional layers matching the first 13 convolutional layers in the VGG16 network structure.

The initial authors who took part in the ISBI Challenge proposed the segmentation network U-Net, which can learn from a relatively short training set (about 30 images.) In order to make better use of the available annotated samples, U-Net suggests a network and training technique that depends on a strong usage of data augmentation. A systolic path that captures context and a symmetric extension path that permits exact localization make up the design. This network surpasses prior top techniques on the ISBI problem of segmenting neural structures in electron microscopy stacks and can be trained end-to-end from a small number of pictures. Because medical tasks typically have limited data availability and models that use tiny quantities of data for learning require hours, U-Net is well-liked for segmenting medical images. Unexpectedly, U-Net not only satisfies all the criteria but also outperforms other approaches. The primary factor is that U-Net employs splicing, a completely different feature fusion strategy than FCN.

Two crucial processes are required in a pixel-level prediction problem like semantic segmentation: first, convolution or pooling is used to decrease the picture size while extending the perceptual field, and second, upsampling is used to increase the image size. A very serious issue arises when downsampling an image using convolution or pooling, though, as visual features are lost and information about small objects cannot be recreated. A novel convolution method called atrous deconvolution addresses the issue of information loss and image resolution decrease brought on by downsampling in the context of image semantic segmentation. Atrous Deconvolution makes it possible to achieve a broader perceptual field for a given size of convolution kernel by introducing the Dilation Rate parameter. Therefore, given the same field size, it is possible to convolve with fewer parameters than standard convolution.

Transformer is a technique that was put forth in 2017 for use in Natural Language Processing (NLP), focusing on sequences as opposed to pixels. The ViT writers made an attempt to immediately apply the standard Transformer to the picture with very minor modifications, motivated by the success of the Transformer extension in NLP [11]. To achieve this, the image is divided into small blocks, and the Transformer is given a series of linear embeddings of these blocks as input. In NLP applications, image chunks are treated similarly to tokens (words) in order to train the model for supervised image classification. By introducing SETR (Segmentation Transformer), which turns semantic segmentation into a sequence-to-sequence prediction problem and does away with the necessity for models to learn local-to-global features by lowering resolution, Zheng et al. created a new perspective for semantic segmentation approaches. The idea of picture chunks (patches) is introduced, drawing on the ViT model. The two-dimensional image is first chunked, with each block of the chunked image being flattened into a one-dimensional vector. Each vector is then transformed by a linear projection, and a position encoding is added to incorporate the position information of the sequence.

With linear computational complexity of the input image size, Swin-Transfomer computational difficulty is significantly reduced [10]. Swin Transformer builds a hierarchical Transformer by gradually merging picture pieces as depth grows. A moving window self-attentive model is what swin Transformer suggests. Due to a reduction in computation from a squared picture size to a linear connection and an increase in model inference speed, Swin Transformer is able to achieve nearly worldwide attention. As a result, a multi-scale feature extraction of the input image is provided, expanding the perceptual field of the next window attention operation on the original image.

2.2. Dataset

CamVid. The Cambridge-driving Labeled Video Database (CamVid), the first collection of movies semantically tagged with target categories, is a publicly accessible dataset of urban road

scenes from the University of Cambridge. The data sample from the CamVid dataset is shown in Figure 1.

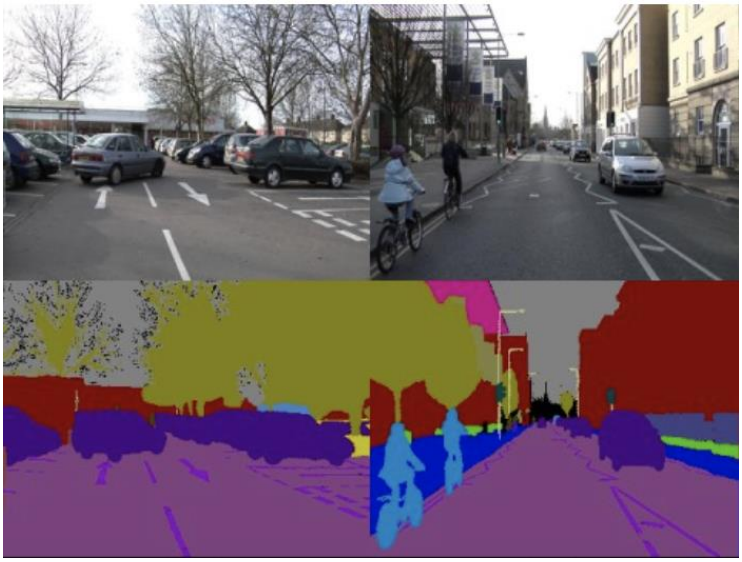

Figure 1. Sample images on the CamVid dataset.

COCO. Microsoft's MS COCO (Microsoft Common Objects in Context) collection is full of diverse image categories, and the majority of the images it contains are from intricate indoor and outdoor settings that are frequently utilized for image identification and semantic segmentation tasks. The collection includes 2500000 instances of objects, 328000 photos, and 80 categories. The COCO dataset's data sample is depicted in Figure 2.

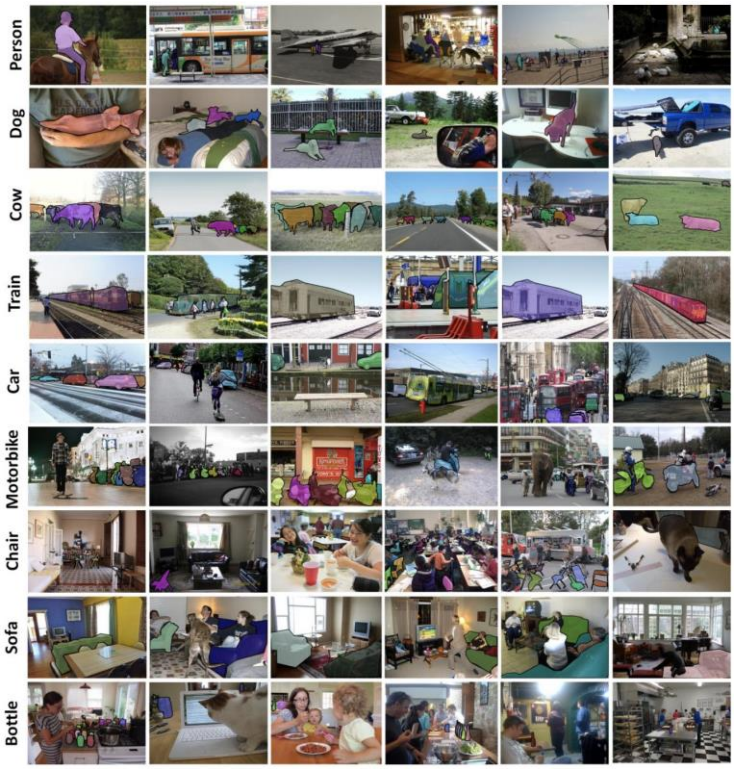

Figure 2. Sample images on the COCO dataset.

PASCAL VOC. From 2005 to 2012, it changed from being an international competition for specific detection tasks to producing a variety of high-quality data, the majority of which are being utilized in PASCAL VOC 2012. The dataset includes humans, animals, cars, indoor objects, and other types in a total of 21 (including context). The data sample from the PASCAL VOC dataset is shown in Figure 3.

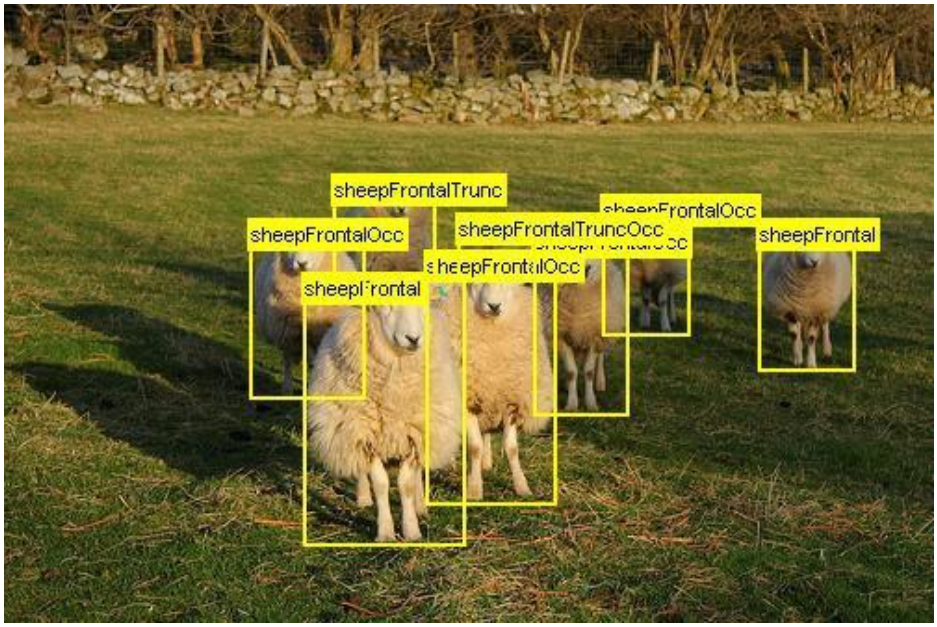

Figure 3. Sample images on the PASCAL VOC dataset.

Cityscapes. A sizable dataset primarily used for research of urban street scenes big collection of pictures from 50 cities with various backdrops, eras, and scene arrangements. For some of the photographs in the dataset, there are semantic and instance annotations. For some of the photos in the dataset, including fine-annotated photographs, there are semantic and instance annotations. For some of the photos in the collection, which includes around 5,000 finely and 20,000 coarsely tagged images, semantic and instance annotations are provided. The data sample from the Cityscapes dataset is shown in Figure 4.

Figure 4. Sample images on the Cityscapes dataset.

PASCAL Context. A PASCAL VOC dataset expansion that includes 10,103 semantically labeled images across 540 classes. The dataset contains numerous classes, but many of them are sparse. As a result, it is usual practice to utilize the 59 classes that occur most frequently as semantic labels when evaluating the effectiveness of semantic segmentation algorithms. Figure 5 shows the data sample of the PASCAL Context dataset.

Figure 5. Sample images on the PASCAL Context dataset.

ADE20k. ADE20k has over 25,000 images (20ktrain, 2k val, 3ktest) that are densely annotated with open dictionary tag sets. Figure 6 shows the data sample of the ADE20k dataset.

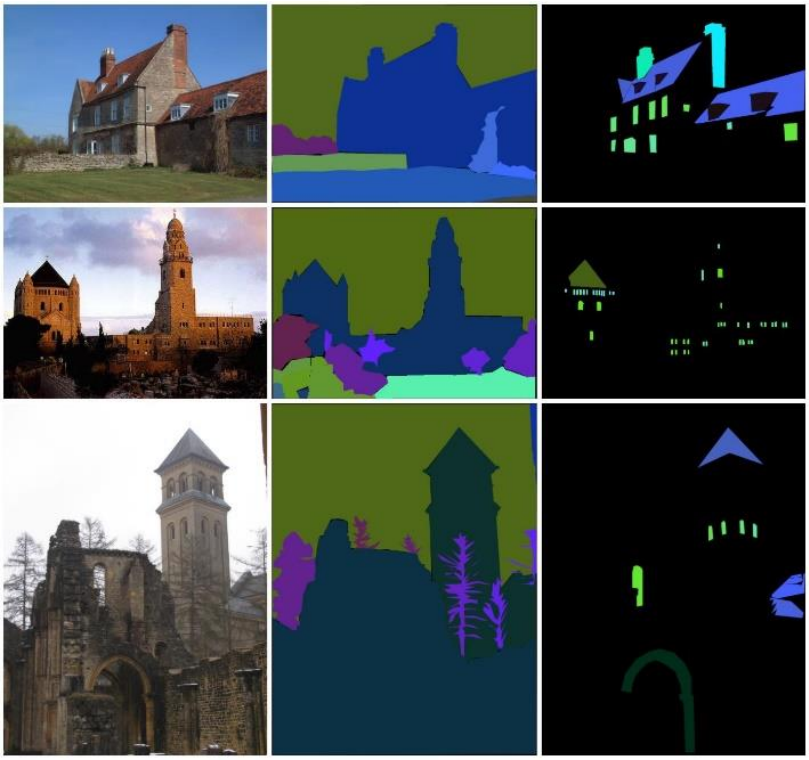

Figure 6. Sample images on the ADE20k dataset.

2.3. Measurement

Mean Intersection over Union (mIoU) and Pixel Accuracy are the two evaluation measures in semantic segmentation that are most frequently utilized (PA). IoU is the ratio of the number of pixels in the common region of the target mask and the predicted mask to the total number of pixels in both. mIoU is the average number of each category for IoU. The PA indicator shows the proportion of accurate predictions across all pixel categories to all pixels.

3. Results and discussion

Table 1. Comparison of experimental results of semantic segmentation methods

Algorithms	Basic Networks	Datasets	mIoU/%
FCN	VGG-16	CituSpaces test ADE20K	65.3 39.91
SegNet	FCN	CitySpace test ADE20K	57.0 21.64
Deeplab-CRF	ResNet	CityScapes test PASCAL VOC 2012 test	70.4 79.7
DeepLab V2	ResNet	PASCAL Context	45.7
SETR	Transformer	ADE20K CityScapes	50.28 82.15
swin-Transformer	Transformer	ADE20K	53.5

Because the feature fusion method employed in FCN had superior results in semantic segmentation networks, it can be seen from the comparison in Table 1 that FCN, as an early suggestion for a basic network, and the FCN-based approach also received better results on the CituSpaces test, ADE20K dataset. Since mIoU values primarily reflect accurate values and SegNet is quick in end-to-end real-time segmentation projects, no improvement can be seen from mIoU values alone, despite the fact that SegNet's results are relatively poor in terms of mIoU values. Instead, the main improvement of SegNet is its ability to improve image boundary segmentation. As a result, Atrous Convolution is successfully applied to ResNet, boosting the output size significantly while retaining the resolution without using more processing resources, demonstrating that DeepLab's innovation is a considerable gain in accuracy. As a result, the feature response is denser, which enhances the detail of the restored original image. In fact, further updates of DeepLab, such DeepLab V3+, have increased accuracy even more. As a pioneer in semantic segmentation, SETR obtained a mIoU value of 50.28 on the ADE20K dataset, the first time a method has ever done so. This is an exceptional outcome. This is built upon and improved upon by the swin-Transformer.

4. Conclusions

CNN-based semantic segmentation, including FCN and FCN-based networks, essentially uses the Convlution plus Deconvlution model, followed by multi-scale feature fusion. FCN uses features to add up point by point, while U-net uses the feature stitching method, and finally obtains the pixel-level segment map. While SegNet adopts an encoder-decoder fully symmetrical structure and uses the index of pooling layer positions during downsampling to facilitate the recovery of image information by upsampling, improving the effect of image boundary segmentation, Atrous Convulution optimizes the downsampling process and achieves better results. Contrarily, the

Transformer-based SETR converts the image into a sequence and applies it to the Transformer, which does little to enhance the outcomes. On the other hand, the swin-Transformer gets superior results because it is much more versatile when applying the Transformer to the image segmentation. Transformer-based approaches can fill the vacuum left by CNN-based methods' poor use of global information, and Transformer-based semantic segmentation methods have also recently become widely used. However, Transformer also has limits because its use is now primarily restricted to NLP. The Transformer has a large computational cost, and there are still certain issues that haven't been entirely resolved, such the recognition rate of various video tasks, which has to be fixed. All of these are awaiting the Transformer's resolution.

References

[1] Long J. et al. Fully Convolutional Networks for Semantic Segmentation [R], Nov. 2014, doi: 10.48550/arXiv.1411.4038.

[2] Ronneberger O, et al. U-net: Convolutional networks for biomedical image segmentation [C], in International Conference on Medical image computing and computer-assisted intervention, 2015, 234–241.

[3] Lecun Y. et al. Gradient-based learning applied to document recognition [C], Proceedings of the IEEE, 1998, 86(11): 2278–2324

[4] Vaswani A. et al., Attention is all you need [J], Advances in neural information processing systems,30, 2017.

[5] Huang G. Weinberger, Densely connected convolutional networks [C], in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 4700–4708.

[6] Badrinarayanan, V. et al. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation [J], IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481–2495

[7] He K. et al. Deep Residual Learning for Image Recognition [C], in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, 770–778

[8] Chen L C. et al. Encoder-decoder with atrous separable convolution for semantic image segmentation [C], in Proceedings of the European conference on computer vision (ECCV), 2018, 801–818.

[9] Zheng S. et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers [C], in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, 6881–6890.

[10] Liu Z. et al., Swin transformer: Hierarchical vision transformer using shifted windows [C], in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 10012–10022.

[11] Dosovitskiy A. et al., An image is worth 16x16 words: Transformers for image recognition at scale [R], arXiv preprint arXiv:2010.11929, 2020.