Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230889

A survey on the method of image segmentation based on deep learning

Zengbin Zhu

Science, The University of Melbourne, Melbourne, Victoria, 3010, Australia

jacobzhu2020@163.com

Abstract. Image segmentation is a widely used technology, such as autonomous driving, smart factories, smart cities, computer vision, medical image segmentation, robot perception, and augmented reality. The success of Convolutional Neural Networks (CNNs) has contributed greatly to the field of recent computer vision, including image segmentation. This article have conducted a review of some of the most important recent literature in the field of semantic segmentation based on deep learning (DL), which is divided into CNNs, Fully Convolutional Models, Encoder-Decoder Based Models, Pyramid Network Based Models, R-CNN Based Models. At the end of the article, the main features of these models are discussed and future directions for development are proposed.

Keywords: Image Segmentation, Computer Vision, CNNs.

1. Introduction

Many computer vision systems include semantic segmentation of images as an essential part of their design. it is important in numerous fields, including the medical field, smart factories, autonomous driving, smart cities, robotics perception, augmented reality, etc. Image segmentation can be divided into segmentation of a class of objects (semantic segmentation), segmentation of a single object (instance segmentation) or both (panoramic segmentation). Image segmentation predicts and labels all pixels of an image, which is usually more difficult and demanding than target detection and target recognition. The development of computer vision has led to the development of many image segmentation algorithms, from the earliest methods, such as K-means clustering and region growing, to more cutting-edge algorithms, such as graph cuts.

DL-based models have made significant progress in recent years, which have attracted widespread attention in the field. Reviewing the different types of DL-based image segmentation literature, this paper focuses mainly on the following types of literature: the main DNN architectures used in computer vision and important segmentation methods based on DL. In the following paragraphs this paper follows: Section 2 describes the main DNN architectures used in computer vision. Section 3 describes the significant segmentation methods based on DL. Section 4 describes a summary of the features of significant DL-based segmentation methods. Section 5 A future outlook is presented and a summary is provided.

2. Major DNN architectures

DNN architectures that are commonly used in computer vision are discussed in this section. A Convolutional Neural Network (CNN) and a Fully Convolutional Model (FCM) are introduced in this paper.

2.1. Convolutional Neural Networks (CNNs)

A major source of inspiration for artificial neural networks was Hubel and Wiesel's research on the cat brain's visual cortex, which established the idea of receptive fields [1]. The first CNN was developed by Fukushima in 1980, relying on Hubel and Wiesel's visual cortex receptive field model as the base for his "Neocognitron" and “learning without a teacher" convolutional neural layer [2]. Then, Based on backpropagation training and temporal receptive fields, Waibel et al. proposed CNNs for recognizing phonemes [3]. By fusing the backpropagation algorithm with a weight-sharing convolutional neural layer, LeCun created a practical CNN and successfully introduced the CNN to the U.S. Post Office's handwritten character recognition system for the first time [4] . LeCun's further enhanced the accuracy of handwritten character recognition, allowing for improved accuracy in recording personal and business checks, but it did not recognize large images [5].

Typically, CNNs have three different kinds of layers:

1.Convolutional layer, which determines a kernel of weights with the intention of identifying certain features in the data by computing a combination of the input values using a discrete convolution operation.

2. Nonlinear layers provide nonlinear interactions between inputs and outputs by applying an activation function to feature maps.

3. By using the pooling operation, a pooling layer gets the feature maps as input and outputs a certain statistical index (mean, max, etc.) of the input. The process is usually carried out to reduce spatial resolution.

The two areas of weight sharing and local connectivity are where CNNs excelled in comparison to other neural networks.

The computational benefit of CNNs is the sharing of weights among all the receptive fields in a layer, which network model is simplified and fewer weights are used. In general, dealing with high-dimensional pictures is more effective because it can take the image directly into account as the input and avoid the complicated feature process of the traditional approach. Layers of neurons receive weighted inputs from certain neuronal units in the previous layer, which explains their local connectivity.

Despite CNN's appealing features, their use on a broad scale for high-resolution photos has remained prohibitively costly [6]. Particularly in medical imaging, a pixel's class label is predicted using the local region surrounding it when localization over large regional areas [7]. Some of the most popular CNN designs, like AlexNet, VGGNet, and ResNet, as well as specialized techniques, such in the YOLO or U-net models, can be used to address these challenges [8],[9],[10].

2.2. Fully Convolutional Models

Using deep learning, Long et al. developed a Fully Convolutional Network (FCN) for semantic segmentation challenges [11]. It produces an input image-sized segmentation map by using only convolutional layers. In this study, the authors substituted all completely connected layers with convolutional layers, resulting in a heat map of class presence instead of classification results for popular CNN models, such as AlexNet, VGGNet, and GoogLeNet. [8][9][12]. In order to provide precise and fine-grained segmentation, the model mixes semantic information with appearance information by using skip connections , which include up-sampling and combining feature-rich lower layers with final prediction layers. Numerous segmentation issues have been addressed with FCNs, including interactive image segmentation [13], instance-aware semantic segmentation [14], and the segmentation of brain tumors [15]. The FCN has some disadvantages, including high computational costs for inference in real-time, inefficient incorporation of information about global context, and difficulty generalizing to three-dimensional images. The FCN's restrictions have been circumvented by a number of researchers. In an end-to-end architecture known as ParseNet, Liu et al. enhanced the FCN's global features by boosting features at each site and averaging feature for one layer [16]. The authors then extracted global contextual information from the entire image using global average pooling to produce a context vector. A ParseNet and FCN are same up until the convolutional feature map extraction [17]. The two feature maps are stitched together to form a new feature map of the same size as the original after normalizing the context vector and leaving it unpooled.

3. Some important DL based segmentation models

This section also provides a review of some important segmentation methods based on DL. This article introduce them into three major categories.

3.1. Encoder-Decoder Based Models

DeConvNet is made up of a multilayer de-convolutional network that uses deconvolution and unpooling to reconstruct the map of pixel-wise class probabilities with original activation size and forecast segmentation masks, and an encoder that uses convolutional layers that are identical to the VGG 16-layer network with the exception of the final classification layer [18]. SegNet features an encoder-decoder architecture with an additional layer of classification is applied pixel-by-pixel, much like the deconvolution network [19]. According to the topological structure, VGG16's basic 13 convolutional layers constitute the encoder network, and the decoder network similarly includes 13 deconvolutional layers. SegNet's primary novelty is that it executes non-linear starting sampling in the decoder using indexes obtained from the appropriate maximalist steps of the encoder feature mapping, resulting in a high-resolution sparse feature mapping. Due to the loss of resolution during the encoding process, an encoder-decoder models have the disadvantage of providing less precise spatial and semantic representations. During HRNet's processing, high-resolution representations are maintained by using parallel high-to-low resolution convolution streams and switching resolutions frequently [20] .

U-shaped network called U-Net, which tackles the issue of images localization of CNN, was presented for effectively segmenting biological microscope images [21]. A U-Net architecture consists of two components: an extraction of features through contracting, which is a standard CNN with no fully connected layers, only layers with maximum pooling, and an expanding path that is symmetric and allows for precise features reconstruction using convolutional layers and upsampling layers. Precise localization is promoted by reconstructing through concatenation using the corresponding feature map from the contracting path. The U-Net training technique enables learning from a small number of annotated inputs by data augmentation.

3.2. Pyramid Network Based Models

There are numerous neural network topologies that use the concept of multiscale analysis in image processing. An example of a famous model of this type is the Feature Pyramid Network (FPN), which was created by Lin et al. and is used for segmentation and object recognition. [22]. Deep CNNs' built-in multiscale, pyramidal hierarchy was utilized to generate feature pyramids for a minimal additional cost. FPN consists of lateral connections, a top-down pathway, and a bottom-up pathway to integrate low and high resolution characteristics. The output of each step is then created by processing the concatenated feature maps via convolution. Finally, the top-down route provides a prediction to find an object at each level. The Pyramid Scene Parsing Network (PSPN), created by Zhao et al., is a multiscale network that uses global contextual information to better learn how to represent a scene for segmentation [23]. In order to extract features, residual networks with dilated networks (ResNet) are used to extract several patterns from the input picture. The pyramid pooling module uses these feature maps as input to identify patterns of sub-regions with various locations and scales. Feature maps are pooled using four distinct scales with bin sizes of 1 × 1, 2 × 2, 3 × 3 and 6 × 6, a pyramid level corresponds to each of them, and the dimensions are reduced by 1 × 1 convolution layers. It is necessary to combine local and global contextual information by up-sampling the outputs of the convolution layers and concatenating them with the original feature maps. A convolutional layer is applied to the outputs once more to create the pixel-wise predictions.

3.3. R-CNN Based Models

R-CNN, which conduct object detection using a two-stage network, used CNN based classifier to each proposal after extracting region proposals using a selective search approach, and its expansions have been effective in the applications of object detection [24]. In the faster R-CNN architecture, a CNN-based region proposal network (RPN) that makes bounding box proposals is used[25]. To determine the bounding box coordinates and object class, a RoIPool layer calculates features from these proposals after the RPN extracts the Region of Interest (RoI). Mask R-CNN has two stages. During the first stage, to retain the spatial position, RoIAlign rather than RoIPool was employed by author[26]. The related classes, the bounding box coordinates, and the binary mask for each unique RoI are all simultaneously predicted in the second step. Three branches of the Mask R-CNN are jointly trained to predict class, bounding box, and segmentation mask for instances inside RoI. The Mask R-CNN and FPN models serve as the base for an algorithm proposed by Liu et al. calls for the formation of a Path Aggregation Network (PANet) [27]. FPN was used as the backbone for PANet's feature extraction. To enhance the propagation of lower-layer features through the network, a novel augmented bottom-up approach is utilized. The previous stage's feature maps are used as input in each level of this third pathway, then a convolutional layer is applied to them. Features from the augmented bottom-up approach are aggregated through using an adaptive feature pooling method. A Multitask Network Cascades (MNC) for instance segmentation task was created by Dai et al. and comprises of three networks for instance differentiation, estimating masks, and object classification [28]. In addition to using a cascaded structure and region proposal networks (RPN) for improved instance segmentation, these networks also employed a cascaded structure to share their convolutional features.

4. Discussion

Table 1 lists some key characteristics of some mile storm state-of-the-art model to provide a clear picture of the effectiveness of each model. Capture local and global information, multi-scale features, deep and shallow feature information, maintain high-resolution representation, less data required for training, faster speed, etc. are the main factors for model success.

Table 1: Some key characteristics of different state-of-the-art semantic segmentation models.

		Pros	Cons
Fully Convolutional Models	FCN		Costly for real-time inference Not account for global context Information not easily generalizable to 3D images.
Encoder-Decoder Based Models	ParseNet	Extra global context Smoother than that of an FCN End-to-end architecture Global feature or global context information	
	DeConvNet	Segmentation based on deconvolution Pixel-wise prediction Edge-box to generate region proposal Two stages of training for simple examples and more challenging examples End to end trainable	Missing of fine-grained image information

Pyramid Network Based Models	HRNet	Maintains high-resolution representations	The calculation cost of the model is too large
	U-Net	Data augmentation to learn effectively from very few annotated images. End to end trainable Inference time for testing was less than 1 sec per image Outputs up-sampled	
	PSPN	Capture both local and global End to end training Used pyramid pooling module for aggregating multi-scale features	
R-CNN Based Models	Mask R-CNN	Simultaneously generating a high-quality segmentation mask for each instance Used RoIAlign instead of RoIPool	Computationally expensive alignment procedures
	PANet	New augmented bottom-up pathway improving the propagation of lower-layer features FPN is used as Backbone network Adaptive feature pooling layer is introduced	
	MNC	Cascaded structure to share their convolutional features. End to end trainable Convolutional feature sharing leads to reduction of test time of 360 ms/image	

5. Conclusion

DL-based models are often referred to as "black boxes," and they are often complex and difficult to interpret, which is an obstacle to improving and approaching reliable AI. The inability to implement reliable AI means that users cannot delegate high-risk things to the AI to manage. This article review and analyze the main features of important architectures and methods, and also summarize the characteristics of important segmentation methods, and finally this article presents some of the main issues currently facing the field and an outlook for the future. At present, it seems that the perception mechanisms of human and AI are very different. Some aggressive changes to the input image can have a malicious effect on the prediction results without the knowledge of humans. How to improve AI or let AI evolve itself to deal with malicious attacks? How to protect personal privacy in case of AI misuse? DL-based models can only cope with types that have been trained in advance. The quality of the training information used to train these models limits its development, for example, in the medical field, where the quality of the training sample annotation plays an important role. Also, these training data are expensive. How to train with a few training samples? Most researchers currently focus only on using metrics to evaluate the accuracy of a model. However, when evaluating an image segmentation model, one should consider many other factors, such as visual quality, speed, and storage requirements.

Acknowledgment

This is to acknowledge and thank all of the organizations and individuals helped me for completing this research. I would like to thank University of Wisconsin – Madison for providing me with the resources available to complete the research, which allows me to collect the cases for writing this paper. I would also like to thank Professor Vipul G from Carnegie Mellon University for giving me some insight on the development of neural networks and suggested the idea of tracing this

development within a specific field of artificial intelligence, which leads to the completion of this paper.

Reference

[1] Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal Of Physiology, 160(1), 106-154. doi: 10.1113/jphysiol.1962.sp006837

[2] Fukushima, K., & Miyake, S. (1982). Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. Competition And Cooperation In Neural Nets, 267-285. doi: 10.1007/978-3-642-46466-9_18

[3] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. IEEE transactions on acoustics, speech, and signal processing, 37(3), 328-339.

[4] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

[5] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[6] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[7] Ciresan, D., Giusti, A., Gambardella, L., & Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. Advances in neural information processing systems, 25.

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.

[9] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[11] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[12] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[13] Hu, Y., Soltoggio, A., Lock, R., & Carter, S. (2019). A fully convolutional two-stream fusion network for interactive image segmentation. Neural Networks, 109, 31-42.

[14] Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In Competition and cooperation in neural nets (pp. 267-285). Springer, Berlin, Heidelberg.

[15] Man, D., & Vision, A. (1982). A computational investigation into the human representation and processing of visual information. WH San Francisco: Freeman and Company, San Francisco.

[16] Liu, W., Rabinovich, A., & Berg, A. C. (2015). Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579.

[17] Visin, F., Kastner, K., Cho, K., Matteucci, M., Courville, A., & Bengio, Y. (2015). Renet: A recurrent neural network based alternative to convolutional networks. arXiv preprint arXiv:1505.00393.

[18] Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE international conference on computer vision (pp. 1520-1528).).

[19] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.

[20] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 43(10), 3349-3364.

[21] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[22] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).

[23] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

[24] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

[25] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

[26] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[27] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8759-8768).

[28] Dai, J., He, K., & Sun, J. (2016). Instance-aware semantic segmentation via multi-task network cascades. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3150-3158).