Image Semantic Segmentation Based on Deep Learning: A Review

Gao Yun

Abstract— **Image semantic segmentation is an important research direction in the field of computer vision, and it is also one of the key technologies for intelligent systems to understand natural scenes. Image semantic segmentation technology has broad application prospects in the fields of auto-driving, intelligent monitoring systems, and unmanned aerial vehicle. In recent years, with the development and application of deep learning technology, image semantic segmentation methods based on deep learning have begun to emerge in large numbers. This paper mainly classifies and summarizes the image semantic segmentation techniques based on deep learning. According to the characteristics of the network architecture, the method of image semantic segmentation based on deep learning is divided into FCN-based method, Encoder-Decoder-based method, RNN-based method, and GAN-based method. We analyze and introduce the representative algorithms of each method, summarize the basic ideas, advantages and disadvantages of each method, systematically explain the contribution of deep learning to image semantic segmentation, and finally predict and analyze possible future development trends and research in this field direction.**

Index Terms—**Image Semantic Segmentation, FCN, GAN, Deep Learning.**

I. INTRODUCTION

Image semantic segmentation is a kind of technology that enables the computer to automatically classify the pixels in the image according to the categories of objects or background. The so-called semantics refers to the label of objects in the image. Compared with the traditional image segmentation, image semantic segmentation adds certain semantic information to each pixel in the image. It can get the information that the image itself needs to express according to the texture, color, scene and other high-level semantic features of the image. So it has more practical value than image segmentation.

In recent years, deep learning technology develops rapidly, and image semantic segmentation methods based on deep learning technology emerge in endlessly. The most representative core technology of deep learning is convolutional Neural Network(CNN) [1]. Because of its high-efficiency learning performance and good application effect, CNN has become the key research field of researchers in various fields around the world. In addition to CNN, Recurrent Neural Network(RNN) [2] is especially suitable for processing sequence related information because of its recursive processing of historical information and modeling of historical memory. It is often used by researchers to capture context information in images. In addition, the General Adversarial Network(GAN) [3] is a new network architecture in the field of deep learning. Because of its unique characteristics of adversarial learning, it can generate realistic images, and has good applicability in the task of image semantic segmentation, so it has been paid more and more attention. In general, deep learning technology uses deep network model to learn image features, which promotes the development of image semantic segmentation technology.

The purpose of this article is to introduce the methods of image semantic segmentation which is based on deep learning. The main content is to review the research and development process of image semantic segmentation technology, focusing on the theory and methods of image semantic segmentation technology based on deep learning, and then analyze the advantages and disadvantages of these algorithm. Finally, the development of the technology is summarized and prospected. We believe that this work is necessary and useful for researchers who are interested in image semantic segmentation based on deep learning.

Manuscript received December 18, 2019

Gao Yun, School of Computer Science and Technology, Tiangong University, Tianjin, 300387, China

II. BASIC THEORY

A. *Deep Learning*

The concept of deep learning was first proposed by Hinton et al. [4] in 2006, and it is a branch of machine learning. Deep learning technology can effectively extract low-level, intermediate-level, and high-level semantic information in images, and combine the results of the pixel classification to improves the segmentation accuracy. Currently, mainstream deep learning models include CNN, RNN and GAN.

The basic structure of CNN consists of input layer, convolutional layer, pooling layer, fully connected layer and output layer. The input image is subjected to multiple CNN convolution operations and pooling operations for feature extraction, gradually transforming low-level rough features into high-level features, high-level features are then classified after the fully connected layer. CNN is particularly suitable for processing image data due to its special network structure, and has high invariance to image deformation such as translation and scaling.

The RNN structure consists of a series of repetitive neural network module sequences, each element in the sequence performs similar tasks, and continuous information between image contexts can be reasonably utilized. Currently, representative RNNs include traditional RNN sequences Model, long short-term memory neural network (LSTM) [5] and gated recurrent unit (GRU) [6].

The GAN structure consists of a generator network and a discriminator network. The basic idea is: let the model obtain a large number of training samples for learning in the training library, the generator network continuously generates artificial samples, and the discriminator network continuously judge for the generated artificial samples. During training, the two sets of networks continuously improve their ability to generate samples and distinguish between true and false during the process of adversarial learning.

B. *Semantic Segmentation*

Semantic Segmentation is a branch of image segmentation and a very important task in computer vision. The difference between image semantic segmentation and image segmentation is that image segmentation refers to dividing the image into different regions, such as dividing the image into foreground and background parts, while image semantic segmentation refers to assigning one label to each pixel, such as separating humans from animals in an image. Furthermore, if you want to represent each person in the image with a different label, it becomes instance segmentation.

III. METHODS

A. *FCN Based Method*

In 2014, Long et al. [7] proposed the fully convolutional network (FCN) which is compatible with images of any size and performs image semantic segmentation in a supervised learning manner. FCN is based on the VGG-16 network, it replace the fully connected layers with convolutional layers, use the skip layer method to combine the feature maps generated by the intermediate convolutional layers, and then use the bilinear interpolation algorithm to perform the up-sampling operate to convert rough segmentation results into fine segmentation results. FCN uses a cross-layer method, which simultaneously considers both global semantic information and local position information, and can recover the category to which a pixel belongs from abstract features. Classification further extends to pixel-level classification, successfully transforming the network originally used for image classification into a network for image segmentation.

FCN can recover the category of pixels during the segmentation process, which has greatly promoted the development of image semantic segmentation. However, there are still some problems. For example, FCN does not consider pixel-to-pixel relationships when performing image semantic segmentation, lacks spatial consistency, and is not sensitive enough to the details in the image, resulting in poor segmentation. To solve this problem, Chen et al. [8] added a fully connected conditional random field (FCCRF) at the end of the FCN to optimize the boundary of the rough segmentation image and use atrous convolution to expand the receptive filed of feature map. This is called DeepLab network. The DeepLab network sends the image to the FCN combined with the hole algorithm for processing to obtain a rough feature map, and then uses the BI algorithm to up-sample the output of the FCN. A rough segmented image is obtained; then, the FCCRF is used to perform structured prediction on the rough segmented image, and the pixels in the image are modeled and solved to smooth the edges of the rough segmented image; finally, a complete image semantic segmentation result is obtained.

B. *Encoder-Decoder Based Method*

The encoder-decoder structure is a mechanism for image semantic parsing using a symmetric network. The essence of the encoder-decoder structure is to use the encoder formed by operations such as convolution and pooling in deep learning technology to encode captured data. The pixel position information and image features are then parsed by using a decoder composed of operations such as deconvolution or un-pooling to restore the spatial dimensions of the image and the pixel position information.

The UNet network proposed in [9] is based on the encoder-decoder structure. UNet performs a down-sampling operation during the encoding process to gradually reduce the resolution of the feature map, and performs an up-sampling operation during the decoding process to gradually restore the object details and image resolution. Badrinarayanan et al. [10] aims to solve the image semantic segmentation problem of self-driving and intelligent robots, and proposed the SegNet-Basic network. The SegNet-Basic network calculates the classification of each pixel based on the prior probability, which is also similar to a symmetric structure network in the encoding-decoding process. On the left side of the network is an encoder composed of a full convolutional network, which performs down-sampling through convolution, pooling and other operations; on the right is a decoder composed of a deconvolution network. The transposition convolution and pooling operations are used for up-sampling. To solve the problem that the prior probability cannot give the confidence of the classification results, Badrinarayanan et al. [11] proposed the Bayesian-SegNet network based on the SegNet-Basic network. A dropout layer is added behind the layer, which can effectively prevent the weight from overfitting and enhance the learning ability of the network; at the same time, it also introduces the Bayesian network and Gaussian processes, and calculate pixel categories based on the posterior probability, so that the network can more reasonably simulate the event probability during the image semantic segmentation process.

C. *RNN-Based Method*

The RNN can process historical information recursively and model the characteristics of historical memory. RNN is used to capture image context information and global features during image segmentation. RNN can not only learn information at the current moment, but also rely on previous sequence information. It is beneficial for modeling global content and saving historical information, and promotes the use of image context information. When RNN-based method is used for image semantic segmentation, the RNN layer is embedded into CNN, and the local spatial features of the image are extracted in the convolution layer, and the convolution layer extracts pixel sequence features. Its general processing flow is as follows:

1) The input image is processed by CNN to obtain a feature map;
2) The feature map is input into the RNN to obtain image context information. The network layer in the RNN is

used to serialize pixels and analyze the dependency relationship of each pixel to obtain global semantic features.

3) Use the deconvolution layer for up-sampling to get the segmentation result.

Pinheiro et al. [12] refers to RNN's cyclic thought and applies the generalized RNN to ISS field. Visin et al. [13] comprehensively utilizes the advantages of CNN and RNN, and uses RNN's derived network ReNet to process image data, and proposes the ReSeg network. ReSeg uses four traditional RNN sequence models to replace the convolution and pooling operations of the convolution layer in CNN, and cuts the image in horizontal and vertical directions respectively and makes space for it based on the model of interdependency, the input image gets the local features of the image after passing through VGG-16 network, and then sends the feature map to the ReNet to extract the global features and context information of the image. Finally, the up-sampling layer composed of deconvolution network is used to recover the resolution of the feature map, and then output the segmentation result. At the same time, ReSeg also uses the gated recurrent unit to balance the memory occupation rate and calculate the load capacity. In paper [14], four RNN subnets with different directions are used to complete the task of image annotation: the input image is divided into multiple non-overlapping windows and sent into four independent and different direction LSTM memory blocks, local and global features are captured without other additional conditions.

D. GAN-Based Method

When using GAN-based methods for image semantic segmentation, segmentation networks such as FCN and SegNet are generally used as the generator. The input image is processed by the generator to obtain the predicted segmented image. The predicted segmented image is used as artificial samples and the ground truth image as a real sample are send into the discriminator, the discriminator learns the difference between real samples and artificial samples, and was training through the adversarial learning. After judging whether the output sample data is true or false, the internal feedback mechanism will adjust the generator network and the discriminator network. After several iterations of training, the segmentation accuracy of the generator network and the discriminator network's discriminating ability will continue to improve.

In 2016, Luc et al. [15] first introduced GAN to the image semantic segmentation field and proposed a new method for image segmentation. The original image was transformed into a segmentation result in a segmentation network composed of CNN. The segmentation result was judged to be true or false after being input into the adversarial network. Two groups of networks conduct adversarial learning and compete with each other. After iterative training, the segmentation accuracy of the segmented network is gradually improved. Hoffman et al. [16] proposed a domain adaptive framework for image semantic segmentation which is based on FCN, combined the idea of GAN with domain adaptation, shared the source domain and the target domain, and optimized the target loss function to reduce the impact of global offset and specific offset. Koziński et al. [17] uses GAN to implement parameter regularization of segmented networks, using Annotated images are used to train segmentation networks (generator networks). Souly et al. [18] uses conditional generative adversarial network (CGAN) [19] to generate artificial samples for adversarial training.

The GAN model has the ability to continuously generate data and discriminate between true and false, to a certain extent, it can reduce the problems caused by CNN, FCN and other networks in the process of image semantic segmentation. At the same time, GAN introduces discriminators to solve the problem of inconsistent data domain distribution. The adversarial learning is used to approximate the unsolvable loss function, which has a good segmentation effect [20]. However, the optimization process of the GAN model is unstable, and it is easy to collapse to a saddle point during training. Its interpretability need to be improved when processing large-scale image data [21].

IV. CONCLUSION

Nowadays, deep learning technology has been widely used in the field of image semantic segmentation. This paper mainly classifies and summarizes the classic methods of image semantic segmentation based on deep learning. According to different network architectures, the method of image semantic segmentation based on deep learning is divided into FCN-based image semantic segmentation method, Encoder-Decoder-based image semantic segmentation method, RNN-based image semantic segmentation method, and GAN-based image semantic segmentation method. The representative algorithms of each method are analyzed and compared, and the technical characteristics of each method are summarized. In the near future, image semantic segmentation is still the research focus in the field of computer vision. We believe that the existing problems of image semantic segmentation will be solved by new methods.

REFERENCES

[1] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[2] Pearlmutter B A. Gradient calculations for dynamic recurrent neural networks: A survey[J]. IEEE Transactions on Neural networks, 1995, 6(5): 1212-1228.

[3] Goodfellow I. NIPS 2016 tutorial: Generative adversarial networks[J]. arXiv preprint arXiv:1701.00160, 2016.

[4] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786): 504-507.

[5] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[6] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

[7] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[8] Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.

[9] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[10] Badrinarayanan V, Handa A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling[J]. arXiv preprint arXiv:1505.07293, 2015.
[11] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.
[12] Pinheiro P H O, Collobert R. Recurrent convolutional neural networks for scene labeling[C]//31st International Conference on Machine Learning (ICML). 2014 (CONF).
[13] Visin F, Ciccone M, Romero A, et al. Reseg: A recurrent neural network-based model for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016: 41-48.
[14] Byeon W, Breuel T M, Raue F, et al. Scene labeling with lstm recurrent neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3547-3555.
[15] Luc P, Couprie C, Chintala S, et al. Semantic segmentation using adversarial networks[J]. arXiv preprint arXiv:1611.08408, 2016.
[16] Hoffman J, Wang D, Yu F, et al. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation[J]. arXiv preprint arXiv:1612.02649, 2016.
[17] Koziński M, Simon L, Jurie F. An adversarial regularisation for semi-supervised training of structured output neural networks[J]. arXiv preprint arXiv:1702.02382, 2017.
[18] Souly N, Spampinato C, Shah M. Semi and weakly supervised semantic segmentation using generative adversarial network[J]. arXiv preprint arXiv:1703.09695, 2017.
[19] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.
[20] Im D J, Kim C D, Jiang H, et al. Generating images with recurrent adversarial networks[J]. arXiv preprint arXiv:1602.05110, 2016.
[21] Denton E L, Chintala S, Fergus R. Deep generative image models using a laplacian pyramid of adversarial networks[C]//Advances in neural information processing systems. 2015: 1486-1494.