Image semantic segmentation using deep learning technique

Yifei Fan

School of advanced technology, Xi'an Jiaotong-liverpool University Xi'an,111 Ren ai Road Suzhou Industrial Park Suzhou China.

Yifei.Fan18@student.xjtlu.edu.cn

Abstract. With the deepening research on image understanding in many application fields, including auto drive system, unmanned aerial vehicle (UAV) landing point judgment, virtual reality wearable devices, etc., computer vision and machine learning researchers are paying more and more attention to image semantic segmentation (ISS). In this paper, according to the different region generation algorithms, the regional classification image semantic segmentation methods are classified into the candidate region method and the segmentation mask method, according to different learning methods, the image semantic segmentation methods based on super pixels are divided into fully supervised learning method and weakly supervised learning method. The typical algorithms in these various categories are summarized and compared. In addition, this paper also systematically expounds the role of DL technology in the field of ISS, and discusses the main challenges and future development prospects in this field.

Keywords: Image semantic segmentation, Region generation algorithms, Learning methods, DL technology.

1. Introduction

ISS is the key direction of computer vision. Among them, semantic segmentation can be applied to still 2D images, videos, even 3D or volume data. Image semantic segmentation needs to pave the way for higher-level semantic segmentation applications. Image semantic segmentation involves related knowledge in many fields and it is an interdisciplinary subject with a wide prospect for the development. Ohta et al [1] proposed ISS firstly and its definition is to assign an image with many pixel-level semantic labels. The different parts of image with semantic information added by ISS could be classified through different semantic information annotations, and can express the texture, scene or other semantic information of images. Usually, the image information is complex, and the semantic segmentation task is easily affected by the complex background information of the image. With the great advancement of DL, the DLISS recently has also improved a lot. Among them, the regional classification ISS method (RCISS) divides the original image into different candidate regions by combining the traditional image segmentation method with semantic classification on each image block by using deep neural network (DNN). Finally, the final segmentation result is obtained by integrating all the image annotations. On the other hand, the pixel classification image semantic segmentation (PCISS) can capture more abundant image features by classifying each pixel. To a certain extent, PCISS solves the shortcomings of RCISS that the segmentation accuracy is not high enough and it is easy to lose the global features of the image.

In 2006, Hinton [2] firstly proposed the concept of DL. The mainstream DL models developed so far include CNN, RNN and GAN. In 2013, researchers [3] tried to use deep learning technology for SS of indoor scenes. However, the image segmentation in early stage took a lot of time and it was difficult to capture the key overall features of image, and the accuracy of semantic segmentation was not satisfactory. After that, researchers thought of split images into different candidate regions, and then using DL to classify these regions, that is, the RCISS that will be specifically introduced in this paper. In 2014, researchers [4] tried to capture and classify visual features on each candidate region of the image by combining candidate region generation algorithm and CNN. After that, researchers put more focus on improving the segmentation performance and optimizing the candidate region generation algorithm. On the other hand, researchers found that segmentation mask is also an effective scheme of RCISS. After that, research on this aspect has been continuously carried out. As image segmentation is an important step of RCISS and it is impossible to completely avoid the performance and time loss caused by image segmentation. Researchers try to find a solution for image semantic segmentation directly on pixels, hoping to increase the overall fit of image segmentation, that is PCISS.

The first part of this article introduces the relevant background of DL and ISS and the development history from the early ISS to the current development of ISS technology, including application of ISS with DL technology. In the second section, RCISS and PCISS are introduced and summarized in detail, including some representative algorithms of each subclass, their basic ideas, progress and shortcomings. The third section summarizes the main viewpoints and opinions of this paper.

2. Main body

2.1. Regional classification image semantic segmentation

RCISS combines the image segmentation algorithm with the deep learning network, and it can be divided into two steps. The procedure of RCISS is displayed in Figure 1.

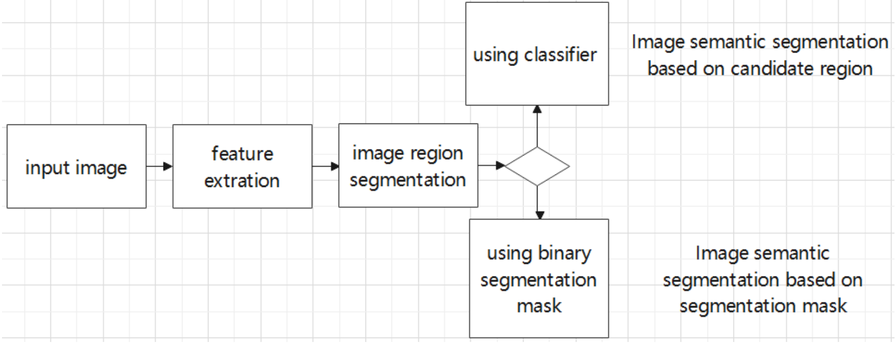
using classifier
Image semantic segmentation based on candidate region
input image
feature extration
image region segmentation
using binary segmentation mask
Image semantic segmentation based on segmentation mask

Figure 1. The processing flow of RCISS.

Firstly, the underlying features are extracted from the image, then the image is segmented according to the extracted features to obtain the candidate regions that may contain the target object. Secondly, the target candidate regions generated in the first step are classified by different classification algorithms. Researchers have relatively mature solutions in feature extraction and image candidate region segmentation. The main difference lies in the different classification methods in the second step. The candidate region ISS uses classifier for semantic classification, while the segmentation mask ISS uses binary segmentation mask. It will be described in detail below.

2.2. Candidate region image semantic segmentation

This kind of method firstly generates and filters candidate regions through specific algorithms. Each candidate region has a certain possibility to contain the target object that needs to be classified. Secondly, CNN is used to candidate region feature extraction, which are prepared for sending to the classifier for image block classification and segmentation result output using the extracted features.

The semantic segmentation performance of this kind of method is mainly affected by two aspects. Because the feature extraction of CNN is closely related to the generation quality of candidate regions, the generation algorithm quality of candidate regions affects the classification speed and performance of this kind of method. On the other hand, the selection of classifiers also has an important impact. Exploring higher performance classifier algorithms is the focus of researchers in this area. The following is an introduction to several important algorithms of this kind of methods:

In 2014, based on CNN researchers [4] proposed a regional CNN(RCNN). The processing flow of RCNN is shown in the figure 2. RCNN combines candidate region generation and image classification, and realizes image semantic segmentation by applying deep learning technology to target detection. The specific processing flow of RCNN is as follows: firstly, 2000 candidate regions are generated by selective search method and resized to facilitate feature extraction; Secondly, CNN extracts the potential features and the extracted features are passed to SVM classifier; Finally, SVM and non-maximum suppression are correspondingly taken to classify the image blocks and correct the segmentation results.

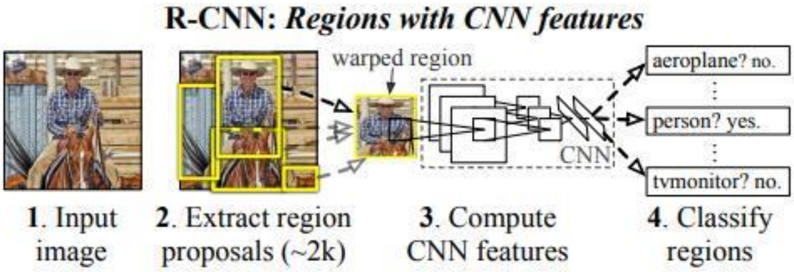

Figure 2. The Overview of RCNN Procedure.

As an early image semantic segmentation method, the comprehensive performance of RCNN is not enough satisfactory because candidate regions have a great influence on the model performance, and it needs to crop and scale the candidate regions which causes the original image is prone to deformation in the segmentation process. Although RCNN has these disadvantages, researchers have developed more excellent variants based on RCNN, including Fast-RCNN, Faster-RCNN and Mask-RCNN. To solve the problem that RCNN needs to convolute each candidate region which results in too many repeated convolution operations affecting the model speed, Fast-RCNN [5] avoids repeated convolution operations between overlapping candidate regions by convoluting the entire image. In addition, Fast-RCNN replaces the resize process in the RCNN model with the ROI-pooling layer, which solves the problem of image deformation to a certain extent. For the classifier, it turns to use SoftMax classifier instead of SVM. These changes significantly improve the model performance which the training speed of Fast-RCNN is nine times as fast as that of RCNN, and the test classification speed is nearly 200 times as fast as that of RCNN. Compared with RCNN, Fast-RCNN still does not solve the problem of long time-consuming image segmentation caused by selective search.

After that, Faster-RCNN [6] further saves the time required to determine the candidate regions by using the RPN instead of the selective search method, thus further improving the model performance. In 2017, Mask-RCNN [7] was expanded on the basis of Fast- RCNN to realize instance level image semantic segmentation. The method uses the merged deep and shallow features to meet the requirements of classification and detection simultaneously, specifically, unlike single-scale feature map adopted by Fast-RCNN, Mask-RCNN efficiently improves the accuracy by constructing a multi-scale feature pyramid from multi-scale input images, while avoiding the increase of time. In addition to RCNN and its variants, simultaneous detection and segmentation (SDS) [8] improves the segmentation performance by using multi-scale combinatorial grouping (MCG) algorithm to extract features from candidate regions and regional foreground for joint training. The multi-scale path aggregation (MPA) method [9] obtains multi-scale feature maps by using sliding windows of different sizes to convolute the images. This method effectively avoids the inherent problems of RCISS that easily depends on the segmentation quality of candidate regions and neglects the global features of the images by synthesizing the diversified image features in these feature maps.

2.2.1. Segmentation mask image semantic segmentation. Different from the selection and use of classifiers in the candidate region ISS method, the main difference of this kind of method is that the segmented candidate regions generate class-unknown binary segmentation masks, and then the model is trained according to multiple generated segmentation masks, and the refined segmentation masks can be regenerated to get the final result. The literature [10] proposes a deep mask model that can perform the task of instance segmentation. This model regards classification as a two-class classification problem based on massive data. In addition, it completes the segmentation task and foreground target recognition based on different branches of the same grid structure. In addition to the use of segmentation mask, optimizing the process of image features extraction can also improve the model performance. On the basis of Deep-Mask, a Sharp-Mask model is proposed in the literature [11], which integrates the features of low-pixel-level and high-object-level, and generates a feature map with better performance. In addition, this model solves the problem that the deep mask only uses a simple forward network and cannot generate a refined target mask, which generates a refined target mask by inputting the coarse target mask generated by the deep mask into the refinement module.

2.3. Pixel classification image semantic segmentation

Unlike RCISS, which first segments the region that may contain the target object, the pixel classification ISS method directly classifies at the pixel level. Specifically, PCISS directly completes the classification task of each pixel in the image through model training, starting from the original image, the finely labeled image, the weakly labeled image and other massive data. This method fundamentally avoids the loss of accuracy and performance caused by RCISS when performing region segmentation. In addition, according to the different levels of image annotation, PCISS can be further divided into a fully supervised learning ISS method that takes pixel-level annotations processed manually as training samples and a weakly supervised learning ISS method that takes weak annotation data as training samples. These two sub methods are described in detail below:

2.3.1. Fully supervised learning image semantic segmentation. ISS aims to label images in the pixel level. After the model is trained by using the labeled image, it can be generalized to semantic segmentation for unknown images. Manual labeling of each pixel can make the model obtain rich image details, so semantic segmentation network training can obtain higher training efficiency and more accurate segmentation results. This is also the main reason why fully supervised learning ISS method is the mainstream of ISS, although the generation of truth value annotation, especially manual annotation, is quite time-consuming and labor-consuming. The following describes the relevant explorations made by researchers in this field so far:

In 2014, the literature [12] proposed a full convolution network (FCN, figure 3) that can accept arbitrary input images without limiting their size. In terms of the implementation details of FCN which shows in the figure3, this network is different from the classical CNN that outputs the feature vector predicted by the probability through the full connection layer after the convolution layer. Instead, it uses the convolution layer to replace the last full connection layer of CNN, and up sample the final feature map with the deconvolution layer to restore the image to the original size to generate a marked semantic segmentation map. As an early fully supervised ISS method, FCN can restore abstract features to semantically segmented images, and transform the CNN network into a network capable of ISS, which promotes the development of ISS.

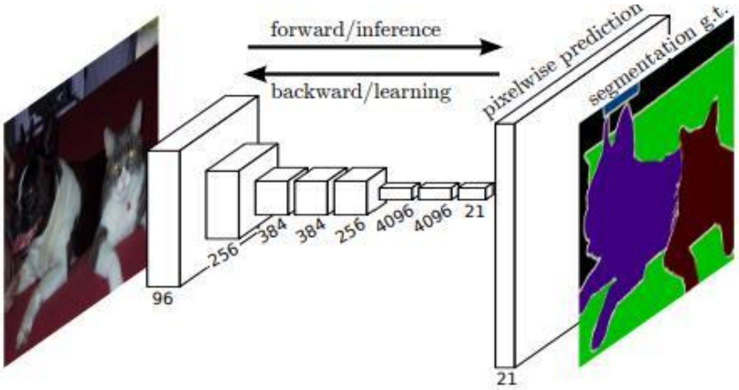

Figure 3. Fully Convolutional Network Architecture.

Based on the problem that the image will lose pixel information in the pooling process of FCN network and it is difficult to recover in the up-sampling process, researchers use different methods to optimize the model performance. Based on the FCN network, DeepLab network [13] replaces the ordinary convolution of FCN with the void convolution that can increase the receptive field. This convolution method can avoid the problem of image resolution reduction caused by the slow growth of receptive field. In addition, the full connection conditional random field performs post-processing optimization on the obtained segmentation map. After that, the DeepLab series of methods have been optimized based on the previous generation of DeepLab.

In view of the disadvantage that traditional convolution needs massive data to provide the adaptability to the geometric deformation of objects, researchers have proposed relevant methods including deformable [14]. These methods optimize the traditional convolution structure through variable convolution, use more flexible receptive fields to adapt the target shape and improve the convolution efficiency. In addition, RefineNet model uses shallow perfect features to strengthen the semantic features, and merges the feature information of multiple aspects in the image to obtain more refined segmentation results.

2.3.2. Weakly supervised learning image semantic segmentation. In view of the fact that the image semantic segmentation model needs to train a certain amount of images to strengthen the generalization ability, and the high cost of truth value annotation images required by fully supervised learning, the cost of database annotation is difficult for some researchers to accept, and they turn to weak supervision to reduce the cost of image annotation. Specifically, weak supervised learning can be classified according to different weak annotation data. The main categories include image-based annotation methods, bounding boxes-based annotation methods, and graffiti-based annotation methods. Image level annotation is the simplest annotation data. The production of this kind of annotation only needs to refer to the category information of the object in the name image and cannot distinguish different instances of the same category. The paper [15] proposes a method to extract the foreground pixel information through a typical encoder decoder structure and ignore the background information. Border level annotation can effectively distinguish the boundaries between different instances, and integrates semantic information and instance information. The researchers realize the training of neural network classifier under a given border level annotation through continuous iteration. Graffiti level annotation is a very convenient annotation method in interactive image segmentation. It marks objects, especially objects without definite shape, by any form of line segments. The ScribbleSup model combines FCN model, which inputs the training image automatically marked by GraphCut algorithm into FCN for training to obtain the final semantic segmentation result. Table 1 shows the method comparison in image semantic segmentation.

Table 1. Method comparison in image semantic segmentation.

Method category	Method subclass	Correlation algorithms	Brief introduction	Advantages and disadvantages
Regional Classification ISS (RCISS)	Candidate Region ISS	RCNN	Candidate regions were generated using Selective Search method and classified by SVM classifier	realizes image semantic segmentation, but comprehensive performance is not satisfactory
		Fast-RCNN	Convolute the whole image and replacing the resize process in RCNN with ROI-Pooling layer	Improve the model performance, but Selective Search method still consume a lot of time
		Faster-RCNN	Replace SS with RPN	Save the time required to determine the candidate region and further improve the model performance
		Mask-RCNN	Construct feature pyramid from multi-scale input image	Efficiently improve accuracy while avoiding time increase
		SDS	Joint training of extracted feature using MCG algorithm	Improve the model performance
		MPA	Convolute through sliding windows of different sizes	Avoid the problem of easily ignoring the global features of the image
	Segmentation Mask ISS	Deep Mask	Regard classification as an optimization problem based on massive data	Complete the segmentation task and foreground target recognition task by different branches
		Sharp Mask	Integrate the low-pixel-level features and the high-object-level features	Solve the problem that Deep-Mask only use a simple forward network and cannot generate a refined target mask

Table 1. (continued).

Pixel Classification ISS (PCISS)	Fully Supervised Learning ISS	FCN	Can accept input images without limiting their size, and use convolution layer to replace the last full connection layer of CNN	The transformation of CNN network used for image classification into a network capable of ISS has promoted the development of ISS
		DeepLab	Replace the ordinary convolution of FCN with the hole convolution that can increase the receptive field	Avoid image resolution reduction caused by slow growth of receptive field
		Deformable,	Optimize the traditional convolution structure by variable convolution	Better adapt the target shape and improve the convolution efficiency
		RefineNet	Use shallow perfect features to strengthen high-level semantic features	Obtain a more precise segmentation result
	Weakly Supervised Learning Image Semantic Segmentation	Image-based annotation	Only need to name the category information to which the object belongs in the image, and cannot distinguish different instances of the same category	Propose a method of extracting foreground pixel information and ignoring background information
		Bounding boxes-based annotation	Can effectively distinguish the boundaries between different instances, and integrate semantic information and instance information	Realize the training of neural network classifier with given frame level labeling
		Graffiti-based annotation	Mark objects with arbitrary line segments, especially those without definite shapes	ScribbleSup model inputs the training image automatically marked by GraphCut algorithm into FCN for training to get the final semantic segmentation result

3. Conclusion

To sum up, in this paper, according to different region generation algorithms, the semantic segmentation methods of region classified images are divided into candidate region methods and segmentation mask methods. According to different learning methods, these methods based on super pixels are divided into fully supervised learning methods and weakly supervised learning methods. The typical algorithms of these different classes are summarized and compared. According to the comparation of these methods which can be further classified into two subclasses: regional classification ISS method and pixel classification ISS method, which can be more detailed segmentation. Regional classification ISS method can be divided into candidate region ISS and segmentation mask ISS according to the difference of regional segmentation methods. Pixel classification ISS method can be divided into fully supervised learning method and weakly supervised learning method according to the difference of image annotation levels. This paper describes the detailed segmentation criteria of these subclasses of ISS methods, and summarizes the development trends and technological breakthroughs of the relevant methods in each subclass.

References

[1] Csurka G, Perronnin F. An efficient approach to semantic segmentation. 2011, Int. J. Com. Vis., 95(2): 198-212.

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

[3] Couprie C, Farabet C, Najman L, et al. Indoor semantic segmentation using depth information. 2013 arXiv preprint arXiv:1301.3572.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation, 2014 Conf. Com. Vis. Pat. Rec. 580-587.

[5] Girshick R. Fast r-cnn 2015 Conf. Com. Vis: 1440-1448.

[6] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. 2015 Adv. Neu. Inf. Pro. Sys., 28.

[7] He K, Gkioxari G, Dollár P, et al. Mask r-cnn. 2017 Conf. Com. Vis. Pat. Rec.: 2961-2969.

[8] Hariharan B, Arbeláez P, Girshick R, et al. Simultaneous detection and segmentation, 2014, Euro. Conf. Com. Vis.: 297-312.

[9] Liu S, Qi X, Shi J, et al. Multi-scale patch aggregation (mpa) for simultaneous detection and segmentation. 2016 Conf. Com. Vis. Pat. Rec.3141-3149.

[10] O Pinheiro P O, Collobert R, Dollár P. Learning to segment object candidates. Adv. Neu. Inf. Pro. Sys., 2015, 28.

[11] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation 2015 Conf. Com. Vis. Pat. Rec. 3431-3440.

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

[13] Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks. 2017 Conf. Com. Vis, 764-773.

[14] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation 2017 Conf. Com. Vis. Pat. Rec. 1925-1934.

[15] Saleh F, Aliakbarian M S, Salzmann M, et al. Built-in foreground/background prior for weakly-supervised semantic segmentation. 2016, Euro. Conf. Com. Vis. 413-432.