Research Progress of Image Segmentation

Jiyuan Ran *

Department of Computer Science, Huazhong University of Science and Technology, Wuhan, China

* Corresponding Author Email: u202215572@hust.edu.cn

Abstract. Image segmentation is an essential and challenging task in CV. It can be used in a massive scope including automatic driving, medical analysis and geological detection. In the early stage of the research, some researchers based on the geometry of the image, color space and other ways of semantic segmentation results are poor. Fortunately, with the flourishing of deep learning, great improvement has been gained in various deep learning-based image segmentation fields. At the same time, in the open world, semantic segmentation research is not limited to fixed categories or known categories, and some open-world semantic segmentation methods and reference semantic segmentation methods based on text description have also made great breakthroughs. This paper summarizes commonly used datasets for segmentation and introduces main ideas of popular algorithms in traditional, open-vocabulary and referring segmentation respectively. In the end, the important contribution of the paper is summarized and some suggestions regarding the future improvement of existing algorithms are given.

Keywords: Image segmentation; open-vocabulary segmentation; referring image segmentation.

1. Introduction

Semantic segmentation is a typical CV task in which computer ought to segment the target area pixel-wisely. The target is given as a class label, a description of spatial, semantic features, or a sequence of natural language, depending on different segments of the task. A well-trained semantic segmentation model can gain comprehensive understanding of each part of the image as well as the given instructions. Therefore, it can be applied to multiple areas related to scene understanding, including automatic driving, medical image analysis, remote sensing, geological detection, etc. With the help of semantic segmentation model, information can be detected quickly and effectively, which is especially meaningful in extreme conditions where human eyes are unfeasible to differentiate objects. Moreover, the digitalized result generated by the model can be shared and processed upon, which is vital in today's Big Data era.

Up to now, there are three segments of image segmenting tasks in general, posing higher and higher requirements to the model. The first segment is traditional semantic or instance segmentation, which learns and predicts given limited label masks. Popular algorithms include Mask R-CNN and it derivatives, which extend faster R-CNN model by adding a paralleled branch to predict mask [1, 2]. The second segment is open-vocabulary semantic segmentation, whose models are required to behave well in front of unlimited arbitrary(zero-shot) classes that may not appear in the training set. SAM is one of the effective models, which applies encoder-decoder framework and transforms to learn comprehensive features [3]. The last segment is referring image segmentation, whose models gain a multimodal understanding of the image and offer mask in accord to rigorous natural language. In realization, the mask can either be predicted directly like mLSTM or generated after object detection block like LTS [4, 5]. But regardless of the steps, fusion of multimodal information is always a deciding essential. Armed with these algorithms, traditional segmenting models can be used to automatic driving, classifying obstacles, people, pavements, etc. since the class is fixed; open-vocabulary models can be used in geological detection system, dealing with trivial real-life images; referring image segmentation model can be used in security system, finding the targeting object only given the language description.

With the rapid improvement of image segmentation algorithm, model accuracy and efficiency variey hugely among different algorithms in various segments, posing challenges to real-life

applications. This paper centers on the main algorithms in each segment and the common datasets they are using. In section 2, datasets related to semantic segmentation are introduced. In section 3, main algorithms in traditional semantic, open-vocabulary and referring image segmentation are introduced. At last, the paper is summarized, and orientation of further study is given.

2. Datasets

2.1. COCO

COCO dataset, put forward by Microsoft in 2014, is used widely in segmentation tasks. It contains 91 different objects, 328,000 images and 2,500,000 labels. The dataset is designed for scene understanding tasks and usually used for precise segmentation tasks, since the images come from intricate environment. Furthermore, there are three derivatives of COCO, RefCOCO, RefCOCO+ and RefCOCOg. In these datasets, the bounding box for each instance are labeled with phrases, instead of a single class label. Those added referring descriptions are used for referring image segmentation. Therefore, COCO has a wide application in object detection, image generation and image description. More detailed information is shown in Table 1.

Table 1. Numbers of image, phrase and instance in RefCOCO, RefCOCO+ and RefCOCOg

Dataset	Image	Phrase	Instance
RefCOCO	19,994	142,209	50,000
RefCOCO+	19,992	141,564	49,856
RefCOCOg	26,711	85,474	54,822

2.2. VOC

PASCAL VOC 2011 and PASCAL VOC 2012 are also popular dataset, used in object detection and semantic segmentation. VOC 2011 and VOC 2012 have 21 classes, which contains 21 popular objects (including 1 background class) and 10 action classes, such as jumping, for more precise segmentation. While VOC 2011 does not provide pixel-wise labels for every image, VOC 2012 complements these labels therefore VOC 2012 is used more in segmentation tasks while VOC 2011 more in object detection task.

2.3. Data Engine

In zero-shot segmentation tasks, however, given data is inadequate since there is not a dataset that covers all the existing objects. Therefore, in these tasks, a semi-automatic data generator is designed to learn masks with insufficient data [3]. After a pre-trained stage involving human annotators, the data engine will output a final dataset SA-1B automatically, which includes more than 1 billion marks and 11 million images. SA-1B contains 400∗ more masks than the past segmentation datasets. The generating details is discussed in section 3.

3. Algorithms

This section is about some core algorithms in the 3 main segments of segmentation tasks. For each algorithm, aside from their basic working mechanism, their main contribution and inspiring ideas are highlighted while their trivial details are ignored in this section.

3.1. Traditional Semantic and Instance Segmentation

3.1.1 Mask R-CNN

As shown in Fig. 1, Mask R-CNN is a two stages method, which is feature extraction and proposal mapping separately. The first stage extracts feature from the original image and usually backboned with ResNet or FPN (feature pyramid network) [6, 7]. The second stage borrows the RPN (region proposal network) from faster R-CNN, where k anchors are generated from each pixel and classified

and regressed through its softmax layer [2]. With feature map and proposals prepared, RoIAlign layer projects the proposals onto the feature map and has its shape fixed at the same time. At last, every projected RoI is sent into 3 paralleled branches, which do classification (Dense and softmax), bounding box regression (Dense and linear regression) and mask generation (FCN) respectively [8].

The loss function is the sum of classification, bounding box regression and mask loss. The loss for classification and bounding box regression are similarly defined as faster R-CNN [2]. Note that only those positive anchors contribute to L_{box} and only the masks that correspond to the ground truth contribute to L_{mask} so that masks across classes do not compete [3]. The architecture of mask R-CNN is simple and effective, making it easy to generalize to multiple tasks and lay a solid foundation for its derivatives.

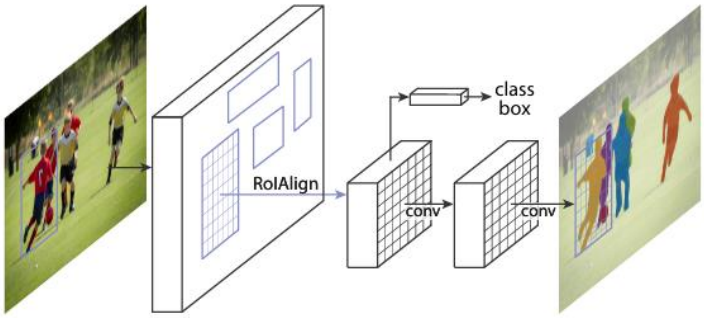


Fig. 1 Architecture of mask R-CNN [1]

3.1.2 Cascade Mask R-CNN

Cascade Mask R-CNN is a derivative of R-CNN family, aiming to have a more precise bounding box and RoI proposal [9]. With a better location of the RoI, the AP of mask can improve simultaneously. Since the naive thought of increasing IoU threshold is not working because of overfitting and mismatching, the Cascade R-CNN model increases the threshold in a step-by-step manner, resulting in a 2-4 percent increment of Accuarcy precision (AP).

As shown in Fig.2, 3 similar block (combination of H, C and B) is cascaded after the conv layer. The author shows that after every block the distribution of positive sample is getting more concentrated (variance is getting smaller) thus the noise becomes harder to exclude [9]. Therefore, in each block their H, C and B are different, and their threshold goes higher gradually to resample the RoI and elimiate the noise. It overcomes the overfitting problem because the number of samples increase as stage goes up as shown by the author [9]. It also helps mitigate the mismatch problem, since RoI is refined in both train and inference stage.

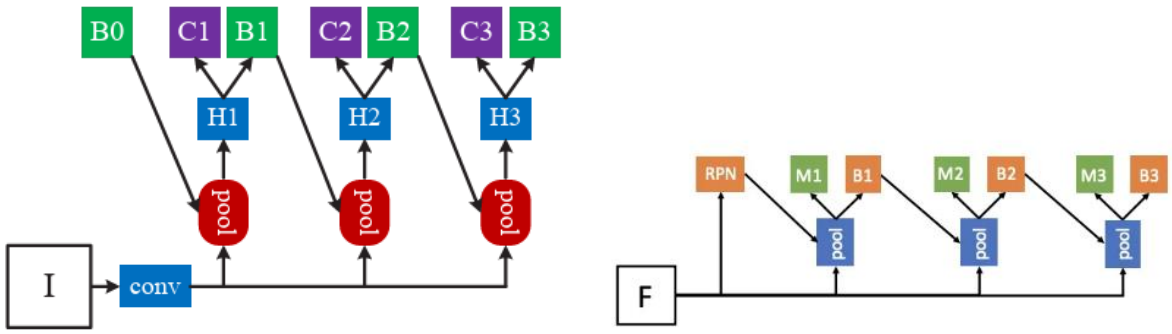


Fig. 2 Architecture of Cascade R-CNN (left) and Cascade Mask R-CNN (right) [9, 10]. "I" is input image, "F" feature map, "conv" backbone, "pool" RoI projection, "H" network head, "B" bounding box regression, "M" mask, and "C" classification.

3.2. Open-Vocabulary Semantic Segmentation

In open-vocabulary semantic segmentation, CNNs become less effective since the sliding window of convolution has an inductive bias on regional relationship and neglects the global property of the

whole image. Therefore, segment anything (SAM) is put forward to address the bias by applying encoder-decoder structure, transforming it to a sequence-to-sequence problem [3]. It proves to perform well given a massive amount of data (generated by its data engine) since it removes the local bias by seeing the whole picture in one look.

Similar to other encoder-decoder model, SAM has 3 parts: image encoder, prompt encoder, and mask decoder. First, the model minimally adapts MAE pre-trained Vision Transformer (ViT) as the image encoder. Second, given the segmentation prompt (as shown in Fig.3), the model encodes them in different ways. Sparse prompt like points and box are positional embedded; Dense prompt like mask is encoded through an FCN block; text prompt is encoded by CLIP's text encoder (refined and trained) [11]. At last, mask decoder combines the image and prompt embedding, and uses a transformer to decode the mask.

The other great contribution of SAM is its data generator: data engine. Its construction includes three stages: assisted-manual stage, semi-automatic stage and a fully automatic stage.

First, in assisted-manual stage, resembling active learning, the model is trained on open dataset while well-trained annotators modify the wrong predicted masks and put them back into the dataset. The procedure is repeated 6 times, resulting in 4.3M masks.

The semi-automatic stage is also in analogy with the refining step in active learning, where masks that the model shows low confident with are refined by annotators. The confident value is given by the assisted bounding box detector. Likewise, this procedure is repeated 5 times and output 5.9M masks.

In fully automatic stage, the model can be used to annotate mask for itself. Then, the resulting masks are judged by the IoU prediction module, where confident and stable masks (change the threshold near 0.5 and output a similar mask) are selected. After eliminating the replicates using non-maximum suppression, 1.1B masks remain, which is known as SA-1B. The data provides SAM with sufficient raw data, which is one of the biggest reasons why it can out-compute traditional CNN model.

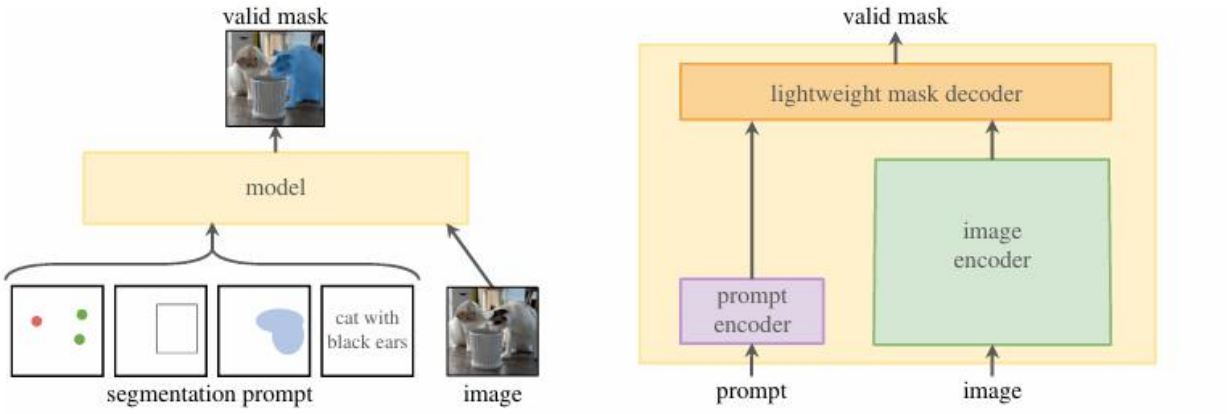


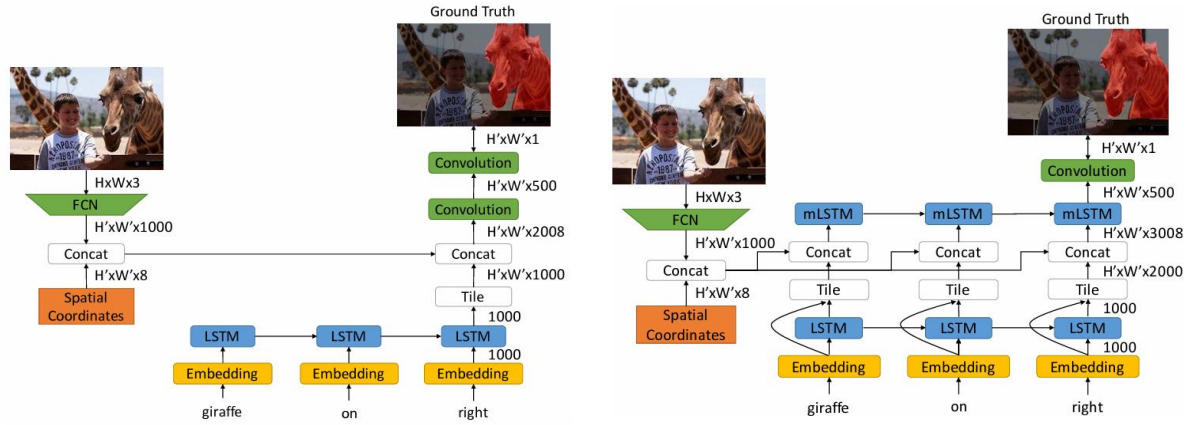
Fig. 3 Architecture of SAM [3]

3.3. Referring Image Segmentation

3.3.1 Recurrent Multimodal LSTM

Given LSTM, a language machine that has long-term memory, it's natural to have a baseline model shown in Fig.4 (a), where the model extracts the features from the image and uses the knowledge learning from LSTM to give a mask proposal [4, 12]. In this case, the model needs to memorize all the key information in order to locate the target area. However, mLSTM offers a sequential method of doing filtering the mask out of the image. Referring to Fig.4 (b), after feature map (included spatial information) is generated, every pixel on the map is concatenated with an identical language information generated from the current state of LSTM. Then, the multimodal LSTM is learned upon these all of the three information and therefore can forget or memorize information at once.

The whole process resembles applying LSTM mechanism to 1*1 convolution where language information, image semantic and spatial information can interact with each other more commonly, instead of only in the last step in the baseline model. This is the key to the LSTM model.



(a) (b)

Fig. 4 Comparison of the baseline model and recurrent multimodal LSTM [4]

3.3.2 Locate and Segment

Aside from generating the mask directly, the process can be broken down into 2 steps: (1) locate the target object with a bounding box; (2) segment the target [5]. More specifically, the model first extracts language and visual features using GRU and CNNs respectively and fuses them to get a cross-model heatmap. Then, the model filters the heatmap with the text features by using multi-head attention mechanism. At last, the model segment the target from the bounding box. The LTS is simple and interpretable but very powerful. It is promising to be the basis of further study.

4. Summary

In this work, 3 kinds of image segmenting tasks are summarized and for each task at least one SOTA algorithm is introduced. In traditional semantic segmentation, mask R-CNN and its derivatives form the baseline of two-stage algorithms. In this type of algorithm, proposal generation and mask prediction are done separately, and the feature map is shared throughout the network. The steps can be cascaded to perform a better mask. In open-vocabulary semantic segmentation, having sufficient data, combining image and prompt information, and fitting flexible prompts are 3 main factors to a successful model. SAM meets the 3 requirements respectively by designing an automatic data engine resembling active learning, applying encoder-decoder structure and transformers to CV and training the model recursively. In referring image segmentation, 2 different strategies are used. MLSTM predicts mask directly by sequentially applying LSTM mechanism to encoded multimodal information while LTS separately predicts bounding box and mask. Both of them emphasize the importance of the fusion of language and visual information in comprehending and segmenting. The datasets for different segmentation tasks are also briefly introduced in section 2.

However, there are still problems that remain. In small object segmenting, the model tends to focus more on large objects and ignore the small ones since its unbalanced distribution in the training sets. Further study can refine the model's backbone, data selection method and loss function to perform better in small objects. In addition, the model is required to respond quickly for automatic driving and military use. Thus, having a lightweight version of the existing model while keeping its precision high is another critical research orientation.

References

[1] He K, Gkioxari G, Dollár P, et al. Mask r-cnn. Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[2] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 2015, 28.

[3] Kirillov A, Mintun E, Ravi N, et al. Segment anything. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 4015-4026.dd.

[4] Liu C, Lin Z, Shen X, et al. Recurrent multimodal interaction for referring image segmentation. Proceedings of the IEEE international conference on computer vision. 2017: 1271-1280.

[5] Jing Y, Kong T, Wang W, et al. Locate then segment: A strong pipeline for referring image segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 9858-9867.

[6] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[7] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

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

[9] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162

[10] Chen K, Pang J, Wang J, et al. Hybrid task cascade for instance segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4974-4983.

[11] Radford A, Kim J W, Hallacy C, et al. Clip: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020, 2021.

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