Spatio-temporal Attention Network for Video Instance Segmentation

Xiaoyu Liu
Moku Lab
Alibaba Group
beilin.lxy@alibaba-inc.com

Haibing Ren
Moku Lab
Alibaba Group
haibing.rhb@alibaba-inc.com

Tingmeng Ye
Moku Lab
Alibaba Group
tingmeng.ytm@alibaba-inc.com

Abstract

In this paper, we propose a method named spatio-temporal attention network for video instance segmentation. The spatio-temporal attention network can estimate the global correlation map between the successive frames and transfers it to the attention map. Added with the attention information, the new features may enhance the response of the instance for pre-defined categories. Therefore, the detection, segmentation and tracking accuracy will be greatly improved. Experimental result shows that combined with MaskTrack R-CNN, it may improve the video instance segmentation accuracy from 0.293 to 0.400@Youtube VIS test dataset with a single model. Our method took the 6th place in the video instance segmentation track of the 2nd Large-scale Video Object Segmentation Challenge.

1. Introduction

Video object segmentation (VOS) is to segment the objects in all images of a video clip. With the increasing of computation power and development of the algorithms, video object segmentation becomes more and more popular in recent years. There are 3 scenarios in video object segmentation: semi-supervised VOS, interactive VOS, and unsupervised VOS.

The semi-supervised VOS is the most basic scenario. The masks of interested objects in the first frame are given and algorithms should segment them in the following video images.

In the interactive VOS, the masks in first frame are not given. Instead, some user interaction will be input to guide the segmentation of interested objects. And these objects should also be segmented in the following video images as same as semi-supervised VOS. The human interaction can be scribbles, bounding box or clicks [1,2].

Unsupervised VOS is also called video instance segmentation (VIS). This scenario is fully automatic video object segmentation. It is more difficult than semi-supervised and interactive VOS. In the unsupervised VOS, no masks or any user interaction are given. Actually there no any information about which objects should be tracked. The algorithms should analyze the image saliency, track and segment the salient objects of pre-defined categories in all the video images.

In the video instance segmentation track of 2nd Large-scale Video Object Segmentation Challenge, there are 40 pre-defined categories, including person, animals, vehicles, etc. All the instances should belong to these categories.

MaskTrack R-CNN[3] is an end-to-end method for video instance segmentation. It can detect, segment and track video instance simultaneously. But it doesn't use spatial information for object detection and segmentation. This paper proposes a method named spatio-temporal attention network for video instance segmentation. The spatio-temporal attention network can estimate the global correlation information between the successive frames and transfers it to the attention map. Experimental result shows that combined with MaskTrack R-CNN, it may improve the video instance segmentation accuracy from 0.293 to 0.400@Youtube VIS test dataset.

2. Related work

In semi-supervised VOS, video object mask propagation is the key technologies which aims to obtain the object region given the segmentation results of previous frames. Most mask propagation technologies can be classified into two categories: with or without online learning. The online learning-based methods fine-tune the model parameters specify for the object instances in the dataset, such as [1,4]. From early this year, some mask propagation methods without online learning appeared with very good performance, for example FEELVOS[5].

For interactive VOS, there are two core technologies interactive image object segmentation and mask propagation. In most papers, the two technologies are independent of each other. But in [6], interaction segmentation and mask propagation are conducted by two convolutional neural networks and the two networks are trained jointly to adapt to each other.

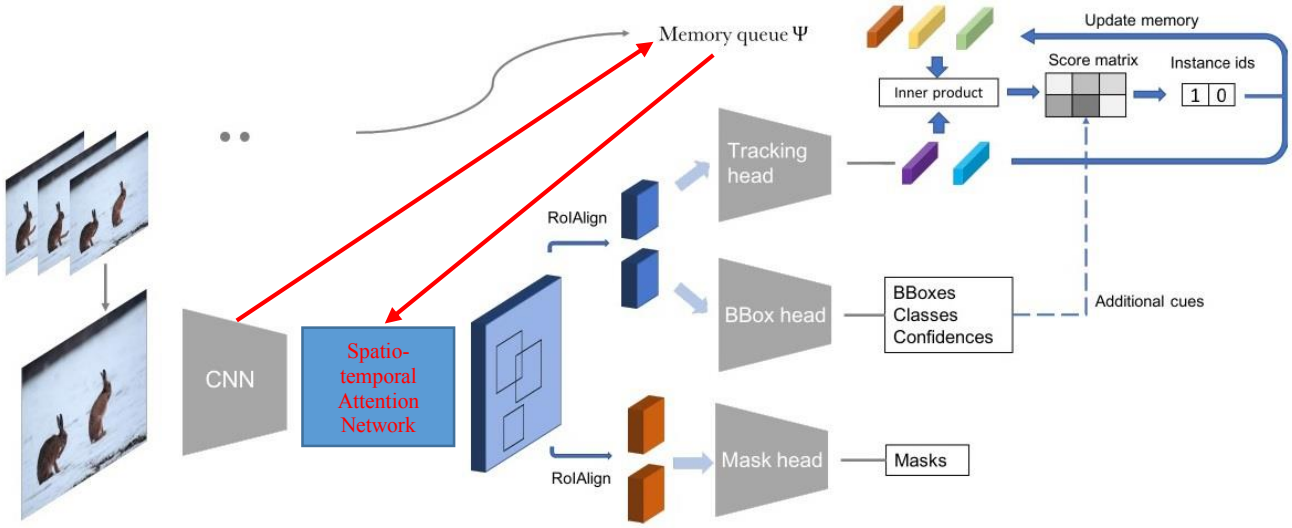

Figure 1. Our framework.

For video instance segmentation, there are two kinds of methods: non end-to-end methods and end-to-end methods. UnOVOST[7] is non end-to-end method. In UnOVOST, object proposal masks of each image are first generated with Mask R-CNN[8] method, then grouped as tracklets with each other with temporally consistent information. Finally, these tracklets are merged into long-term consistent object tracks using their temporal consistency and an appearance similarity metric. Those tracklets with long temporal track length and detection confidence scores are selected as final length. The solution-based method has several separate steps which can not be optimized together.

MaskTrack R-CNN is an end-to-end video instance segmentation method. It introduces a new tracking branch to Mask R-CNN to jointly perform the detection, segmentation and tracking tasks simultaneously. In MaskTrack R-CNN, spatial information is only used in the tracking head. For bounding box and mask estimation, the method is as same as Mask R-CNN which only utilizes current image information. Therefore, the accuracy is not so satisfying.

3. . Our approach

The framework of our approach is as Figure 1. The difference between our approach and MaskTrack R-CNN is that there is spatio-temporal attention network in our structure. It is located between the backbone and region proposal network.

It can utilize the spatial information to calculate the attention map which will enhance the response of the instance for pre-defined categories.

3.1. Spatio-temporal attention network

The network structure of spatio-temporal attention module is as Figure 2. The network has two inputs: x0 and x1. They are the image feature maps of previous frame and current frame. These feature maps are the output of backbone part. Each feature map has the dimension c*w*h where c is the channel number, w is the width and h is the height.

The left side of the network is the correlation map module. It is as the following function:

$$Cor(x0, x1) = S[Conv1(x0) \otimes Conv1(x1)] \quad (1)$$

Where x0 and x1 are the feature maps of previous and current frame. $Conv1$ is a shallow convolution network, $\otimes$ is matrix multiply operator and S is the softmax function. In the function, $Conv1(x0)$ and $Conv1(x1)$ has the dimension c1*w*h where c1 is the channel number. During the matrix multiply operation, x0 is reshaped to a matrix with the dimension (w*h)*c and x1 is reshaped to c* (w*h). So correlation map is a matrix of dimension (w*h)* (w*h).

The network structure inside the green box is the attention estimation module. With the correlation map, the attention map can be calculated via the attention estimation function $A(x0, x1)$:

$$A(x0, x1) = Conv2(x1) \otimes Cor(x0, x1) \quad (2)$$

Where $Cor(x0, x1)$ is the correlation map, $Conv2$ is another convolution network and $\otimes$ is matrix multiply operator. In this equation, the result of $Conv2(x1)$ is of dimension c*w*h. Before multiplication, it is reshaped to c*(w*h). As the correlation map has the dimension (w*h)* (w*h), the attention map is of the dimension c*(w*h).

The final output of spatio-temporal attention module is $STA(x0, x1)$. It can be described as the following equation:

$$STA(x0, x1) = A(x0, x1) * Gamma + x1 \quad (3)$$

where Gamma is a 1D coefficient and $A(x0, x1)$ is the attention estimation function. Here, Gamma is not a hyper-

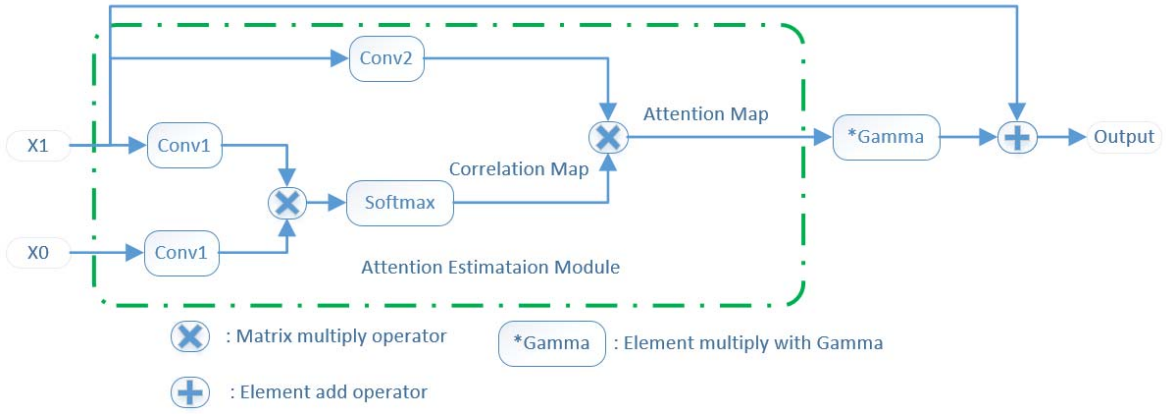

Figure 2. Spatio-temporal attention module

parameter and it should be achieved in training procedure.

3.2. Memory queue

Besides the instance tracking embedding, the new memory queue also stores the feature map of previous frame. The feature map will be used for attention calculation in next frame.

4. Experiment

With single model, our method took the 6th place in the video instance segmentation track of the 2nd Large-scale Video Object Segmentation Challenge. The accuracy is as following table:

mAP	AP50	AP75	AR1	AR10
0.400	0.578	0.449	0.396	0.452

Table 1. Accuracy on You-tube VIS test dataset.

Some of the result images are as follows:

Figure 4. Some segmentation result on test dataset.

5. Conclusion

MaskTrack R-CNN is an end-to-end method for simultaneous video instance detection, segmentation and tracking. In this paper, spatio-temporal attention network is proposed for utilize the previous frame information. Combined with MaskTrack R-CNN, It can improve the video instance detection, segmentation and tracking accuracy greatly.

References

[1] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, Thomas Huang. Deep GrabCut for Object Selection. arXiv preprint arXiv:1707.00243, 2017

[2] Zhuwen Li, Qifeng Chen, Vladlen Koltun. Interactive Image Segmentation With Latent Diversity. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 577-585

[3] Linjie Yang, Yuchen Fan, Ning Xu. Video Instance Segmentation. In arXiv preprint arXiv: 1905.04804, 2019

[4] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS: Proposal-generation, refinement and merging for video object segmentation. arXiv preprint arXiv:1807.09190, 2018.

[5] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam. FEELVOS: Fast End-to-End Embedding Learning for Video Object Segmentation. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[6] Seoung Wug Oh, Joon-Young Lee, Ning Xu, Seon Joo Kim. Fast User-Guided Video Object Segmentation by Deep Networks. CVPR 2018 Workshops of DAVIS Challenge on Video Object Segmentation.

[7] Idil Esen Zulfikar, Jonathon Luiten, Bastian Leibe. UnOVOST: Unsupervised Offline Video Object Segmentation and Tracking for the 2019 Unsupervised DAVIS Challenge. 2019 DAVIS challenge on video object segmentation - CVPR 2019 workshops.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross Girshick. Mask r-cnn. In ICCV, 2017. 2, 4, 6