Video Object Segmentation Based on Location and RoIAlign in Weight Modulated Multi-Scale Network

Wenqing Luo Yongzhao Zhan Qianling Wu
Jiangsu University School of Computer Science and Communication Engineering

优胜奖

Abstract

As a binary labeling problem, video object segmentation targets at segmenting one or more specific objects throughout a video sequence. Since 2016, many video object segmentation methods based on deep learning have been developed. Most of these methods use the ground truth segmentation mask of the first frame to fine-tune the subsequent segmentation frames of the video. Satisfactory segmentation accuracy can be achieved ultimately. But the fine-tuning process fails to meet the requirements of real-time applications. In this work, we propose a fast and effective method which does not rely on fine-tuning. Our network is mainly divided into two parts: location and segmentation. We also extract the semantic information of the annotation object in the first frame to generate corresponding channel-wise weights so as to re-target the network to locate and segment the specific object accurately. The experimental results show that on the DAVIS 2016 dataset, our method without fine-tuning exhibits more competitive results than the most advanced methods using online fine-tuning. The J&F average index reached 79.7%, and the running time per frame was only 0.11s.

Proposed Method

Architecture

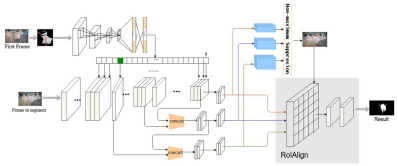

Fig. 1. We propose a two-stage framework. Feature maps are shared with location and segmentation. And we adjust channel-wise weights of the network feature map by extracting the semantic information of the annotated object in the first frame and generating the scale parameter of the channel-wise.

Object Location

The input image is divided into S×S grids in the location network. If the center of the object falls into the grid cell, that grid cell is responsible for detecting the object. Each grid cell needs to predict k bounding boxes, where k = 3. Each bounding box consists of 5 predictions: x, y, w, h and confidence. The (x, y) coordinates represent the center of the box relative to the bounds of the grid cell. w and h are the height and width predicted relative to the whole image. The confidence prediction the intersection over union (IoU) between the predicted box and any ground truth box. Formally we define confidence as:

$$c = P(\text{Object}) * IoU_{\text{pred}}^{\text{truth}}.$$

The confidence score should be zero if no object exists. By multiplying this cross-union ratio, the accuracy of the predicted bounding box is reflected.

Object Segmentation

Based on the location results of the object, the next step is to segment the object within bounding box. The FCN architecture is utilized as the base network for segmentation. FCN predicts the segmentation mask of each object bounding box. We align the box with the feature map through the Feature Pyramid RoIAlign layer, making the per-pixel spatial maintains a good correspondence. This makes the segmentation more accurate.

From Fig. 1, we can see that the feature map shared by segmentation network and location network. The features are inputted to the segmentation network also have multiple scales. First, we map the obtained object bounding box to the corresponding feature map. We use the mapping formula in fastrcnn to find the feature map of which scale the object bounding box corresponds to. After obtaining the mapped feature map, RoIAlign pools the corresponding area in the feature map into a fixed-size feature map according to the position coordinates of the object bounding box. RoIAlign cancels the quantization operation and uses the bilinear interpolation method to obtain the image value on the pixel with the floating point number, thereby converting the entire feature aggregation process into a continuous operation. RoIAlign avoids the negative impact of quantization on predicting pixel-accurate masks, and makes the segmentation result more accurate. Finally, the feature map is inputted into the FCN to get the final segmentation mask.

Loss Function

During training we optimize the following, multi-part loss function:

$$L = L_{conf} + L_{loc} + L_{seg},$$

the object loss Lconf and the location offset loss Lloc are identical as those defined in fastrcnn. Similar to semantic segmentation, Lseg use weighted cross-entropy loss.

$$L_{seg}(B) = -\sum_{i,j \in fg} \log(y_{ij} = 1;\theta) - \omega \sum_{i,j \in bg} \log(y_{ij} = 0;\theta),$$

where θ denotes CNN parameters, yij denotes the network prediction for the input box B at pixel (i, j) and ω is foreground-background pixel-number ratio used to balance the weights.

Experiment Result

We evaluate our proposed method on the validations set of DAVIS 2016 and compare with some state-of-the-art methods. TABLE I shows the results of different approaches on DAVIS 2016. We compared with three approaches using model fine-tuning on target video (OSVOS, MaskTrack and OnAVOS). And we also compared with four approaches without fine-tuning (VPN, OSMN, FAVOS, and OnAVOS-).

Method	with FT	*J&F*	*J*-mean	*F*-mean	f/t
OnAVOS	✓	85.50	86.1	84.9	13s
OSVOS	✓	80.20	79.8	80.6	10s
MaskTrack	✓	77.55	79.7	75.4	12s
Ours	✓	**81.30**	**79.9**	**82.7**	**10s**
VPN	✗	67.85	70.2	65.5	0.63s
OSMN	✗	73.45	74.0	72.9	0.14s
FAVOS	✗	76.95	77.9	76.0	0.60s
OnAVOS-	✗	—	73.6	—	3.55s
Ours	✗	**79.7**	**78.5**	**80.9**	**0.11s**

Table 1. Evaluation Results of Different Methods on DAVIS 2016 Dataset

Fig. 2. Some qualitative results of our approach compare with OSVOS on DAVIS 2016.

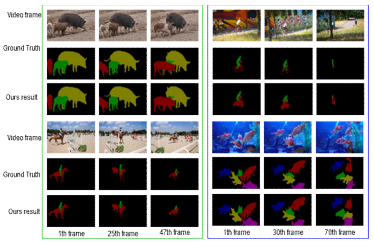

Fig. 3. Our segmentation results of DAVIS 2017.

Conclusion

In this paper, we propose a novel framework to process video object segmentation which is fast and accurate. Different from existing methods that heavily rely on the fine-tuning base on the object mask in the first frame. We propose a two-stage framework, locate and then segment. Especially, we extract the semantic information of the annotation object in the first frame to generate corresponding channel-wise weights so as to re-target the network. Compared with other video object segmentation algorithms, our method has obtained competitive results on both DAVIS 2016 and DAVIS 2017 datasets, and can better solve complex background problems. But there are rooms for improvement, we will continue to study. The next step we will be to introduce more spatial information to learn more powerful feature representations, so as to deal with the problem of similar object instances occluding each other.

References

[1]R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, "Video Object Segmentation and Tracking: A Survey," In arXiv preprint arXiv: 1904.19172v3, 2019.
[2]J. Redmon, A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv: Computer Vision and Pattern Recognition, 2018.
[3]L. Chen, J. Shen, W. Wang and B. Ni, "Video Object Segmentation Via Dense Trajectories," in IEEE Transactions on Multimedia, vol. 17, no. 12, pp. 2225-2234, Dec. 2015.
[4]K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2980-2988.

Contact

Contact Person：Wenqing Luo
Tel：18856893831
E-mail：1585648384@qq.com