Understanding of Low-latency Segmentation

Lina Zhu

25 May 2018

1 Introduction

This article describes the semantic segmentation of video and has achieved significant success. Despite this, the application is still a challenging task segmentation technology to video-based applications. Semantic segmentation, the task of segmenting observational scenes into the semantic region, has been an active research topic in computer vision. In recent years, in-depth progress in learning, especially the development of the full convolutional network (FCN) [1] has brought the performance of this task to a new level. However, many existing methods designed for semantic segmentation analyze the image [2]. How to extend the segmentation technique successfully applied to video applications is still a challenging challenge for video-based semantic segmentation, including two aspects. On the one hand, video usually involves a significant increase in the amount of data compared to images. In particular, video usually contains 15 to 30 frames second per frame. Therefore, analyzing video requires more computing resources. On the other hand, many real-world systems require video segmentation, such as autopilot, so there are strict requirements on response delays that make the problem more challenging. Figure 1 compares the performance delay tradeoffs of various methods. We can see that using the previous method is not sufficient in any one area. Our main goal in this work is not only to reduce the overall cost of this issue but also to maximize the delay, while maintaining a scene of performance in complex and ever-changing competition. In order to achieve this goal, we explore a new framework. Here, we adopt the idea of functional sharing but surpass the limitations of previous methods in two important aspects. (1) We introduce an adaptive feature propagation component that combines the preceding functions with spatially varying convolutional frames. By adapting the combination to local weighting, the result is more efficient use of previous features and therefore improved segmentation accuracy. (2) We adaptively allocate keyframes as needed based on accuracy prediction and incorporate a parallel scheme to coordinate keyframe calculations and feature propagation. This approach not only leads to more efficient use of computing resources, but also reduces the maximum latency. These components are integrated into the network.

2 Comparison of feature transmission

We start from the evaluation of the effectiveness of our adaptive propagation module. Evaluating specific performance gains In this module, we fix the scheduling scheme method as before, that is, select a keyframe every 5 frames. Table 1 shows the quantitative comparison. We compared our

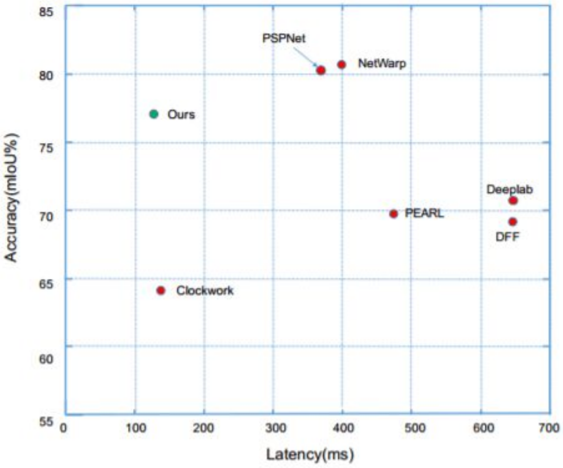

Figure 1: **Latency and mIoU performance on Cityscapes dataset. Methods involved are NetWarp, PSPNet, Deeplab, PEARL, DFF [3], Clockework and Ours. Our method achieves the lowest latency while maintaining competitive performance.**

approach to adopting global learning and unified communication, and the results were not satisfactory. It reveals this point in our solution, spatial variation weight is very important. We also compare it with the baseline of weights by setting the difference in the input pixel values directly, the learning weights we learn also perform better. Our experiments also showed that the fuse has low-level features from *AdaptNet* current framework to achieve significant functional performance improvement (from 68.41% to 75.26%), which may be attributed to the powerful complementary information learned through the fuse model. Compared with the recently proposed feature method, our adaptive propagation module still performs significantly better. In particular, the direct springs replace some of the current functionality with the previous functions, so derived representations do not apply to the current changes, resulting in poor performance. The DFF based on the optical flow method [31] relies on the optical flow propagation function and the display effect is more accurate than the clockwork because it is more adaptive. However, its performance is very sensitive to quality and it ignores the spatial relationship feature space of Flownet. These factors limit its performance gain, and therefore it is still not as good as the proposed method.

Table 1: **Comparison of different feature propagation modules.**

Method	mIOU
Clockwork Propagation	56.53%
Optical Flow Propagation [3]	69.2%
Unified Propagation	58.73%
Weight By Image Difference	60.12%
Adaptive Propagation Module (no fuse)	68.41%
Adaptive Propagation Module (with fuse)	75.26%

3 Conclusion

We propose a framework for efficient video semantic segmentation with two key components: adaptive function propagation and adaptive keyframe scheduling. In particular, our specially designed schedule scheme enables low-latency online settings. In the future, we will explore more model compression methods that can further reduce the overall computational cost and delay of the actual system.

References

[1] E. Shelhamer and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[2] H. Noh and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, pages 1520–1528, 2015.

[3] X. Zhu and Y. Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2349–2358, 2017.