Applying SlowFast Networks to Video Object Segmentation

Chantal Pellegrini
Technical University of Munich
chantal.pellegrini@tum.de

Ege Özsoy
Technical University of Munich
ege.oezsoy@tum.de

1. Introduction

The SlowFast architecture [2], which was successfully applied for video action classification, splits the workload between two branches, allowing it to use less computational resources. Our main goal is to apply SlowFast Networks [2] to Video Object Segmentation and to understand if this concept is beneficial for VOS; both in the unsupervised and semi-supervised setting. To this end, we build a SlowFast inspired VOS architecture, and evaluate different configurations of it. You can find the Code at `https://bit.ly/slowfast_segmentations_repository` and the video at Video: `https://youtu.be/GPHSpEV5wbQ`.

2. Related Work

The main work of consideration is the SlowFast Network [2], which has one fast and one slow pathway, allowing it to concentrate on different aspects of the video while keeping the performance high. The slow pathway is computationally much more expensive than the fast pathway and works with less frames.

OSVOS [1] focuses on the task of one-shot video object segmentation, where in test time, the network is fine-tuned on the first frame of the video, which has a manually annotated mask. This fine-tuning allows the network to adapt to the object in that video.

Mask R-CNN [3] is one of the best known architectures for image segmentation. They extend the popular Faster R-CNN architecture with a masking layer at the end.

3. Approach

3.1. Dataset

We use both of the DAVIS datasets [4, 5]. These datasets include videos with segmentation ground truths for every frame.

3.2. Network Design

Our architecture is build upon Mask R-CNN, which we extend with a SlowFast module between the backbone and the head. An overview is shown in Fig. 1. We fine-tune both the backbone and head of Mask R-CNN on DAVIS17 [5].

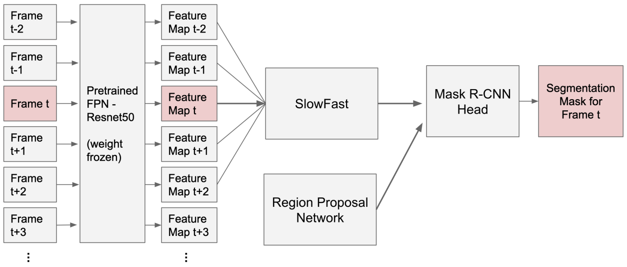

Figure 1: Architecture Overview.

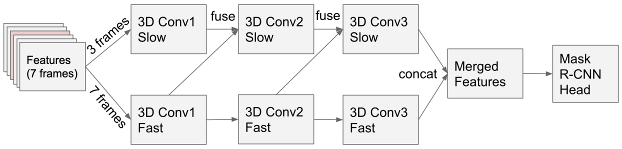

Figure 2: Overview of SlowFast Layers.

The computed feature maps of several frames are fed into the SlowFast module, which can be seen in Fig. 2. It consists of two pathways, which can get a different number of frames as input. Both are built out of three 3D convolutions, followed by batch norm and for the first two layers a ReLU activation. After both first two layers the outputs of the fast pathway are fused into the slow pathway.

The final outputs of both pathways are concatenated and used as input for the Mask R-CNN head. The head also receives region proposals computed by a RPN.

3.3. Training

For the unsupervised case we are training for 20 epochs on the training data of DAVIS17 [5]. We are using SGD with momentum as optimizer and our learning rate is set to 0.001. The semi-supervised training starts with a parent model trained on the task of unsupervised VOS and fine-tunes this model for 2000 iterations on the first frame for each sequence. We are using different augmentations, including Random Horizontal Flip, Rotation of up to 30 degrees, and Scaling of the image.

4. Experiments

We conducted several experiments for both hyperparameter tuning and the evaluation of different configurations. The qualitative results can be found here `https://www.youtube.com/watch?v=Exw_RgEA07w&feature=youtu.be`

4.1. Hyperparameter Search

As our resources were limited, we relied on previous works for some decisions such as which optimizer to use. Nonetheless, for the unsupervised case, we experimented with 3 different learning rates and also freezing the backbone or not. In the end we settled on 0.001 as lr and freezing the backbone, as training it did not provide much improvement but came at a big cost of speed. For the semi-supervised case we ran 36 experiments for every pathway configuration, which tested different settings for data augmentation, learning rate and freezing different parts of the network. Following the results we freeze only the SlowFast part of the network and use a random scale of 0.25 or 0.4 (only for 1-1) and learning rate of 0.001.

4.2. Pathway Configurations

The main goal of the pathway configurations is to show the benefit of more temporal context and the SlowFast concept. We denote our configurations as m-n, where m/n refers to the number of frames given to the slow/fast pathway. We create a baseline configuration, 1-1, which does not use any temporal context and two additional architectures 3-3 and 7-7 which progressively use more temporal context. These three configurations are not using a SlowFast inspired architecture, as both of the pathways have the same size. In addition to these three, we test 1-7 and 3-7, both utilizing the concept of SlowFast.

4.3. Results

Table 1 and 2 show the results for the unsupervised, respectively semi-supervised experiments. As evaluation metric we use J&F Mean, like described in the DAVIS16[4].

5. Evaluation

Our first observation is that in our case temporal context improves the unsupervised segmentation results, but only up to three frames, as can be seen by comparing the results of 1-1, 3-3 and 7-7. As SlowFast networks are mainly designed to see a lot of temporal context in the fast pathway and this does not seem to be beneficial here, the potential benefit is small. When we compare 1-1 to 1-7 and 3-3 to 3-7, we only see an improvement through a larger fast pathway size if the slow pathway size is one.

Configuration	J & F Mean	Param. Count	Eval. Time
1-1	0.645	45,421,851	477 sec
3-3	0.679	46,398,747	544 sec
7-7	0.673	48,407,835	853 sec
1-7	0.655	45,618,459	528 sec
3-7	0.676	46,570,779	584 sec

Table 1: Unsupervised VOS results on DAVIS16 validation set. The Evaluation Time refers to computation of masks for all validation sequences.

Configuration	J & F Mean
1-1	0.747
3-3	0.747
7-7	0.755
1-7	0.741
3-7	0.758

Table 2: Semi-supervised VOS results on DAVIS16 validation set.

In the semi-supervised experiments we only see little temporal benefit between 3-3 and 7-7 and none between 1-1 and 3-3. We also see a small improvement between 3-3 and 3-7 but a small regression from 1-1 to 1-7. Overall the small differences indicate in our case neither temporal context nor SlowFast have significant influence on the OSVOS results.

Performance wise, with bigger pathway sizes, both the parameter count and especially the evaluation time increase. The slow pathway has significantly more influence on both, as can be seen by comparing 3-3, 3-7 and 7-7.

Overall for our tasks the benefit of SlowFast appears to be limited.

6. Conclusion

In this project our goal was to implement a SlowFast inspired architecture for video object segmentation, and thereby investigate the potential benefit of SlowFast Networks for VOS. The takeaway from our results is that there was no substantial benefit observed from applying SlowFast, especially because the benefit of having more and more temporal context vanished quickly. In the cases where a temporal benefit was observed, SlowFast improved the results with little performance overhead.

References

[1] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. *CoRR*, abs/1611.05198, 2016.

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *CoRR*, abs/1812.03982, 2018.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[4] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[5] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.