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

Abstract

1. Introduction

EGE - General Idea - Why

2. Related Work

The main work of consideration is the SlowFast Network [2], which introduces the idea of using two pathways to analyze a video. This architecture 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 fps. In the original task of video action classification they show this two pathway approach is especially beneficial if there is a high speed action such as clapping or dancing involved.

OSVOS [1] focuses on the task of one-shot video 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 scene.

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

In the following we present the details of our approach.

3.1. Dataset

We used both of the DAVIS datasets [4, 5], . These datasets include videos with segmentation ground truths for every frame. We trained on DAVIS17, as it contains more video sequences, but evaluated on DAVIS16, which is a subset of DAVIS17, since there only exists one annotation per frame.

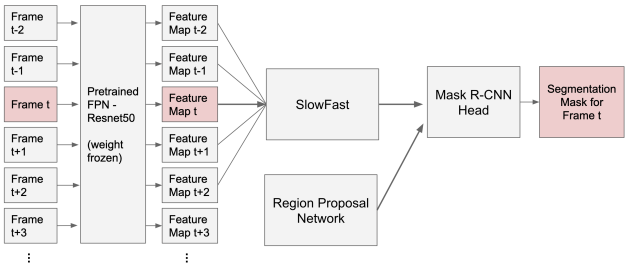

Figure 1. Architecture Overview.

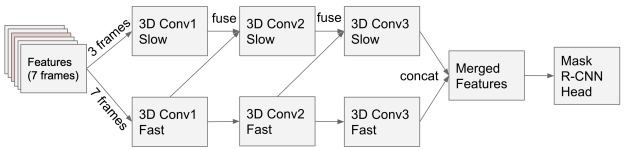

Figure 2. Overview of SlowFast Layers.

3.2. Network Design

Our architecture mainly uses MaskRCNN, which we extended with a SlowFast module between the backbone and the head. An overview is shown in Fig. 1. The backbone is a ResNet50 Feature Pyramid Network, which was pretrained on Coco and fine-tuned on DAVIS17 [5].

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, both with three 3D convolutions, followed by batch norm and for the first two layers a ReLu activation. After the first two layers the outputs of the fast pathway are fused into the slow pathway using another combination of Convolution, Batchnorm and ReLu.

The final outputs of the slow and the fast pathway are then concatenated and used as input for the Mask R-CNN head. Additionally the head also receives region proposals computed by a RPN, which works with the original image features computed by the backbone.

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 0.0001.

The semi-supervised training starts with a parent model trained on the task of unsupervised VOS and finetunes this model on specific sequences, resulting in one model per sequence. We are using different augmentations, including Random Horizontal Flip, Rotation of up to 30 degrees, and Scaling of the image. We experimented with different scaling strengths and learning rates, that are described in the experiment section.

4. Experiments

EGE

4.1. Hyperparameter Tuning

4.2. Pathway Configurations

4.3. Qualitative Results

5. Evaluation

EGE

5.1. Unsupervised

5.2. Semi-Supervised

6. Conclusion

EGE - Benefit of SlowFast

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.