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

TALLE

1. Introduction

VOS is a computationally expansive task, especially requiring a lot of GPU memory to look into many frames at once. SlowFast architecture, which was successfully applied for video action classification, splits the workload between two branches, allowing it to use less memory. Our main goal is to apply SlowFast Networks to Video Object Segmentation and to understand if this concept is beneficial for VOS; both in the unsupervised and semi-supervised settings. To this end, we build a SlowFast inspired VOS architecture, and evaluate different configurations on it.

2. Related Work

TALLE - SlowFast Networks - MaskRCNN - OSVOS - maybe track-RCNN

3. Approach

TALLE

3.1. Dataset

3.2. Network Design

- also for what configuration of freezing we decided (default)

3.3. Training

3.3.1 Unsupervised Training

3.3.2 Semi-supervised Training

4. Experiments

As our network has can be applied in many different configurations, both in terms of pathway size but also in terms of freezing certain layers etc, we conducted several experiments. The qualitative results can be found here TODO

4.1. Hyperparameter Tuning

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 didn't provide much improvement but came at a big cost of speed. For the semi-supervised case we actually ran in total 36 different experiments for every configuration, which tested different amount of data augmentations, learning rate and freezing different parts of the network. For every pathway configuration except for pathway size 1-1, the best performing configuration was only freezing the SlowFast part of the network, a random scale of 0.25 and learning rate of 0.001. For 1-1, everything is the same except random scale which was 0.4 for the best configuration. We used the best performing configurations in the following experiments.

4.2. Pathway Configurations

The main goal of the pathway configurations are two fold, first two show that using more temporal context provides a benefit, and second to show the benefit of SlowFast. For both of these, we created a baseline configuration, 1-1, which doesn't use any temporal context. We then created two additional architectures 3-3 and 7-7 which progressively use more temporal context, but also work get slower with the increased capacity. It is important to note that these 3 configurations are not using a SlowFast inspired architecture, as both of the pathways have the same size. In addition to these 3, we tested 1-7 and 3-7 configurations, both utilizing the concept of SlowFast.

The first three experiments are to show the benefit of seeing more temporal context. The last two are to see the benefit of SlowFast.

5. Evaluation

EGE

5.1. Unsupervised

5.2. Semi-Supervised

6. Conclusion

EGE - Benefit of SlowFast

References