Pyramid Hierarchies for Multi-scale Temporal Action Detection

Jiayu He, Guohui Li, Jun Lei*

National University of Defense Technology, Science and Technology on Information Systems Engineering Laboratory

Introduction

Temporal action detection is a main task in visual content understanding which is aimed at detecting human action instances from untrimmed video clips, classifying the instance into one of several action classes, and precisely predicting starting and ending time points of the action instance.

In this paper, we introduce Feature Pyramid Convolutional 3D Networks, named FPC3D. The network is designed to enhance the ability of detecting actions across a large range of temporal scales.

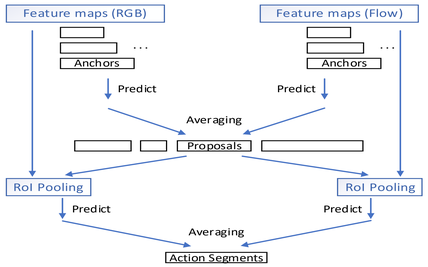
Purpose

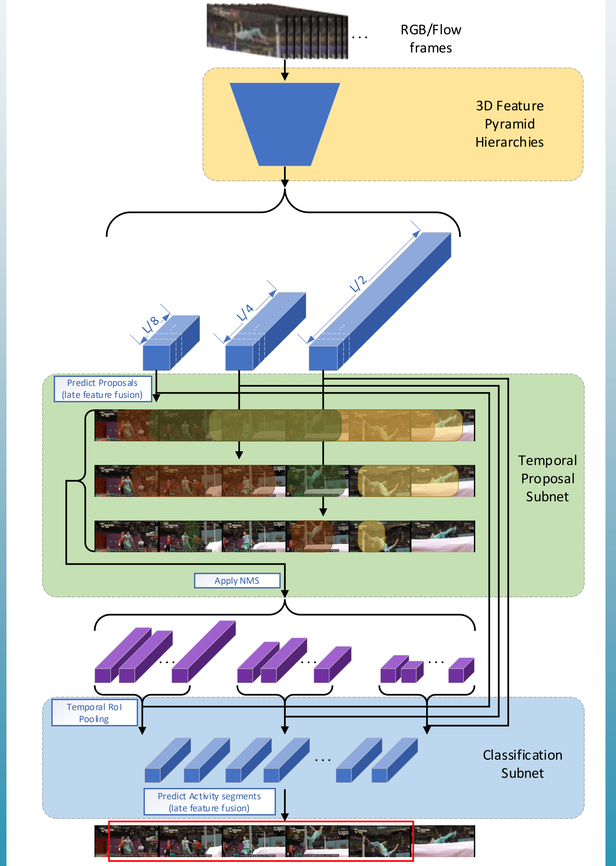
In practical application, most of the videos which need to be detected are untrimmed, long-lasting videos with multiple different action segments. For example, we might need to monitor the behaviors of prisoners held in prisons by detecting surveillance videos in real time, or we need to filter videos with nasty clips on YouTube. And these videos are exactly long-lasting and they always contain complicated action segments.

In this paper we propose Feature Pyramid Convolutional 3D Networks (FPC3D), an end-to-end framework which consists of three subnets. The network aims to improve its ability to detect actions of different temporal lengths.

Methods

The network is designed to enhance the ability of detecting actions across a large range of temporal scales. In our work, we built a 3D feature pyramid hierarchical feature to get multi-scale semantic information. Specifically, input RGB/optical flow frames of a certain video are scale-invariant, these frames are encoded through a finetuned C3D network and output a base feature map. After this, the base feature map would go through the top-down pathway and generate three novel feature maps of different temporal scale. These feature maps are higher-resolution feature maps combined with high-level semantic feature which will be shared by the following two subnets. Temporal proposing subnet is aimed to generate proposals via anchor mechanism. Feature maps utilized in this subnet are used to set positive or negative label to anchors and initially adjust the boundaries of anchors. Prediction results of RGB and optical flow features are averaged for the first time in this subnet, and it is a late fusion scheme. Then as its name says, the classification





Literature Cited

[1] Yancheng Bai, Huijuan Xu, Kate Saenko, and Bernard Ghanem. Contextual multi-scale region convolutional 3d network for activity detection. arXiv preprint arXiv:1801.09184, 2018.

[2] Yupan Huang, Qi Dai, and Yutong Lu. Decoupling localization and classification in single shot temporal action detection. In 2019 IEEE International Conference on Multimedia and Expo (ICME), pages 1288–1293. IEEE, 2019.

[3] Tingting Xie, Xiaoshan Yang, Tianzhu Zhang, Changsheng Xu, and Ioannis Patras. Exploring feature representation and training strategies in temporal action localization. In 2019 IEEE International Conference on Image Processing (ICIP), pages 1605–1609. IEEE, 2019.

[4] Abhimanyu Sahu and Ananda S Chowdhury. Multiscale summarization and action ranking in egocentric videos. Pattern Recognition Letters, 2020

[5] Jiagang Zhu, Zheng Zhu, and Wei Zou. End-to-end video-level representation learning for action recognition. In 2018 24th International Conference on Pattern Recognition (ICPR), pages 645–650. IEEE, 2018.

Acknowledgment

This study was funded by National Natural Dcience Foundation of China (Grant No: 61806215) and National Natural Dcience Foundation of China (Grant No: 71673293).