Action and Object Detection for TRECVID

Yogesh Singh Rawat, Aayush Rana, Praveen Tirupattur, and Mubarak Shah
Center for Research in Computer Vision, University of Central Florida

October 2018

Abstract

We present a deep architecture for the 2018 TRECVID [1] Video Activity Detection. We evaluated our proposed architecture for two different tasks: action detection (AD task), and activity object detection (AOD task). Our proposed framework consists of two components: activity tube detection, and activity classification. We implemented our activity tube detection network on Keras with Tensorflow backend and the classification network on PyTorch.

1 Introduction

In this work we developed a spatio-temporal action localization network in untrimmed videos. Our main focus is on untrimmed videos where we aim to localize the actions in the long videos both in the spatial and the temporal domain. We developed an end-to-end deep network which can perform joint spatial-temporal action localization and recognition. The network is based on our recent work on joint semantic segmentation and action detection [2], which we extended for multiple activity detection in untrimmed videos. The TRECVID dataset is much more challenging than the existing action detection datasets in various aspects. The first challenge is its untrimmed nature where we can have video segments of no action in a video. This requires the need of temporal localization of actions apart from action classification. The next challenge is the presence of multiple activities at the same time step within a video, which makes the action localization much more difficult. The activities in the TRECVID dataset can have multiple actors and also may have an actor-action association which opens up another interesting research direction.

2 Method

In our approach, we have extended our work from single-activity localization on trimmed videos [3] to multiple-activity localization on untrimmed videos. The proposed network has a two-step approach where activity is localized in the first step and activity detection is performed in the second step. The activity localization network takes in a 8-frames clip and predicts a localization segmentation map for presence of activities in the clip. The predicted segmentation map is then utilized to generate action tubes which are passed on to the action

classification network. The action classification network is a multi-label prediction network, which classifies the action tubes as one or more of the 19 action classes. The reason we solve this as a multi-label classification is that we can have multiple action classes assigned to the same activity. For example, activity carrying is also annotated as activity walking and similarly activity standing and activity talking can co-occur together.

2.1 AD Task

For the AD task we take the classification predictions from the classification network and stitch short clips to form long activity tubes for temporal activity evaluation.

2.2 AOD Task

For the AOD task we utilized the priors from detected activities to identify the object present in the tube location and evaluated the detections. This approach will have issue when we have multiple objects associated with the activity and we plan to explore this further with separate object localizations.

3 Conclusion

In this work we presented a deep learning framework for activity detection in untrimmed videos. The proposed framework consists of two components: activity tube detection and activity identification. The results obtained show the effectiveness of the proposed framework.

References

[1] George Awad, Asad Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, Joao Magalhaes, David Semedo, and Saverio Blasi. Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *Proceedings of TRECVID 2018*. NIST, USA, 2018.

[2] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos.

[3] Rui Hou, Chen Chen, and Mubarak Shah. An end-to-end 3d convolutional neural network for action detection and segmentation in videos. *arXiv preprint arXiv:1712.01111*, 2017.