BSN: Boundary Sensitive Network for Temporal Action Proposal Generation

Cheng Guan

August 13, 2018

Abstract

Temporal action proposal generation is an important yet challeng- ing problem, since temporal proposals with rich action content are indispensable for analysing real-world videos with long duration and high proportion ir- relevant content. This problem requires methods not only generating proposals with precise temporal boundaries, but also retrieving proposals to cover truth action instances with high recall and high overlap using relatively fewer proposals. To address these difficulties, the authors introduce an effective proposal generation method, named Boundary-Sensitive Network (BSN), which adopts local to global fashion. Locally, BSN first locates temporal boundaries with high probabilities, then directly combines these boundaries as proposals. Globally, with Boundary-Sensitive Proposal feature, BSN retrieves proposals by evaluating the confidence of whether a proposal contains an action within its region.

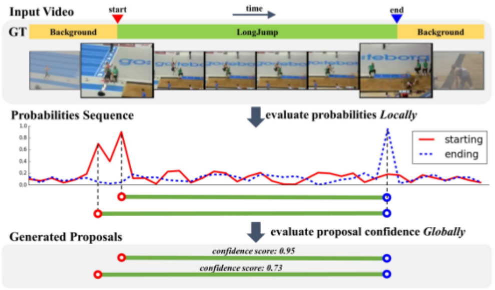

Figure 1. Overview of their approach. Given an untrimmed video, (1) The authors evaluate bound- aries and actionness probabilities of each temporal location and generate proposals based on boundary probabilities, and (2) they evaluate the confidence scores of proposals with proposal-level feature to get retrieved proposals.

1. Introduce

Nowadays, with fast development of digital cameras and Internet, the number of videos is continuously booming, making automatic video content analysis methods widely re- quired. One major branch of video analysis is action recognition, which aims to classify manually trimmed video clips containing only one action instance. However, videos in real scenarios are usually long, untrimmed and contain multiple action instances along with irrelevant contents. This problem requires algorithms for another challenging task: temporal action detection, which aims to detect action instances in untrimmed video including both temporal boundaries and action classes.

To achieve high proposal quality, a proposal generation method should generate pro- posals with flexible temporal durations and precise temporal boundaries, then retrieve proposals with reliable confidence scores, which indicate the probability of a proposal containing an action instance. Most recently proposal generation methods [1, 2, 3] generate proposals via sliding temporal windows of multiple durations in video with regular interval, then train a model to evaluate the confidence scores of generated pro- posals for proposals retrieving, while there is also method making external bound- aries regression.

They propose the Boundary- Sensitive Network (BSN), which adopts "local to global" fashion to locally combine high probability boundaries as proposals and globally retrieve candidate proposals us- ing proposal-level feature as shown in Fig. 1.

2. Related Work

Action recognition. Action recognition is an important branch of video related research areas and has been extensively studied. Earlier methods such as improved Dense Trajec- tory (iDT) mainly adopt hand-crafted features such as HOF, HOG and MBH. In recent years, convolutional networks are widely adopted in many works [4, 5, 7, 8] and have achieved great performance.

Object detection and proposals. Recent years, the performance of object detection has been significantly improved with deep learning methods. R-CNN and its variations construct an important branch of object detection methods, which adopt "detection by classifying proposals" framework. For proposal generation stage, besides sliding windows , earlier works also attempt to generate proposals by exploiting low-level cues such as HOG and Canny edge.

Temporal action detection and proposals. Temporal action detection task aims to detect action instances in untrimmed videos including temporal boundaries and action classes, and can be divided into proposal and classification stages. Most detection methods [6] take these two stages separately, while there is also method [1] taking these two stages jointly.

References

[1] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles. SST: Single-stream temporal action proposals. In *ICCV*, 2017. 1, 2

[2] F. Caba Heilbron, J. Carlos Niebles, and B. Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *CVPR*, 2016. 1

[3] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem. Daps: Deep action proposals for action understanding. In *ECCV*, pages 768–784, 2016. 1

[4] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 1

[5] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1

[6] G. Singh and F. Cuzzolin. Untrimmed video classification for activity detection: submission to activitynet challenge. *arXiv preprint arXiv:1607.01979*, 2016. 2

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015. 1

[8] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015. 1