Temporal Convolution Based Action Proposal: Submission to ActivityNet 2017

Tianwei Lin[1], Xu Zhao[1*], Zheng Shou[2]

[1]Computer Vision Laboratory, Shanghai Jiao Tong University, China. [2]Columbia University, USA

{wzmsltw, zhaoxu}@sjtu.edu.cn, zs2262@columbia.edu

Abstract

In this notebook paper, we describe our approach in the submission to the **temporal action proposal (task 3)** *and* **temporal action localization (task 4)** *of ActivityNet Challenge hosted at CVPR 2017. Since the accuracy in action classification task is already very high (nearly* 90% *in ActivityNet dataset), we believe that the main bottleneck for temporal action localization is the quality of action proposals. Therefore, we mainly focus on the temporal action proposal task and propose a new proposal model based on temporal convolutional network. Our approach achieves the state-of-the-art performances on both temporal action proposal task and temporal action localization task.*

1. Introduction

Action recognition and temporal action localization are both important branches of video content analysis. The temporal action localization or detection task aims to detect action instances in untrimmed video, containing categories and temporal boundaries of action instances.

Temporal action localization task can be divided into two main parts. (1) Temporal action proposal, which means we need generate some temporal boundaries of action instances without classifying their categories. (2) Action recognition (or we can say action classification), in this part we need to decide the categories of temporal action proposals. Most of previous works [11, 15] address these two parts separately. There are also works [3, 1] focusing on temporal action proposal. For action recognition, there are already many algorithms [12, 4] with great performance. However, the localization accuracy (mean average precision) is still very low in multiple benchmarks such as THUMOS'14 [6] and ActivityNet [2], comparing with the situation in object localization. We think the main constraint on accuracy of temporal action localization is the quality of action proposals. Therefore, we mainly focus on the temporal action proposal task in this challenge and our high quality proposals

also lead to state-of-the-art performance in temporal action localization task.

2. Our Approach

The framework of our approach is shown in Fig 1. In this section, we introduce each part of the framework, which consists of feature extraction, temporal action proposal and temporal action localization.

2.1. Feature Extraction

The first step of our framework is feature extraction. We extract two-stream features in a similar way described in [5]. We adopt two-stream network [14] which is pre-trained on ActivityNet v1.3 training set. First we segment video into 16-frames snippets without overlap. In each snippet, we use spatial network to extract appearance feature with central frame, and we use the output of "Flatten-673" layer in ResNet network as feature. For motion feature, we compute optical flows using 6 consecutive frames around the center frame of a snippet, then these optical flows are used for extracting motion feature with temporal network, where the output of "global-pool" layer in BN-Inception network is used as feature. Then, we concatenate appearance and motion feature to form the snippet-level features, which are 3072-dimensional vectors. So after feature extraction, we can transfer a video into a sequence of snippet-level feature vectors. Finally, we resize the feature sequence to new length 256 by linear interpolation.

Since we only use two-stream network trained on ActivityNet v1.3 training set to extract features, there is no external data used in our approach.

2.2. Temporal Action Proposal

Prop-SSAD. In our previous work [7] [1], we design a model called **Single Shot Action Detector (SSAD)** network which simultaneously conducts temporal action proposal and recognition. A core idea of SSAD is applying anchor mechanism to temporal action localization task based on temporal convolutional layers, which is similar

---

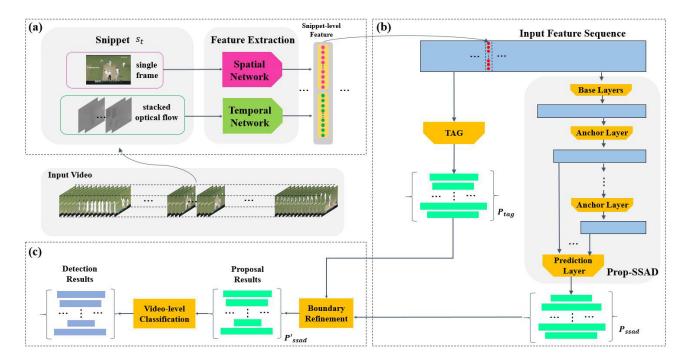[1]This paper can be found at: https://wzmsltw.github.io/

Figure 1: The framework of our approach. (a) Two-stream networks are used to extract snippet-level features. (b) Prop-SSAD model and TAG method are used for proposal generation separately. (c) Proposals generated by TAG are used for refining the boundaries of proposals generated by Prop-SSAD model. We use video-level action classification result as the category of temporal action proposals to get temporal action localization result.

with YOLO [9] and SSD [8] network for object localization task. In detail, we associate multiple temporal anchor instances with multi-scale temporal feature maps, then use temporal convolutional layers to predict information of anchor instances, including action categories, overlap score and location offsets. So SSAD can directly detect temporal action instances using feature sequence of untrimmed video.

In this challenge, we use SSAD network to make temporal action proposal without action recognition and we call it **Prop-SSAD**. The main differences of network configuration between Prop-SSAD and SSAD are listed below.

- Type of input features. In SSAD, we use two-stream network and C3D network to extract feature of video; in Prop-SSAD, only two-stream networks are used.

- Number of anchor layers. In SSAD, we only associate temporal anchors with 3 temporal feature maps using anchor layers with length 4, 8 and 16; in Prop-SSAD, we associate temporal anchors with 7 temporal feature maps with length 1, 2, 4, 8, 16, 32, 64.

- Loss function. In SSAD, we use classification, overlap and location loss jointly to train network; in Prop-SSAD, only overlap loss is used.

TAG [15]. We also implement **Temporal Actionness Grouping (TAG)** method to generate temporal action proposals, which is proposed in [15]. Since the code of TAG is not released yet, we implement TAG by ourselves. First we train a multi-layer perceptron (MLP) model with one hidden layer to predict the actionness score for each snippet, then we use grouping method described in [15] with multiple threshold to generate temporal action proposals. Proposals generated by TAG are used for refining the proposals' boundaries generated by Prop-SSAD.

Boundaries Refinement. Given feature vector sequence of a video, we can get temporal action proposals set P_{ssad} usng Prop-SSAD and temporal action proposals set P_{tag} using TAG. For each proposal p_t in P_{tag}, we calculate its IoU with all proposals in P_{ssad}. If the maximum IoU is higher than threshold 0.75, we replace the boundaries of corresponding proposal p_s in P_{ssad} with boundaries of p_t. After refinement procedure, we get refined proposals set P'_{ssad}, which is the final proposal results.

2.3. Temporal Action Localization

Since most videos in ActivityNet dataset only contain one action category, we use video-level action classification result as the category of temporal action proposals to get temporal action localization result.
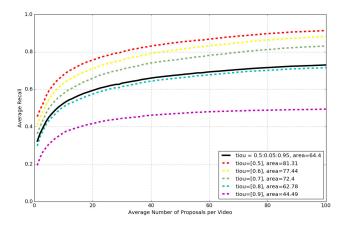
Figure 2: AR-AN curve of our proposal results in validation set. The area under black curve is the AR-AN score.

Table 1: Proposal Results on validation set of ActivityNet.

Method	AR@10	AR@100	AR-AN
Uniform Random (baseline)	29.02	55.71	44.88
Prop-SSAD	50.44	69.54	61.52
Refined Prop-SSAD	52.50	73.01	64.40

3. Experimental Results

3.1. Evaluation Metrics

Localization. In temporal action localization task, mean Average Precision (mAP) is used as the metric for evaluating result, which is similar with metrics used in object localization task. In detail, the official metric used in this task is the average mAP computed with tIoU thresholds between 0.5 and 0.95 with the step size of 0.05.

Proposal. In temporal action proposal task, the area under the Average Recall vs. Average Number of Proposals per Video (**AR-AN**) curve is used as the evaluation metric, where **AR** is defined as the mean of all recall values using tIoU thresholds between 0.5 and 0.95 with a step size of 0.05. In this notebook, we call **AR** with a certain number of **AN** as **AR@AN**. For example, AR@100 means average recall with 100 proposals.

3.2. Temporal Action Proposal

The proposal performance on validation set of our approach are shown in Table 1 and Figure 2. Our approach significantly outperform the baseline method and refined Prop-SSAD has better performance than Prop-SSAD. The boundaries refinement mainly improve the average recall with high tIoU.

Table 2: Action localization results on validation set. Results are evaluated by mAP with different IoU thresholds α and average mAP of IoU thresholds from 0.5 to 0.95. Ours@n means first n proposals used for localization.

mAP	0.5	0.75	0.95	Average mAP
Wang et al. [10]	42.28	3.76	0.05	14.85
Shou et al. [10]	43.83	25.88	0.21	22.77
Xiong et. al. [15]	39.12	23.48	5.49	23.98
Ours@1	42.14	27.17	6.54	27.00
Ours@5	46.56	30.94	7.53	30.49
Ours@10	47.84	31.90	7.76	31.41
Ours@25	48.56	32.53	7.83	31.93
Ours@100	48.99	32.91	7.87	32.26

Table 3: Action localization results on testing set. Only average mAP is provided in evaluation server, which is calculated with IoU thresholds from 0.5 to 0.95.

Method	Average mAP
Wang et. al. [13]	14.62
Xiong et. al. [15]	26.05
Zhao et. al. [16]	28.28
Ours result	33.40

3.3. Temporal Action Localization

In the temporal action localization task, we directly use proposals submitted in temporal action proposal task. For action categories, we use the video-level classification results of [13], which has 87.7% Top-1 classification accuracy and obtains 2rd place in untrimmed video classification task of ActivityNet Challenge 2016.

Evaluation results in validation set are shown in Table 2. These results suggest that localization mAP mainly depends on first several proposals. Therefore, we think AR-AN may not be the best evaluation metric for temporal action proposal task. AR with small proposals amount should has higher weight in evaluation metric.

Evaluation results in testing set are shown in Table 3. Our approach significantly outperform other state-of-the-art approaches. We think the main contributor is our high quality temporal action proposals.

4. Conclusion

In this challenge, we mainly focus on the temporal action proposal task and obtains the salient performance in both temporal action proposal and temporal action localization task. Our results suggested that anchor mechanisms and temporal convolution can work well in temporal action proposal task. In the future, we will improve our framework such as training the whole networks end-to-end.

References

[1] F. Caba Heilbron, J. Carlos Niebles, and B. Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1914–1923, 2016.

[2] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[3] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*, pages 768–784. Springer, 2016.

[4] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1933–1941, 2016.

[5] J. Gao, Z. Yang, and R. Nevatia. Cascaded boundary regression for temporal action detection. *arXiv preprint arXiv:1705.01180*, 2017.

[6] Y. G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. Thumos challenge: Action recognition with a large number of classes. In *ECCV Workshop*, 2014.

[7] T. Lin, X. Zhao, and Z. Shou. Single shot temporal action detection. *25nd ACM international conference on Multimedia*, 2017.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[10] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. *arXiv preprint arXiv:1703.01515*, 2017.

[11] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016.

[12] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[13] R. Wang and D. Tao. Uts at activitynet 2016. *AcitivityNet Large Scale Activity Recognition Challenge*, 2016:8, 2016.

[14] Y. Xiong, L. Wang, Z. Wang, B. Zhang, H. Song, W. Li, D. Lin, Y. Qiao, L. V. Gool, and X. Tang. Cuhk & ethz & siat submission to activitynet challenge 2016. *arXiv preprint arXiv:1608.00797*, 2016.

[15] Y. Xiong, Y. Zhao, L. Wang, D. Lin, and X. Tang. A pursuit of temporal accuracy in general activity detection. *arXiv preprint arXiv:1703.02716*, 2017.

[16] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, D. Lin, and X. Tang. Temporal action detection with structured segment networks. *arXiv preprint arXiv:1704.06228*, 2017.