An Empirical Study of Feature Representation for Actionness based Temporal Action Localization

Qiang Wang and Rongliang Cheng
Institute of Automation, Chinese Academy of Sciences.
qiang.wang@nlpr.ia.ac.cn, chengrongliang@hotmail.com

Abstract

In developing a practical temporal action localization framework, we utilize the Boundary-Matching Network(BMN) to efficiently generate temporal proposals, and incorporate it with a simple clip-classification model to assign fine-grained label for each proposal. After that, we thoroughly investigate the feature representation for proposal learning and proposal classification. Our approach improves over the BMN baseline by an absolute 4% *mAP in the validation set of FineAction and acheives* 12.52% *mAP in the final test set.*

1. Introduction

Temporal Action Localization is a basic component in practical video stream analytics applications, which aims to localize the start time and end time of multiple instance and recognize their action categories from a *long*, *untrimmed* video with complex background contents. The vision community has shown an increasing degree of interest in the problem, with recent methods becoming increasingly sophisticated and accurate [6, 10, 8]. Besides, the emergence of datasets has also helped the field to establish a unified standard for rapid development, from THUMOS14 [5], ActivityNet [4] to HACS [11]. However building a practical temporal action localization framework in the wild, is still challenging due to two reasons: (1)the smoothing temporal transition makes the boundary of each instance very unclear, and (2)the temporal scales for each actions are extremely different from seconds to few minutes. Recently, a newly collected large-scale dataset, FineAtion [7] was proposed, which plays more attention to practical challenges, such as fine-grained actions and short-term segments.

During the recent years, the actionness methods [6] have shown dominant and superior performance for temporal action localization. The BMN [6] achieves action detection in a two-stage fashion: first it enumerates all discrete segments as pre-defined anchors and extracts a boundary-sensitive feature for each proposal, and then predicts a category-agnostic tIoU score for each proposal. In the second stage, a video-level classifier is used to determine the action category for all segments in one video.

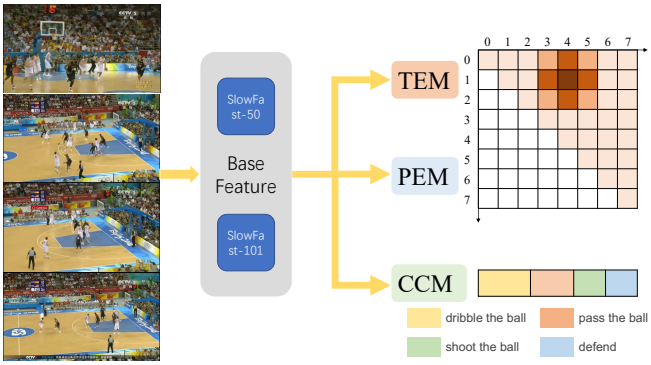

Figure 1. The entire pipeline for temporal action localization.

This separate design has achieved excellent performance in some datasets [4]. While, for a more fine-grained dataset [7], there are multiple classes of atomic actions in one video (10.74% videos on FineAction, while only 0.15% videos on ActivityNet), and the design of action classification module for each proposal is crucial. We propose a simple and efficient clip-classification module (CCM) to realize the independent predictions for each proposal, which helps to improve the overall mAP of the detection framework.

2. Our Approach

We apply BMN [6] as the baseline to generate action proposals. In addition, we use an efficient clip-level classification algorithm to help classify the proposal.

Specifically, we address the problem of finding actions from video input $V \in R^{t\times 3\times h\times w}$, where t is the length of the video frames, h and w are the height and width of the frame. We use the pre-trained SlowFast [3] to perform feature extraction on the video clip, and the sampling interval is τ. In order to facilitate a unified training dataloader,

we resize the feature to a fixed length $f \in R^{L\times d}$. The BMN algorithm uses 2d-grid to enumerate all discrete proposals to perform classification and regression of tIoU for the ground-truth action segments. We input this feature 1d feature f into the Temporal Evaluation Module (TEM) and Proposal Evaluation Module (PEM) for proposal detection learning.

Following [6], Soft-NMS [1] is used to remove the redundant temporal proposals. We finally select the top 100 proposals according to their scores as the final detection results for evaluation.

Unlike 99.85% videos in ActivityNet [4], which has only a unique category, 10.74% of the videos in FineAction [7] contain multiple categories. It is very necessary to classify different proposals independently. In order to get the classification label of the proposal, a corresponding classifier is needed to be trained. However, directly performing the proposal-level classification on the 2d-grid will bring extremely high memory consumption. We simplify it to predict each clip in the 1d time dimension. The classification result on the clip-level is denoted as $p \in R^{T\times c}$, and then the action label of each proposal is obtained by the average category probability of the proposal interval.

Different from the proposal detection branch, the classification branch requires more global context. We need to separate the video features for detection branch and the classification branch. Therefore, we add multiple non-local [9] modules to enhance the representation for clip-level classification.

3. Experiments

In this section, we firstly describe the implementation details. Then the proposed approach will be decomposed step-by-step to reveal the effect of each component.

3.1. Implementation Details

FineAction [7]. The temporal action localization task in FineAction involves 106 action categories. At the submission, we use 8440 untrimmed training videos and 4174 untrimmed validation videos to train our model and inference 4118 untrimmed testing videos. For all ablation study, we only train on the training set.

Evaluation metrics. Following conventional metrics [4], we use the mean Average Precision (mAP) as the performance metric, which is defined as the mean of all mAP values computed with tIoU thresholds between 0.5 and 0.95 with a step size of 0.05.

Training Parameter. Following traditional protocols [6], the features are re-scaled to 100 clips ($L = 100$) for the following experiments. We employ the step decay schedule with an initial learning rate 0.001 and drop gamma 0.1 at 7 epochs . The networks are optimized for 9 epochs using Adam optimizer with a weight decay of $1e-4$. We construct each mini-batch for training from 16 random videos.

Feature	AR@AN	AP@0.5	AP@0.75	AP@0.95	mAP
I3D +Video CLS	19.47%	12.72%	7.84%	2.78%	8.12%
I3D +Clip CLS	19.05%	16.00%	9.74%	3.30%	10.09%
SF50	18.78%	16.52%	10.31%	3.44%	10.60%
SF101	18.57%	16.59%	10.26%	3.41%	10.59%
SF50+SF101	19.02%	17.35%	10.57%	3.62%	11.06%
+3crop	19.22%	17.50%	10.70%	3.70%	11.13%
+NL	18.78%	16.65%	10.23%	3.61%	10.61%
+LGTE	19.42%	18.51%	11.27%	4.13%	11.81%
+TemporalShift	19.78%	19.04%	11.72%	4.26%	12.22%

Table 1. Compare feature representation for BMN module on the FineAction validation set.

3.2. Ablation Study

We will introduce the improvement components of the algorithm in detail below.

Features Enhancement. Robust video features play an important role to improve the performance of temporal action localization. We first compare the official I3D[2] feature with a pre-trained SlowFast [3] network for feature extraction. From Table 3.2, both SlowFast-50 (SF50) and SlowFast-101 (SF101) can significantly improve the mAP with more than 0.5%. Therefore, we further concatenate these two features, and the ensemble model achieves 0.97% mAP improvement.

In addition, we uniformly sample 3 crops of 256×256 to cover the spatial dimensions and average the feature. The improvement is not obvious.

We adopt a temporal shift strategy [8] for data augmentation. The overall mAP increase is around 0.3%.

We improve the representation with more temporal context. We compared the non-local [9] and LGTE [8] module, and found that the LGTE module can significantly improve the performance of BMN with 1.2% mAP. Non-local features show inferior improvement for the BMN model, since it will smooth the video features, which will hinder the precise detection edge.

Clip-level Classification Module. For the improvement of the classification model, we directly stack 4 1D-Conv-ReLu blocks as the baseline, and then insert the Non-Local [9] module in each block. Experiments shows that non-local is very effective for clip-level classification. We finally adopted the 4-layer non-local design.

Feature	w/o NL	+1 NL	+2NL	+3NL	+4NL
I3D	65.77%	69.92%	72.22%	74.15%	75.61%
SF50	81.56%	83.99%	85.64%	85.76%	86.44%
SF101	83.66%	85.11%	87.06%	87.53%	87.74%

Table 2. Compare feature representation for clip-level classification module on the FineAction validation set.

4. Conclusion

The main contribution to the competition is still feature representation. In addition, clip-level classification is also very important for fine-grained TAL tasks.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 961–970. IEEE Computer Society, 2015.

[5] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/`, 2014.

[6] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: boundary-matching network for temporal action proposal generation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3888–3897. IEEE, 2019.

[7] Yi Liu, Limin Wang, Xiao Ma, Yali Wang, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. 2021.

[8] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 485–494. Computer Vision Foundation / IEEE, 2021.

[9] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[10] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali K. Thabet, and Bernard Ghanem. G-TAD: sub-graph localization for temporal action detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10153–10162. Computer Vision Foundation / IEEE, 2020.

[11] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. HACS: human action clips and segments dataset for recognition and temporal localization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8667–8677. IEEE, 2019.