ByteDance AI Lab AVA Challenge 2019 Technical Report

Wei Li[1], Zehuan Yuan[2], An Zhao[2], Jie Shao[2], and Changhu Wang[2]

[1]Shanghai Jiao Tong University
[2]ByteDance AI Lab
{*liweihfyz*}*@sjtu.edu.cn*
{*yuanzehuan,zhaoan,shaojie.mail,wangchanghu*}*@bytedance.com*

Abstract

In this technical report, we will introduce our solution on Spatial-temporal localization (AVA datasets) of ActivityNet Challenge 2019. To this end, we propose an novel two-stage training algorithm to integrate temporal context spanning multiple seconds. Firstly, we utilize 3D ConvNet pretrained on kinetics dataset to exploit short-term visual contents of video clips centered at key frames. Secondly, we propose to link region proposals of several key frames with dynamic programming to form deformable action tubes spaning multiple seconds. In both stages, additional 3D ConvNets are introduced after ROI-pooling to integrate contents of a single clip or multiple clips into a compact representation for further classification and regression.

1. Our method

In our method, we split the whole training process into two-stages: baseline training and multi-second training. The baseline stage exploits short-term visual contents while multi-second training integrates contents of multiple seconds with linked action tubes to exploit long-term information.

1.1. Baseline

We follow the training strategy of Faster-RCNN [3] for end-to-end localization. We train out RPN Network with 2D resnet50 backbone on key frames to generate off-the-shelf region proposals. We utilize SlowFast-50[1] pretrained on kinetics dataset as our backbone to exploit visual contents of each clips centered at key frames. Following [1], We input 32 frames with a temporal stride of 2 into our backbone. Also, the spatial stride of res_5 is set to 1 to increase the spatial resolution. We rescale the shorter side of each image to 300 pixels due to GPU memory limit. Following [2], we replicate region proposals along the temporal axis to generate feature volume $V \in R^{4\times256\times7\times7}$ with 2D ROI-pooling. We utilize additionnal point-wise 2d Convolution layer to reduce the channel dimension into 256. The output feature volume V is further forwarded into an additional 2-layer 3D ConvNet to generate a compact representation. The region proposal is considered as positive if its Iou with any ground-truth box of the key frame is higher than 0.5. We choose 40 proposals with a ratio of 1:3 for position and negative proposals, respectively.

	Frame-mAP
baseline	24.0
multi-sec (T=5)	25.45
multi-sec (T=9)	**25.83**

Table 1. Comparison results of baseline model and corresponding multi-sec model.

1.2. Multi-sec training

After training our baseline model, we link region proposals of key frames of multiple seconds into an action tubes with dynamic programming algorithm. Due to memory limits, we precompute res_4 feature volumes off-the-shelf and finetune res_5 stage parameters loaded from baseline model. We uniformly sample 5 clips centered at key frames from T seconds and use average pooling or max pooling to reduce the temporal dimension of feature volumes of each clip into 1 and stack them together into a feature volume $V_m \in R^{5\times256\times7\times7}$. Similarly, an additional 2-layer ConvNet is used to aggregate stacked feature volume into a compact representation. The comparison results of baseline model and multi-sec models are summarized in Table1.

1.3. Ensemble with LFB

In order to better combine the results of long-term information, we ensemble our results with LFB[4]. For overlap-

	Frame-mAP
multi-sec (mean)	25.83
multi-sec (max)	25.3
LFB (single-crop)[4]	26.98
ensemble	**29.4**

Table 2. Comparison results of each single model and ensemble model on validation set.

ping boxes of the same class, we average their positions and adding weighted confidence scores. The weights of each methods sums to 1. We use different pooling strategies to increase the diversity of our trained models. We summarize our results on the AVA validation dataset in Table2.

References

[1] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *arXiv preprint arXiv:1812.03982*, 2018.

[2] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[4] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross Girshick. Long-Term Feature Banks for Detailed Video Understanding. In *CVPR*, 2019.