Multi-shot Temporal Event Localization: a Benchmark Supplementary Material

Xiaolong Liu[1*] Yao Hu[2] Song Bai[2,3†] Fei Ding[2] Xiang Bai[1] Philip H.S. Torr[3]
[1]Huazhong University of Science and Technology
[2]Alibaba Group [3] University of Oxford
{liuxl, xbai}@hust.edu.cn, songbai.site@gmail.com
{feifei.df, yaoohu}@alibaba-inc.com, philip.torr@eng.ox.ac.uk

In this supplementary material, we present more experimental details.

Video Feature Extraction. The interval for feature extraction is 8. Layers after the global pooling layer of I3D [7] are discarded during feature extraction. The I3D network is pre-trained on Kinetics-400 [1]. Since many categories on MUSES are unique to MUSES, we finetune the I3D network on MUSES. Besides, optical flow is not used for feature extraction due to excessive computation cost. On THUMOS14, we adopt the two-stream strategy [9] and extract features from both RGB and optical flow frames with two-stream I3D models pre-trained on Kinetics-400. It's worth noting that the categories on THUMOS14 highly correlate with those on Kinetics-400. On ActivityNet-1.3, we employ the two-stream networks trained on ActivityNet-1.3 by Xiong *et al*. [10] for features extraction.

Proposal Generation. On THUMOS14, we use proposals generated by [6] for its excellent performance. On MUSES, we find it predicts too many false boundaries and achieves low recall rates, probability due to the difficulty in detecting event boundaries in multi-shot scenarios. Therefore, a different proposal generation method is employed. Following [8], we generate sliding windows proposals of multiple lengths and employ a binary classifier to rank the proposals. The window is slid with a stride of 25% of its length and the lengths are 10, 25, 40, 55, 70, 85, 100, 130, 160, and 190 seconds. The binary classifier is composed of a convolutional stage and a fully connected stage. The convolutional stage stacks 4 1D convolutional layers with 128, 256, 512, and 1024 filters of kernel size 3 respectively, each followed by a ReLU layer and a max-pooling layer with kernel size 2. The fully connected stage includes 2 fully connected layers of 512 and 3 units respectively. After proposal ranking and non-maximal suppression (NMS) with a threshold 0.8, the top 100 scored proposals are kept for the event localization model.

Network Architecture. By default, we use two multi-scale blocks for Temporal Aggregation. In each multi-scale block, there are $K = 4$ branches. The kernel sizes of these branches are $\{1 \times 3, 3 \times 3, 3 \times 3, 3 \times 3\}$ and the corresponding unit sizes (W) are $\{3, 3, 6, 9\}$ respectively. The output channels of the first and the second block are 384 and 512 respectively.

For proposal feature extraction, we follow [3, 2, 11] to extend the boundaries of each proposal by 50% of its length on both the left and right sides before RoI Pooling [4], which is helpful for exploiting contextual information.

Loss Function. For proposal classification, completeness classification and boundary regression, the cross-entropy loss, the hinge loss and the Smooth L1 loss are used respectively. The total loss is the weighted sum of the three losses, with weights of 1, 0.5, and 0.5 respectively.

Training and Inference. During training, we set the initial learning rate to 0.01, mini-batch size to 32 and train the models with SGD optimizer with momentum 0.9. The models are trained for 20 epochs on THUMOS14 and 30 epochs on MUSES. After training for 15 epochs, the learning rate is divided by 10. For post-processing, we apply NMS with a threshold 0.4 to remove redundant detections. On THUMOS14, the predictions of the RGB and flow streams are fused using a ratio of 2 : 3 during inference. On ActivityNet, the training and inference details are the same as [5].

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 1

[2] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Re-

*Work done during an internship at Alibaba Group.
†Corresponding author

thinking the faster r-cnn architecture for temporal action localization. In *CVPR*, pages 1130–1139, 2018. 1

[3] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *ICCV*, pages 5727–5736, 2017. 1

[4] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 1

[5] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3889–3898, 2019. 1

[6] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*, September 2018. 1

[7] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, pages 5533–5541, 2017. 1

[8] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, pages 1049–1058, 2016. 1

[9] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014. 1

[10] Yuanjun Xiong, Limin Wang, Zhe Wang, Bowen Zhang, Hang Song, Wei Li, Dahua Lin, Yu Qiao, Luc Van Gool, and Xiaoou Tang. CUHK & ETHZ & SIAT submission to ActivityNet challenge 2016. *arXiv preprint arXiv:1608.00797*, 2016. 1

[11] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, pages 7094–7103, 2019. 1