– Supplementary Material –
Distilling Vision-Language Pre-training to Collaborate with Weakly-Supervised Temporal Action Localization

Chen Ju[1*], Kunhao Zheng[1*], Jinxiang Liu[1], Peisen Zhao[3], Ya Zhang[1,2], Jianlong Chang[3], Qi Tian[3], Yanfeng Wang[1,2✉]

[1]CMIC, Shanghai Jiao Tong University [2]Shanghai AI Laboratory [3]Huawei Cloud

{ju_chen, dyekuu, jinxliu, ya_zhang, wangyanfeng}@sjtu.edu.cn, {pszhao93, jianlong.chang, tian.qi1}@huawei.com

Framework Architectures

For the CBP branch, we employ one Transformer pyramid encoder as the backbone architecture to model multi-scale information. It contains alternating blocks of Multi-head Self-attention and MLPs. Layer Norm is used before each block, and residual connection is used after each block. Then, to generate frame-level action probabilities $\mathbf{P}^{\text{cb}}$, we utilize one lightweight convolutional decoder, which contains 3-layer 1D convolutional networks.

For the VLP branch, both the CLIP image and text encoders are adopted from ViT-B/16. For the textual stream, we prepend and append 16 learnable prompt vectors to the tokenized action category names. For the visual stream, we attach two temporal Transformer layers to the frame-level features $\mathbf{F}_{\text{vis}}$. And each Transformer layer is composed of Multi-head Self-attention, Layer Norm, and MLPs. During training, we freeze both encoders, and only optimize these prompt vectors and temporal Transformer layers.

Implementation Details

Training. Our framework is implemented by PyTorch, all experiments are done on one 24G GeForce RTX 3090 GPU. On both datasets, the model is optimized with Adam using a learning rate of 10^{-4}, and a batch size of 64 videos.

We first warm up the CBP branch for 10 epochs using only action category supervision, to initialize reliable background frames. Then, we alternately train the B/F step: in B step, we optimize the VLP branch for 20 epochs, to suppress false positives; while in F step, we re-train the CBP branch for 20 epochs, to suppress false negatives.

For the CBP branch, we use a fixed FPS of 25 on both datasets, following the literature [5, 6, 9, 10, 14]. To handle the large variety in video durations, we randomly sample T consecutive snippets for each video. T is set to 1000 on THUMOS14, and 400 on ActivityNet1.2. We use the TV-L1 algorithm [13] to extract optical flow from RGB data.

For the VLP branch, we set T to 1000 on THUMOS14, 400 on ActivityNet1.2. For spatial resolution, we perform center cropping to give each video frame a size of 224×224, following a convention [2, 3]. Prompt vectors and temporal Transformer have the same dimension, and are both initialized with $\mathcal{N}(0, 0.01)$. The temperature τ is set to 0.07.

Inference. During testing, we adopt the results from the CBP branch for post-processing to ensure efficiency, as no VLP branch can save computing costs. And we believe an ensemble of two branches will further boost performance. Given one video, we first obtain video-level category probabilities and frame-level localization scores. For action classification, we select the categories with probabilities greater than θ_{cls}. For action localization, we threshold localization scores with θ_{loc}, concatenate consecutive snippets as action proposals, and eliminate redundant proposals with soft non-maximum suppression (NMS). Each proposal is scored by the localization maximum within the proposal interval.

Limitations and Future Work

For the CBP branch, we freeze the I3D architecture pre-trained on Kinetics [1], to extract RGB and Flow features, following the consensus [4, 7, 8, 12, 15]. Such one frozen extractor could save computing resources, but may limit the TAL performance ceiling somewhat.

For the VLP branch, we leverage the CLIP [11], which is pre-trained with 400M image-text pairs collected from web, thus could potentially bias towards web data.

As the future work, we expect more computing resources available, to further optimize our distillation-collaboration framework into the end-to-end training setup, also rendering the asynchronous online training.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recog.*, 2017. 1

[2] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *Proc. of Eur. Conf. Comput. Vis.*, 2022. 1

[3] Chen Ju, Zeqian Li, Peisen Zhao, Ya Zhang, Xiaopeng Zhang, Qi Tian, Yanfeng Wang, and Weidi Xie. Multi-modal prompting for low-shot temporal action localization. *arXiv preprint arXiv:2303.11732*, 2023. 1

[4] Chen Ju, Haicheng Wang, Jinxiang Liu, Chaofan Ma, Ya Zhang, Peisen Zhao, Jianlong Chang, and Qi Tian. Constraint and union for partially-supervised temporal sentence grounding. *arXiv preprint arXiv:2302.09850*, 2023. 1

[5] Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Divide and conquer for single-frame temporal action localization. In *Proc. of the IEEE Int. Conf. on Comput. Vis.*, 2021. 1

[6] Chen Ju, Peisen Zhao, Siheng Chen, Ya Zhang, Xiaoyun Zhang, and Qi Tian. Adaptive mutual supervision for weakly-supervised temporal action localization. *IEEE Trans. Multimedia*, 2022. 1

[7] Chen Ju, Peisen Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Point-level temporal action localization: Bridging fully-supervised proposals to weakly-supervised losses. *arXiv preprint arXiv:2012.08236*, 2020. 1

[8] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proc. of Eur. Conf. Comput. Vis.*, 2018. 1

[9] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recog.*, 2018. 1

[10] Sujoy Paul, Sourya Roy, and AmitK Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proc. of Eur. Conf. Comput. Vis.*, 2018. 1

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. on Mach. Learn.*, 2021. 1

[12] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proc. of the IEEE Conf. on Comput. Vis. and Pattern Recog.*, 2016. 1

[13] Andreas Wedel, Thomas Pock, Christopher Zach, Horst Bischof, and Daniel Cremers. An improved algorithm for tv-l 1 optical flow. In *Statistical and geometrical approaches to visual motion analysis*. 2009. 1

[14] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *Proc. of Eur. Conf. Comput. Vis.*, 2020. 1

[15] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proc. of the IEEE Int. Conf. on Comput. Vis.*, 2017. 1