Weakly-Supervised Temporal Action Localization with Multi-Modal Plateau Transformers

Supplementary Material

1. Experiments

1.1. Ablation Study on THUMOS14

1.1.1 Impact of training strategy

In Table 1, we explore the impact of different training strategies. First, we pre-train a multi-modal Transformer model with reconstruction loss $\mathcal{L}_{\text{rec}}$. Then we use different schemes of obtaining pseudo labels to fine-tune the pre-trained network. Note that "Cross" and "Self" on the left side of the arrow mean we use cross-attention and self-attention to generate corresponding pseudo labels. The right is the training model with cross-attention or self-attention. The model for pseudo-label generation and training is the same, we just change the attention style. We can find that "Cross → Self" shows better than only "Cross → Cross" because this strategy can utilize the supplementary information between self and cross-attention, which also proves the effectiveness of cross-supervision. Our proposed "Cross-Supervision" highly outperforms other combinations.

Table 1. Comparison of different training strategies **A** → **B**, where **A** represents the way to generate pseudo labels and **B** stands for the trained model. **Note that the plateau refinement is not included.**

Strategy	**mAP@IoU (%)**				
	0.1	0.3	0.5	0.7	AVG
Cross → Cross	70.4	54.5	37.3	12.5	44.1
Self → Self	69.4	55.8	39.4	14.5	45.2
Cross → Self	70.3	55.5	37.7	14.1	44.8
Self → Cross	70.9	55.4	37.3	12.3	44.4
Cross-Supervision	70.8	56.5	39.4	14.3	**45.7**

1.1.2 Impact of inference methods

For inference, we follow [1, 2] and use the mixed attention weight (since it incorporates information from both modalities) to get action proposals and TCAM to classify proposals. Both are generated by self-attention. While we can also use the cross-attention results or a hybrid of self-attention and cross-attention(with shared weights) results for inference, we observe that there is little difference among these choices, as shown in Table 2. To make it simple, we keep using self-attention to evaluate our methods during the whole development.

Table 2. Comparison of different inference settings.

Combination		**mAP@IoU (%)**				
Attn	**TCAM**	0.1	0.3	0.5	0.7	AVG
Self	Self	74.1	60.0	41.1	15.1	**48.18**
Cross	Cross	73.8	59.7	40.9	14.9	48.08
Self	Cross	74.0	59.6	40.8	15.0	48.07
Cross	Self	74.0	59.8	41.1	15.1	48.17

1.1.3 Impact of Hyper-parameters

We conduct ablation study on the hyper-parameters of loss functions. Table 3 shows the effect of pseudo labels. We observe that model performance degrades heavily when λ_0 is smaller, which denotes the emphasises on pseudo label effect makes more contribution.

Table 3. Effect of pseudo label loss

Various λ_0	**mAP@IoU (%)**				
	0.1	0.3	0.5	0.7	AVG
1	72.9	56.9	36.7	12.4	45.1
2	72.3	57.1	37.9	13.1	45.7
5	73.2	58.4	40.8	14.3	47.1
10	74.1	60.0	41.1	15.1	**48.2**
20	74.3	59.6	40.2	14.4	47.6

Table 4 reports the results of the comparison with different values of λ_1, which aims to achieve the trade-off for the last two terms, as these two items are both constraints on attention weights. We observe that the model obtains better performance when choosing $\lambda_1 = 0.8$. That is why we simple set it to 0.8 by default.

Table 4. Effect of regularization (sparsity and opposite loss)

Various λ_1	**mAP@IoU (%)**				
	0.1	0.3	0.5	0.7	AVG
0.5	73.5	58.1	39.1	14.7	46.8
0.8	74.1	60.0	41.1	15.1	**48.2**
1	73.7	58.2	38.7	14.3	46.7

1.1.4 Qualitative results on multi-action categories

Figure 1 shows results for a multi-action instance in THUMOS14. There are very fewer multi-action videos(less than 10%) in both THUMOS14 and ActivityNet1.2 datasets, and different action segments are very close to others, as Figure 1 shows. Since attention weights are class agnostic and only represent existence of action. Figure 1 shows that our plateau method covers almost every segment for all action categories, proving the effectiveness of the plateau module.

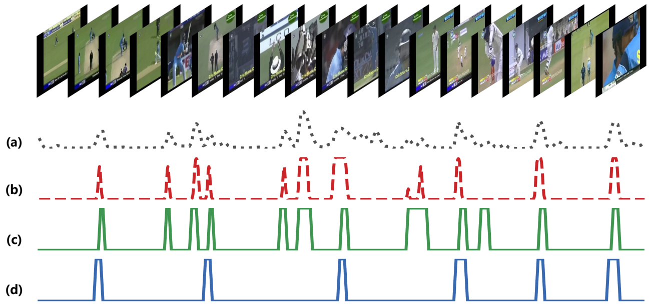


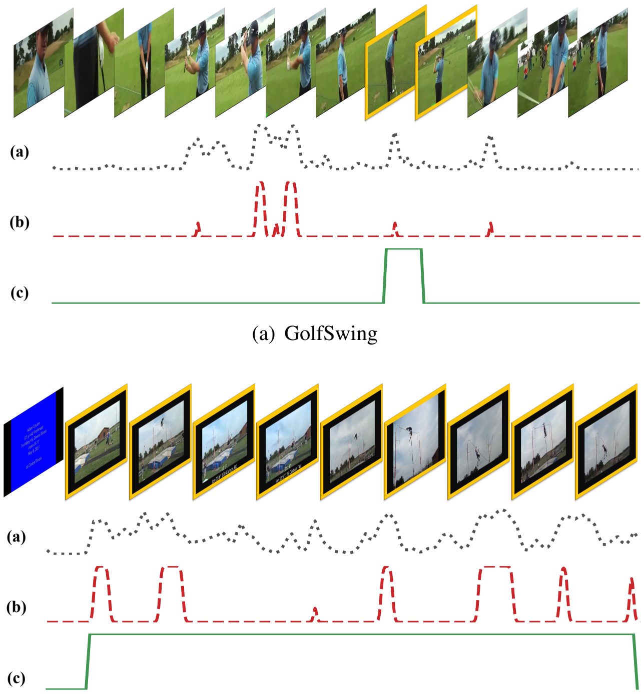
Figure 1. Results for multi-action instance. Row-(a) denotes the original attention weights, row-(b) shows the output of our plateau refinement, row-(c) means the ground-truth temporal localization for action ”CricketShot”, and row-(d) represents ground-truth for action ”CricketBowling”.

1.1.5 Failure examples

Considering the difficulty of WS-TAL task, the learned attention weights would not well discover the intrinsic ground-truth temporal action segments. In this sense, we represent the results of two THUMOS14 failure examples shown in Figure 2. Since plateau optimization is applied to origin attention weights, it could fail when attention weights are too confident in negative samples. Also, we observe that plateau functions might suppress possible positive parts as shown in the ground truth parts.

1.2. ActivityNet1.2

In Figure 3 and 4, we illustrate the success and failure examples in ActivityNet1.2. It is noticed that plateau functions might introduce extra noise to attention weights in crowded actions. Note that some videos in ActivityNet1.2 are not available to download, so we just post refined results here.



(a) GolfSwing

(b) PoleVault

Figure 2. Two failure cases from action “GolfSwing” and “PoleVault” in THUMOS14. Row-(a) denotes the original attention weights, row-(b) shows the output of our plateau refinement, and row-(c) means the ground-truth temporal localization.

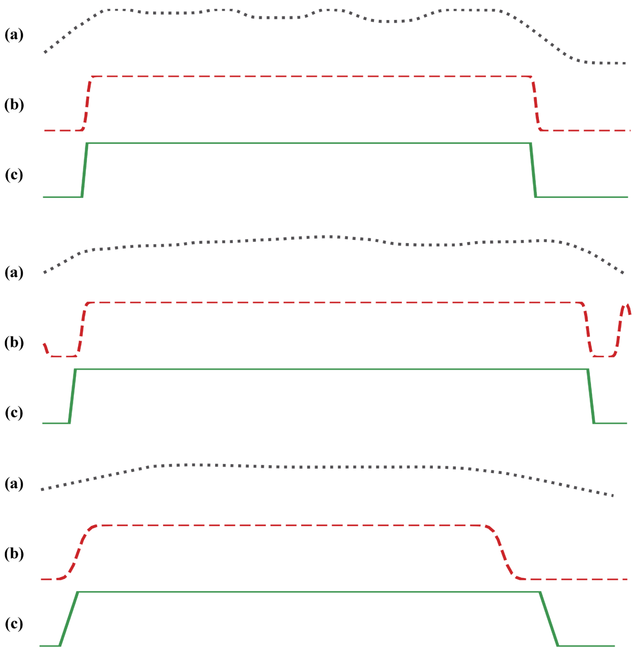


Figure 3. Illustration of three success qualitative cases in ActivityNet1.2, where the actions are “Tumbling”, “Tai chi” and “Getting a haircut” from up to bottom.

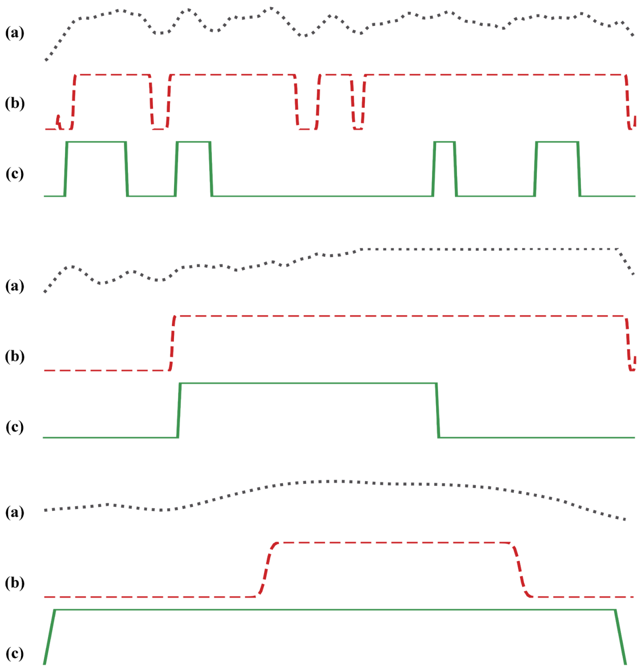


Figure 4. Illustration of three failure cases in ActivityNet1.2, where the actions are “Ping-pong”, “Starting a campfire” and “Polishing shoes” respectively.

References

[1] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *European Conference on Computer Vision*, pages 192–208. Springer, 2022. 1

[2] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1591–1599, 2021. 1