Probing Visual-Audio Representation for Video Highlight Detection via Hard-Pairs Guided Contrastive Learning

Shuaicheng Li*[1]
lishuaicheng@sensetime.com
Feng Zhang*[1]
zhangfeng4@sensetime.com
Kunlin Yang[1]
yangkunlin@sensetime.com
Lingbo Liu[2]
liulingbo918@gmail.com
Shinan Liu[1]
liushinan@sensetime.com
Jun Hou†[1]
houjun@sensetime.com
Shuai Yi[1]
yishuai@sensetime.com

[1] Sensetime Research
[2] The Hong Kong Polytechnic University
Hong Kong, China

Recall that the visual-audio framework is introduced for video highlight detection, the key idea of our proposed method is to adopt intra- and inter-modality encoding and hard-pairs guided contrastive learning scheme to model visual-audio representation. In this supplementary material, we include the following materials:

1) Details about datasets and implementations.

2) additional ablation studies.

3) more qualitative results.

1 Details about datasets, metric and implementations

1.1 Datasets and Metric

YouTube Highlights [4] is composed of six video categories, *i.e.* dog, gymnastics, parkour, skating, skiing and surfing, and each category has approximately 100 videos. Segment-level

* Equal contribution
† Corresponding author

annotations are provided to indicate whether a segment is a highlight moment. We follow the training-test split of previous works [1, 5] for model training and evaluation.

TVSum [3] is a video summarization dataset containing 10 categories with 5 videos of each category and an average of minutes per video. Since the ground truth annotations in TVSum are given as frame-level scores, we first need to aggregate frame-level scores to obtain segment-level scores, and then select the higher 50% of the scores as the segment-level highlight annotations. For each category of 5 videos, following [5], we choose the longer two videos for training and the remaining three videos for testing.

Evaluation Metrics: On the YouTube Highlights dataset, we adopt mean Average Precision(mAP), which is widely adopted in previous methods [1, 2, 5], as the evaluation metric. Unlike in image object detection, where mAP is obtained by accumulating and computing the average precision over all images, highlight detection compute mAP for each video separately, because highlight moments in one video are not necessarily more interesting than non-highlight moments in other videos. On the TVSum dataset, we refer to previous works [1, 5, 6] and compute the mAP at top-5 scores for every video.

sampling ratios	Average Results
	YouTube Highlights
1 : 3	0.718
1 : 2	0.747
1 : 1	0.741

Table 1: Ablation Study for Various Positive and Negative Sampling ratios.

1.2 Implementations

In data pre-process, we split the untrimmed videos into several segments following the previous works [1, 5]. Since the highlight detection datasets are highly imbalanced, we make sure the T selected video segments in a video sequence contain both highlight (*positive*) and non-highlight (*negative*) segments for contrastive learning and hard-pairs ranking. Specifically, in the sampling process, the training samples need to meet certain conditions. First, we sample $T = 20$ segments for a video sequence where the segment sequence must contain both positive samples and negative samples, and the ratio is not less than 1:2. Secondly, if the above is not satisfied, we re-sample additional samples in the same video, and the final sampling results must satisfy the temporal order. In our model, we use $n = 2$ transformer encoder/decoder layers with 8 attention heads blocks for intra-modality and cross-modality co-occurrence encoding and set dropout probability to 0.5. The embedding dimension $d = 256, d_k = 512, d_v = 512$. Adam optimization algorithm is applied for training the models with 20 epochs on both Youtube Highlights and TVSum benchmarks. The learning rate is set to 1e-4. Finally, all our experiments are conducted on a single NVIDIA Tesla V100 GPU. And our model is implemented with the Pytorch deep learning framework.

2 Additional Ablation Studies

Similar to the main paper, We conduct extensive experiments using I3D and PANN to extract visual and audio features respectively.

Encoder Layers	Average Results	
	YouTube Highlights	TVSum
1	0.723	0.781
2	0.747	0.801
3	0.711	0.765
4	0.705	0.733

Table 2: Ablation study for different intra-modality encoder layers.

Eecoder Layers	Average Results	
	YouTube Highlights	TVSum
1	0.728	0.797
2	0.747	0.801
3	0.722	0.763
4	0.716	0.737

Table 3: Ablation study for different cross-modality squence-wise encoder layers.

2.1 Ratio of Positive Samples

The effect of different positive and negative sampling ratios is studied. It is worth noting that this setting is not available on the TVSum dataset since the highlight moments are determined through selecting the top 50% shots (segments) of all segments in a video following recent methods [1, 5]. As shown in Table 1, when the ratio of positive and negative samples is increased from 1:3 to 1:2, the performance is significantly improved while it drops slightly when increasing from 1:2 to 1:1.

2.2 Number of Encoders / Decoders

In our model, we use transformer encoders / decoders with attention heads blocks for intra-modality and cross-modality co-occurrence encoding. We first show the effect of different encoder layers in intra-modality encoding in Table 2. when the number of encoders in intra-modality encoding $n = 2$, our methods perform best. The accuracy drops from best 0.747 to 0.705 when the encoder layer $n = 4$. The effect of cross-modality sequence-wise encoders is also studied in Table 3. We achieve the best results when adopting 2 sequence-wise decoders.

3 More Qualitative Results

In the main paper, we show the visual results on the Youtube highlight dataset. We present additional convincing visualization results of our proposed method on the TVSum dataset. As shown in Figure 1, our proposed method can perfectly detect the highlight moment and clearly separate the highlight from the background.

References

[1] Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. Joint visual and audio learning for video highlight detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8127–8137, 2021.

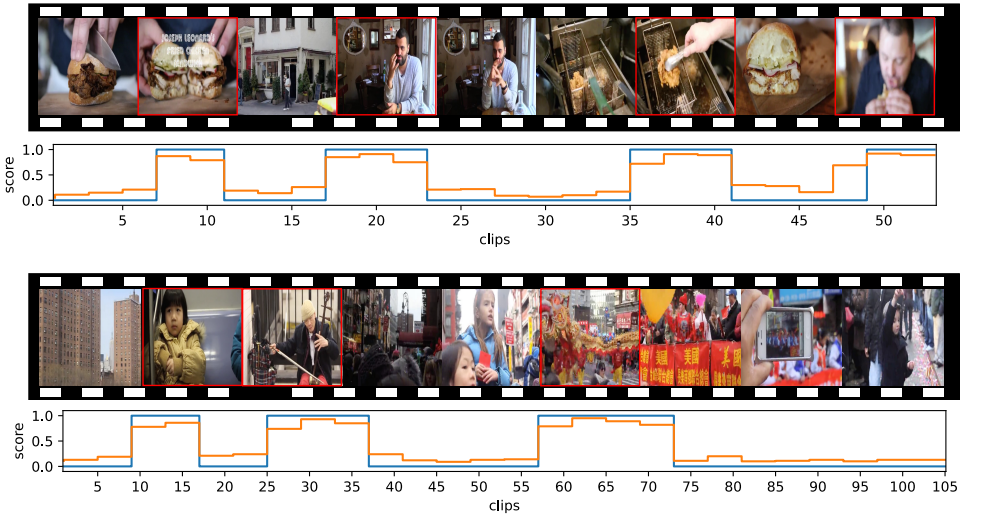

Figure 1: Qualitative results on TVSum dataset. We show highlight detection results on the test set of YouTube Highlights. The red box represent the ground truth segments.

[2] Fa-Ting Hong, Xuanteng Huang, Wei-Hong Li, and Wei-Shi Zheng. Mini-net: Multiple instance ranking network for video highlight detection. In *European Conference on Computer Vision (ECCV)*, pages 345–360. Springer, 2020.

[3] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision (ECCV)*, pages 787–802. Springer, 2014.

[5] Minghao Xu, Hang Wang, Bingbing Ni, Riheng Zhu, Zhenbang Sun, and Changhu Wang. Cross-category video highlight detection via set-based learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7970–7979, 2021.

[6] Qinghao Ye, Xiyue Shen, Yuan Gao, Zirui Wang, Qi Bi, Ping Li, and Guang Yang. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7950–7959, 2021.