Supplementary Material for "MINI-Net: Multiple Instance Ranking Network for Video Highlight Detection"

Fa-Ting Hong[1,4,5], Xuanteng Huang[1], Wei-Hong Li[3], and Wei-Shi Zheng[1,2,5] ⋆

[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Peng Cheng Laboratory, Shenzhen 518005, China
[3] VICO Group, University of Edinburgh, United Kingdom
[4] Pazhou Lab
[5] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
{hongft3,huangxt57}@mail2.sysu.edu.cn, w.h.li@ed.ac.uk, wszheng@ieee.org

1 Implementation Details

In this work, we implement the proposed method in Pytorch, using SGD as optimizer. The learning rate is initialized as 0.005 and scaled by a factor of 0.7 every 20 epochs. Additionally, we set the weight decay and momentum as 0.0005 and 0.9, respectively, for all experiments. During the training, to be more likely to sample the positive instance, we take the videos that are shorter than τ in the interest event as *positive videos* and videos that are longer than τ in non-interest events as *negative videos*, inspired by [5], and we set τ as 60. We set ϵ as 1 and the bag size as 60. We show that our method is not too much sensitive to τ and the bag size by reporting results of using various value for τ and the bag size in supplemental experiments. To form each bag, we simply break a video up uniformly into 1-second segments and randomly sample a bag size number of segments. If the total number of segments in the video is less than the bag size, we repeat the sampling. We follow the standard evaluation metric in [5], *i.e.*, the mean average precision is reported to measure the performance of all of the methods on YouTube Highlights dataset, and top-5 mean average precision for TVSum dataset and CoSum dataset. We adopt the C3D network [3] pretrained on Kinetics [1] to extract a 512-dimensional feature as vision feature for each segment, and the VGGish model [4] pretrained on AudioSet [2] for extracting a 128-dimensional feature as audio feature.

2 More Detail about Datasets

- TVSum consists of 50 videos grouped by 10 categories (5 videos per category), including changing a Vehicle Tire (VT), getting a Vehicle Unstuck (VU), Grooming an Animal (GA), Making a Sandwich (MS), ParKour (PK), PaRade (PR),

⋆ Corresponding author

Flash Mob gathering (FM), BeeKeeping (BK), attempting a Bike Trick (BT) and Dog Show (DS).
- **CoSum** The dataset consists of 50 videos grouped by 10 categories (5 videos per category), including Base Jumping (BJ), Bike Polo (BP), Eiffel Tower (ET), Excavators River Cross (ERC), Kids Playing in leaves (KP), Major League Baseball (MLB), National Football League (NFL), Notre Dame Cathedral (NDC), Statue of Liberty (SL) and SurFing (SF)

3 Additional Experimental Results

3.1 Variants of max-max ranking loss

In this work, we exploit a max-max ranking loss (MM-RL) to acquire a reliable relative comparison between the most likely positive segment instance and the most hard negative segment instance. To verify the effectiveness of our proposed MM-RL, we evaluate several variants of our MM-RL.
- **Min-Min Ranking Loss**. This variant picks the minimum value from the highlight scores of all segments in both positive bag and negative bag, *i.e.*, $\min_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i$ and $\min_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i$. After that, the min-min ranking loss ensures that $\min_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i$ is larger than $\min_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i$ with a margin of ϵ as follows:

$$\mathcal{L}_{min-min}(\mathcal{B}_p, \mathcal{B}_n) = \max(0, \epsilon - \min_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i + \min_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i) \tag{1}$$

- **Min-Max Ranking Loss**. Differently, min-max ranking loss, a variant of our max-max ranking loss, picks the minimum value and maximum from the highlight scores of all segments in the positive bag and negative bag, respectively (*i.e.*, $\min_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i$ and $\max_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i$). After that, Min-min ranking loss ensures that $\min_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i$ is larger than $\max_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i$ with a margin of ϵ as follows:

$$\mathcal{L}_{min-max}(\mathcal{B}_p, \mathcal{B}_n) = \max(0, \epsilon - \min_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i + \max_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i) \tag{2}$$

- **Max-Min Ranking Loss**. Moreover, we also evaluate the max-min ranking loss variant that picks the maximum value and minimum from the highlight scores of all segments in the positive bag and negative bag, respectively (*i.e.*, $\max_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i$ and $\min_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i$) before ensuring that $\max_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i$ is larger than $\min_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i$ with a margin of ϵ:

$$\mathcal{L}_{max-min}(\mathcal{B}_p, \mathcal{B}_n) = \max(0, \epsilon - \max_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i + \min_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i) \tag{3}$$

We adopt three variants mentioned above into our MINI-Net by replacing max-max ranking loss (*i.e.*, MINI-Net$^{\text{Min-Min}}$ for $\mathcal{L}_{min-min}$, MINI-Net$^{\text{Min-Max}}$ for $\mathcal{L}_{min-max}$ and MINI-Net$^{\text{Max-Min}}$ for $\mathcal{L}_{max-min}$). We evaluate these variants for highlight detection on three datasets (*i.e.*, YouTube Highlights dataset, TVSum dataset and CoSum dataset.) and report our experimental results on Table 1.

Dataset	MINI-Net$^{\text{Min-Min}}$	MINI-Net$^{\text{Min-Max}}$	MINI-Net$^{\text{Max-Min}}$	MINI-Net
YouTube	0.5884	0.6165	0.6186	0.6436
TVSum	0.6469	0.6747	0.7103	0.7324
CoSum	0.7863	0.8004	0.8338	0.9278

Table 1: Ablation study for ranking loss on three datasets.

From Table 1, our proposed MINI-Net with max-max ranking loss performs the best, followed by MINI-Net$^{\text{Max-Min}}$, for the reason that picking $\max_{\mathcal{I}_p^i \in \mathcal{B}_p} \mathcal{E}_p^i$ can ensure that the highlight segment of interest event is selected with the highest probability, and picking $\max_{\mathcal{I}_n^i \in \mathcal{B}_n} \mathcal{E}_n^i$ means that all segments from non-interest events is non-highlights.

3.2 Evaluation of hyperparameters

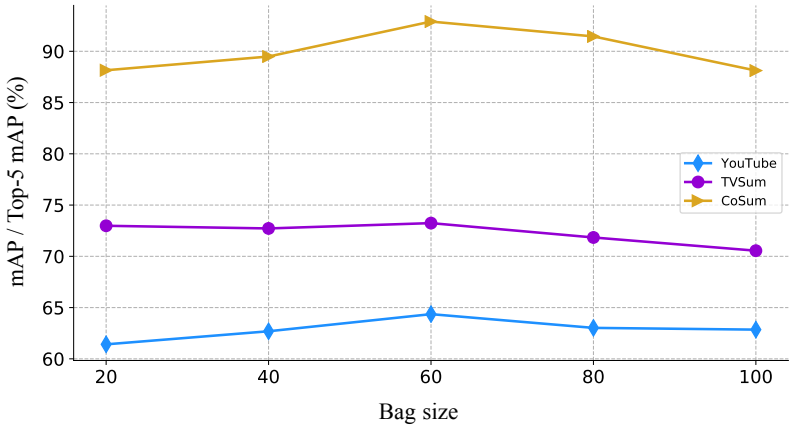

Fig. 1: Accuracy vs. bag size of our multiple instance learning framework on three datasets.

Figure 1 shows highlight detection accuracy as a function of bag size. We conduct this ablation on three datasets, *i.e.*, YouTube Highlights dataset, TVSum dataset and CoSum dataset. It can be seen that our method has little performance variance on the three datasets as increasing the number of bag size.

In this work, we take the videos that are shorter than τ in the interest event as positive videos and videos that are longer than τ in non-interest events as negative videos. We also conduct the experiments to evaluate threshold τ. Here, we report the experimental results on Table 2. It can be found that our method is not too much sensitive to τ (*e.g.*, we obtain 1.16% among implementation of τ

Dataset	$\tau = 40$	$\tau = 60$	$\tau = 80$
YouTube	0.6150	0.6436	0.6290
TVSum	0.7003	0.7324	0.6981
Cosum	0.8622	0.9278	0.8825

Table 2: Evaluation of different τ set in training process on three datasets.

Dataset	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 1.5$
YouTube	0.6345	0.6436	0.6372
TVSum	0.7174	0.7324	0.7291
Cosum	0.8999	0.9278	0.9127

Table 3: Evaluation of different ϵ set in training process on three datasets.

= 40, 60, 80 on YouTube Highlights dataset) and we get the best performance when $\tau = 60$.

We enforce the maximum value of highlight score in positive bag larger than that in negative bag with a margin of ϵ. We test varying value of ϵ, *i.e.*, 0.5, 1, 1.5 and report their results in Table 3, it can be found that our method is not too much sensitive to both ϵ (*e.g.*, we obtain 0.39% among implementation of ϵ = 0.5, 1, 1.5 on YouTube Highlights dataset).

4 Visual Examples

Moreover, we also illustrate the highlight detection results on three datasets (*i.e.*, YoutTube Highlights dataset, TVSum dataset and CoSum dataset) in Figure 2

5 Future Discussion

While we focus on the even-specific highlight detection in this work, our method could be extended to various topics where only weak supervision is provided, including event-agnostic highlight detection. One straightforward idea is treating videos which are annotated as highlight-worthy as positive bags and videos with non-highlight-worthy as negative bags for training.

Fig. 2: Examples of highlight detection results for three datasets.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Computer Vision and Pattern Recognition (2017)
2. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: International Conference on Acoustics, Speech and Signal Processing (2017)
3. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Computer Vision and Pattern Recognition (2018)
4. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: International Conference on Acoustics, Speech and Signal Processing (2017)
5. Xiong, B., Kalantidis, Y., Ghadiyaram, D., Grauman, K.: Less is more: Learning highlight detection from video duration. In: Computer Vision and Pattern Recognition (2019)