Cross-category Video Highlight Detection via Set-based Learning

Minghao Xu[1] Hang Wang[1] Bingbing Ni[1*] Riheng Zhu[2] Zhenbang Sun[2] Changhu Wang[2]
[1]Shanghai Jiao Tong University, Shanghai 200240, China [2]ByteDance AI Lab
{xuminghao118, wang–hang, nibingbing}@sjtu.edu.cn
{zhuriheng, sunzhenbang, wangchanghu}@bytedance.com

Table 1. Statistics of the ActivityNet dataset in our experiments.

Split	eat&drink	personal care	household	sport	social	Total
Training	140	186	458	1289	447	2520
Test	65	95	212	672	216	1260

1. More Experimental Setups

Combining SL-module with UDA methods. For the sake of fair comparison, we combine five Unsupervised Domain Adaptation (UDA) algorithms, *i.e.* DAN [3], DeepCORAL [5], RevGrad [1], MCD [4] and AFN [7], with the proposed SL-module and compare these combinations with the DL-VHD method. DAN, DeepCORAL and AFN align the feature distributions of source and target domain by minimizing specific domain discrepancy metrics, and we exert these metric-induced alignment losses on the contextualized segment embeddings (*i.e.* outputs of Transformer encoder) to narrow the distributional gap between source and target category video segments in the latent space. For RevGrad, we append a domain discriminator on the top of contextualized segment embeddings to conduct adversarial domain adaptation. For MCD, we train two scoring models on the source video category in a supervised way, and a minimax game is performed between Transformer encoder and two scoring models to derive more reliable highlight predictions on target video category.

Dataset statistics of ActivityNet. In our experiments, we employ a subset of ActivityNet [2] for model evaluation. The number of videos in the training and test split for each video category is shown in Tab. 1. Note that, each of these videos contains at least one highlight moment of the corresponding video category.

2. More Results of Cross-category Video Highlight Detection

In Tab. 2, we evaluate different methods on five cross-category video highlight detection tasks of YouTube Highlights [6], in which *dog* serves as the source video category. This setting is more difficult than the one employing *surfing* as the source category, since it intends to transfer the highlight patterns of dog to human. Source-only (target-oracle) method denotes the SL-module trained on the source (target) video category in a supervised way. From the table, we can observe that the full model of DL-VHD outperforms five UDA approaches with a clear margin, and it surpasses the target-oracle model on the dog $\rightarrow$ gymnastics task. When the coarse-grained or fine-grained learner is individually applied (*i.e.* the configuration $\mathcal{L}_{\mathrm{coarse}}$ only and $\mathcal{L}_{\mathrm{fine}}$ only), their performance is apparently lower than their combination (*i.e.* the configuration w/o $\mathcal{L}_{\mathrm{distill}}$ and full model). After integrating the knowledge of two learners, the full model can derive more precise highlight predictions on target video category than the model without applying knowledge distillation.

Table 2. Cross-category highlight detection results (mAP) on the YouTube Highlights dataset. (source video category: dog; the underlined result surpasses the target-oracle.)

Methods	$\rightarrow$gymnastics	$\rightarrow$parkour	$\rightarrow$skating	$\rightarrow$skiing	$\rightarrow$surfing
Source-only	0.486	0.480	0.535	0.564	0.531
DAN [3]	0.520	0.674	0.632	0.613	0.575
DeepCORAL [5]	0.518	0.615	0.615	0.609	0.517
RevGrad [1]	0.514	0.630	0.629	0.618	0.587
MCD [4]	0.479	0.587	0.658	0.614	0.625
AFN [7]	0.498	0.594	0.607	0.620	0.589
DL-VHD ($\mathcal{L}_{\mathrm{coarse}}$ only)	0.489	0.495	0.571	0.608	0.559
DL-VHD ($\mathcal{L}_{\mathrm{fine}}$ only)	0.486	0.480	0.535	0.564	0.531
DL-VHD (w/o $\mathcal{L}_{\mathrm{distill}}$)	0.525	0.686	0.654	0.630	0.649
DL-VHD (full model)	**0.556**	**0.734**	**0.692**	**0.653**	**0.676**
Target-oracle	0.532	0.772	0.725	0.661	0.762

3. More Visualization Results

Fig. 1 visualizes the highlight prediction results of three approaches on the target category video of two more difficult tasks, *i.e.* *dog* $\rightarrow$ *surfing* and *surfing* $\rightarrow$ *dog*. Compared to source-only and AFN [7], the proposed DL-VHD method can better acquire the concepts about the highlight moments on target video category, *e.g.* the segments describing an athlete surfing on the wave or the ones containing dogs.

*Corresponding author: Bingbing Ni.

(a) dog → surfing

(b) surfing → dog

Figure 1. Highlight predictions of three methods on two cross-category highlight detection tasks, *i.e.* *dog* → *surfing* and *surfing* → *dog*. (Each video segment is represented by its first and last frames.)

References

[1] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 2015.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[3] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 2015.

[4] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018.

[5] Baochen Sun and Kate Saenko. Deep CORAL: correlation alignment for deep domain adaptation. In *ECCV Workshop*, 2016.

[6] Min Sun, Ali Farhadi, and Steven M. Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European Conference on Computer Vision*, 2014.

[7] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *International Conference on Computer Vision*, 2019.