Collaborative Noisy Label Cleaner:
Learning Scene-aware Trailers for Multi-modal Highlight Detection in Movies

Tencent YouTu Lab

Bei Gan, Xiujun Shu*, Ruizhi Qiao*, Haoqian Wu, Keyu Chen, Hanjun Li, Bo Ren

* Corresponding Author

Tencent YouTu Lab

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Abstract

Movie highlights stand out of the screenplay for efficient browsing and play a crucial role on social media platforms. Based on existing efforts, this work has two observations: (1) For different annotators, labeling highlight has uncertainty, which leads to inaccurate and time-consuming annotations. (2) Besides previous supervised or unsupervised settings, some existing video corpora can be useful, e.g., trailers, but they are often noisy and incomplete to cover the full highlights. In this work, we study a more practical and promising setting, i.e., reformulating highlight detection as "learning with noisy labels". This setting does not require time-consuming manual annotations and can fully utilize existing abundant video corpora. First, based on movie trailers, we leverage scene segmentation to obtain complete shots, which are regarded as noisy labels. Then, we propose a Collaborative noisy Label Cleaner (CLC) framework to learn from noisy highlight moments. CLC consists of two modules: augmented cross-propagation (ACP) and multi-modality cleaning (MMC). The former aims to exploit the closely related audio-visual signals and fuse them to learn unified multi-modal representations. The latter aims to achieve cleaner highlight labels by observing the changes in losses among different modalities. To verify the effectiveness of CLC, we further collect a large-scale highlight dataset named MovieLights. Comprehensive experiments on MovieLights and YouTube Highlights datasets demonstrate the effectiveness of our approach. Code has been made available at https: / / github . com / TencentYoutuResearch / HighlightDetection-CLC.

Motivation

Trailers are usually composed with shots sparsely selected from movies to avoid spoilers, and the audience cannot get complete highlight information. Some trailer clips convey the artistic style of the film only and lack movie storylines, disturbing the audience's impressions.In addition, different audiences may be interested in different styles of clips, which makes it challenging to learn highlights from them.

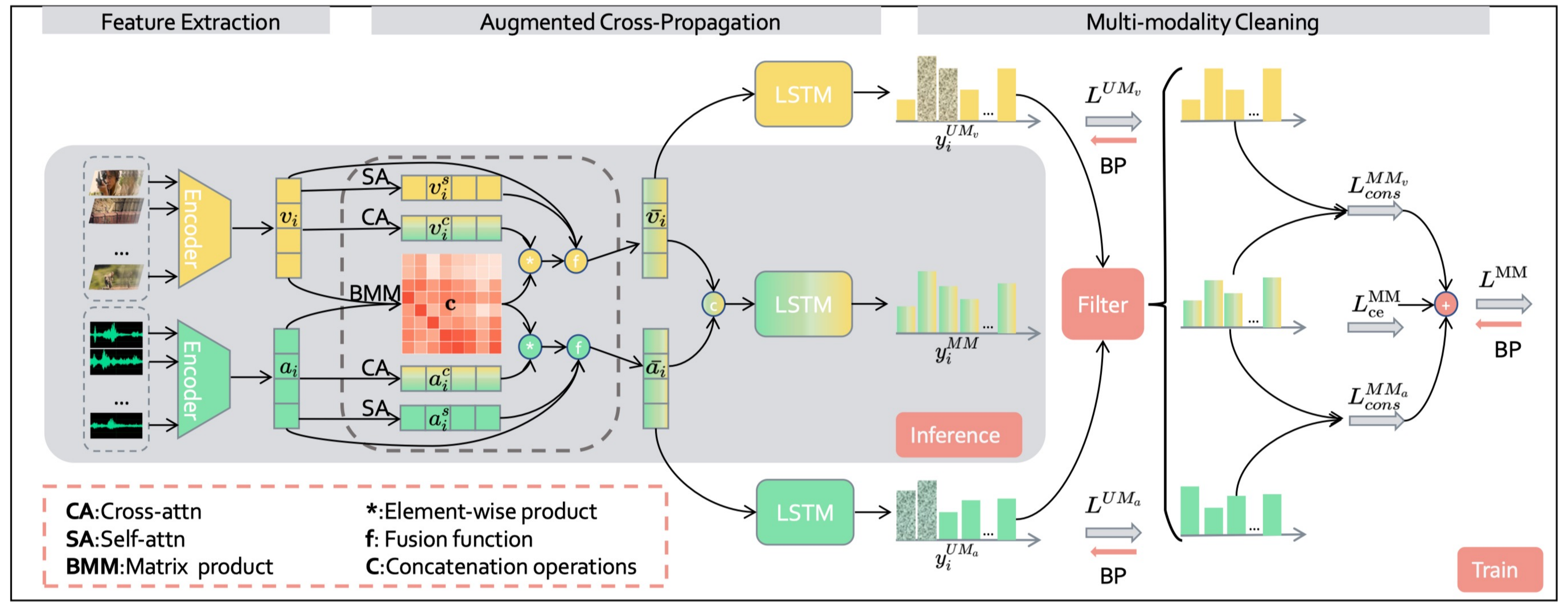
Dataset Summary

	Train	Test
Movie Number	144	30
Avg Durations per Movie	2.19h	2.14h
Avg Shot Number per Movie	1852	1940
Avg Scene Number per Movie	207	193
Annotator1 Positive sample Proportion	-	0.27
Annotator2 Positive sample Proportion	-	0.30
Positive sample proportion	0.35	0.21

We construct MovieLights, a Movie Highlight Detection Dataset. MovieLights contains 174 movies and the highlight moments are all from officially released trailers.

To collect a large amount of training data efficiently, we introduce a scene-aware paradigm to obtain the highlight moments label without any manual annotation. Since the trailer shots may contain some less important moments, the acquired highlight labels are still noisy.

Approach



CLC includes three modules.The visual and audio modalities of the input video are represented as vectors by the feature extraction module. Then the features are augmented by ACP module to capture semantic associations acorss modalities. MMC is used to filter outs noisy and incomplete labeling with additional uni-modal branches.

Experimental Results

Results on MovieLights.

Methods	Modality	mAP
GIFs [14]	V	25.48
SL-Module [50]	V	32.34
SL-Module [50]	VA	34.27
UMT [29]	VA	38.7
CLC–	VA	39.65
CLC– w/ SCE [44]	VA	39.83
CLC– w/ LS [40]	VA	40.49
CLC	VA	**43.88**

Results on YouTube Highlights.

Methods	dog	gym.	park.	ska.	ski.	surf.	Avg.
GIFs [14]	30.8	33.5	54	55.4	32.8	54.1	46.4
LSVM [37]	60.0	41.0	61.0	62.0	36.0	61.0	53.6
HighlightMe [5]	63	73	72	64	52	62	64
MINI-Net [16]	58.2	61.7	70.2	72.2	58.7	60.1	64.4
CHD [1]	60.6	71.1	74.2	49.8	68.2	68.5	65.4
Trail [43]	63.3	82.5	62.3	52.9	74.5	79.3	69.1
SL-Module [50]	70.8	53.2	77.2	72.5	66.1	76.2	69.3
Joint-VA [2]	64.5	71.9	80.8	62	73.2	78.3	71.8
PLD [45]	74.9	70.2	77.9	57.5	70.7	79	73
CO-AV [27]	60.9	66	89	74.1	69	81.1	74.7
UMT [29]	65.9	75.2	81.6	71.8	72.3	82.7	74.9
CLC(ours)	70.5	79.4	83.9	83.5	79.5	83.6	**80.1**

CLC outperforms the baseline methods by a notable margin on MovieLights and YouTube Highlights

Results on YouTube Highlights with Noisy Label.

Annotation	Noise	Methods	dog	gym.	park.	ska.	ski.	surf.	Avg.
Harvested matched	clean	UMT [29]	65.90	75.20	81.60	71.80	72.30	82.70	74.90
Harvested matched	clean	CLC	70.51	79.43	83.85	83.51	79.46	83.56	80.05 (↑ 5.15)
Harvested borderline	slight noise	UMT [29]	65.93	74.31	81.58	71.84	70.24	82.46	74.39
Harvested borderline	slight noise	CLC	69.41	80.73	78.50	85.36	81.11	83.16	79.71 (↑ 5.32)
Mturk	severe noise	UMT [29]	63.78	76.16	75.02	73.62	69.99	81.59	73.36
Mturk	severe noise	CLC	66.92	80.44	85.92	82.33	78.05	81.72	79.22 (↑ 5.86)

As the noise level increases, the VHD task becomes more difficult, but the performance superiority of our CLC over UMT becomes even more obvious.

Visualization

The highlight moments selected by our CLC