基于压缩视频学习的端到端通用事件边界检测

End-to-End Compressed Video Representation Learning for Generic Event Boundary Detection

Congcong Li, Tiejian Luo, Libo Zhang, Yanjun Wu

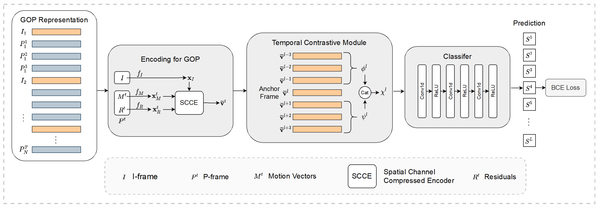
CVPR 2022，张立波 18655882017

Motivation

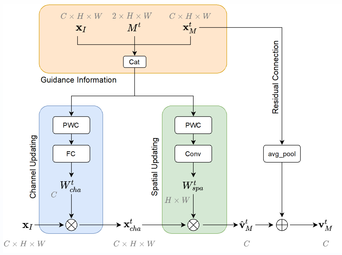
Existing methods for GEBD typically require video frames to be decoded before feeding into the network, which demands considerable computational power and storage space. To that end, we propose a new end-to-end compressed video representation learning for event boundary detection that leverages the rich information in the compressed domain. Our method can save inference time in both decoding and model forwarding stages.

Method

➢ Framework



➢ Spatial-Channel Compressed Encoder(SCCE)



The SCCE module integrates the features of the reference I-frame when computing the features of P-frames in both channel and spatial dimen-sions. This module projects I-frames and P-frames into the same feature space.

➢ Temporal Contrastive Module

we compute the contrastive features before and after the candidate boundary frames in the temporal domain. The simple linear weighted summation can be efficiently implemented using the 1D convolutional operation.

$$\phi^l = \sum_{j=1}^{k} W_j \cdot \tilde{\mathbf{v}}^{l-j}$$

➢ Loss Function

we use the Gaussian kernel to preprocess the ground-truth event boundaries to obtain the soft labels instead of using the "hard labels" of boundaries.

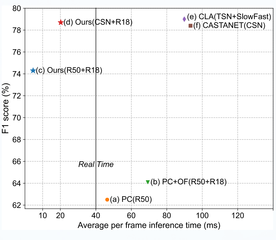
$$g_i^l = \exp\left(-\frac{(l-i)^2}{2\alpha^2}\right)$$

Experiments

➢ Results on Kinetics-GEBD validation set

Ours vs. PC + Optical Flow
Accuracy F1@0.05 : ↑ 9.7%
Speed 1.12ms vs. 3.23ms (per frame)

Rel.Dis. Threshold	0.05	0.1	0.15	0.45	0.5	avg
BMN [23]	0.186	0.204	0.213	0.239	0.241	0.223
BMN-StartEnd [28]	0.491	0.589	0.627	0.681	0.683	0.640
TCN-TAPOS [28]	0.464	0.560	0.602	0.682	0.687	0.627
TCN [21]	0.588	0.657	0.679	0.710	0.712	0.685
PC [28]	0.625	0.758	0.804	0.867	0.870	0.817
PC + Optical Flow	0.646	0.776	0.818	0.877	0.879	0.830
Ours	**0.743**	**0.830**	**0.857**	**0.896**	**0.898**	**0.865**



➢ Speed Comparison

Our method achieves competitive F1 score with extremely fast running speed by directly leveraging motion vectors and residuals in the compressed domain.

➢ Results on HMDB-51 and UCF-101

Results on UCF101 and HMDB51 show the effectiveness of our method.

	HMDB-51	UCF-101
CoViAR [45]	59.1	90.4
DMC-Net(ResNet-18) [29]	62.8	90.9
DMC-Net(I3D) [29]	71.8	92.3
Ours (ResNet-18)	**63.3**	**91.0**
Ours (I3D)	**72.1**	**92.5**

Summary

- We propose an fully end-to-end compressed video representation learning method for GEBD.
- We achieves competitive F1 score with extremely fast running speed by directly leveraging motion vectors and residuals in the compressed domain.
- Our method also achieves competitive results comparing with the state-of-the-art methods in the action recognition task.