Hierarchical Multimodal Attention for Deep Video Summarization

Melissa Sanabria[1,2,3]; Frédéric Precioso[1,2,3]; Thomas Menguy[4]
[1]Université Cote d'Azur, [2]I3S Laboratory, [3]Maasai, [4]Wildmoka Company

Abstract

This paper explores the problem of summarizing professional soccer matches as automatically as possible using both event-stream data collected from the field and the content broadcasted on TV. We have designed an architecture, introducing first (1) a Multiple Instance Learning method that considers the sequential dependency among events and then (2) a hierarchical multimodal attention layer that grasps the importance of each event in an action. We evaluate our approach on matches from two professional European soccer leagues, showing its capability to identify the best actions for automatic summarization by comparing with real summaries made by human operators.

Our Approach

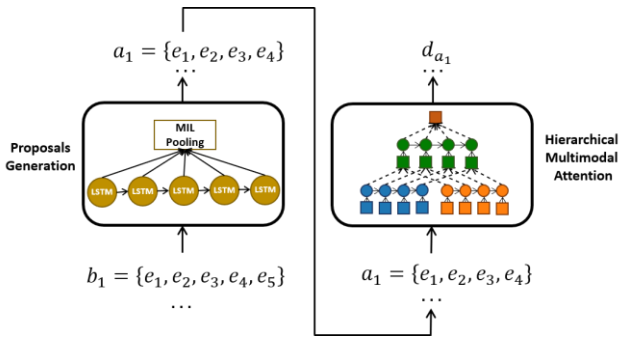
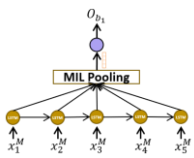


Figure 1. General schema of our approach

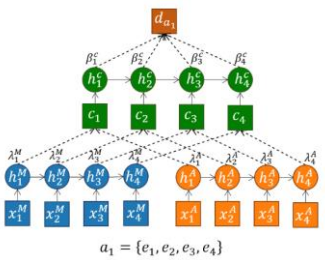
LSTM MIL Pooling

We tackle the similarity of inter-categorical actions with a Multiple Instance Learning (MIL) approach. We propose an LSTM network followed by a MIL Pooling to get the bag representation.



Hierarchical Multimodal Attention

We propose a hierarchical multimodal attention mechanism that in the first stage learns the importance of each modality at the event level and in the second stage learns the importance of each event inside the action



Results

Method	Missing Intervals	Recall
SST	39.79	60.11
MI-Net	18.62	81.33
MI-Net Attention	16.07	83.89
LSTM MIL Pooling	13.01	86.96

Table I. Performance comparison of Proposal Generation methods.

Method	Missing Intervals	F-score
Sanabria et al.	47.95	64.30
Naive Fusion	36.19	71.23
Hori et al.	32.99	72.03
Ours	27.38	74.09

Table II. Performance comparison of Multimodal Attention methods

Method	Precision	Recall	F-score
Only Goals	99.55	28.29	44.18
All Shots-on-Target	40.77	75.71	52.99
Random	41.87	48.72	45.03
Ours	75.46	72.76	74.09

Table III. Performance comparison of Soccer Baselines

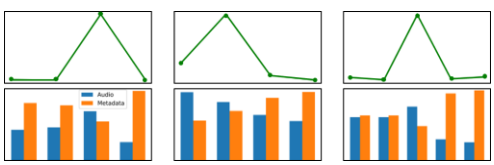


Figure 2. Examples of attention in different actions learned by our model. Blue and orange represent the audio and the metadata respectively.

Sanabria et al. (2019). A Deep Architecture for Multimodal Summarization of Soccer Games. ACM Sports.
Hori et al. (2017). Attention-based multimodal fusion for video description. ICCV.
Buch et al. (2017). Sst: Single-stream temporal action proposals. CVPR.
Wang et al. (2018). Revisiting multiple instance neural networks. *Pattern Recognition*, *74*, 15-24.
Ilse et al. (2018). Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*.