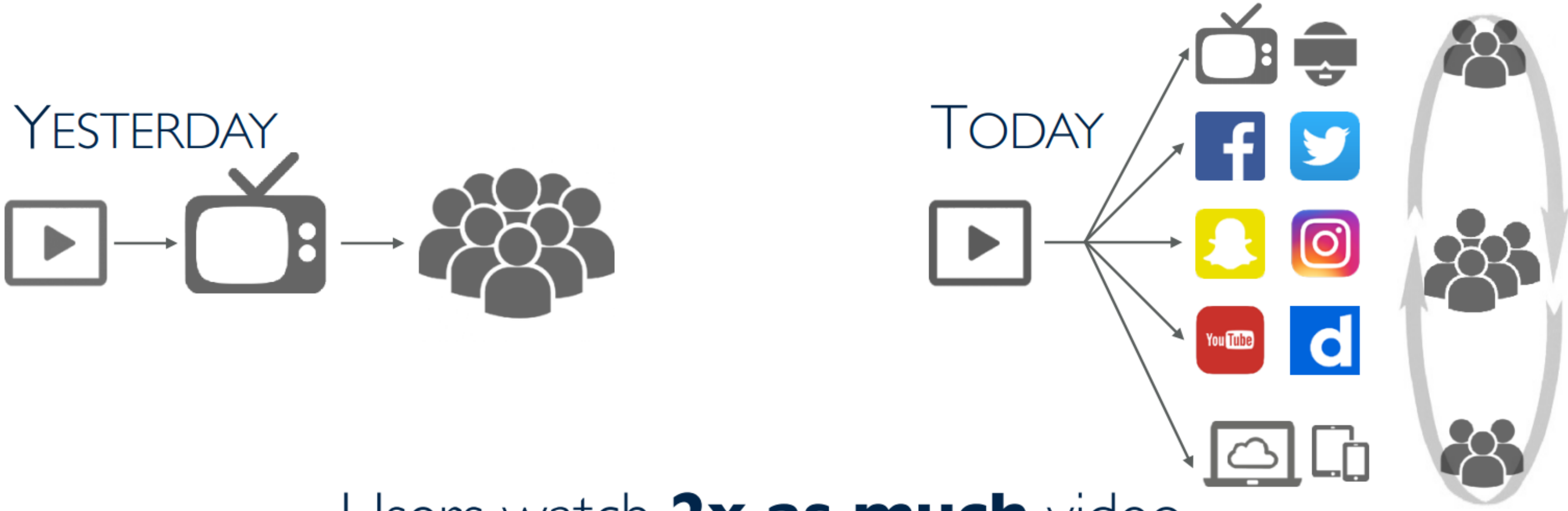
Hierarchical Multimodal Attention for Deep Video Summarization

Melissa Sanabria, Frédéric Precioso, Thomas Menguy

ICPR 2020 – 25th International Conference on Pattern Recognition

Video Consumption evolves

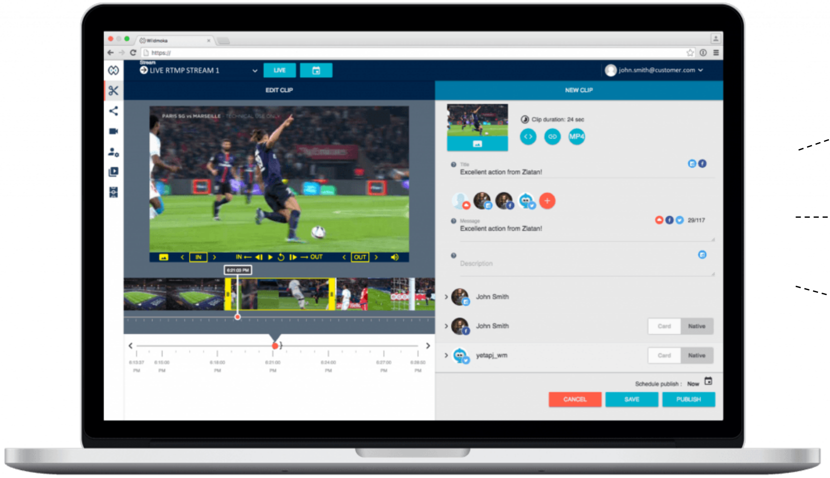


YESTERDAY

TODAY

Users watch **2x as much** video
on **10x as many** platforms

WILDMOKA



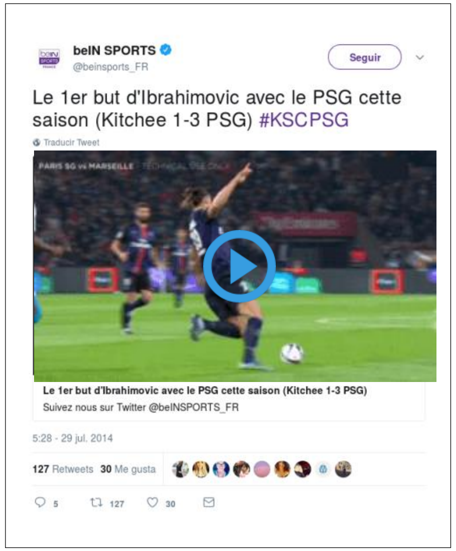
Clip duration: 24 sec

Excellent action from Zlatan!

Excellent action from Zlatan!

John Smith

John Smith



beIN SPORTS
@beinsports_FR

Le 1er but d'Ibrahimovic avec le PSG cette saison (Kitchee 1-3 PSG) #KSCPSG

Traducir Tweet

Le 1er but d'Ibrahimovic avec le PSG cette saison (Kitchee 1-3 PSG)
Suivez nous sur Twitter @beINSPORTS_FR

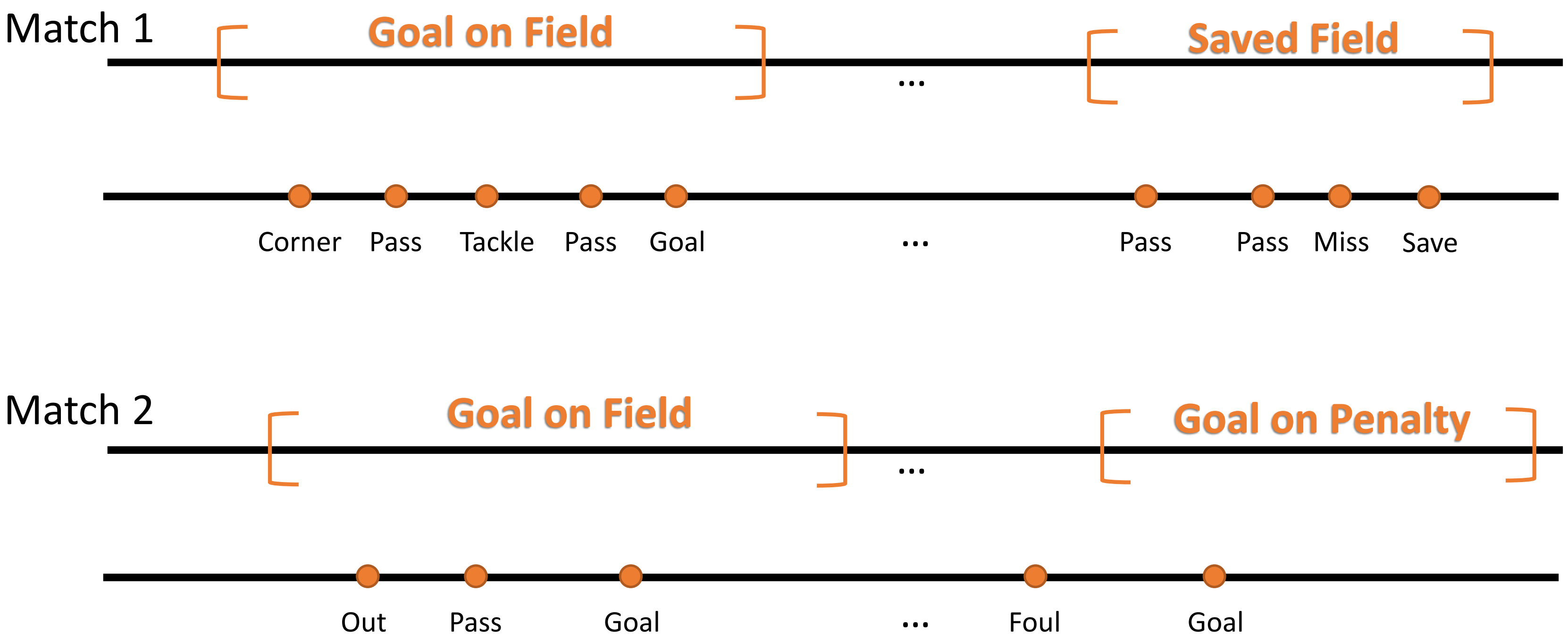
5:28 - 29 jul. 2014

127 Retweets 30 Me gusta

5 127 30

-
-
-

Video Actions - Events



Match 1

Goal on Field … Saved Field

Corner Pass Tackle Pass Goal … Pass Pass Miss Save

Match 2

Goal on Field … Goal on Penalty

Out Pass Goal … Foul Goal

Video Actions

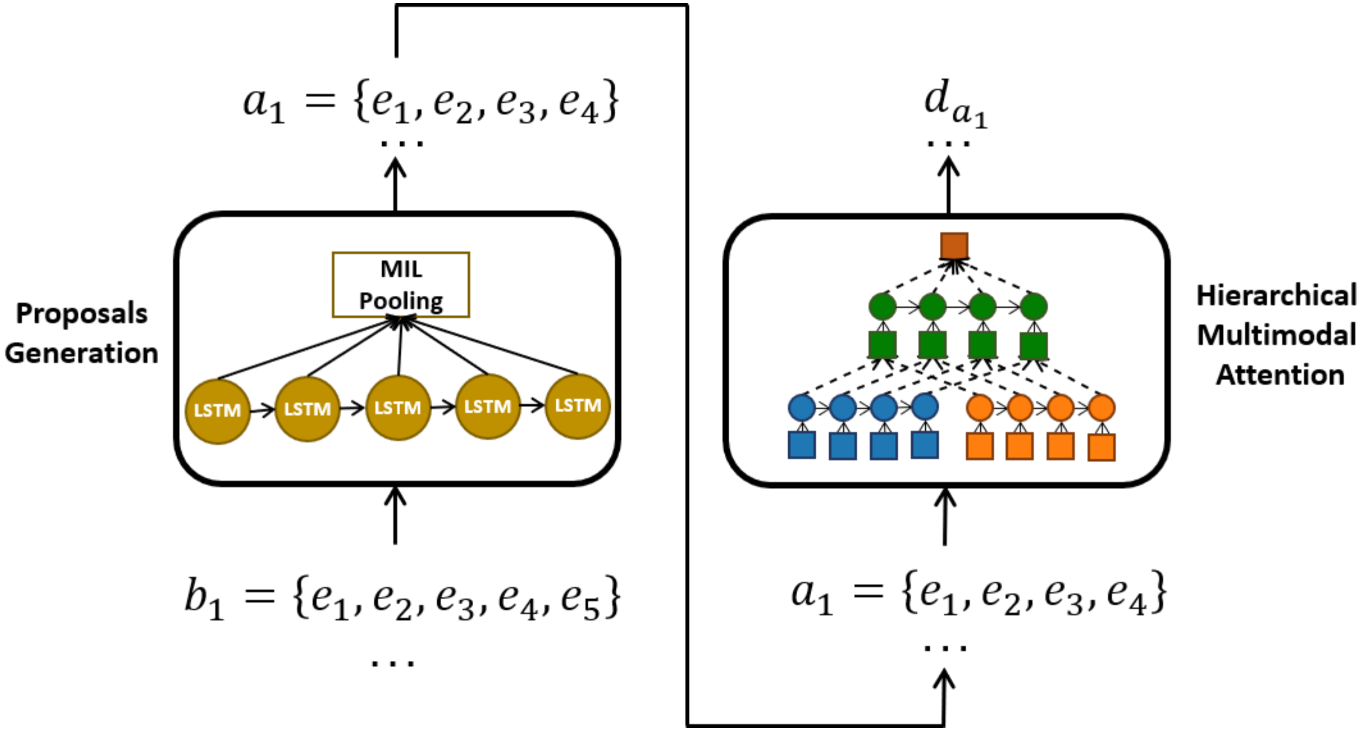
Video annotations made by human operators from broadcasted videos

Events

opta sportradar wyscout

InStat CATAPULT METRICA SPORTS

Our Approach


$a_1 = \{e_1, e_2, e_3, e_4\}$
$\cdots$
MIL Pooling
Proposals Generation
LSTM
LSTM
LSTM
LSTM
LSTM
$b_1 = \{e_1, e_2, e_3, e_4, e_5\}$
$\cdots$
d_{a_1}
$\cdots$
Hierarchical Multimodal Attention
$a_1 = \{e_1, e_2, e_3, e_4\}$
$\cdots$

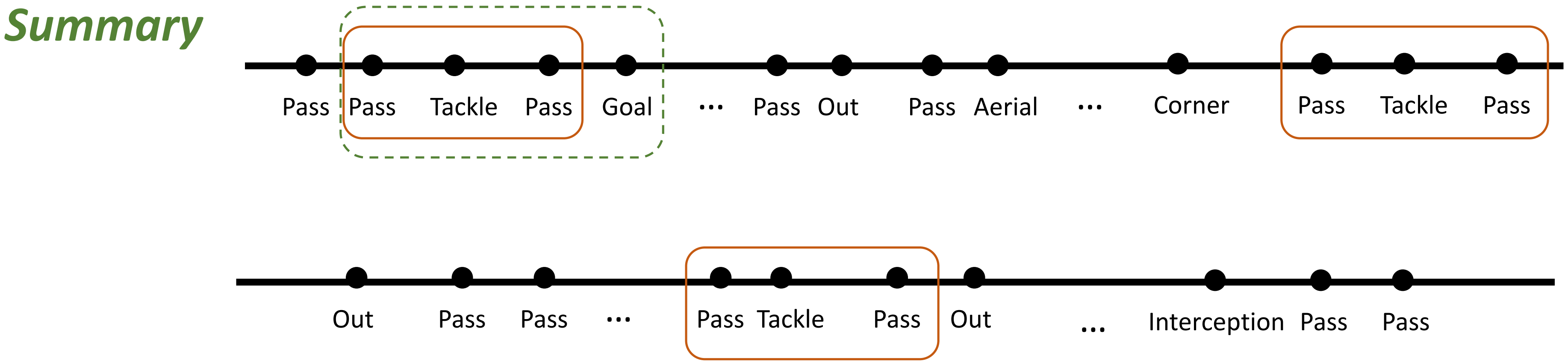
Proposals Generation

Proposal
Parts (consecutive relevant events) of the match that might belong to the summary

Inspired by:

- Object Detection: Region Proposal Network, Faster RCNN, ..
- Action Detection: SST, R-C3D, …

Proposals Generation

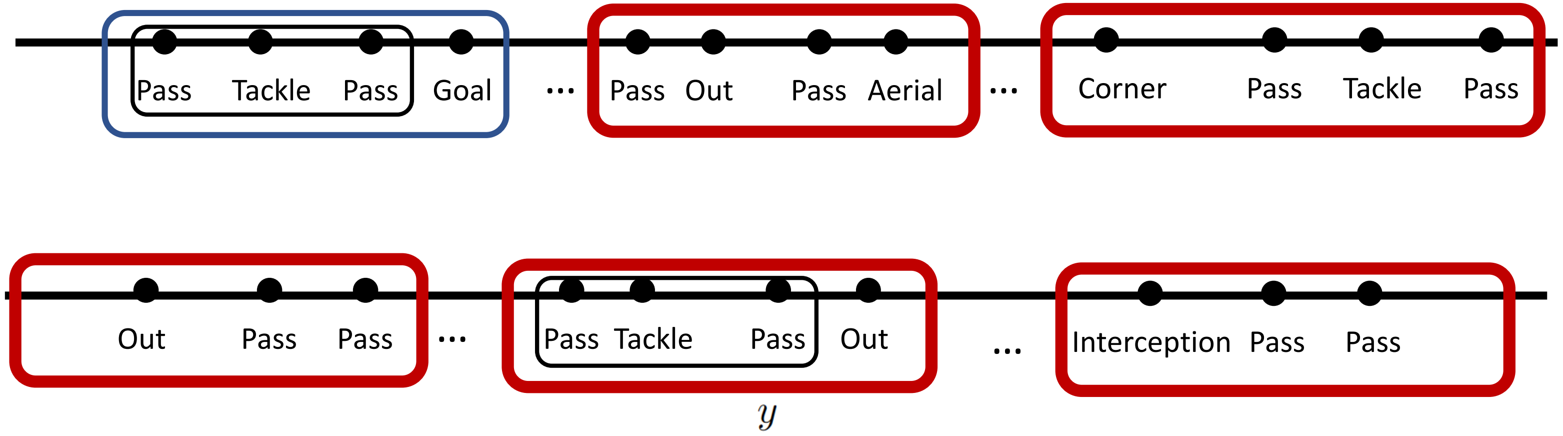


Similarity of inter-categorical actions

Very similar sets of events belong to different classes

Proposals Generation

MIL: Multiple Instance Learning



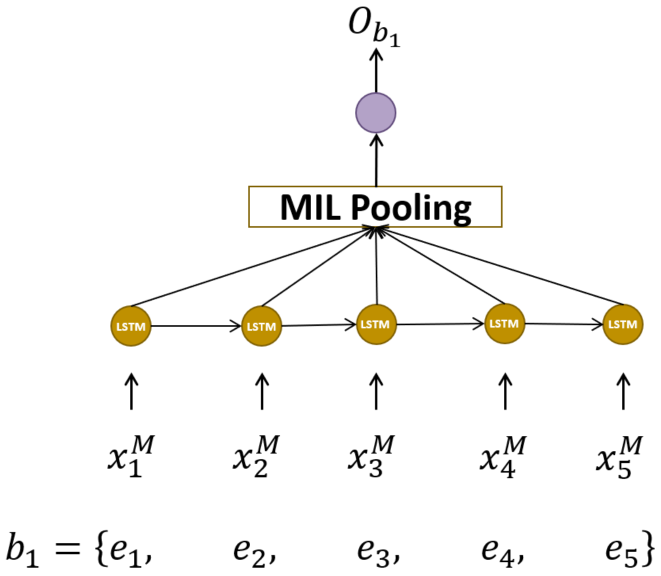
$$Y = \begin{cases} +1 & \text{if} \quad \exists y_i : \; y_i = +1; \\ -1 & \text{if} \quad \forall y_i : \; y_i = -1. \end{cases}$$

where Y is the label of a bag and
is the label of the instance

Negative: All the instances inside the bag are negative

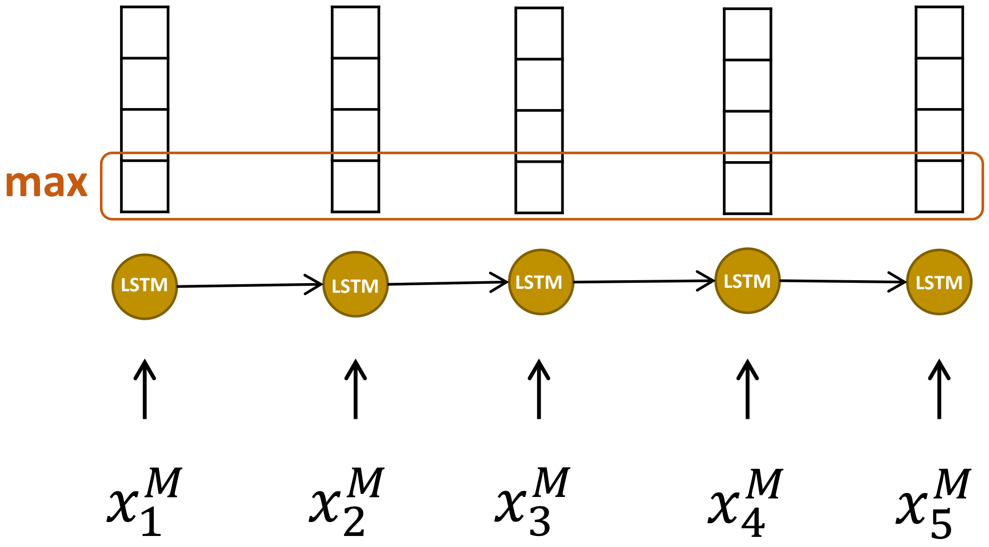
Positive: If there is at least on instance inside the bag which is positive

LSTM MIL Pooling



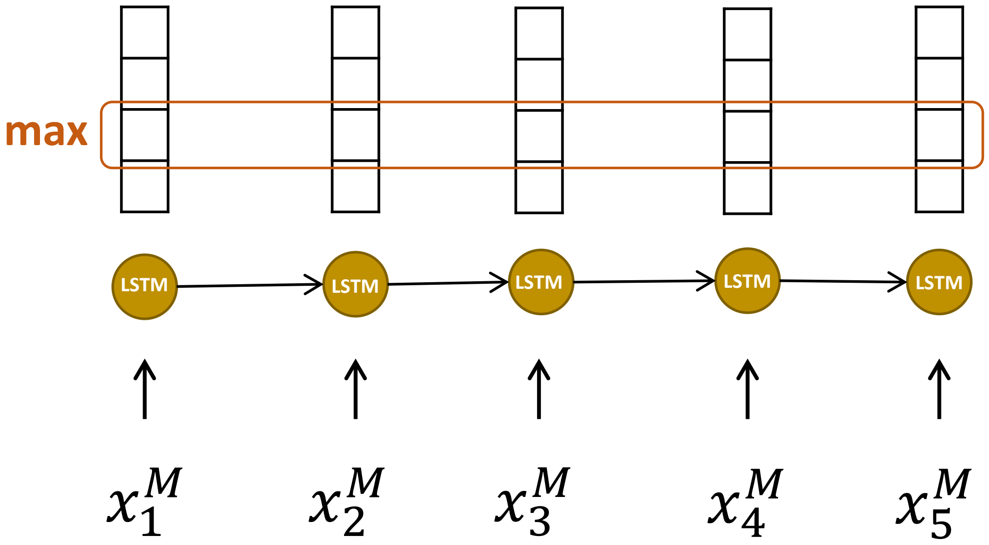
- Traditional MIL paradigm **assumes neither ordering nor dependency** of instances within a bag
- However, the selection of an action to be part of a summary **is highly dependent on the sequence of its events**

LSTM MIL Pooling



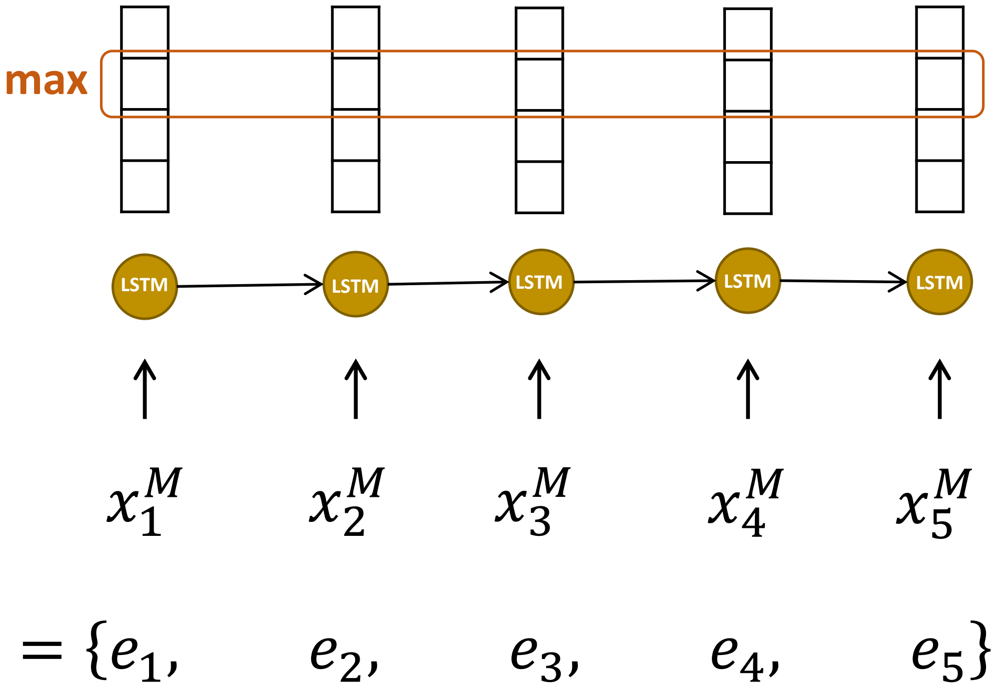
$$b_1 = \{e_1, \quad e_2, \quad e_3, \quad e_4, \quad e_5\}$$

LSTM MIL Pooling

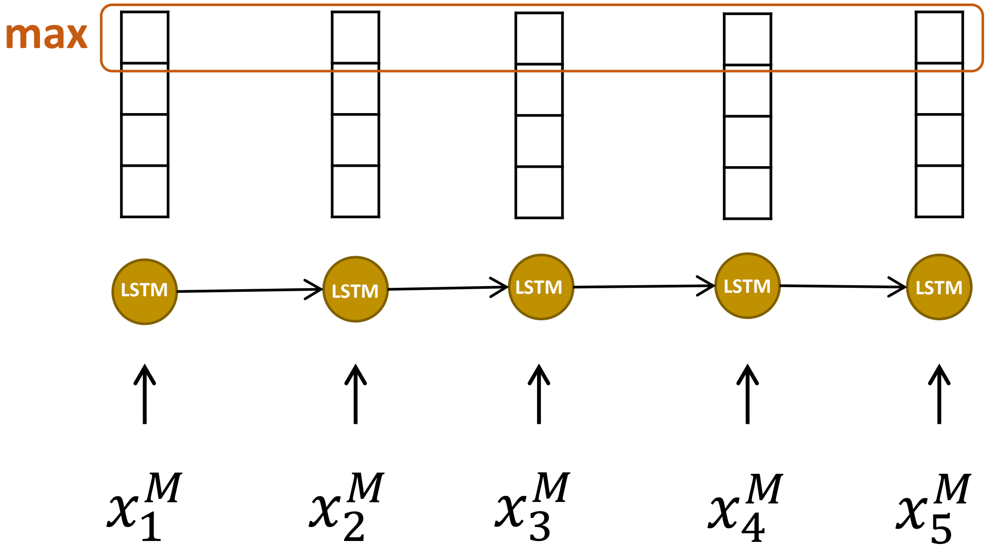


$$b_1 = \{e_1, \quad e_2, \quad e_3, \quad e_4, \quad e_5\}$$

LSTM MIL Pooling

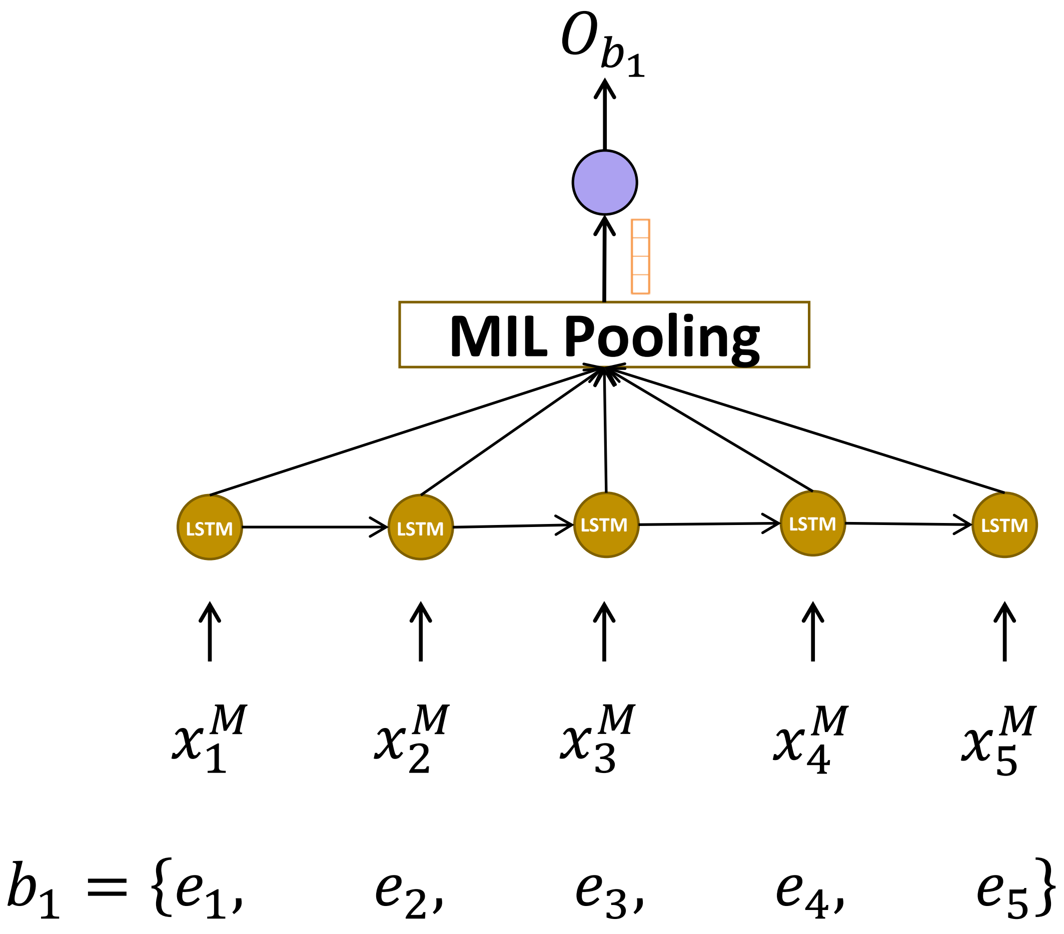


LSTM MIL Pooling



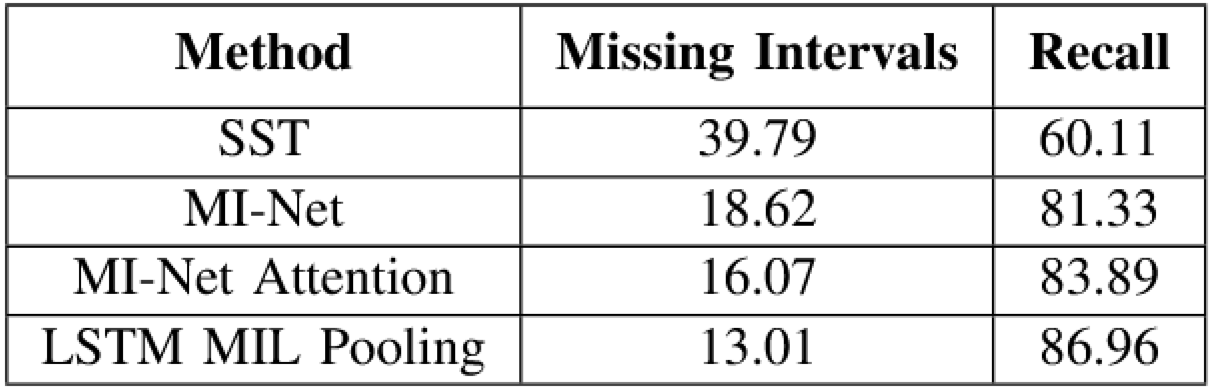
$$b_1 = \{e_1, \quad e_2, \quad e_3, \quad e_4, \quad e_5\}$$

LSTM MIL Pooling



LSTM MIL Pooling

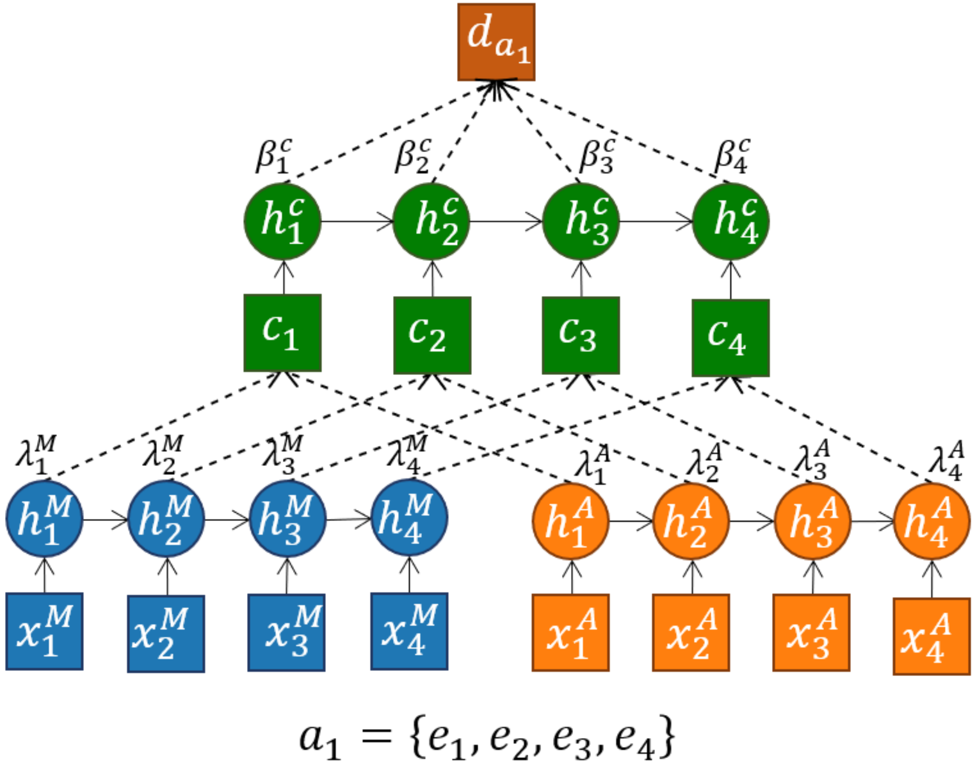
Comparison with State of the Art



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

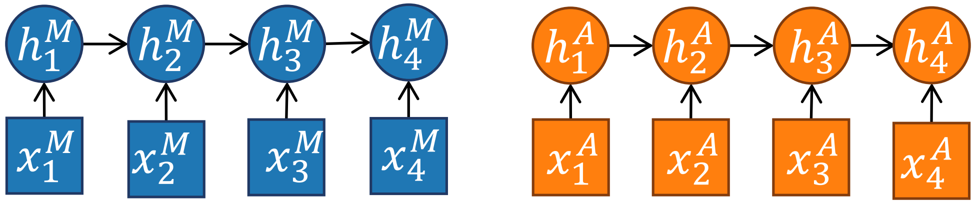
SST: Buch, S., Escorcia, V., Shen, C., Ghanem, B., & Carlos Niebles, J. (2017). Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 2911-2920).
MI-Net: Wang, X., Yan, Y., Tang, P., Bai, X., & Liu, W. (2018). Revisiting multiple instance neural networks. *Pattern Recognition*, *74*, 15-24.
MI-Net Attention: Ilse, M., Tomczak, J. M., & Welling, M. (2018). Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*.

Summarization:
Hierarchical Multimodal Attention



Hierarchical Multimodal Attention

First Stage

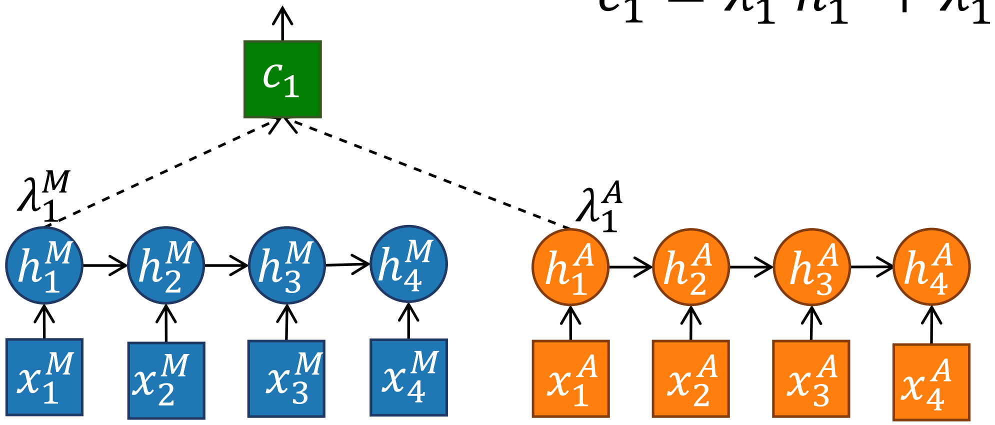


$$a_1 = \{e_1, e_2, e_3, e_4\}$$

Hierarchical Multimodal Attention

First Stage

$$c_1 = \lambda_1^M h_1^M + \lambda_1^A h_1^A$$

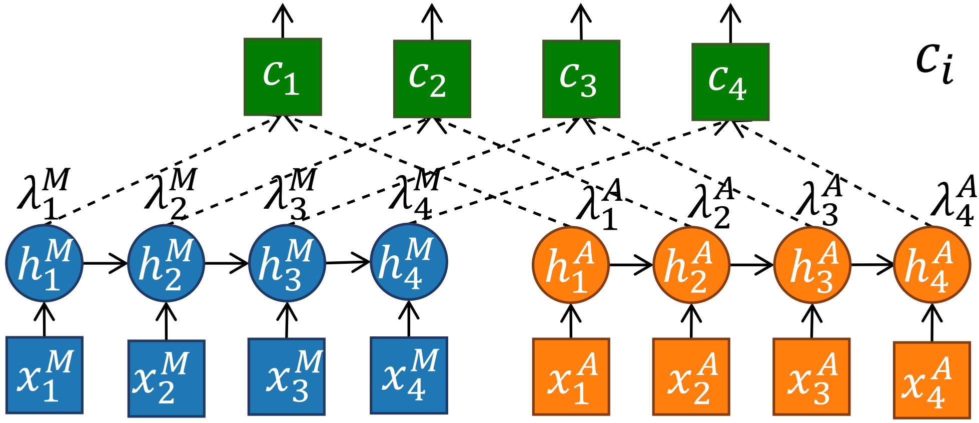


$$a_1 = \{e_1, e_2, e_3, e_4\}$$

Hierarchical Multimodal Attention

First Stage

Learn the importance of each modality at the event level



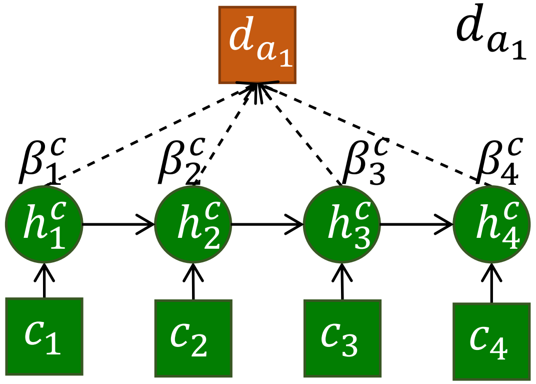
$$c_i = \lambda_i^M h_i^M + \lambda_i^A h_i^A$$

$$a_1 = \{e_1, e_2, e_3, e_4\}$$

Hierarchical Multimodal Attention

Second Stage

Learn the importance of each event inside the action

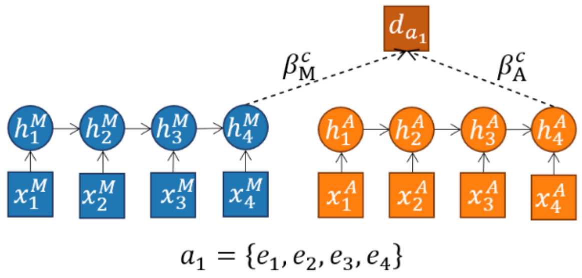


$$d_{a_1} = \beta_1^c h_1^c + \beta_2^c h_2^c + \beta_2^c h_2^c + \beta_2^c h_2^c$$

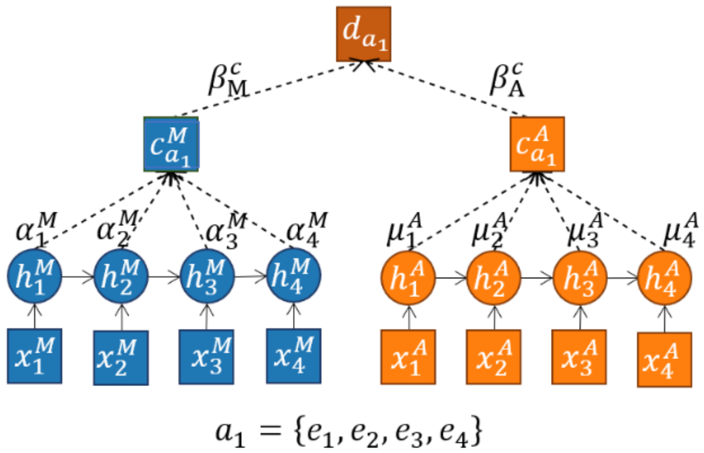
$$a_1 = \{e_1, e_2, e_3, e_4\}$$

Hierarchical Multimodal Attention

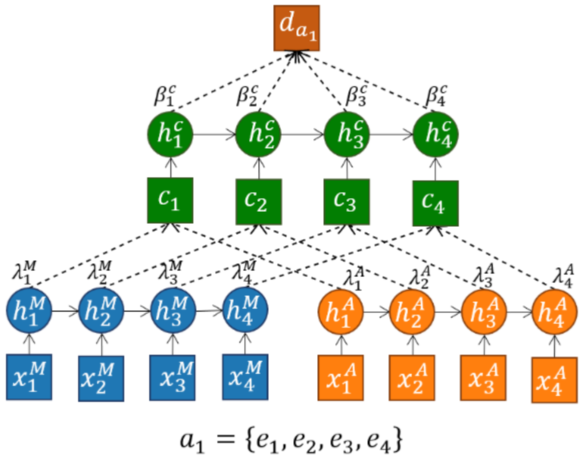
Comparison with the State of the Art


d_{a_1}
β^c_M
β^c_A
h^M_1 h^M_2 h^M_3 h^M_4
h^A_1 h^A_2 h^A_3 h^A_4
x^M_1 x^M_2 x^M_3 x^M_4
x^A_1 x^A_2 x^A_3 x^A_4
$a_1 = \{e_1, e_2, e_3, e_4\}$

Naive


d_{a_1}
β^c_M
β^c_A
$c^M_{a_1}$
$c^A_{a_1}$
α^M_1 α^M_2 α^M_3 α^M_4
μ^A_1 μ^A_2 μ^A_3 μ^A_4
h^M_1 h^M_2 h^M_3 h^M_4
h^A_1 h^A_2 h^A_3 h^A_4
x^M_1 x^M_2 x^M_3 x^M_4
x^A_1 x^A_2 x^A_3 x^A_4
$a_1 = \{e_1, e_2, e_3, e_4\}$

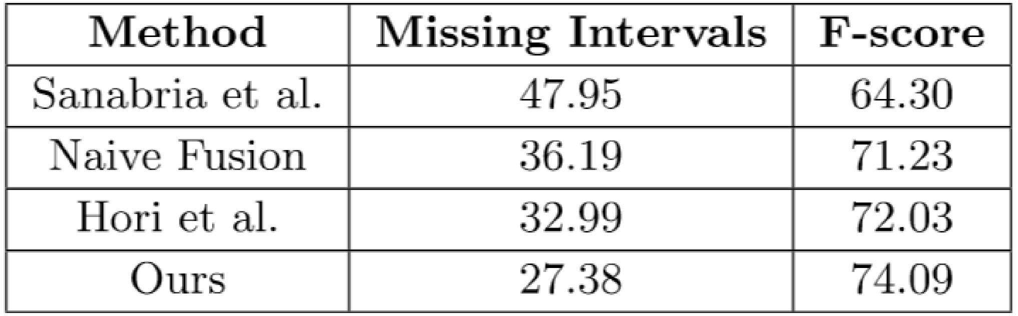
Hori et al.


d_{a_1}
β^c_1 β^c_2 β^c_3 β^c_4
h^c_1 h^c_2 h^c_3 h^c_4
c_1 c_2 c_3 c_4
λ^M_1 λ^M_2 λ^M_3 λ^M_4
λ^A_1 λ^A_2 λ^A_3 λ^A_4
h^M_1 h^M_2 h^M_3 h^M_4
h^A_1 h^A_2 h^A_3 h^A_4
x^M_1 x^M_2 x^M_3 x^M_4
x^A_1 x^A_2 x^A_3 x^A_4
$a_1 = \{e_1, e_2, e_3, e_4\}$

Ours

Hierarchical Multimodal Attention

Comparison with the State of the Art



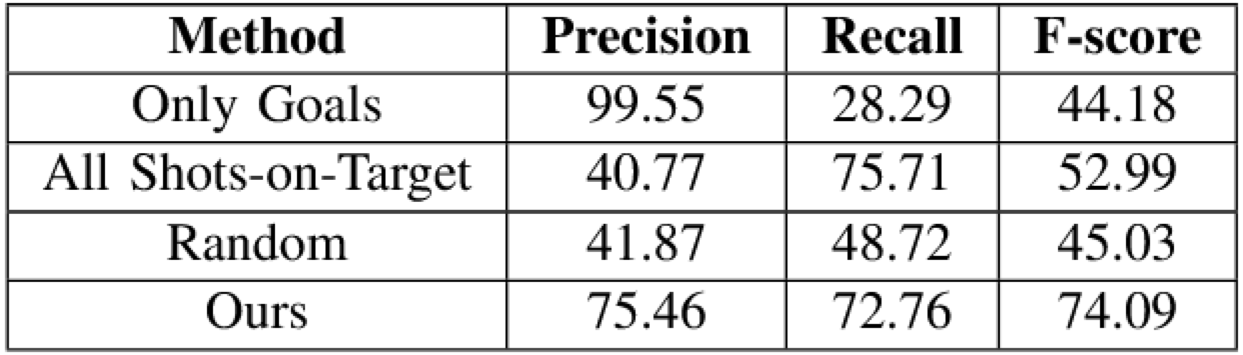
Method	Missing Intervals	F-score
Sanabria et al.	47.95	64.30
Naive Fusion	36.19	71.23
Hori et al.	32.99	72.03
Ours	27.38	74.09

Sanabria et al: Sanabria, M., Precioso, F., & Menguy, T. (2019, October). A Deep Architecture for Multimodal Summarization of Soccer Games. In Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports (pp. 16-24). ACM.

Hori et al: Hori, C., Hori, T., Lee, T. Y., Zhang, Z., Harsham, B., Hershey, J. R., ... & Sumi, K. (2017). Attention-based multimodal fusion for video description. In Proceedings of the IEEE international conference on computer vision (pp. 4193-4202).

Hierarchical Multimodal Attention

Comparison with Soccer Baselines



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

Thank you!

Hierarchical Multimodal Attention for Deep Video Summarization

Melissa SANABRIA

sanabria@unice.fr

Région
Provence
Alpes
Côte d'Azur

UNIVERSITÉ
CÔTE D'AZUR

i3S
sophia antipolis

WILDMOKA

Maasai