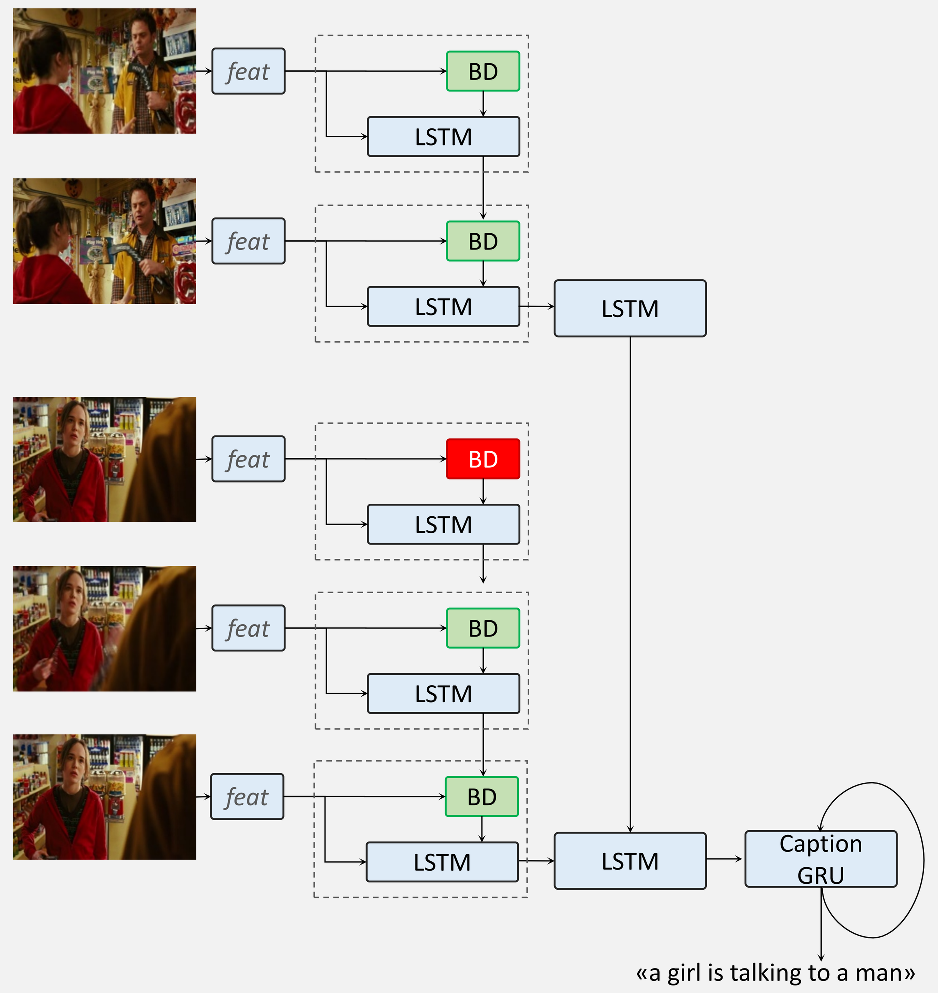
Hierarchical Boundary-Aware Neural Encoder for Video Captioning

Lorenzo Baraldi, Costantino Grana and Rita Cucchiara
Dipartimento di Ingegneria “Enzo Ferrari” – Università degli Studi di Modena e Reggio Emilia

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

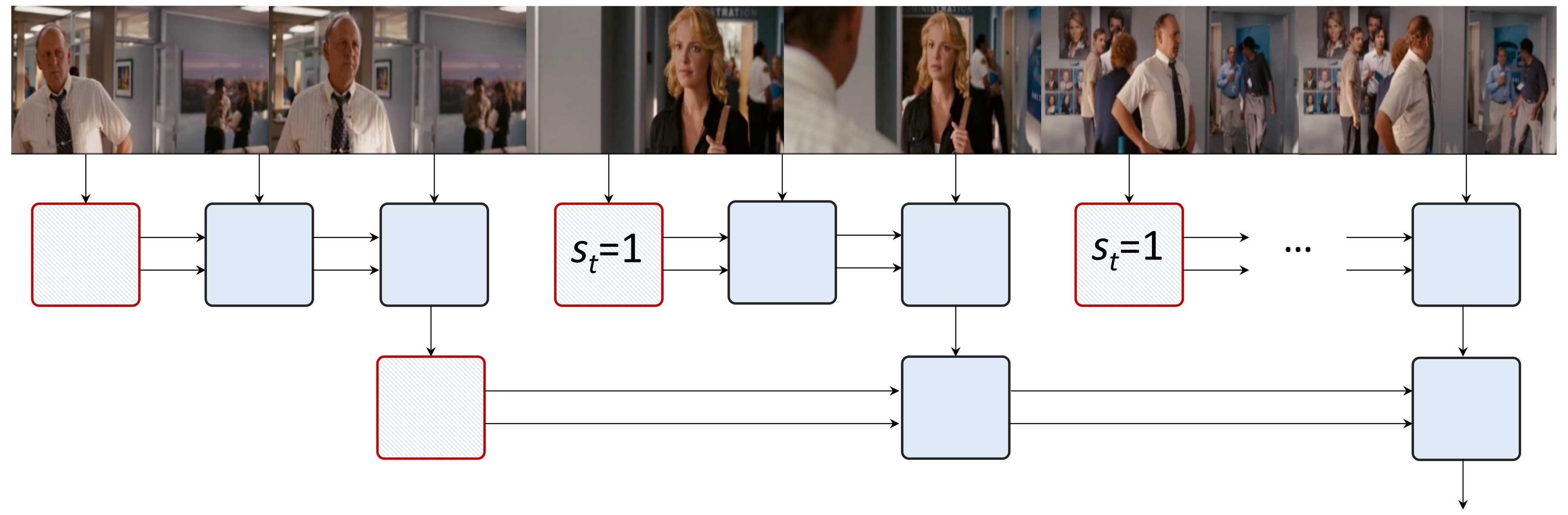
CVPR July 21-26 HONOLULU 2017

LSTMs for Video Captioning

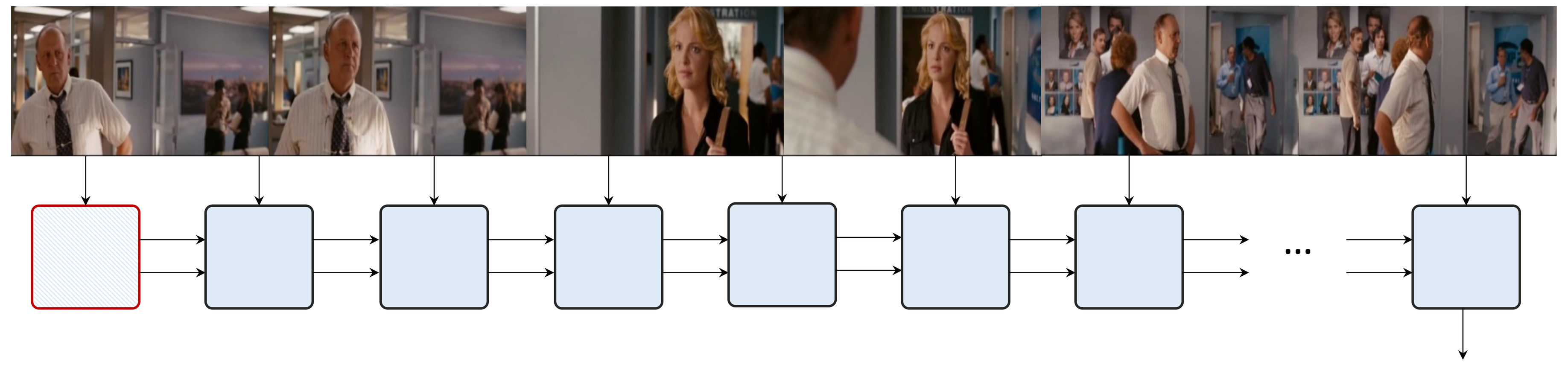


Recurrent networks are a popular choice as video encoders for captioning.
However, they can not optimally deal with long video sequences, especially when they have a layered structure.

The memory of the LSTM mixes representations computed while attending at different actions and appearances.



(a) Boundary-Aware Video Encoder



(b) Classic LSTM Video Encoder

Experimental Results

Performance improvements on movie description datasets over 1- and 2-layers LSTM encoders and when forcing the boundary detector to fire on camera changes or after small video chunks.

Model	METEOR
SA-GoogleNet+3D-CNN [1]	4.1
S2VT-RGB(VGG) [2]	6.7
HRNE with attention [3]	6.8
LSTM encoder (C3D+ResNet)	6.7
Double-layer LSTM encoder (C3D+ResNet)	6.7
Boundary encoder on shots	7.1
Boundary-aware encoder (C3D+ResNet)	**7.3**

(a) M-VAD dataset

Model	CIDEr	B@4	R_L	M
SMT (best variant) [4]	8.1	0.5	13.2	5.6
Rohrbach *et al.* [5]	10.0	**0.8**	16.0	**7.0**
LSTM encoder (C3D+ResNet)	10.5	0.7	16.1	6.4
Double-layer LSTM encoder (C3D+ResNet)	10.6	0.6	16.5	6.7
Boundary encoder on shots	10.3	0.7	16.3	6.6
Boundary-aware encoder (C3D+ResNet)	**10.8**	**0.8**	**16.7**	**7.0**

(b) MPII-MD dataset

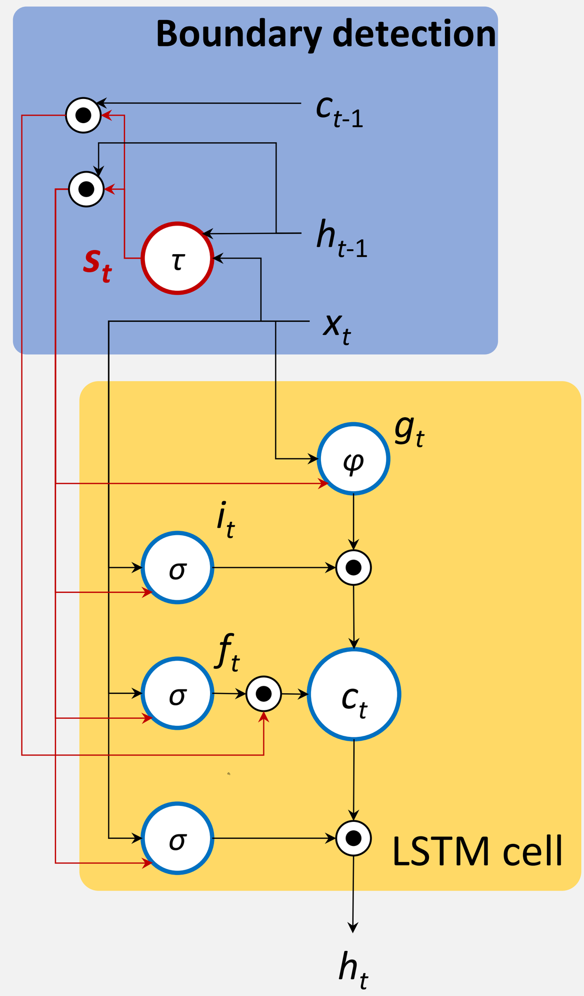
Boundary-Aware Cell

A video encoding cell capable of identifying discontinuity points and modify the layer connectivity through time.

$$s_t = \tau(\mathbf{v}_s^T \cdot (W_{si}\mathbf{x}_t + W_{sh}\mathbf{h}_{t-1} + \mathbf{b}_s))$$

$$\tau(x) = \begin{cases} 1, & \text{if } \sigma(x) > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

During training: stochastic version of the step function in the forward pass, and a differentiable estimator in the backward pass.



Connectivity through time

When a boundary is estimated, the hidden state and memory cell are reinitialized, and the previous hidden state is given to the output, as a summary of the detected segment.

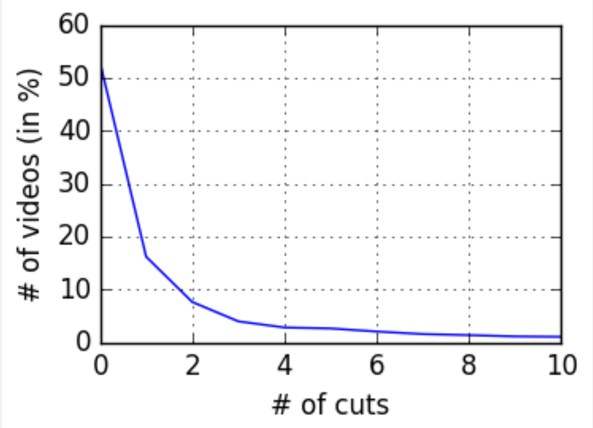
$$\mathbf{h}_{t-1} \leftarrow \mathbf{h}_{t-1} \cdot (1 - s_t)$$
$$\mathbf{c}_{t-1} \leftarrow \mathbf{c}_{t-1} \cdot (1 - s_t)$$

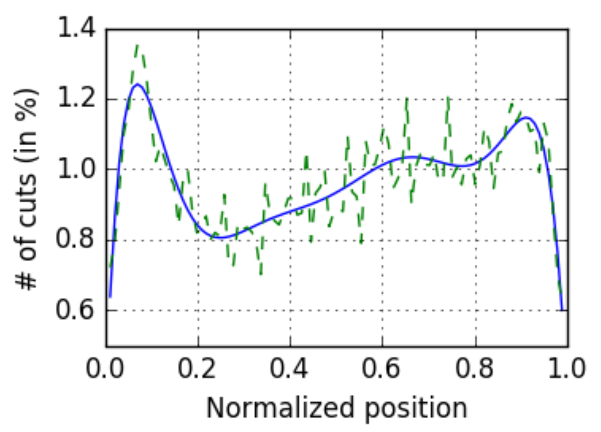
The connectivity schema of the layer is thought as an activation rather than as a non-learnable hyperparameter.

Then, a second recurrent layer encodes this variable-length representation into a feature vector for the overall video.

Analysis of learned boundaries

Video are split in large, very significant chunks, some corresponding to camera changes and others to more soft action or appearance boundaries.





Also, boundaries help to tackle alignment defects in the groundtruth.

References

[1] L. Yao, et al. Describing videos by exploiting temporal Structure, CVPR 2015
[2] S. Venugopalan, et al. Sequence to sequence-video to text, CVPR 2015
[3] P. Pan, et al. Hierarchical recurrent neural encoder for video representation with application to captioning, CVPR 2016
[4] A. Rohrbach, et al. A dataset for movie description, CVPR 2015
[5] A. Rohrbach, et al. The long-short story of movie description, GCPR 2015