Une grande école d'ingénieur au coeur de la société numérique

EURECOM
Sophia Antipolis

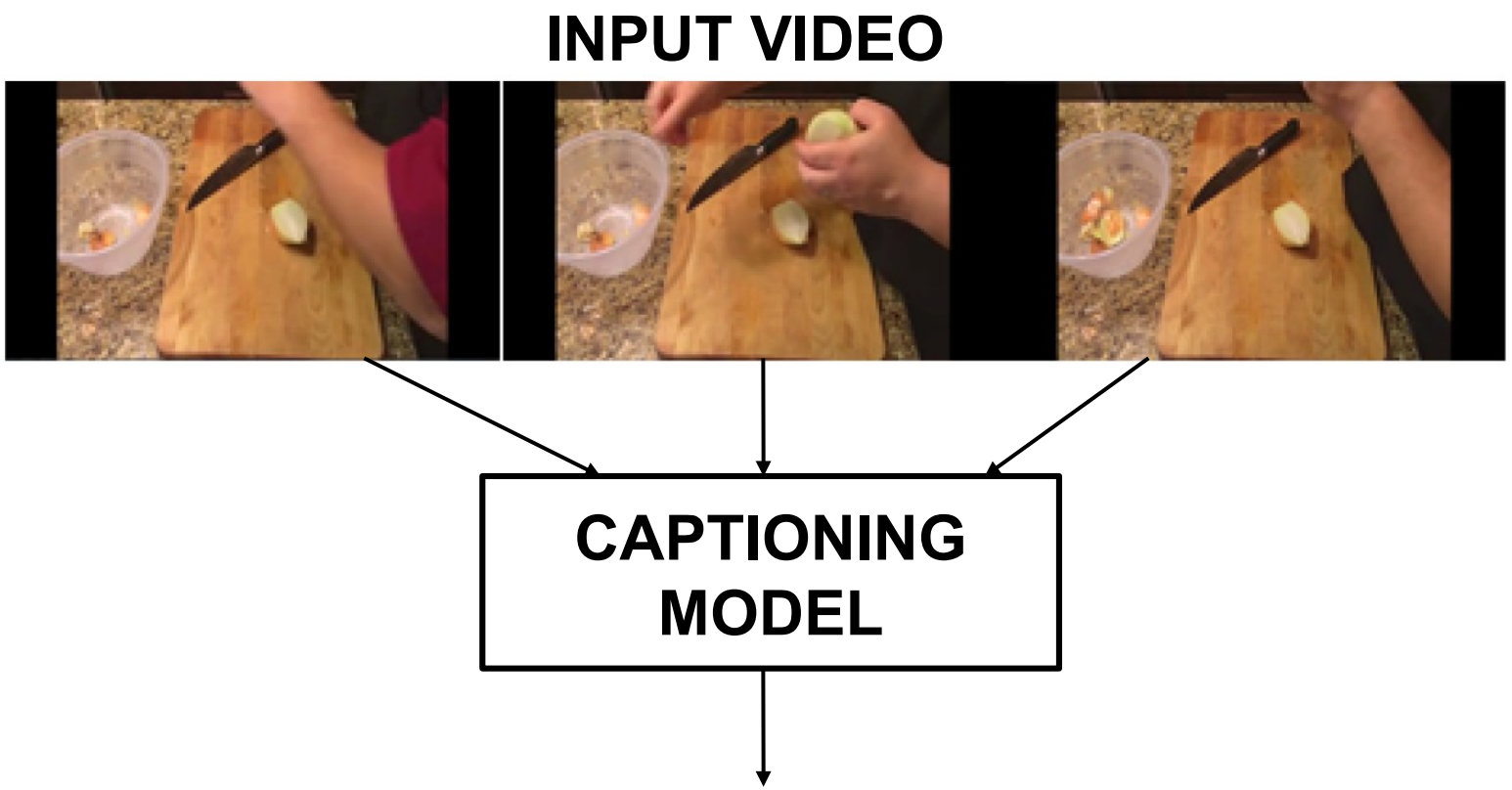
Learned Spatio-Temporal Adaptive Pooling for Video Captioning

Danny FRANCIS and Benoit HUET

AI4TV 2019, Nice

Cti
Commission des Titres d'Ingénieur

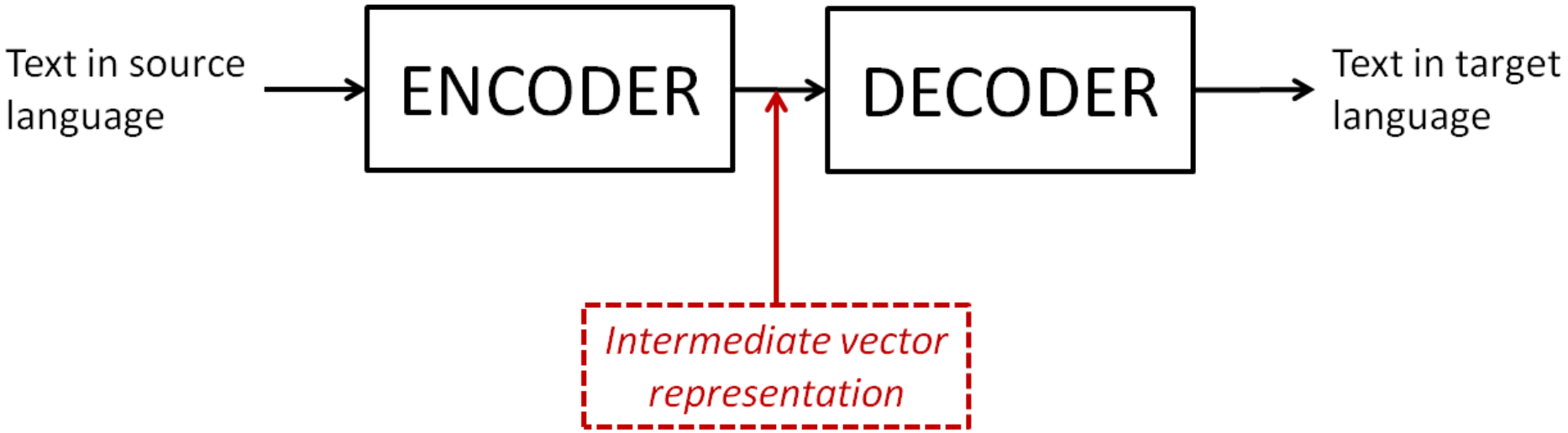
Video Captioning in a Nutshell



Video Captioning for TV

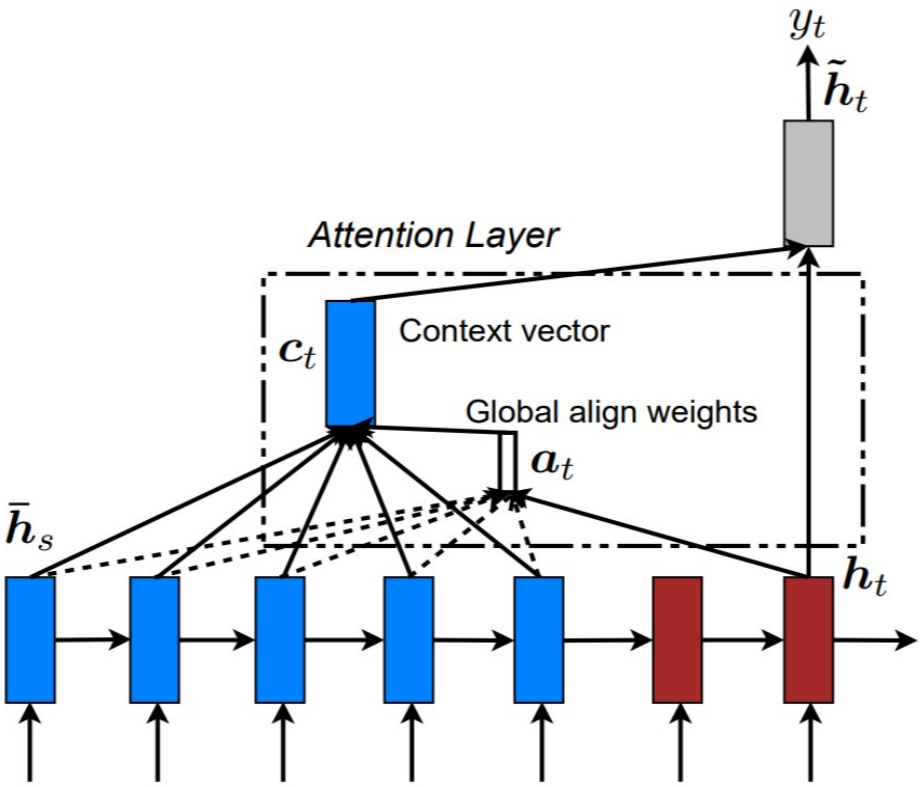
- **Annotations for impaired people**
- **Old TV archives need to be annotated**
- **Textual indexing**
- **Summarization with shot detection**
- **…**

The Encoder-Decoder Scheme for NMT



Reference: Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to sequence learning with neural networks*. In Advances in neural information processing systems (pp. 3104-3112).

Attention for Encoder-Decoder



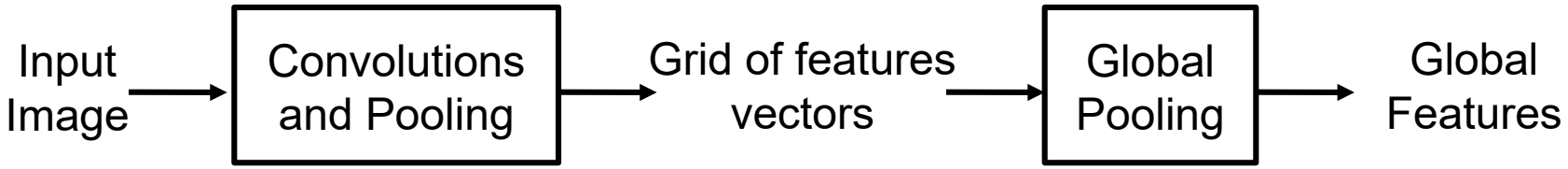
Reference: Luong, M. T., Pham, H., & Manning, C. D. *Effective Approaches to Attention-based Neural Machine Translation*.

Encoder-Decoder for Video Captioning

- **Frame sequences vs word sequences**
- **Visual features vectors vs word embeddings**
- **The Encoder-Decoder scheme can be easily extended to Video captioning:**
  - **Source language = Video**
  - **Target language = unchanged**

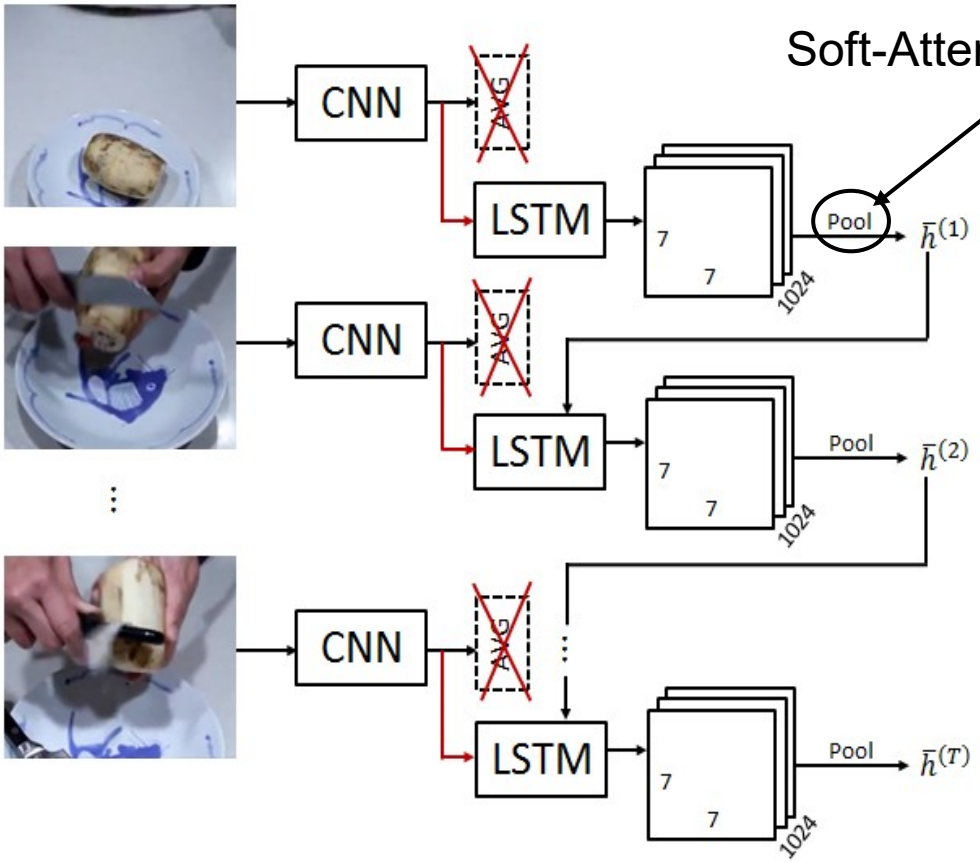
Problems with the Naive Approach

- **Visual features usually extracted from a CNN**



- **Loss of space information**
  - **How to relate an object with another one in a frame?**
  - **How to relate an object with another in another frame?**

Our L-STAP Method (1)



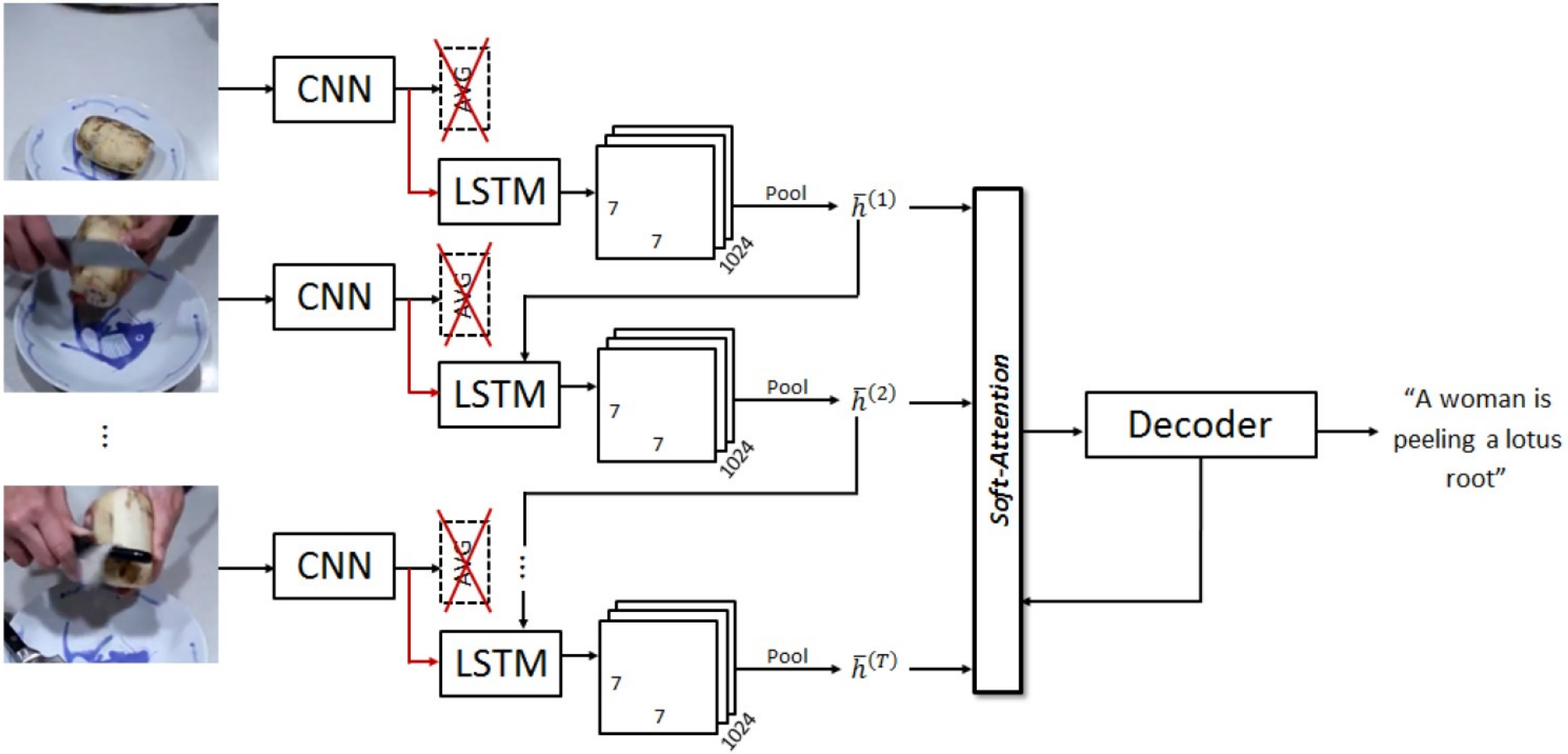
Instead of the global pooling:
- use an LSTM to compute local embeddings
- change computation of LSTM hidden state: soft-attention pooling on local embeddings based on previous hidden state

Our L-STAP Method (2)

Learned Spatio-Temporal Adaptive Pooling

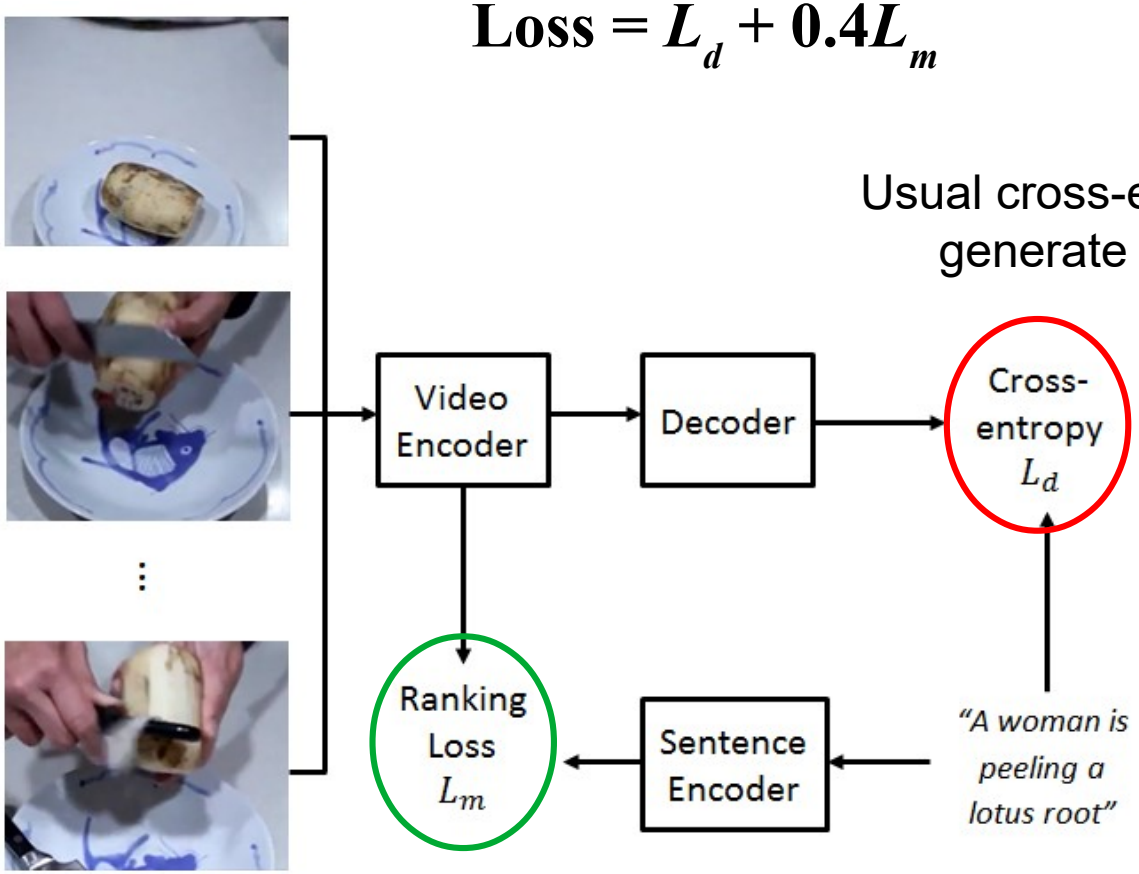
- **It is Learned: pooling depends on training data**
- **It is Spatio-Temporal: LSTM hidden states contain temporal information based on local features**
- **It is Adaptive: the soft-attention pooling of local embeddings makes it adaptive to input data**

Video Captioning with L-STAP



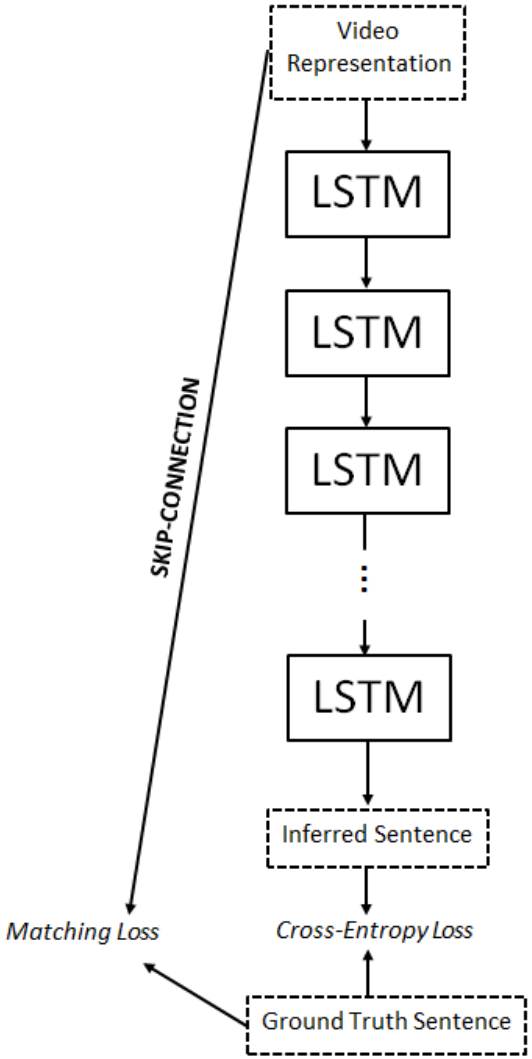
Optimization

$$\text{Loss} = L_d + 0.4L_m$$



Make sentence embeddings and visual embeddings match

Matching Component



Improves results by directly matching video embeddings with the ground-truth sentence

Results on MSVD

Model	BLEU-4	ROUGE	METEOR	CIDEr
TSL	51.7	-	34.0	74.9
RecNet	52.3	69.8	34.1	80.3
MGRU	53.8	-	34.5	81.2
AGHA	55.1	-	35.3	83.3
SAM	54.0	-	35.3	87.4
E2E	50.3	70.8	34.1	87.5
SibNet	54.2	71.7	34.8	88.2
Ours	**55.1**	**72.7**	**35.4**	**86.7**

Ablation Study

Model	BLEU-4	ROUGE	METEOR	CIDEr
Baseline	52.7	71.4	34.1	79.5
Baseline + matching	53.3	71.2	34.5	82.2
L-STAP (avg) + matching	55.1	72.3	35.4	84.3
L-STAP (attention) + matching	**55,1**	**72,7**	**35,4**	**86,7**

Conclusion

- **We proposed a Learned Spatio-Temporal Adaptive Pooling method to replace global pooling in CNNs in the context of video captioning**
- **This method leads to significant improvements with respect to the naive approach**
- **Video Captioning is one promising direction for improving TV user experience and TV archives management**

The End

Thank you!