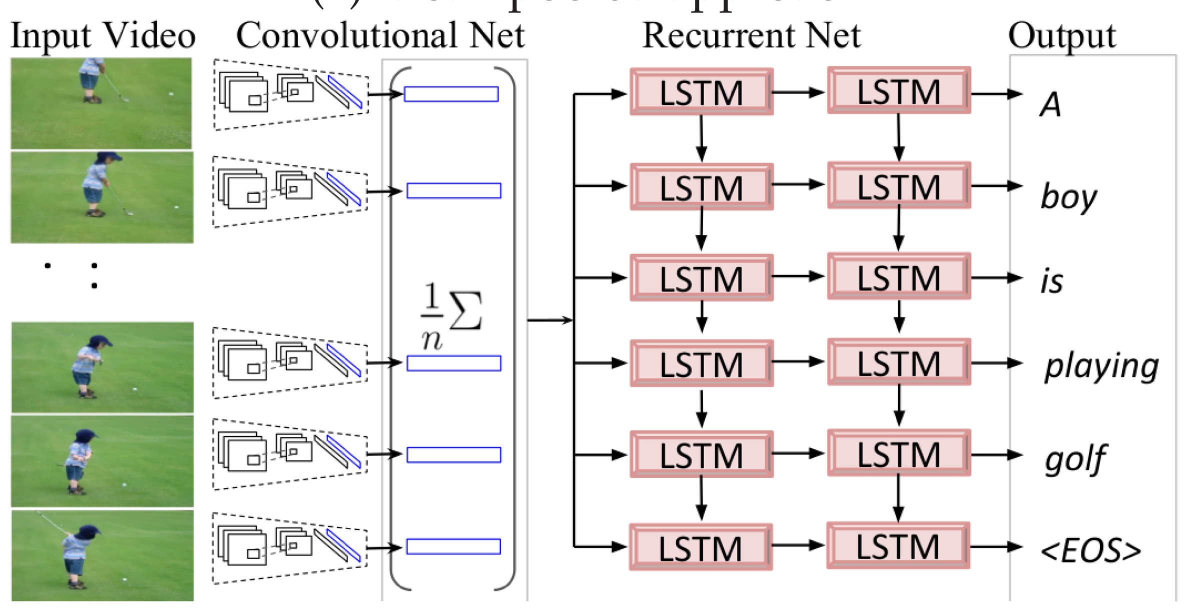
Translating Videos to Natural Language using Deep Recurrent Neural Networks

Subhashini Venugopalan[1], Huijuan Xu[3], Jeff Donahue[2],
Marcus Rohrbach[2], Raymond Mooney[1], Trevor Darrell[2], Kate Saenko[3]
[1] UT-Austin [2] UC-Berkeley [3] UMass-Lowell

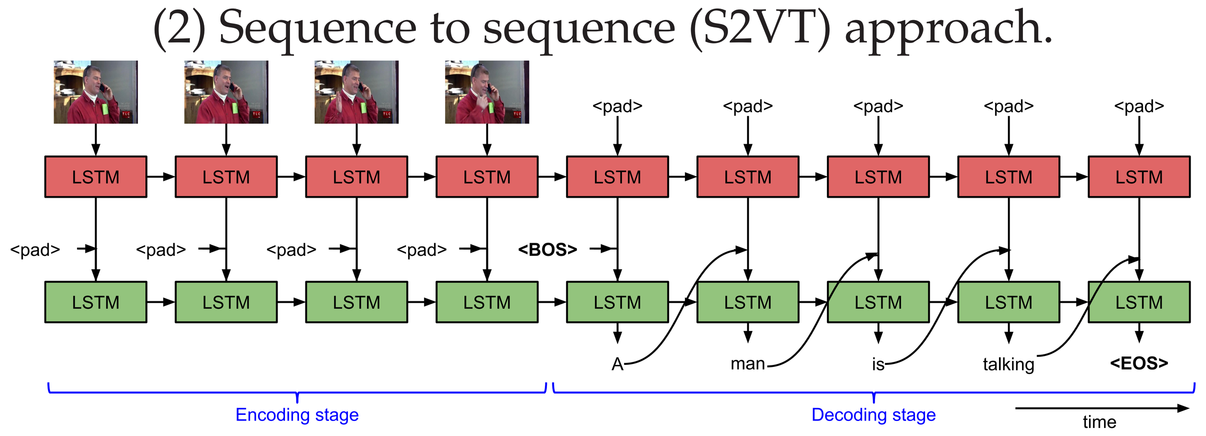
Goals

Given a short YouTube video, output a natural language sentence that describes the event depicted in the video.

(1) Mean pooled approach.



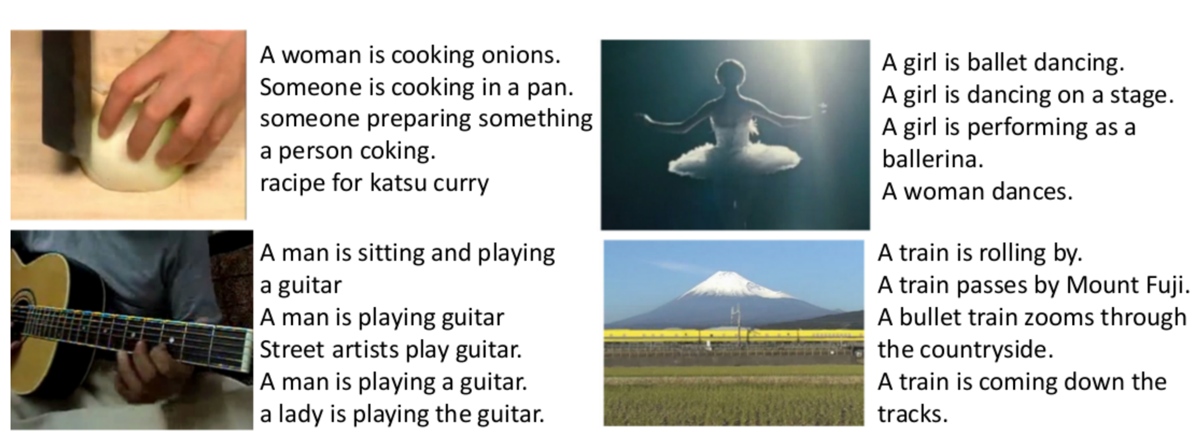
(2) Sequence to sequence (S2VT) approach.



We present methods to generate descriptions for events depicted in videos using Convolutional Neural Networks (CNNs) and Long Short Term Memory (LSTM) networks.

Datasets

We demonstrate our approach on a large, realistic collection of YouTube videos and movies.

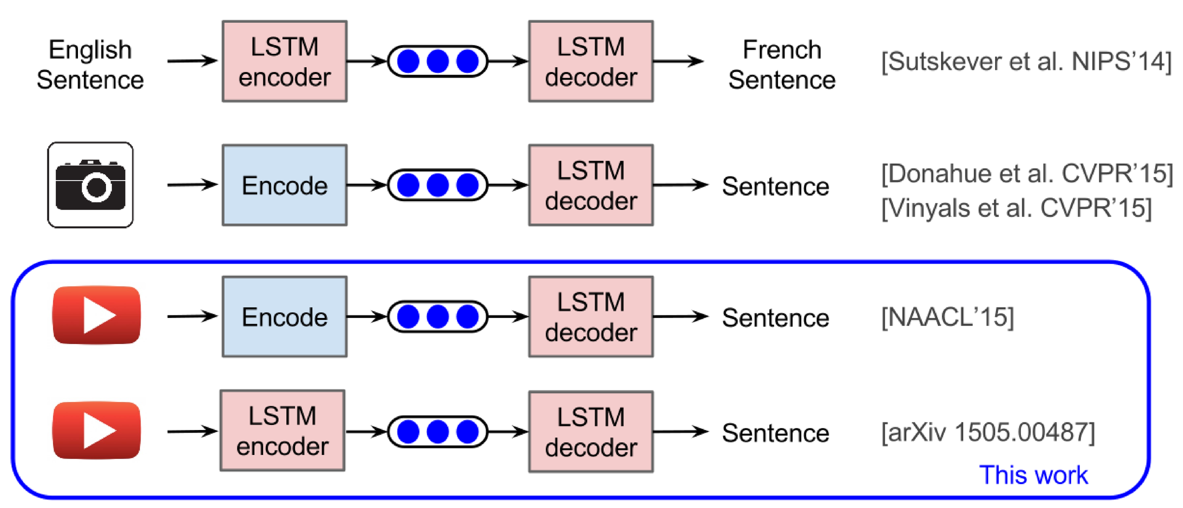


(a) YouTube Video corpus

(b) MPII Movie Description Dataset

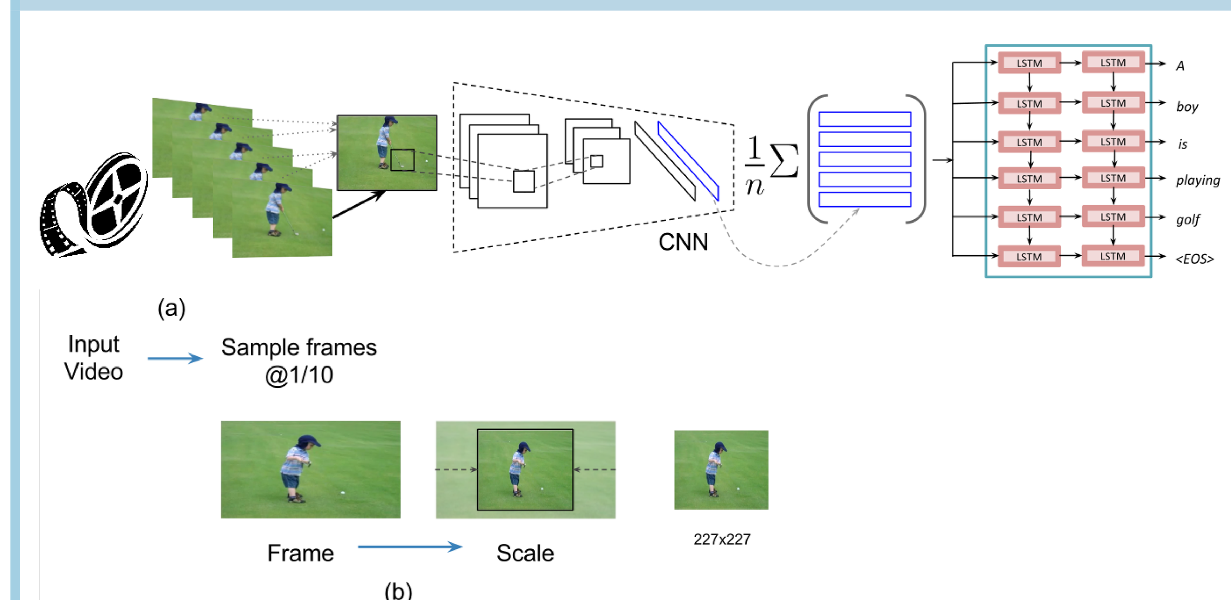
The YouTube dataset, collected by (Chen and Dolan, ACL 2011) consists of 1970 videos, where each video is accompanied by about 41 human descriptions (sentences), see (a) above. We also show results on large movie description corpora like the Montreal and MPII movie description datasets, see (b) above.

Insight

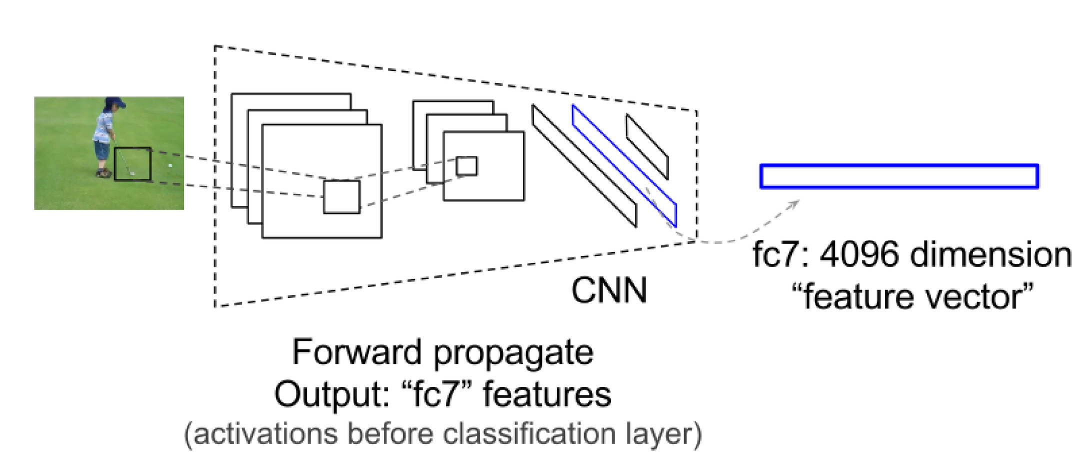


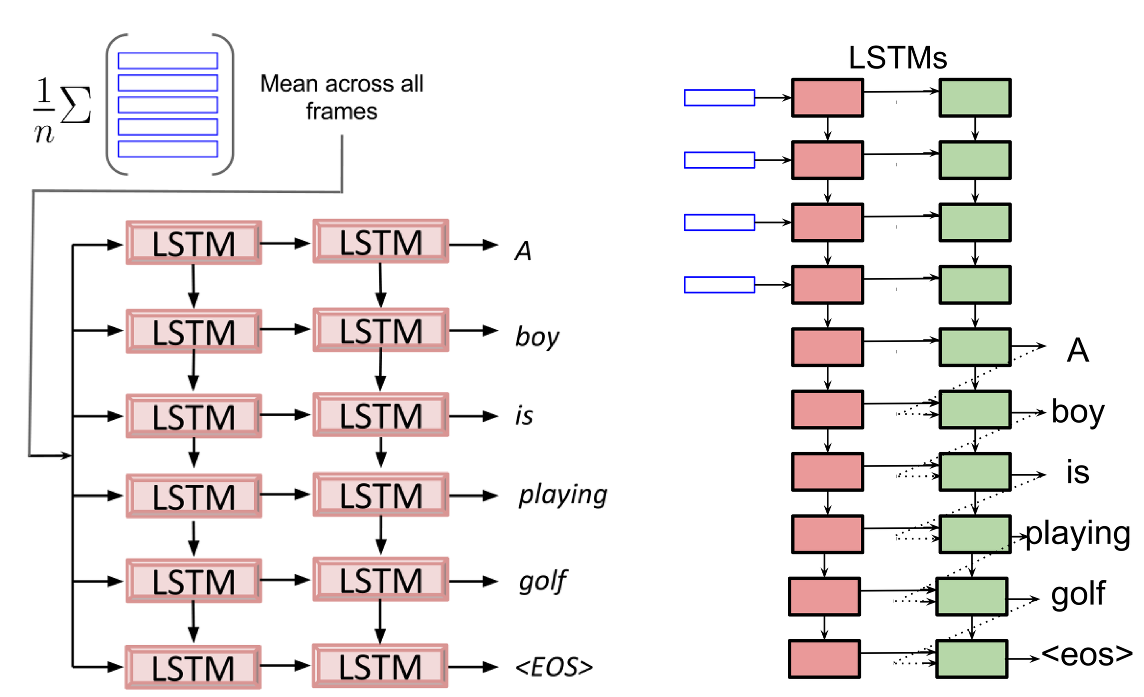
The broad idea of our approach is to encode a video frame sequence and decode it to a sequence of english words (sentence) using LSTMs.

Overview



The image is forward propagated through a convolutional neural network. The activations of the fully connected layer just before the classification is considered as the image feature which is then mean-pooled (as in 1) or is directly provided as input to the LSTM network in the sequence to sequence (S2VT) models (in 2).





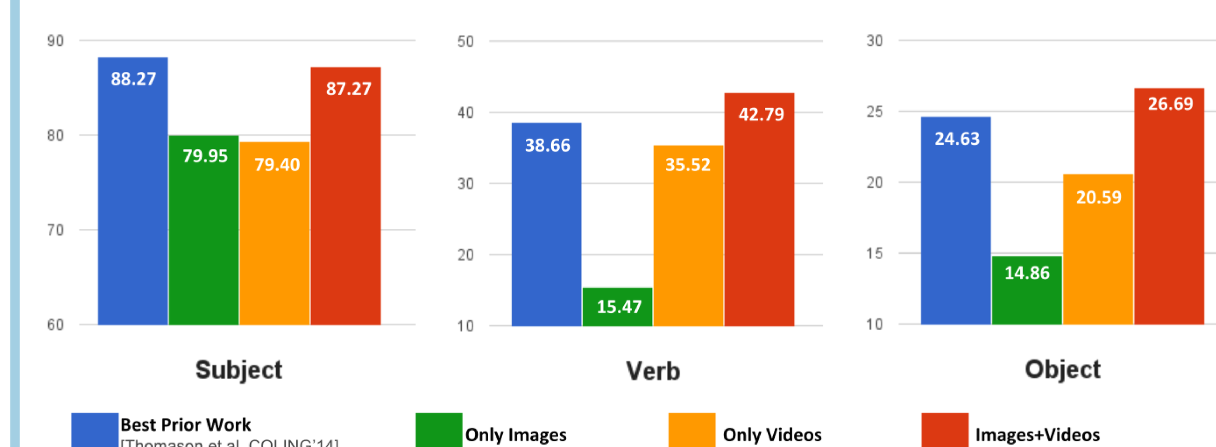
(1) Mean pooled (2) S2VT

Image Pre-Training

Annotated video data is scarce. We can optionally initialize the weights of our network by pre-training on image-caption datasets. Flickr30k and MSCOCO datasets together have over 150,000 images and 750,000 caption (5 descriptions per image).

SVO Accuracy

Accuracy of Subject, Verb and Object of the Mean-Pooled models.



The subject, verb and object are extracted from the generated sentence and compared against all valid subjects, verbs, and objects amongst the ground truth descriptions.

Sentence Evaluation

We use the machine translation metric METEOR to compare the quality of the generated description against the multiple ground truth reference sentences.

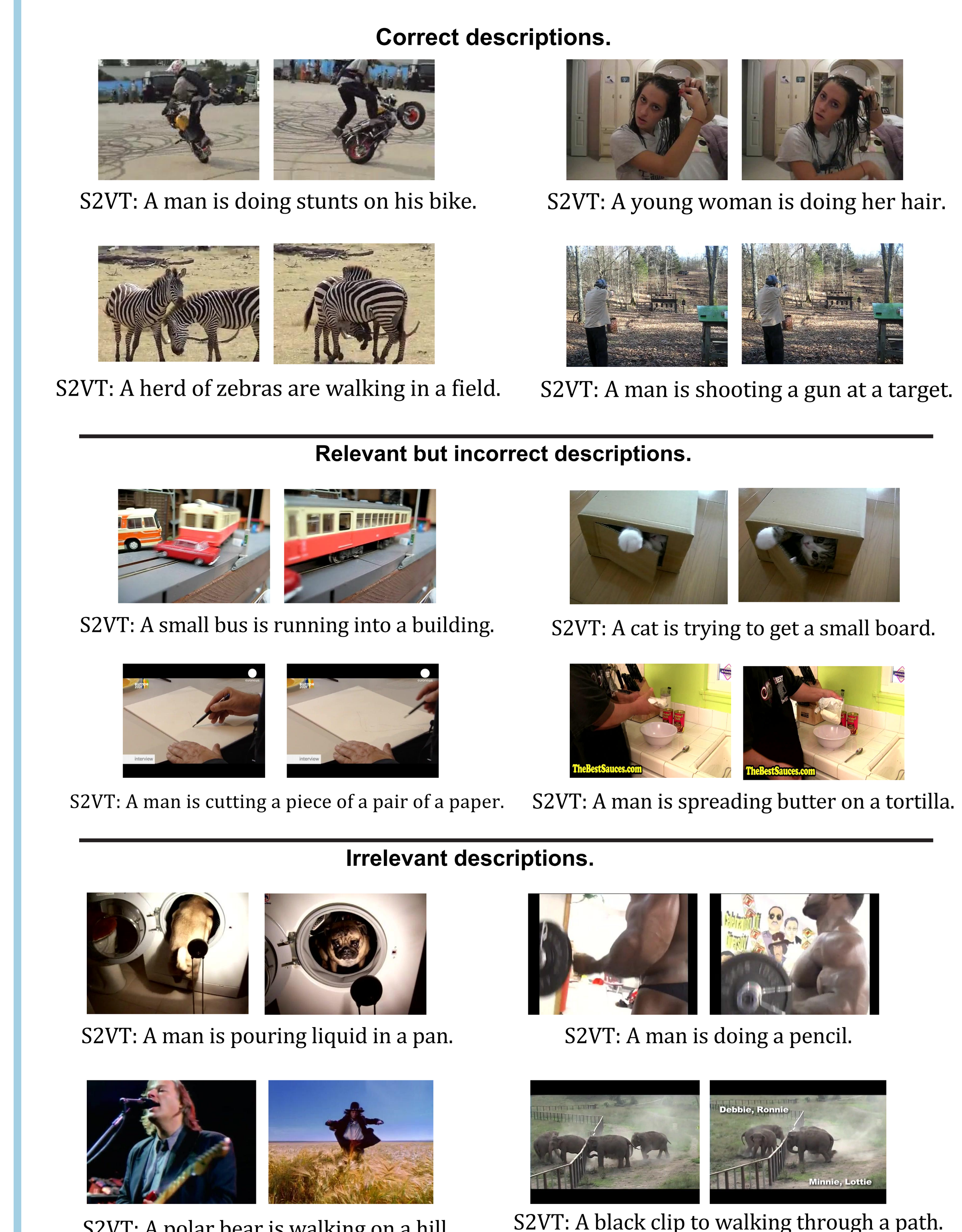
Model	METEOR
FGM (best prior work) [1]	23.9
Mean pool	
- AlexNet [3]	26.9
- VGG	27.7
- AlexNet COCO pre-trained [3]	29.1
- GNet [4]	28.7
Soft-attention	
- GoogleNet [4]	29.0
- GoogleNet + 3D-CNN [4]	29.6
S2VT [2]	
- RGB (AlexNet)	27.9
- Flow (AlexNet)	24.3
- RGB (VGG)	29.2
- RGB (VGG) + Flow (AlexNet)	29.8

Movie Dataset Results

On MPII Movie Corpus	METEOR
SMT (best variant)	5.6
S2VT: RGB (VGG) [ours]	6.3

On Montreal M-VAD Corpus	METEOR
Soft-attention (GNet + 3D-CNN) [4]*	4.1
S2VT: RGB (VGG) [ours]	
- trained on M-VAD	5.6
- trained on MPII-MD & M-VAD	6.7

Qualitative Results



References

[1] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. J. Mooney. Integrating language and vision to generate natural language descriptions of videos in the wild. In *COLING*, 2014.

[2] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, K. Saenko, and T. Darrell. Sequence to sequence – video to text. *arXiv:1505.00487*, 2015.

[3] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. *NAACL-HLT*, 2015.

[4] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. *arXiv:1502.08029v4*, 2015.