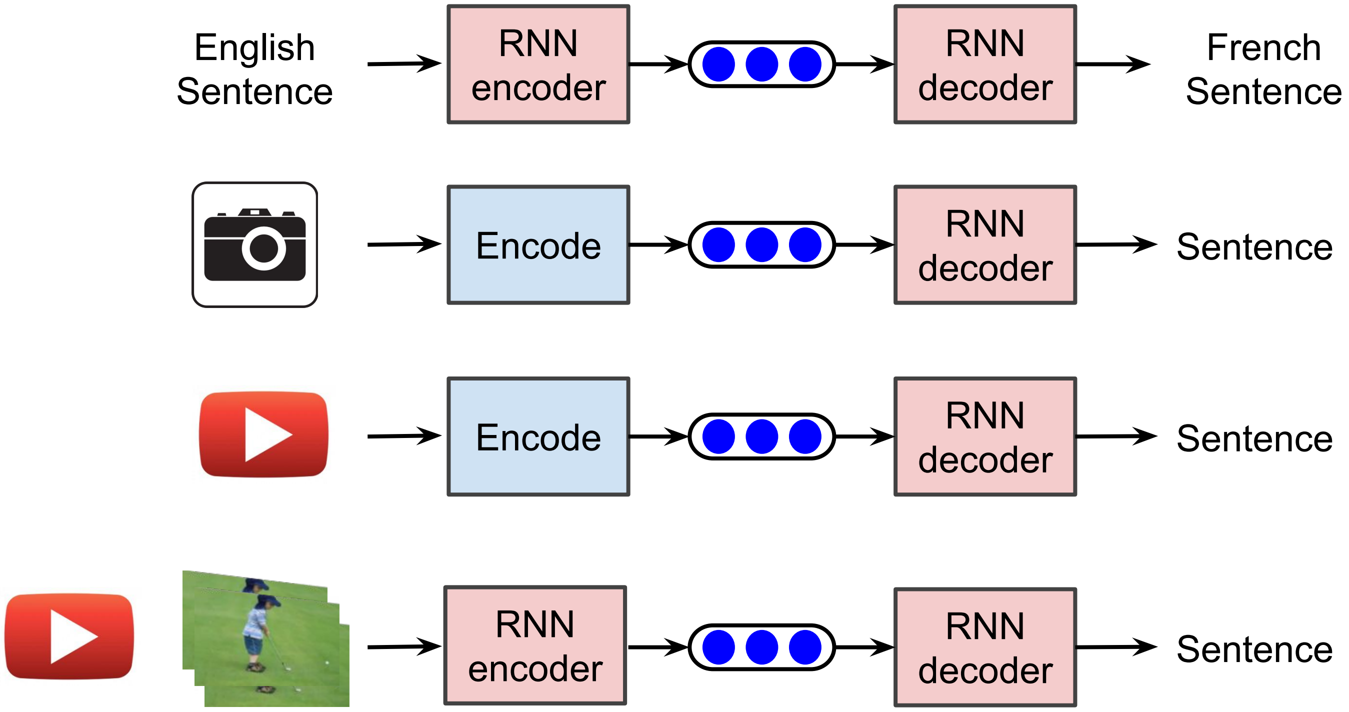
Sequence to Sequence
Video to Text

Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue
Raymond Mooney, Trevor Darrell, Kate Saenko

Objective

A monkey is pulling a dog’s tail and is chased by the dog.

Recurrent Neural Networks (RNNs) can map a vector to a sequence.



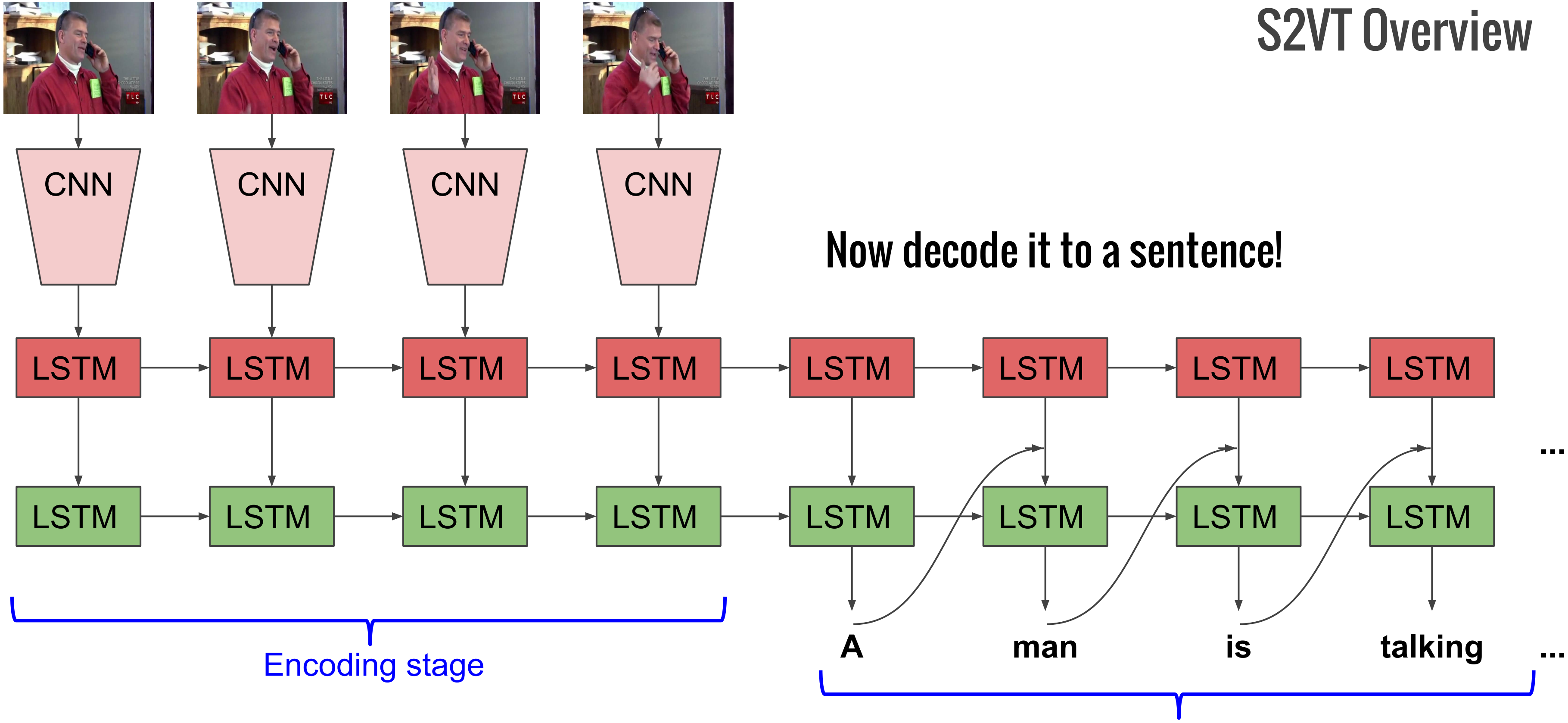
[Sutskever et al. NIPS’14]

[Donahue et al. CVPR’15]
[Vinyals et al. CVPR’15]

[Venugopalan et. al. NAACL’15]

[Venugopalan et. al. ICCV’ 15] (this work)

S2VT Overview

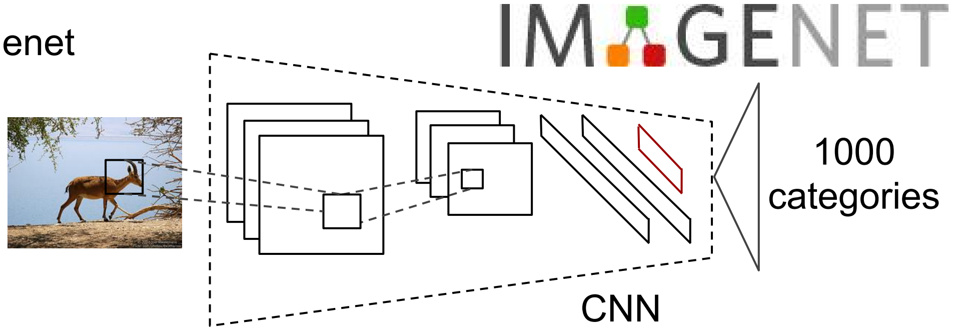


Now decode it to a sentence!

Sequence to Sequence - Video to Text (S2VT)
S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko

Frames: RGB

1. Train on Imagenet

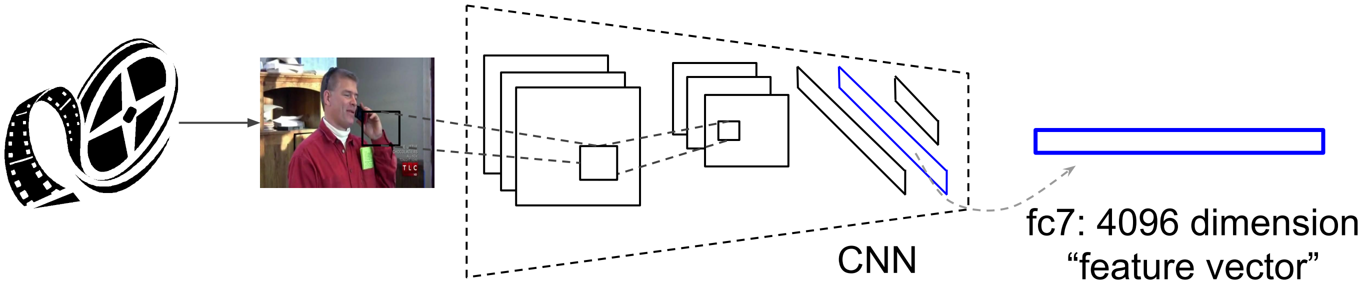


IMAGENET

1000 categories

CNN

2. Take activations from layer before classification



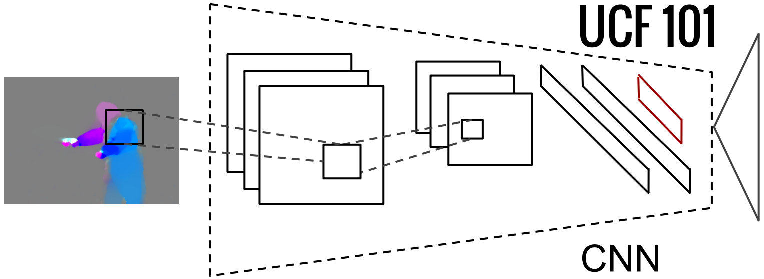
CNN

fc7: 4096 dimension “feature vector”

Forward propagate
Output: “fc7” features
(activations before classification layer)

Frames: Flow

1. Train CNN on Activity classes



UCF 101

101 Action Classes

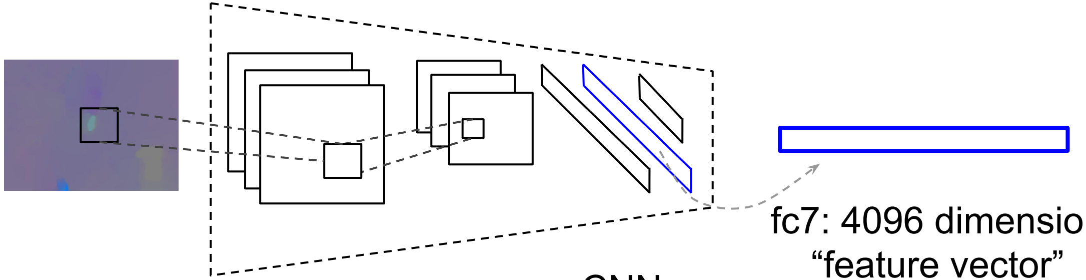
CNN

(modified AlexNet)

2. Use optical flow to extract flow images.

[T. Brox et. al. ECCV ‘04]

3. Take activations from layer before classification

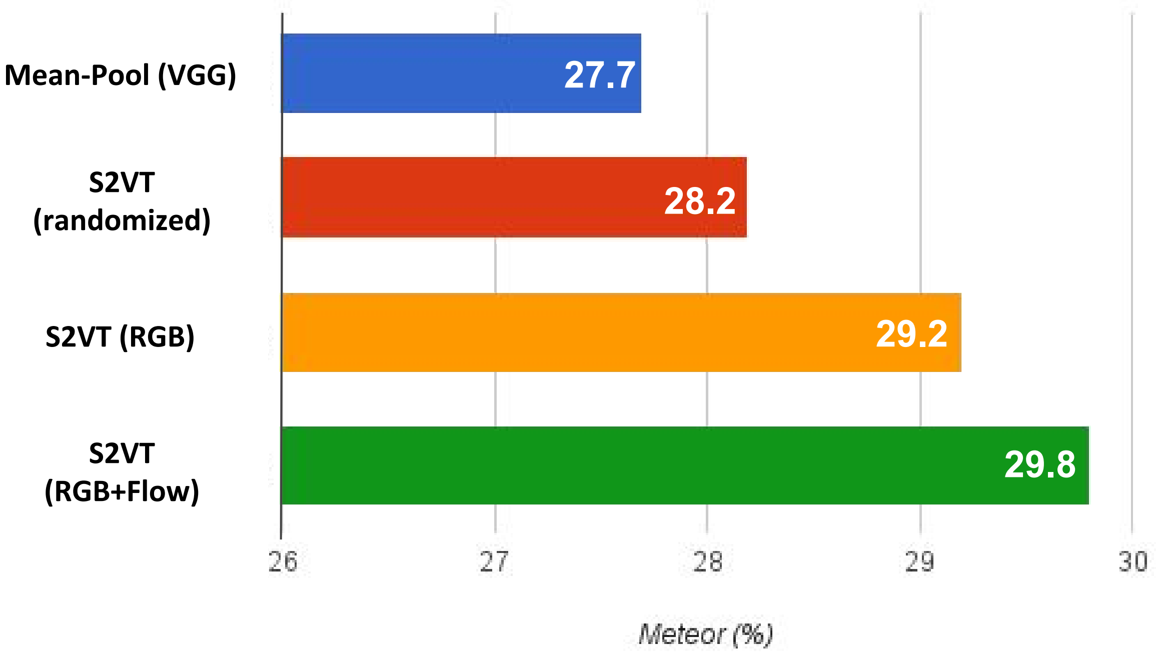


fc7: 4096 dimension “feature vector”

CNN

Forward propagate
Output: “fc7” features
(activations before classification layer)

Results (Youtube)



Evaluation: Movie Corpora

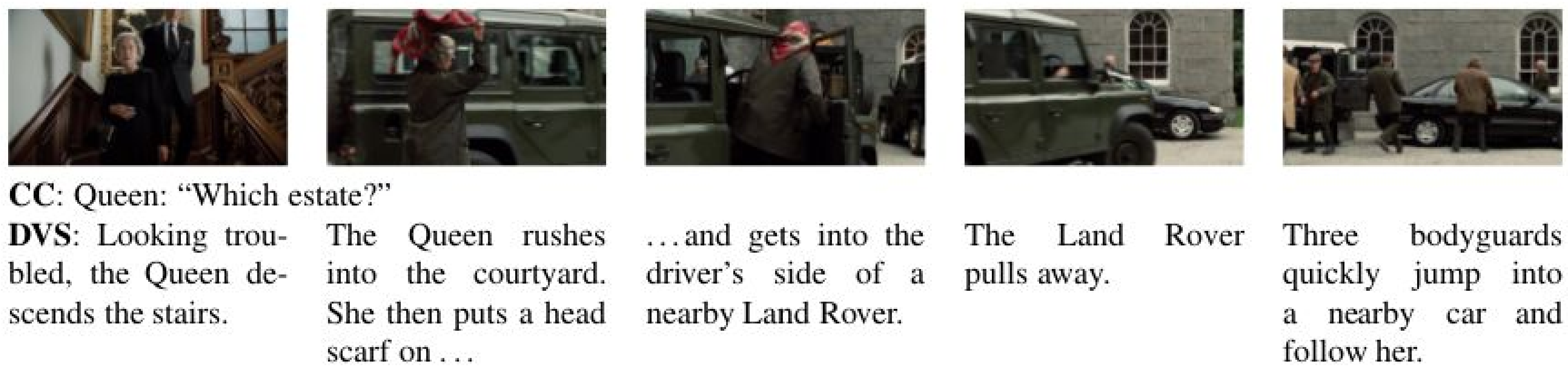
MPII-MD

- MPII, Germany
- DVS alignment: semi-automated and crowdsourced
- 94 movies
- 68,000 clips
- Avg. length: 3.9s per clip
- **~1 sentence per clip**
- 68,375 sentences

M-VAD

- Univ. of Montreal
- DVS alignment: automated speech extraction
- 92 movies
- 46,009 clips
- Avg. length: 6.2s per clip
- **1-2 sentences per clip**
- 56,634 sentences

Movie Corpus - DVS



CC: Queen: “Which estate?”

DVS: Looking troubled, the Queen descends the stairs.

The Queen rushes into the courtyard. She then puts a head scarf on . . .

. . . and gets into the driver’s side of a nearby Land Rover.

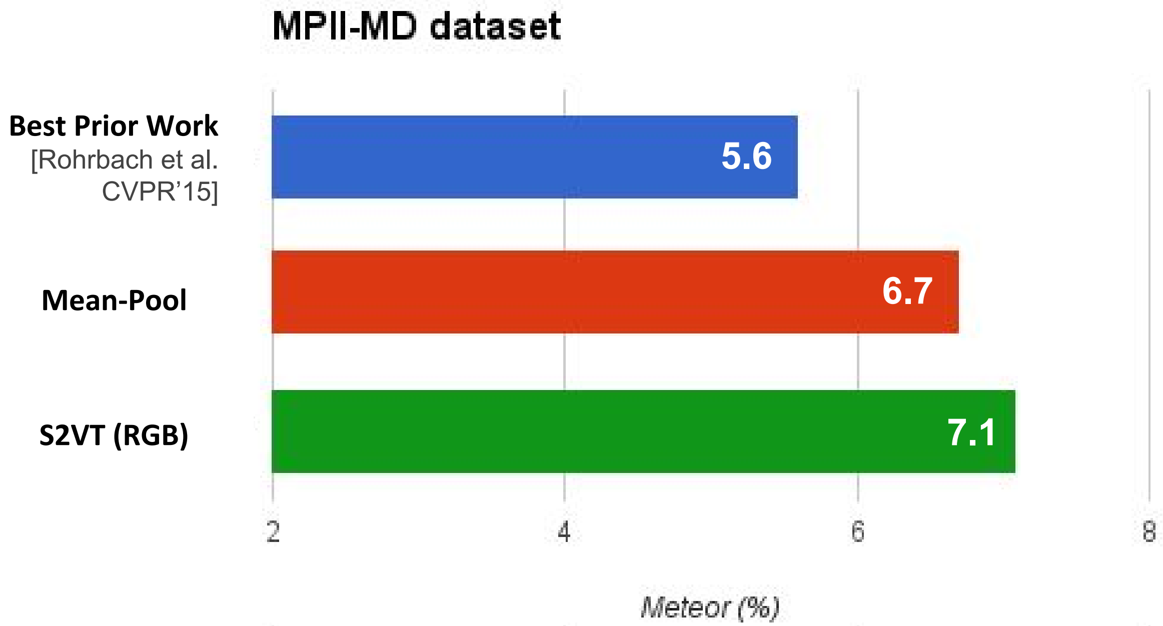
The Land Rover pulls away.

Three bodyguards quickly jump into a nearby car and follow her.

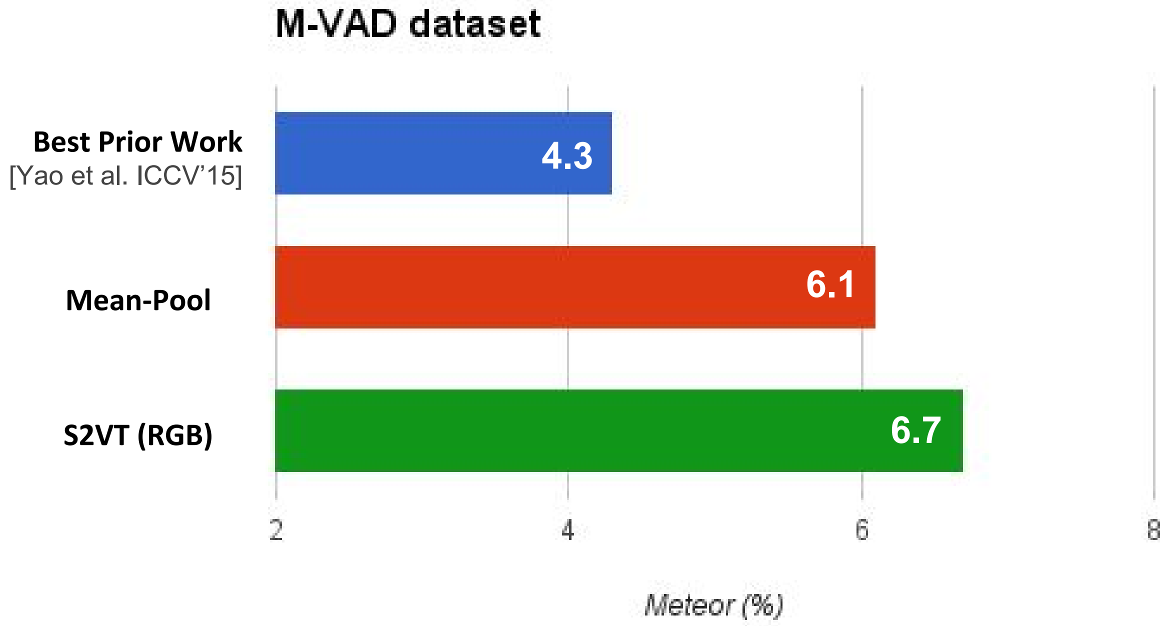
Processed: Looking troubled, someone descends the stairs.

Someone rushes into the courtyard. She then puts a head scarf on ...

Results (MPII-MD Movie Corpus)

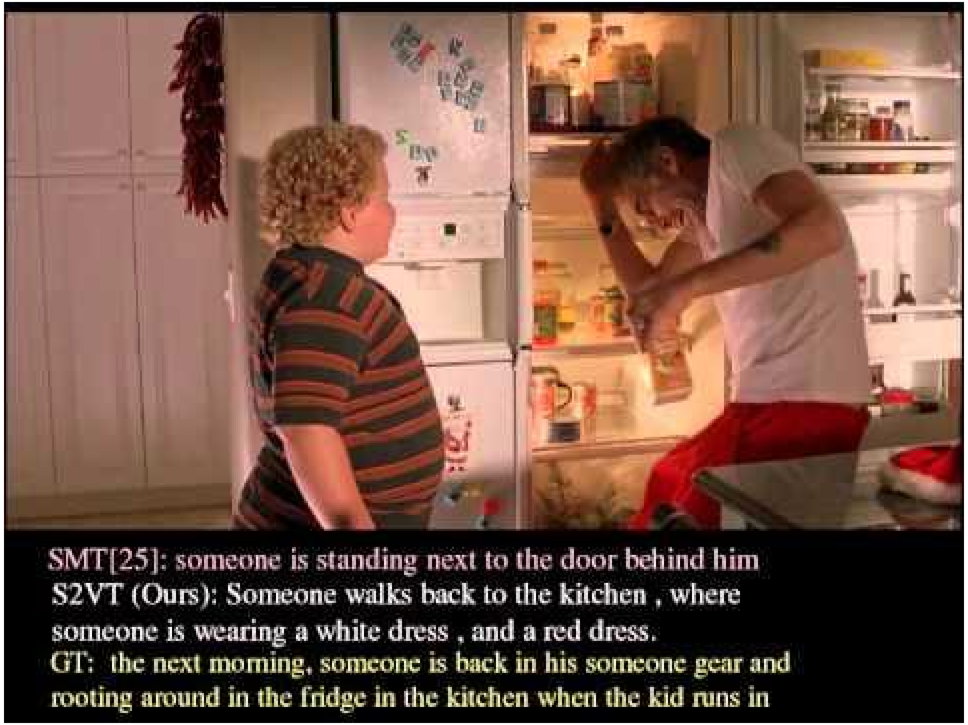


Results (M-VAD Movie Corpus)



Example Movie Clips
MPII-MD & M-VAD

Sequence to Sequence - Video to Text (S2VT)
S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko

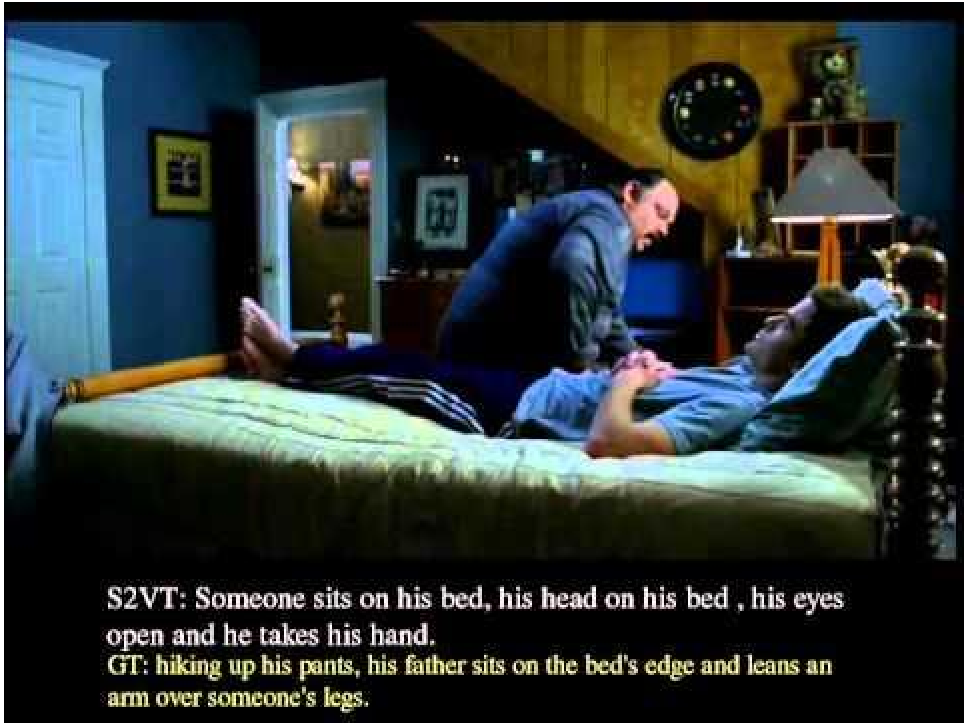


SMT[25]: someone is standing next to the door behind him

S2VT (Ours): Someone walks back to the kitchen , where someone is wearing a white dress , and a red dress.

GT: the next morning, someone is back in his someone gear and rooting around in the fridge in the kitchen when the kid runs in

MPII-MD: https://youtu.be/XTqOhuTXj1M



S2VT: Someone sits on his bed, his head on his bed , his eyes open and he takes his hand.

GT: hiking up his pants, his father sits on the bed's edge and leans an arm over someone's legs.

M-VAD: https://youtu.be/pEROmjzSYaM

Code and more examples

http://vsubhashini.github.io/s2vt.html