Sequence to Sequence Video to Text

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

[Donahue et al. CVPR’15]
[Venugopalan et al. ICCV’ 15] (this work)
S2VT Overview

**Encoding stage**

**Decoding stage**

Now decode it to a sentence!

A man is talking...

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

2. Take activations from layer before classification

Forward propagate
Output: “fc7” features
(activations before classification layer)
1. Train CNN on Activity classes

2. Use optical flow to extract flow images.

3. Take activations from layer before classification

Frames: Flow

CNN (modified AlexNet)

UCF 101
101 Action Classes

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

fc7: 4096 dimension “feature vector”

[T. Brox et. al. ECCV ’04]
Results (Youtube)

- Mean-Pool (VGG): 27.7
- S2VT (randomized): 28.2
- S2VT (RGB): 29.2
- S2VT (RGB+Flow): 29.8
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
Looking troubled, the Queen descends the stairs. Someone rushes into the courtyard. She then puts a head scarf on. Someone rushes into the courtyard. She then puts a head scarf on. The Queen rushes into the courtyard. She then puts a head scarf on. The Land Rover pulls away. Three bodyguards quickly jump into a nearby car and follow her.
Results (MPII-MD Movie Corpus)

![Bar chart showing meteor scores for different methods on MPII-MD dataset]

- **Best Prior Work** [Rohrbach et al. CVPR’15] 5.6
- **Mean-Pool** 6.7
- **S2VT (RGB)** 7.1

*Meteor (%)*
Results (M-VAD Movie Corpus)

<table>
<thead>
<tr>
<th>Method</th>
<th>Meteor (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best Prior Work</td>
<td>4.3</td>
</tr>
<tr>
<td>Mean-Pool</td>
<td>6.1</td>
</tr>
<tr>
<td>S2VT (RGB)</td>
<td>6.7</td>
</tr>
</tbody>
</table>

Source: Yao et al. ICCV'15
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
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/pER0mjzSYaM
Code and more examples
http://vsubhashini.github.io/s2vt.html