Sequence to Sequence
Video to Text
Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue Raymond Mooney, Trevor Darrell, Kate Saenko
Outline

● Objective
● Experimental Setup
● Current model.
● A Simple Extension.
● How is information distributed within the video?
● Does model capture temporal information?
● Conclusions & Future Work
Objective

Generate video descriptions.
S2VT Overview

Encoding stage

Decoding stage

A man is talking...

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

**Code:** Forked from author’s [github account](https://github.com)

**Frame Sampling:** 1 in 10 (unless otherwise mentioned)

**Network Architecture:** VGG CNN + 2 layer LSTM

**Dataset:** MSVD Youtube dataset (Avg Length 10.2 s, #sentences per video = 41)

**Vocabulary:** MSVD + MPII-MD + MVAD

**Performance Metric:** METEOR

**Evaluation Tool:** coco_evaluation
Forward Model

- Able to learn abstract attributes like young etc to reasonable extent.
- Able to capture main content of video in most cases.

PROBLEMS:

- Long sentences repeat words multiple times leading to lower quality sentences
  - The boys are playing with a group of a group of a group of people is sitting on a group of a group of people are watching a gym
  - A woman is cutting a piece of a piece of a pair of a pair of a pair.
  - A man is cutting a large of a large large large large floor.
Backward Model

- Process frames in reverse order !!
- Seems to perform better than forward model on validation set but almost similar performance on test set.
- How to choose best backward model ?
Bidirectional Model

● Motivated from Bidirectional N gram models used for Language Modelling in NLP
● Combine forward and backward models.
  - How do we select forward and backward model?
  - Combining strategy?
  - How are weights selected?
Performance variation of Bidirectional Model with interpolation weight on Validation Set

<table>
<thead>
<tr>
<th>Forward Model Weight</th>
<th>METEOR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>31.7</td>
</tr>
<tr>
<td>0.1</td>
<td>31.5</td>
</tr>
<tr>
<td>0.2</td>
<td>32</td>
</tr>
<tr>
<td>0.3</td>
<td>32.5</td>
</tr>
<tr>
<td>0.4</td>
<td>32.5</td>
</tr>
<tr>
<td>0.5</td>
<td>32.3</td>
</tr>
<tr>
<td>0.6</td>
<td>32.1</td>
</tr>
<tr>
<td>0.7</td>
<td>30.4</td>
</tr>
<tr>
<td>0.8</td>
<td>30.9</td>
</tr>
<tr>
<td>0.9</td>
<td>29.7</td>
</tr>
<tr>
<td>1</td>
<td>29</td>
</tr>
</tbody>
</table>
FORWARD: The boys are playing with a group of a group of a group of people is sitting on a group of a group of people are watching a gym!!

BACKWARD: Two boys are dancing.

BIDIRECTIONAL: The boys are playing.

LABEL: Three men are dancing in beach towels.

This eg shows utility of Bidirectional Model.
FORWARD: A man is using a piece of a sharp.

BACKWARD: A person is cutting a piece of a brush.

BIDIRECTIONAL: A man is cutting a piece of a brush.

LABEL: A person is performing some card tricks.

All Fail :(
How is information distributed within video?

Conjecture: Central part of video contains more relevant information than frames at beginning and end for most videos.
Does Model Capture Temporal Information?

Performance Comparison of Random Models

- Random...
- Random...
- Forward
- Backward
- BiRando...
- Bidirectional...
Conclusions

- Bidirectional model is more powerful than forward or backward model.
- Frames at start and end contain less information.
Future Work

- Try combining bidirectional with optical flow model.
- Try using gaussian sampling centred on video’s centre.
- Is it more suitable for specific kinds of videos? Like generating sports commentary?
References

Sequence to Sequence Video to Text - Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, Kate Saenko
Thank You :}