Comparing Attention-based Neural Architectures for Video Captioning

Jason Li, Helen Qiu
jasonkli@stanford.edu, shiqiu21@stanford.edu
CS224N-Winter2019, Stanford University

Introduction & Related Work

- Video captioning is the task of understanding the visual content of video sequence and translating that understanding to an appropriate caption.
- Automatic video description generation can have many practical applications in daily scenarios such as video retrieval, blind navigation, and automatic video subtitling [1].
- Early works in video captioning extracted features from 2D-CNNs for each frame and averaged them before inputting into an LSTM.
- Venugopalan [2] et al. used a sequence-to-sequence model with per-frame 2D-CNN features as the input sequence.
- Later works incorporate attention, with a particular emphasis on temporal attention to attend to different features over time.

In our work, we explore spatio-temporal attention over features extracted from a pre-trained P3D (Pseudo-3D ResNet). We also compare the performance of LSTMs to Transformers.

Main Findings & Future Work

Main findings:
- Attention over P3D features did not perform as well as 2D-CNN features -- pre-training on action recognition may limit what features are captured, as shown by “man” vs “woman” analysis
- Ensemble model shows promise in providing different, but relevant, information for video captioning
- Transformers outperformed LSTMs

Future Work:
- Further experiment with combining P3D features with LSTM features, such as with Transformer
- Substitute ResNet-152 with Faster-RCNN for feature extra
- Compare with other datasets – image features seem to be sufficient for the MSVD dataset, but what about others?
- Incorporate additional features, such as optical flow, as done in other works
- Look into ways to address bias towards more frequent words

Analysis

- Skew in dataset for “man” vs “woman” (541 vs. 152 in training and 291 vs. 87 in validation)
- We compared performance on subset that included “man” or some close variation, but not “woman”, and vice versa
- Performance significantly better for “man” subset
- Particularly large discrepancy for P3D – possible that it does not discern “man” vs “woman” since it was pre-trained on activity recognition

Approach

1. ResNet-LSTM Encoder + LSTM Decoder

2. P3D Encoder + LSTM Decoder

3. Combined P3D/ResNet-LSTM Encoder with LSTM Decoder

4. Transformer with ResNet Features

5. Transformer with P3D Features

Dataset: MSVD (Microsoft Video Dataset)

- 1969 videos – average of 10 seconds
- Average of 40 captions each
- We used 5 captions per video for training
- 1200 training, 100 validation, 669 training split
- Captions are usually 5-10 words long – lack of diversity with many similar captions across videos

Results

<table>
<thead>
<tr>
<th>BLEU Score</th>
<th>2D-CNN/LSTM Encoder + LSTM Decoder</th>
<th>P3D Encoder + LSTM Decoder</th>
<th>Ensemble Encoder + LSTM Decoder</th>
<th>Transformer + 2D-CNN Features</th>
<th>Transformer + P3D Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>71.85</td>
<td>64.11</td>
<td>66.73</td>
<td>61.47</td>
<td>54.23</td>
</tr>
<tr>
<td>Test</td>
<td>37.75</td>
<td>33.51</td>
<td>38.55</td>
<td>40.78</td>
<td>35.4</td>
</tr>
</tbody>
</table>

Of the 5 models we experimented with, Transformers with ResNet features performed best, followed by combined P3D/ResNet features with an LSTM.

References