Context Visual Information-based Deliberation Network for Video Captioning

Min Lu, Xueyong Li and Caihua Liu

College of Computer Science and Technology, Civil Aviation University of China
Outline

- Introduction
- Related Work
- Our Method
- Experiments and Results
Problem Statement

Video Captioning: automatically describing a video in natural language.

Generated sentence:
A man is cutting a piece of paper.
Outline

- Introduction
- Related Work
- Our Method
- Experiments and Results
Encoder-decoder framework with attention mechanisms.

Related Work

- *Deliberation Networks for Neural Machine Translation*

Outline

- Introduction
- Related Work
- Our Method
- Experiments and Results
**Motivations**

- The hidden states with inaccurate semantic information are not amended before word prediction, which will cause a cascade of errors in predicting words.

- The attention weights for the current word should be calculated based on the current hidden state rather than the previous state.
**Our Method**

- **Model Architecture**

  - **Encoder**
    - CNN
    - GroundTruth: A dog is jumping on a trampoline.

  - **Decoder**
    - LSTM
    - LSTM
    - LSTM
    - LSTM
    - Attn
    - GroundTruth: A dog is jumping on a trampoline.

  - **Deliberator**
    - LSTM
    - LSTM
    - LSTM
    - LSTM
    - Attn
Our Method

![Diagram of model architecture](image)

**Model Architecture**

Ground Truth: A dog is jumping on a trampoline.

Mathematical equations:

- \( h_t = LSTM \left( h_{t-1}, \text{inp}_t, h_t \right) \)
- \( \text{inp}_t = \left[ E[y_{t-1}], \varphi_t(F) \right] \)
- \( \varphi_t(F) = f_{\text{attn}}(h_{t-1}, F) \)

- \( h_t = LSTM \left( h_{t-1}, \text{inp}_t, h_t \right) \)
- \( \text{inp}_t = \left[ E[y_{t-1}], \varphi_t(F) \right] \)
- \( \varphi_t(F) = f_{\text{attn}}(h_{t-1}, F) \)
Outline

- Introduction
- Related Work
- Our Method
- Experiments and Results
Experiments and Results

- **MSVD**
  - It contains 1,200, 100 and 670 videos for training, validation and testing. And each video has 40 English captions annotated by human beings.

- **MSR-VTT**
  - It contains 6,513, 497 and 2,990 videos for training, validation and testing. And each video is associated with 20 descriptions.
## Experiments and Results

### Quantitative Analysis

#### MSVD

<table>
<thead>
<tr>
<th>Models</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2VT(I)[19]</td>
<td>39.6</td>
<td>31.2</td>
<td>67.5</td>
<td>66.7</td>
</tr>
<tr>
<td>RecNet(I)[23]</td>
<td>52.3</td>
<td>34.1</td>
<td>69.8</td>
<td>80.3</td>
</tr>
<tr>
<td>hLSTM(R-152)[10]</td>
<td>53.0</td>
<td>33.6</td>
<td>-</td>
<td>73.8</td>
</tr>
<tr>
<td>TSA-ED(R-152)[39]</td>
<td>51.7</td>
<td>34.0</td>
<td>-</td>
<td>74.9</td>
</tr>
<tr>
<td>PickNet(R-152)[13]</td>
<td>46.1</td>
<td>33.1</td>
<td>69.2</td>
<td>76.0</td>
</tr>
<tr>
<td>TDConvED(R-152)[40]</td>
<td>53.3</td>
<td>33.8</td>
<td>-</td>
<td>76.4</td>
</tr>
<tr>
<td>SCN-LSTM(R-152+C)[12]</td>
<td>50.2</td>
<td>33.4</td>
<td>-</td>
<td>80.5</td>
</tr>
<tr>
<td>SA(R-101+RX-101)[7]</td>
<td>52.4</td>
<td>34.3</td>
<td>71.7</td>
<td>89.5</td>
</tr>
<tr>
<td>MARN(R-101+RX-101)[11]</td>
<td>48.6</td>
<td>35.1</td>
<td>71.9</td>
<td>92.2</td>
</tr>
<tr>
<td><strong>Ours(I)</strong></td>
<td>53.0</td>
<td>33.9</td>
<td>70.8</td>
<td>84.0</td>
</tr>
<tr>
<td><strong>Ours(R-152)</strong></td>
<td>53.3</td>
<td>34.1</td>
<td>70.7</td>
<td>84.4</td>
</tr>
<tr>
<td><strong>Ours(R-101+RX-101)</strong></td>
<td><strong>53.8</strong></td>
<td><strong>35.1</strong></td>
<td><strong>72.4</strong></td>
<td><strong>94.5</strong></td>
</tr>
</tbody>
</table>

#### MSR-VTT

<table>
<thead>
<tr>
<th>Models</th>
<th>BLEU@4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>RecNet(I)[23]</td>
<td>39.1</td>
<td>26.6</td>
<td>59.3</td>
<td>42.7</td>
</tr>
<tr>
<td>hLSTM(R-152)[10]</td>
<td>38.3</td>
<td>26.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TSA-ED(R-152)[39]</td>
<td>39.5</td>
<td>27.5</td>
<td>-</td>
<td>42.8</td>
</tr>
<tr>
<td>VideoLab(R-152+C+A)[41]</td>
<td>39.1</td>
<td>27.7</td>
<td>60.6</td>
<td>44.1</td>
</tr>
<tr>
<td>Aalto(G+C)[42]</td>
<td>39.8</td>
<td>26.9</td>
<td>59.8</td>
<td>45.7</td>
</tr>
<tr>
<td>v2t_navigator(C+A)[43]</td>
<td>40.8</td>
<td>28.2</td>
<td>60.9</td>
<td>44.8</td>
</tr>
<tr>
<td>SA(R-101+RX-101)[7]</td>
<td>39.5</td>
<td>26.4</td>
<td>59.4</td>
<td>45.9</td>
</tr>
<tr>
<td>MARN(R-101+RX-101)[11]</td>
<td>40.4</td>
<td>28.1</td>
<td>60.7</td>
<td>47.1</td>
</tr>
<tr>
<td><strong>Ours(R-101+RX-101)</strong></td>
<td><strong>41.6</strong></td>
<td><strong>28.4</strong></td>
<td><strong>61.3</strong></td>
<td><strong>48.5</strong></td>
</tr>
</tbody>
</table>
## Experiments and Results

- **Qualitative Analysis**

<table>
<thead>
<tr>
<th>GroundTruth</th>
<th>Baseline</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>a man is eating spaghetti.</td>
<td>a man is cooking his kitchen.</td>
<td>a man is eating spaghetti.</td>
</tr>
<tr>
<td>a woman is mixing ingredients in a bowl</td>
<td>a man is cooking something</td>
<td>a woman is mixing ingredients in a bowl</td>
</tr>
<tr>
<td>a man is sliding down a railing of the stairs</td>
<td>a man is walking down the street</td>
<td>a man is running down the stairs</td>
</tr>
<tr>
<td>a dog is jumping on a trampoline</td>
<td>a dog is walking in a pool</td>
<td>a dog is jumping on a trampoline</td>
</tr>
</tbody>
</table>
We propose a novel architecture, Context Visual Information-based Deliberation Network for Video Captioning.

The proposed method can not only amend the inappropriate hidden state in time but also strengthen the semantic coherence of the adjacent words.

Experiments on real datasets show that our approach outperforms the state-of-the-art methods.
Q&A