Text Synopsis Generation for Egocentric Videos

Aidean Sharghi, Niels da Vitoria Lobo, Mubarak Shah
University of Central Florida
Center for Research in Computer Vision
Video Summarization

• Objective: enable efficient browsing in large video datasets

• Conventional video summarization:
  • Select important segments from a given video

• Users still need to watch the summary videos
  • Browsing large datasets remains a challenge

• How to make summaries easier to browse?

• Generate text summaries:
  • Easy to read a short report
  • Search in text summary using natural language
Naïve Approach

My friend and I sat at the table and ate a meal together. I drove in the car.
I washed the dishes.
Issues

- Imperfect summarization
  - Bad or repetitive selection
- Amplified through the text generation
  - Generate correct descriptions for egocentric videos is challenging
Alternative Approach

My friend and I sat at the table and ate a meal together. I drove in the car.

I washed the dishes.
Advantages

• Generate as much text knowledge as possible

• Identify correct descriptions

• Select most representative ones to generate text summary
Framework

Shot 1

2D CNN

2D CNN

Encoder LSTM

Encoder LSTM

Decoder LSTM

Decoder LSTM

Generated Sentence: “I looked at my friend.”

Generated Sentence: “I washed the dishes.”

VLCMU

VLCMU

Purport Network

Purport Network

impact score $\gamma$

shot number

\[ f_1^1 \rightarrow \cdots \rightarrow f_k^1 \]

\[ f_1^N \rightarrow \cdots \rightarrow f_k^N \]
Caption Generation Network

\[ \mathcal{L}^c = - \sum \log p_r(w_i) \]
Visual-Language Content Matching Network

\[ L^n = - \sum \tilde{\eta}_p \log \alpha_p + (1 - \tilde{\eta}_p) \log(1 - \alpha_p) \]
Purport Network

\[ \mathcal{L}^{\psi} = - \sum \psi_p \log \beta_p + (1 - \psi_p) \log (1 - \beta_p) \]
Experiments

• Dataset: UT-Egocentric

• Evaluation:
  • Visual Domain
    • AUC of IOU
  • Text Domain
    • ROUGE
    • BLEU
    • METEOR


Quantitative Results – Text Domain

• Baselines:
  • State-of-the-art Methods
  • Video Caption Generation

<table>
<thead>
<tr>
<th></th>
<th>SeqDP</th>
<th>Superframes</th>
<th>SubMod</th>
<th>SubMod$\text{vis+lang}$</th>
<th>LM-SeqGDPP</th>
<th>LSTMMDPP</th>
<th>TSG</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGESU4</td>
<td>11.95</td>
<td>15.59</td>
<td>16.19</td>
<td>14.10</td>
<td>16.24</td>
<td>13.98</td>
<td><strong>17.33</strong></td>
</tr>
<tr>
<td>METEOR</td>
<td>19.02</td>
<td>18.36</td>
<td>18.90</td>
<td>18.02</td>
<td><strong>19.90</strong></td>
<td>17.32</td>
<td>19.27</td>
</tr>
<tr>
<td>BLEU2</td>
<td>16.69</td>
<td>34.68</td>
<td>35.28</td>
<td>35.38</td>
<td>35.11</td>
<td>31.72</td>
<td><strong>46.90</strong></td>
</tr>
</tbody>
</table>
Quantitative Results – Visual Domain

<table>
<thead>
<tr>
<th>Restaurant ~39 min</th>
<th>I looked around the restaurant.</th>
<th>I walked through the parking lot.</th>
<th>I walked through the store.</th>
<th>I looked at my cell phone.</th>
<th>I walked through the store.</th>
<th>I looked at the products.</th>
<th>I washed the dishes.</th>
<th>I added the ingredients to the pot.</th>
</tr>
</thead>
<tbody>
<tr>
<td>My friend and I sat at the table.</td>
<td></td>
<td>I looked at the bookstore.</td>
<td>My friend drove the car.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SeqDPP</th>
<th>Superframes</th>
<th>SubMod</th>
<th>SubMod_{vis+lang}</th>
<th>LM-SeqGDPP</th>
<th>LSTMDP</th>
<th>TSG</th>
</tr>
</thead>
</table>
Thank you!