Exploiting Spatio-temporal Graphs Generation for Open-ended VQA
Kim Chheu, Aisha Urooj Khan, Dr. Mubarak Shah, Dr. Niels Lobo

Project Overview

**Project Goal:**
Solve the task of open-ended video question answering using the AGQA benchmark.

Previous video question answering benchmarks:
- Rely on additional text
- Bias from human-generated questions
- Models can exploit linguistic biases

AGQA was created to address these problems

Dataset

**Action Genome Question Answering Benchmark:**
- Purely vision-based
- Hand-crafted 2.2M question-answer pairs
- Various techniques to eliminate bias
- 3 metrics to test reasoning capabilities:
  - Novel Compositions
  - More Compositional Steps
  - Indirect References

Approach

We propose a graph-based approach for the task of video question answering.

- 3 main components: Visual Transformer, Graph Decoder, and Text Transformer

Model Architecture

Test Metrics Results

<table>
<thead>
<tr>
<th>Test Metric</th>
<th>PSAC</th>
<th>HME</th>
<th>HCRN</th>
<th>SHG-VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel Compositions</td>
<td>34.71</td>
<td>33.15</td>
<td>34.13</td>
<td>36.01</td>
</tr>
<tr>
<td>More Compositional Steps</td>
<td>47.19</td>
<td>47.72</td>
<td>46.63</td>
<td>46.97</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Metric</th>
<th>PSAC</th>
<th>HME</th>
<th>HCRN</th>
<th>SHG-VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indirect References</td>
<td>38.80</td>
<td>39.23</td>
<td>41.52</td>
<td>43.84</td>
</tr>
<tr>
<td>Object</td>
<td>25.96</td>
<td>25.67</td>
<td>33.17</td>
<td>39.61</td>
</tr>
<tr>
<td>Action</td>
<td>30.64</td>
<td>31.80</td>
<td>34.71</td>
<td>42.21</td>
</tr>
</tbody>
</table>

Conclusions

- We only trained on 100K questions, so results are not directly comparable to these baselines.
- We achieve a higher overall VQA accuracy than all baselines despite training on only 100K subset of questions.
- We outperform all baselines when evaluating on novel compositions and indirect references.
- Our model struggles to generalize as questions get more complex.