Heterogeneous Graph Learning for Visual Commonsense Reasoning

Weijiang Yu, Jingwen Zhou, Weihao Yu, Xiaodan Liang, Nong Xiao

Sun Yat-sen University (SYSU), China

NeurIPS 2019 Spotlight Presentation
Vancouver, Canada

December 2019
Previous Works

➢ Powerful Backbone, such as resnet152, bert-large.

➢ Graph-based Methods

(a) Answer-to-Answer Homogeneous Graph
(b) Vision-to-Vision Homogeneous Graph
(c) Vision-to-Answer Heterogeneous Graph
Previous Works

➢ Powerful Backbone, such as resnet152, bert-large.

➢ Graph-based Methods

(a) Answer-to-Answer Homogeneous Graph

(b) Vision-to-Vision Homogeneous Graph

(c) Vision-to-Answer Heterogeneous Graph
The goal of heterogeneous graph is to explore proper semantic alignment between and linguistic domains and knowledge reasoning to generate persuasive reasoning paths.

The contextual voting module is for visual scene understanding with a global perspective at the low-level features.
The implementation details of our heterogeneous graphs by taking the representation of image, question and answer as inputs.
## Experimental Results

<table>
<thead>
<tr>
<th>Text Only</th>
<th>Model</th>
<th>$Q \rightarrow A$ Val</th>
<th>$Q \rightarrow A$ Test</th>
<th>$QA \rightarrow R$ Val</th>
<th>$QA \rightarrow R$ Test</th>
<th>$Q \rightarrow AR$ Val</th>
<th>$Q \rightarrow AR$ Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance</td>
<td></td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>6.2</td>
<td>6.2</td>
</tr>
<tr>
<td>BERT [12]</td>
<td></td>
<td>53.8</td>
<td>53.9</td>
<td>64.1</td>
<td>64.5</td>
<td>34.8</td>
<td>35.0</td>
</tr>
<tr>
<td>BERT (response only) [44]</td>
<td></td>
<td>27.6</td>
<td>27.7</td>
<td>26.3</td>
<td>26.2</td>
<td>7.6</td>
<td>7.3</td>
</tr>
<tr>
<td>ESIM+ELMo [8]</td>
<td></td>
<td>45.8</td>
<td>45.9</td>
<td>55.0</td>
<td>55.1</td>
<td>25.3</td>
<td>25.6</td>
</tr>
<tr>
<td>LSTM+ELMo [34]</td>
<td></td>
<td>28.1</td>
<td>28.3</td>
<td>28.7</td>
<td>28.5</td>
<td>8.3</td>
<td>8.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VQA</th>
<th>Model</th>
<th>$Q \rightarrow A$</th>
<th>$QA \rightarrow R$</th>
<th>$Q \rightarrow AR$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RevisitedVQA [19]</td>
<td>Baseline</td>
<td>63.8</td>
<td>67.2</td>
<td>43.1</td>
</tr>
<tr>
<td>BottomUpTopDown [2]</td>
<td>Baseline w/ CVM</td>
<td>65.6</td>
<td>68.4</td>
<td>45.4</td>
</tr>
<tr>
<td>MLB [22]</td>
<td>Baseline w/ QAHG</td>
<td>66.1</td>
<td>68.2</td>
<td>45.8</td>
</tr>
<tr>
<td>MUTAN [4]</td>
<td>Baseline w/ VAHG</td>
<td>66.4</td>
<td>69.1</td>
<td>46.4</td>
</tr>
<tr>
<td>R2C [44]</td>
<td>HGL w/o CVM</td>
<td>68.4</td>
<td>69.7</td>
<td>48.3</td>
</tr>
<tr>
<td>HGL (Ours)</td>
<td>HGL w/o QAHG</td>
<td>67.8</td>
<td>69.9</td>
<td>48.2</td>
</tr>
<tr>
<td>Human</td>
<td>HGL w/o VAHG</td>
<td>68.0</td>
<td>68.8</td>
<td>48.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>$Q \rightarrow A$</th>
<th>$QA \rightarrow R$</th>
<th>$Q \rightarrow AR$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGL</td>
<td>69.4</td>
<td>70.6</td>
<td>49.1</td>
</tr>
</tbody>
</table>

### Table 1: Main results of validation and test dataset on VCR with respect to three tasks. Note that we do not need any extra information such as additional data or features.

### Table 2: Ablation studies for our HGL on three tasks over the validation set.
Q: How is [person1] feeling?

a) [person5] is feeling very apprehensive and scared.
b) [person3] is feeling happy.
c) [person1] is feeling tired from the trip.
d) [person1] is getting angry at the witness. ✓

R: d) is right because...

a) [person1]'s glaring eyes and the tight set of his jaw and mouth suggest anger.
b) This is a courtroom and [person3] is probably a lawyer. He is looking towards the middle and not the side which means he is probably talking to the judge and not the witness.
c) [person1] has an angry look on his face, and is moving his mouth in a way that looks like he is shouting, this look is typical of one who is angry at another and is verbally challenging them.
d) [person1] is gritting his teeth. [person1] has a look of pure anger on his face. ✓

➢ The predicted result is shown as bold font, and the ground truth (GT) is shown as ✓.
Experimental Results

Q: What if [person2] fell?
A: Person2 would get wet.
R: Person2 is surrounded by water.

Q: Is it snowing outside?
A: Yes, it is snowing.
R: [person4] is dressed in a hat, scarf and a big jacket, his hat and shoulders are covered in white snowflakes.

(a) Baseline    (b) our HGL
Conclusion & Future Work

The key merits of our work lie in four aspects:
- a framework called HGL is introduced to seamlessly integrate the intra-graph and inter-graph in order to bridge vision and linguistic domain, which consists of a heterogeneous graph module and a CVM;
- a heterogeneous graph module is proposed including a primal VAHG and a dual QAHG to collaborate with each other via heterogeneous graph reasoning and guidance mechanism;
- a CVM is presented to provide a new perspective for global reasoning;
- extensive experiments have demonstrated the state-of-the-art performance of our proposed HGL on three cognition-level tasks.

Several thoughts:
- Characteristics of natural language, such causal relationship.
- The reasoning for the specific number, such as 2 > 1.
- The interaction between visual instance relationships and linguistic contextual semantics

Our code is available in https://github.com/yuweijiang/HGL-pytorch
thank you for your listening