Multi-modal Contextual Graph Neural Network for TextVQA

Yaoyuan Liang
lyy8ztc@outlook.com
Illustration of TextVQA and MCG Model

Brief description of TextVQA problem, and an illustration of our MCG model structure, which contains a GNN-based contextual information propagation mechanism.
Encoding Component

- Non-textual object features are extracted with a pre-trained Faster-RCNN model.
- For the scene texts in the image, we apply scene text detector Rosetta to identify tokens in the image. We get tokens, visual bounding box, and visual feature of scene texts. The visual feature is extracted through feeding the bounding box into the Faster-RCNN model.
- For the question, we follow the common practice as in other VQA works.
Relation Modeling Component

- Spatial Relationship modeling:
Contextual GNN Propagation Mechanism

\[ v_i^{(q)} = \sigma(q \cdot W_q v_i), \quad i = 1, 2, \cdots, K + M, \]

\[ v_i^{h+1} = \sigma \left( \sum_{j \in N_i} \alpha_{ij} \cdot W_h v_j^h \right), \]

\[ \alpha_{ij} = \frac{\exp \left( (U_i^l v_i^h)^T \cdot V_i^l v_j^h \right)}{\sum_{j \in N_i} \exp \left( (U_i^l v_i^h)^T \cdot V_i^l v_j^h \right)}, \]

\[ v_i^{h+1} = \prod_{l=1}^{L} \sigma \left( \sum_{j \in N_i} \alpha_{ij}^l \cdot W_h^l v_j^h \right) \]
### Results

*Overall model performance comparison. The Validation set accuracy (Val) is computed locally, while the Test set accuracy (Test) is obtained through the online judging system.*

<table>
<thead>
<tr>
<th>Model</th>
<th>Object Combine</th>
<th>OCR Combine</th>
<th>No. of GNN Layer</th>
<th>Rich OCR Feature</th>
<th>Acc. on Val</th>
<th>Acc. on Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>LoRRA [29]</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>26.56%</td>
<td>27.63%</td>
</tr>
<tr>
<td>MCG(max-pooling)</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td>yes</td>
<td>17.85%</td>
<td>17.34%</td>
</tr>
<tr>
<td>MCG</td>
<td>residual</td>
<td>residual</td>
<td>1</td>
<td>yes</td>
<td>29.29%</td>
<td>29.29%</td>
</tr>
<tr>
<td>MCG</td>
<td>2 att.</td>
<td>concat.</td>
<td>1</td>
<td>yes</td>
<td>27.68%</td>
<td>27.91%</td>
</tr>
<tr>
<td>MCG</td>
<td>2 att.</td>
<td>residual</td>
<td>1</td>
<td>no</td>
<td>27.81%</td>
<td>27.98%</td>
</tr>
<tr>
<td>MCG</td>
<td>2 att.</td>
<td>residual</td>
<td>2</td>
<td>yes</td>
<td>28.71%</td>
<td>29.06%</td>
</tr>
<tr>
<td>MCG</td>
<td>2 att.</td>
<td>residual</td>
<td>1</td>
<td>yes</td>
<td><strong>29.40%</strong></td>
<td><strong>29.61%</strong></td>
</tr>
</tbody>
</table>
Results - Qualitative

What is the name of the hotspot?
LoRRA: gates
MCG: vodafone

What company is on the advert?
LoRRA: zemel
MCG: nationwide

What kind of GPS logger is it?
LoRRA: peceoi
MCG: wireless

What brand is the yellow box?
LoRRA: cauksng
MCG: triscuit

How much time is left on the washing machine?
LoRRA: 0
MCG: 120

What city is named?
LoRRA: new york
MCG: martinborough
Results - Faulty

*How many way stop is this sign for?*
- **LoRRA:** 3
- **MCG:** all
- **Human:** 4

*What is the largest number on the top row of this ruler?*
- **LoRRA:** 22
- **MCG:** 27
- **Human:** 28

*What does it say in blue?*
- **LoRRA:** kullik
- **MCG:** ilihakvik
- **Human:** kullik ilihaakovik