Introducing “Sense” in Visual Common”sense” Reasoning

Hammad A Ayyubi
Md. Mehrab Tanjim
-CSE 291G
The Task

Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

Rationale: I chose a) because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.
Formal Statement of the task

A model is given an image \( I \) and:

- A sequence \( o \) of object detections. Each object detection \( o_i \) consists of a bounding box \( b \), a segmentation mask \( m_1 \), and a class label \( i \in \mathbb{L} \).
- A query \( q \), posed using a mix of natural language and pointing. Each word \( q_i \) in the query is either a word in a vocabulary \( V \), or is a tag referring to an object in \( o \).
- A set of responses, where each response \( r(i) \) is written in the same manner as the query - a mixture of natural language and pointing. In this task, \( N=4 \) responses is used, of which exactly one is correct.
- 4 responses for answer and 4 responses for rationales are used.
Application

- Social Robots
- Healthcare Robotics
- Moving to next level of perception - cognition.

Task Difficulty

- Open-ended answers and rationales.
- Needs to not only predict correct answer but also the correct reason for it.
Baseline Model

Image I + objects o

Query q

Why is [person4] pointing at [person1]?

Response r(q)

He is telling [person3] that [person1] ordered pancakes.
Our Approach

We try to force the network to choose correct answers because of correct rationales. Essentially join (condition) the answer and rationale network.

We employ three approaches:

Answer prediction network: produces answer choice probabilities,
\[ p \ (N \times 4) = \text{softmax}(h) \]

Rationale prediction network: Question conditioned on the predicted answer as query.

1. **Softmax**

   Input to the Rationale network = \( p_1 \times QA_1 + p_2 \times QA_2 + p_3 \times QA_3 + p_4 \times QA_4 \)
2. **Gumbel-Softmax**

   The weight probabilities for QA pair (input to rationale) is calculated as:
   \[ p = \text{softmax}(\frac{1}{\tau} \cdot (h + g)) \]

3. **Cross Entropy**

   Train answer and rationale together, and do cross entropy. The network now has 20 predictions
   4 for Q->A, and 16 for QA->R, from which only 2 are correct.

   \[ \text{Loss} = \frac{4}{20} \cdot L_1 + \frac{16}{20} \cdot L_2 \]

   \( L_1 \): Loss Q->A  
   \( L_2 \): Loss QA->R

   Unlike their approach, here, we are training for all question answer pairs for rationale.
Experiments

Dataset

- VCR 1.0
- 110k movie scenes
- 290k multiple choice QA
- Images provided with objects segmentation maps, bounding box and labels.

Baseline Results (R2C):

<table>
<thead>
<tr>
<th>Q-&gt;A</th>
<th>QA-&gt;R</th>
<th>Q-&gt;AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>63.8</td>
<td>67.2</td>
<td>43.1</td>
</tr>
</tbody>
</table>
Cross Entropy

What we did:

- Five loaders which loads Q->A and QA->R
- 5 branches of the same model, one for Q->A, four for QA->R
- Calculate the loss as shown earlier

Results:
- After 12 epochs, the validation accuracies are:

<table>
<thead>
<tr>
<th></th>
<th>Q-&gt;A</th>
<th>QA-&gt;R</th>
<th>Q-&gt;AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q-&gt;A</td>
<td>61.55%</td>
<td>13.83%</td>
<td>8.42%</td>
</tr>
</tbody>
</table>
What we did:

- Previous Five loaders are merged into Single loader which saves some memory and time
- 2 models, one for Q->A, one for QA->R.
- Get the logits from Q->A model and pass this to QA->R model to calculate the loss as shown earlier

Results:
- After 6 epochs, the validation accuracies are:

<table>
<thead>
<tr>
<th>Q-&gt;A</th>
<th>QA-&gt;R</th>
<th>Q-&gt;AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>54.6%</td>
<td>59.6%</td>
<td>33.45%</td>
</tr>
</tbody>
</table>
Gumbel-Softmax

What we did:

- Pretty much same as before, but now we calculate gumble softmax as shown before
- Annealing Temperature (reduce temperature for first two epochs from 5 to 1, and keep it constant when it is 1)

Results:

- After 3 epochs, the validation accuracies are:

<table>
<thead>
<tr>
<th>Q-&gt;A</th>
<th>QA-&gt;R</th>
<th>Q-&gt;AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>54.18%</td>
<td>56.57%</td>
<td>32.09%</td>
</tr>
</tbody>
</table>
# Overall Results

<table>
<thead>
<tr>
<th>Approach</th>
<th>Epochs</th>
<th>Q-&gt;A</th>
<th>QA-&gt;R</th>
<th>Q-&gt;AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross Entropy</td>
<td>12</td>
<td>61.55%</td>
<td>13.83%</td>
<td>8.42%</td>
</tr>
<tr>
<td>Softmax</td>
<td>6</td>
<td>54.6%</td>
<td>59.6%</td>
<td>33.45%</td>
</tr>
<tr>
<td>Gumbel-Softmax</td>
<td>3</td>
<td>54.18%</td>
<td>56.57%</td>
<td>32.09%</td>
</tr>
<tr>
<td>R2C</td>
<td>20</td>
<td>63.8%</td>
<td>67.2%</td>
<td>43.1%</td>
</tr>
</tbody>
</table>
Conclusion

- We trained an end to end network conditioning the rationale on predicted answer. Thus, enforcing the network to “actually” consider the rationale while predicting the answer. We presented three approaches for this, showing promising results.

- This was some baselines approaches to tackle the cognition task - plucking the low hanging fruits. The current leaders in the task (FAIR) have achieved only 46.3% accuracy, while humans have 85% accuracy.

- Need to imbibe contextual knowledge (experience, expert knowledge etc.) which humans are so expert at.

- We scaled up the provided codebase to calculate bert embeddings for each question/answer pair, rather than just the question/correct answer pair. We have submitted pull request with the fix. We hope it will help the research community iterate faster.
Future Work

**RL Based sampling (2 weeks)**
Sample the predict answer according to the probability, $p$: $a \sim p$
The loss is then calculated using expectation loss.

**Use Transformer attention (3 weeks)**
Use transformer networks to calculate attention in the network.

**Generate Reason (2 months)**
Of course, we need to choose the right answer for right reason but the current model is not actually learning to reason rather it is learning to map. We can try to generate reasonable reasons, and calculate BLEU score with the right reason in order to train.