Motivation

- Large pretrained language models excel on a variety of NLP tasks, but often suffer from a fundamental weakness: logical inconsistency

- Example of logical inconsistency:
  - Is the apple a fruit? Yes.
  - Is a plant a fruit? No.
  - Is the apple a plant? No.

- Can we use past predictions and NLI model output to build a self-reflecting, consistent pretrained QA model?

Our approach:
- Augmenting a pre-trained QA model with an external memory for storing past model prediction
- Integrate supervisory signals from a large pretrained NLI model to encourage consistency between past and future QA model predictions
- Evaluate model with adversarially sampled batches to increase conflict probability

Background

- Existing approaches for enforcing logical consistency rely on constraint solving algorithms like MassAT, which operate on hand-engineered constraint graphs and are confined to a finite set of entities and facts
- BeliefBank adds a novel memory layer on top of pretrained T5 Macao QA model to track model beliefs over time and modify raw PTLM answers to improve consistency

Methods

- RELATION -- QUESTION + ANSWER PREPROCESSING:
  1. IsA: abstraction, (HasHead; hasTail and Yes)
  2. Related: abstraction, (HasFather: yes) -->
  3. HasPart: abstraction, (HasPart: face yes) -->
  5. HasProperty: abstraction, (HasProperty: above yes) -->

- CONSIDERATION DEFINITION:
  Define a constraint \( c_i \) as a tuple of the form \( (x_i, \neg x_i, \text{context} \), where \( x_i \) are sentences \( \text{context} \), \( \neg x_i \) are sentences \( \text{context} \), \( \text{context} \), and \( x_i \) denotes the weight of the constraint \( c_i \)

- Logic CONSISTENCY DEFINITION:
  If \( x_i \) and \( \neg x_i \) are entailed by the same context, then \( x_i \) and \( \neg x_i \) are in conflict.

Experiments & Analysis

We evaluate whether our method can outperform a Macao QA baseline on BeliefBank1k yes/no questions across a variety of configurations (4 entities, 4 batches, batch size), in a streaming setting.

<table>
<thead>
<tr>
<th>Batch Size</th>
<th>Entities</th>
<th>Exact Match</th>
<th>F1</th>
<th>Avg Consistency</th>
<th>Avg Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20</td>
<td>0.85</td>
<td>0.87</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>0.86</td>
<td>0.88</td>
<td>0.86</td>
<td>0.86</td>
</tr>
</tbody>
</table>

Conclusion

- IntrospectQA performance suggests that augmenting a QA model with NLI model + memory management can improve logical consistency and F1 accuracy in a streaming setting on any topic

References