Introduction

In light of large, pre-trained language models (LMs) nearing and surpassing human performance on a breadth of language understanding tasks [1, 2, 3], we propose Tiered Reasoning for Intuitive Physics (TRIP), a more challenging evaluation targetizing physical commonsense in a densely annotated, tiered reasoning setting.

Baseline Approach

We propose a tiered architecture powered by large, pre-trained LMs and their contextual embeddings:

Baseline Results

We evaluate systems with three metrics:

1. Accuracy: requires story choice to be correct.
2. Consistency: additionally requires conflicting sentences in the implausible story to be correct.
3. Verifiability: additionally requires some physical states to be predicted for the conflicting sentences, and all predicted states must be correct.

Table 1: End and tiered tasks metrics for tiered classifiers on the validation set of TRIP trained on varied combinations of loss functions. Random baseline averaged over 10 runs.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Consistency (%)</th>
<th>Verifiability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>47.8</td>
<td>11.3</td>
<td>0.0</td>
</tr>
<tr>
<td>BERT</td>
<td>78.3</td>
<td>2.8</td>
<td>0.0</td>
</tr>
<tr>
<td>ROBERTA</td>
<td>75.2</td>
<td>6.8</td>
<td>0.9</td>
</tr>
<tr>
<td>DeBERTA</td>
<td>74.8</td>
<td>2.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Omit Story Choice Loss $\mathcal{L}_c$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT</td>
<td>73.9</td>
<td>28.0</td>
<td>9.0</td>
</tr>
<tr>
<td>ROBERTA</td>
<td>73.6</td>
<td>22.4</td>
<td>10.6</td>
</tr>
<tr>
<td>DeBERTA</td>
<td>75.8</td>
<td>24.8</td>
<td>7.5</td>
</tr>
</tbody>
</table>

A consistent but not verifiable prediction.

1. Tom brought a box to the table.
2. Tom opened the box.
3. Tom took scissors out of the box.
4. Tom cut up the line with the scissors.
5. Tom put the scissors in the box.

A consistent and verifiable prediction.

1. Tom brought a box to the table.
2. Tom opened the box.
3. Tom took scissors out of the box.
4. Tom put up his book with the scissors.
5. Tom put the scissors in the box.

Reasoning Breakdown

Figure 4: Distribution of ROBERTA successes and failures on TRIP. SC (sentence conflict) and PS (physical state) denote whether the predicted conflicting sentences or physical states are correct (+) or not (-).

Figure 5: Utility of physical state predictions for selected attributes. Among correctly predicted physical states, bar regions indicate how many contribute to consistent end task predictions (i.e., with successfully detected conflicts). Blue stars indicate macro-F1 score of state prediction.

Conclusion

Our results show that supervising large LMs based on high-level classification tasks in order to learn commonsense language understanding leads to inconsistent and unverifiable reasoning. In order to solve tasks like these coherently, we should directly train systems to incorporate multiple types of lower-level evidence. Our work provides an important first step toward this goal and strong intuition for future progress.

References


Links

Scan the QR codes for related websites:

PREPRINT · CODE · LAB SITE