

**Can BERT Learn Logical Reasoning?**

1. **Deductive Reasoning**: the ability to draw conclusions only based on given facts and rules.
2. We say a model can reason if it can reliably emulate logical reasoning steps.

**SimpleLogic**

- **Facts**: Alice is fast, Alice is smart.
- **Rules**: If Alice is fast and smart, then Alice is test. If Alice is normal and happy, then Alice is not sad.
- **Query 1**: Alice is fast. **Answer**: True.
- **Query 2**: Alice is sad. **Answer**: False.

**Sampling Data from SimpleLogic**

1. **Randomly sample facts & rules**.
   - Facts: B, C
   - Rules: A \( \Rightarrow \) D, B \( \Rightarrow \) E, B, C \( \Rightarrow \) F.

2. **Compute the correct labels for all predicates given the facts and rules**.

**Paradox**

We construct two datasets RP and LP, each with 280k examples, sampled from Rule-Priority and Label-Priority.

<table>
<thead>
<tr>
<th>Priority</th>
<th>Label-Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>P</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>

**What is Logical Reasoning?**

- a. **bounded vocabulary (≤ 150)** and bounded number of rules/facts (≤ 120).
- b. **bounded reasoning steps (≤ 6)**.
- c. **finite domain (≈ 10^{66} examples)**.
- d. **only definite clauses**.
- e. **predicates are purely symbolic**.
- f. **No language variance**: templated language.
- g. **Examples are self-contained and require no prior knowledge**.
- h. **Transformers can solve SimpleLogic**.

**What is Statistical Feature?**

If a certain statistic of examples has a strong correlation with their labels but cannot be used to fully determine the labels, we call it a statistical feature.

**Statistical Features are Inherent**

- **Monotonicity of entailment**: any facts and rules can be freely added to the hypothesis of any proven fact.
- The more rules given, the more likely a predicate is proved.
- **Pr(label = True | rule# = x) should increase (roughly)** monotonically with x.

**Removing Statistical Feature is Hard**

We down-sample from RP to obtain RP_b such that:

1. Pr(label = True | rule# = x) = 0.5 for all x
2. Pr(rule# = x) stays the same as RP

**Main Message**

- **Claim**: BERT trained on RP fails to generalize to RP_b, suggesting that BERT leverages rule# to make predictions.
- **Claim**: BERT trained on RP_b generalizes slightly better, indicating that statistical features inhibit model generalization.

**Statistical Features Explain the Paradox**

While though statistical features are strong signals for in-distribution examples, they vary as the distribution changes.

**Reference**


**Acknowledgments**

This work is partially supported by DARPA/PTC grant, NSF grants IIS-1848061, and NERCIS. JCT is with DEC, Germany, CUHK, and a Sloan Fellowship. This work is supported in part by Amazon scholarships.