Transformers as Soft Reasoners over Language
Peter Clark, Oyvind Tafjord, Kyle Richardson

Francesco Di Stefano
What is the aim of the authors?
Are Transformers models really able to reason?

Bob is big. Big people are rough.
All rough people are green.
Bob is green?

The conclusion "Bob is green" cannot be logically inferred from the given premises.

While it is stated that "Big people are rough" and "All rough people are green," there is no direct statement that "Bob is rough." Therefore, we cannot make a logical deduction that Bob is green based on the given premises alone.

It is important to note that making assumptions or drawing conclusions without sufficient evidence can lead to inaccurate conclusions. So, in this case, we cannot assume that Bob is green just because he is big.
Or they only remember a lot of things?

Given a closed word assumption: Bob is big. Bob is round. If someone is young and round then they are kind. Bob is kind?

Based on the closed word assumption, we only have information about Bob being big and round. We also know that if someone is young and round, then they are kind. However, we do not know if Bob is young or not, so we cannot conclude whether he is kind or not based solely on the information given.

Therefore, we cannot definitively say whether Bob is kind or not based on the information provided.
What do we want from the transformer?

Here we use a formal language, but we would like to use our human language!

So we want to express facts and rules in natural language.
How to create a dataset from which the model can learn how to “reason“?
What is the structure of a sample?

(context, question, answer)

Composed by some facts and some rules

A question that requires reasoning over the facts with the use of the rules

A binary answer, ‘true’ or ‘false’
How we can obtain these samples? (part 1)

Generate a context (so random facts and rules) and then use a forward inference to obtain all the implications from this latter. But it’s that enough?

No, we need a way to derive even the false questions, the authors solved this using the Closed-World Assumption (CWA), i.e. ‘everything that is not derived from the context is false’

But now we have to understand how to generate the context…
How we can obtain these samples? (part 2)

How to generate the facts?

Structure:

- attributes $is(e_i, a_j)$
- relations $r_k(e_i, e_k)$

Example:

- $is(Alan, Big)$
- $eats(Dog, Rabbit)$

How to generate the rules?

Structure:

\[ condition \land condition \] \( \rightarrow \) conclusion.

Example:

// If someone is young and round then they are kind.
\[ is(?X, Young) \land is(?X, Round) \rightarrow is(?X, Kind). \]
Structure of the whole dataset

Five datasets with 5 different maximum depth of inference (0, 1, 2, 3, 5)

For each of these we have two dataset’s type:

- Type 1: uses only the \textit{is()} predicate
- Type 2: uses \textit{is()} and 3 other predicates

For each of these we have a standard version and a version in which a negation (not) has been added in facts and rules conditions/conclusions

Lastly, false questions at each depth have been generated in two ways:

- Negating conclusion from the forward inference
- Random drawing from unproven facts
But we want everything expressed in our (not so formal) language!

\[ \text{condition} \land \text{condition} \ast \rightarrow \text{conclusion}. \]

- If \text{condition} [and \text{condition}]\ast then \text{conclusion}.
- All \text{attribute}\ast people|things are \text{attribute}.
- \text{attribute}\ast people|things are \text{attribute}.

Three natural language templates have been used.
How we can send one of these samples through a Transformer?

All the experiments have been conducted using RoBERTa-large model fine-tuned on the RACE dataset.
What are the results?
We can begin by testing on the standard generated dataset

Trained with binary cross entropy, evaluated measuring accuracy

<table>
<thead>
<tr>
<th>Training</th>
<th>Num Q</th>
<th>Mod0</th>
<th>Mod1</th>
<th>Mod2</th>
<th>Mod3</th>
<th>MMax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test (own)</td>
<td>~ 20000</td>
<td>100</td>
<td>99.8</td>
<td>99.5</td>
<td>99.3</td>
<td>99.2</td>
</tr>
<tr>
<td>Test (DMax)</td>
<td>20192</td>
<td>53.5</td>
<td>63.5</td>
<td>83.9</td>
<td>98.9</td>
<td>99.2</td>
</tr>
<tr>
<td>Depth=0</td>
<td>6299</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Depth=1</td>
<td>4434</td>
<td>57.9</td>
<td>99.0</td>
<td>98.8</td>
<td>98.5</td>
<td>98.4</td>
</tr>
<tr>
<td>Depth=2</td>
<td>2915</td>
<td>34.3</td>
<td>36.8</td>
<td>98.8</td>
<td>98.8</td>
<td>98.4</td>
</tr>
<tr>
<td>Depth=3</td>
<td>2396</td>
<td>20.4</td>
<td>23.1</td>
<td>71.1</td>
<td>98.5</td>
<td>98.8</td>
</tr>
<tr>
<td>Depth=4</td>
<td>2134</td>
<td>10.2</td>
<td>11.4</td>
<td>43.4</td>
<td>98.8</td>
<td>99.2</td>
</tr>
<tr>
<td>Depth=5</td>
<td>2003</td>
<td>11.2</td>
<td>12.3</td>
<td>37.2</td>
<td>97.6</td>
<td>99.8</td>
</tr>
</tbody>
</table>

Out-of-distribution tests (reasoning depth unseen in training)
And if we try to challenge the model with other hand-authored problems?

If someone is a bird and not abnormal then they can fly.
If someone is an ostrich then they are a bird.
If someone is an ostrich then they are abnormal.
If someone is an ostrich then they cannot fly.
If someone is a bird and wounded then they are abnormal.
If someone is wounded then they cannot fly.

Arthur is a bird. Arthur is not wounded. Bill is an ostrich.
Colin is a bird. Colin is wounded.
Dave is not an ostrich. Dave is wounded.


"is/is not flying" in Birds1, "can/cannot fly" in Birds2

increasing complexity with increasing number of rules

<table>
<thead>
<tr>
<th>Test ↓; Train →</th>
<th>Num Q</th>
<th>Mod0</th>
<th>Mod1</th>
<th>Mod2</th>
<th>Mod3</th>
<th>MMax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birds1</td>
<td>40</td>
<td>80.0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>97.5</td>
</tr>
<tr>
<td>Birds2</td>
<td>40</td>
<td>80.0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Electricity1</td>
<td>162</td>
<td>77.8</td>
<td>88.9</td>
<td>100</td>
<td>100</td>
<td>96.9</td>
</tr>
<tr>
<td>Electricity2</td>
<td>180</td>
<td>70.0</td>
<td>80.0</td>
<td>97.2</td>
<td>100</td>
<td>98.3</td>
</tr>
<tr>
<td>Electricity3</td>
<td>624</td>
<td>80.8</td>
<td>93.9</td>
<td>92.8</td>
<td>90.5</td>
<td>91.8</td>
</tr>
<tr>
<td>Electricity4</td>
<td>4224</td>
<td>91.9</td>
<td>97.5</td>
<td>93.6</td>
<td>86.0</td>
<td>76.7</td>
</tr>
</tbody>
</table>

All results are zero-shot (these rulebases completely unseen during training)
Do we really a Transformer for this?

Transformer architectures are **not strictly necessary** even if results show that other architectures do not perform as well as these latter.

Transformer architectures are not strictly necessary even if results show that other architectures do not perform as well as these latter.

an LSTM-based model for natural language inference

decomposable attention model
But, we had a simple question at the beginning, are Transformer able to reason? Let’s return on our main question
What if we try to remove facts from the context?

- A sentence is **critical** for a proof if removing it from this latter causes the prediction to flip from True to False.
- A sentence that is not critical for a proof it’s defined as **irrelevant** for this latter.

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Remove Irrelevant</th>
<th>Remove Critical</th>
<th>Remove Any</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (test)</td>
<td>99.4</td>
<td>99.6</td>
<td>81.2</td>
<td>96.3</td>
</tr>
</tbody>
</table>

(tested on the no-negation half of the DMax test set)

Average accuracy between ‘Remove Irrelevant’ and ‘Remove Critical’
”Why did you answer in this way to me?”

Using the data about which removed sentence in the context caused a flip in the answer, we can observe which sentences the model considers critical for the answer. In this way we can try to ask to the model what it used to answer our questions.
And if we try with a ‘more natural‘ natural language?

Alan, who is round, red, kind, and also green, tends to be rather blue. In the snow sits Bob, crying from being cold. Charlie has green teeth and rough skin. People also notice his blue eyes.  
A quite nice person who is red and green is also big.  
Any big, kind person that turns red is cold to the touch.  
Young, kind people have a habit of being nice.  
A kind person will certainly be young.  
Q1. Dave is nice. True/false? [F]  
Q2. Charlie is big. True/false? [F]  
Q3. Alan is nice. True/false? [T]

A new dataset of 40k examples, using crowdworkers to paraphrase our theories. Only Type 1 theories without negation has been used.
What is happening here? Some final conclusions

The results from the previous slide seems to suggest nothing too much different from what we observed from our initial example with ChatGPT, it seems that the Transformer architecture it’s relying on memory more than reasoning even in this case.

In fact we observe that changing only the ‘syntactic sugar’ of our contexts the model is not able to reason properly.