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?

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 $[\land \text{condition}]^* \rightarrow \text{conclusion}$.

Example:

// If someone is young and round then they are kind.

$\text{is}(?X, \text{Young}) \land \text{is}(?X, \text{Round}) \rightarrow \text{is}(?X, \text{Kind})$. 
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!

condition \( \land \) condition\(^* \rightarrow \) conclusion.

If condition [and condition]\(^* \) then conclusion.

All attribute\(^* \) people|things are attribute.

attribute\(^* \) people|things are 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

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

Out-of-distribution tests (reasoning depth unseen in training)

Trained with binary cross entropy, evaluated measuring accuracy
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.

Q1. Arthur can fly. True/false? [T]  Q2. Bill can fly. True/false? [F]
Q3. Colin can fly. True/false? [F]  Q4. Dave can fly. True/false? [F]

The circuit has a switch.
The switch is on.
The circuit has a light bulb.
If a circuit has a switch and the switch is on then the circuit is complete.
If a circuit does not have a switch then the circuit is incomplete.
If a circuit is complete then a current runs through the circuit.
If a current runs through a circuit and the circuit has a light bulb then the light bulb is glowing.
If a current runs through a circuit and the circuit has a bell then the bell is ringing.
If a current runs through a circuit and the circuit has a radio then the radio is playing.

Q1. The circuit is not complete. True/false? [F]
Q2. The light bulb is glowing. True/false? [T]
Q3. The radio is playing. True/false? [F]

```
| Num Q | Mod0 | Mod1 | Mod2 | Mod3 | MMax |
|-------|------|------|------|------|------|
|       | D = 0 D <= 1 D <= 2 D <= 3 DMax |
| Birds1 | 40   | 80.0 | 100  | 100  | 100  | 97.5 |
| Birds2 | 40   | 80.0 | 100  | 100  | 100  | 100  |
| Electricity1 | 162 | 77.8 | 88.9 | 100  | 100  | 96.9 |
| Electricity2 | 180 | 70.0 | 80.0 | 97.2 | 100  | 98.3 |
| Electricity3 | 624 | 80.8 | 93.9 | 92.8 | 90.5 | 91.8 |
| Electricity4 | 4224 | 91.9 | 97.5 | 93.6 | 86.0 | 76.7 |
```

All results are zero-shot (these rulebases completely unseen during training)

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

increasing complexity with increasing number of rules
Do we really need 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. An LSTM-based model for natural language inference and a decompositional attention model are alternatives.

| Training | Mod0 | Mod1 | Mod2 | Mod3 | MMax |
|----------|------|------|------|------|------|
| Test (own): | | | | | |
| RoBERTa | 100 | 99.8 | 99.5 | 99.3 | 99.2 |
| BERT | 100 | 99.3 | 98.2 | 97.0 | 96.9 |
| ESIM | 100 | 90.3 | 87.8 | 84.2 | 80.0 |
| DECOMP | 72.5 | 68.2 | 58.6 | 57.8 | 64.1 |

| Test (DMax): | | | | | |
| RoBERTa | 53.5 | 63.5 | 83.9 | 98.9 |
| BERT | 53.5 | 64.1 | 90.6 | 95.3 |
| ESIM | 53.5 | 66.4 | 73.2 | 79.6 |
| DECOMP | 56.5 | 58.1 | 56.4 | 57.4 |

(Includes questions at depths unseen during training)
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.

|                  | Original | Remove Irrelevant | Remove Critical | Remove Any |
|------------------|----------|-------------------|-----------------|------------|
| Accuracy (test)  | 99.4     | 99.6              | 81.2            | 96.3       |

(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.

**Statement:** The lion visits the rabbit. (TRUE)  **Depth:** 2

**Context:** If something visits the lion then it chases the rabbit. **The lion is red.**
- If something sees the squirrel and the squirrel is young then the squirrel chases the rabbit.
- If something is red and it visits the rabbit then the rabbit chases the lion.
- If the squirrel chases the lion and the squirrel visits the cat then the squirrel visits the lion.
- If something chases the lion then it is red. The lion is green. The squirrel visits the cat. The rabbit is big.
- The cat visits the rabbit. If something chases the lion and it visits the squirrel then the squirrel visits the cat. **The lion is cold.**
- The squirrel sees the rabbit. The rabbit chases the squirrel. The lion chases the cat. **Red things are young.**
- The lion sees the rabbit. The cat is young. If **something is cold and young then it visits the rabbit.** The squirrel is big.

**Statement:** The squirrel is green. (TRUE)  **Depth:** 3

**Context:** All rough things are young. The squirrel eats the tiger.
- **If something sees the squirrel and it is young then the squirrel is green.**
- If something eats the squirrel then it is rough. The tiger visits the bear. The tiger eats the squirrel.
- The lion eats the tiger. If the squirrel sees the bear and the bear sees the tiger then the bear sees the lion.
- Green things are rough. **The lion eats the squirrel.** If **something is green then it is big.** **The lion sees the squirrel.**
- The tiger is young. If something visits the squirrel and it visits the bear then the squirrel visits the bear.
- The tiger eats the bear. The bear is green. If something sees the squirrel then it eats the tiger. The tiger sees the bear.
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 seem 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.