Adversarial Evaluation of Transformers as Soft Reasoners

Anonymous ACL submission

Abstract

The RuleTaker models (Clark et al., 2020; Tafjord et al., 2020) have recently shown that transformers can be capable of learning to deductively inference over facts and rules provided as natural language sentences. This is a significant achievement, since they can eliminate the need for express the knowledge in a formal representation. In this paper, we evaluate the robustness of these models to adversarial attacks. We first investigate the availability of dataset biases: superficial cues which can be exploited by the models to obtain high accuracies without solving the task. We train a model on partial inputs, ignoring some parts that are essential for true reasoning. High accuracy obtained by this model reveals the existence of dataset biases. To examine possible inattention of the models to the necessary preconditions for valid reasoning, we present three adversarial attacks on the test set: ReplaceMid that replaces a word in the theory, AddMid which adds a new word to the theory, and ChangePolarity that negates one sentence. In our adversarial settings, the accuracy drops from an average of 97.55% to 67.10%. This highlights the need for development of more robust models in both logic and language complexity scopes.

1 Introduction

The task of enabling machines to reason over provided knowledge has been classically approached by explicitly presenting the knowledge in a formal language (Davis, 1979). This approach has the obvious drawback of requiring to formally present the knowledge which can be challenging. Thanks to the successes of Transformers (Vaswani et al., 2017) in natural language processing (NLP) tasks, the alternative approach of presenting the knowledge in natural languages has been recently pursued (Bhagavatula et al., 2019; Clark et al., 2020; Tafjord et al., 2020; Aghahadi and Talebpour, 2021). They have reported promising results in a narrow setting showing that BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) can emulate reasoning over theories authored in English.

In this paper, we evaluate the performance of these transformer-based models in adversarial settings to investigate their *true* reasoning. The base of this evaluation is the assumption that many NLP systems exploit some superficial cues in the dataset, known as dataset biases, to obtain high accuracy in the in-distribution settings without truly solving the task (Jia and Liang, 2017; Gururangan et al., 2018; Manjunatha et al., 2019).

To evaluate this assumption, we train a transformers-based model on partial inputs of the RuleTaker dataset (introduced in §2), ignoring some parts that are essential for reasoning (§3.1). High accuracy of the trained model shows the existence of some biases in the data that are likely to be exploited by the models. We then introduce three adversaries that minimally change the test inputs to evaluate the robustness of the trained models against attacks (§3.2). The significant drop in the performance of the models (see §4) show that even though transformers are fairly capable of extracting patterns to obtain high in-distribution accuracies, they are very fragile against small changes in the input distributions, which is likely a consequence of relying on superficial cues instead of truly solving the reasoning task.

2 Task and Data

RuleTaker (Clark et al., 2020; Tafjord et al., 2020) is a dataset for studying deductive reasoning over narrative texts, formulated as a question-answering problem. The inference rules are explicitly provided instead of implicitly inferred as in reading comprehension datasets (Baradaran et al., 2020). Each example in the dataset is a quadruple $(theory, question, answer, proof)$, where theory

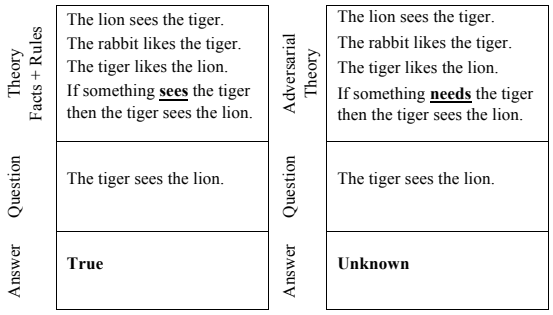

Figure 1: Left: an example of a syllogism (depth-1) in the RuleTaker dataset. Right: adversarial theory replaces the term "sees" with the term "needs"; this changes the label to unknown (an example of ReplaceMid adversary).

includes a number of facts and rules; question is a declarative sentence to prove; answer is true if question deductively follows from the theory, false if it is wrong based on the theory, and unknown under the open-world assumption (OWA) if it cannot be inferred to be true or false; and proof includes the statements in theory that contribute to the answer.

The dataset contains up to 5 depths of inference to prove the answers. Each theory has at least one question in each depth. Depth-0 refers to a lookup question and depth-1 refers to a syllogism. A syllogism is a common form of deductive reasoning which involves exactly two premises (Khemlani and Johnson-Laird, 2012). The anchor term that is present in both premises is called the *middle-term*. Figure 1 (left) shows a syllogism from the RuleTaker dataset.

Table 1 shows the statistics of depth-0 and depth-1 samples in the dataset. In the next section, we introduce the settings to evaluate the robustness of the transformers-based models train on this data in adversarial situations.

3 Evaluation Settings

We first introduce the *partial input* setting in which a model is trained on some parts of the input which are not informative-enough for reasoning. It is used to investigate the existence of biases in the RuleTaker dataset. We then introduce three different adversaries to evaluate the performance of transformer-based reasoners in adversarial settings.

3.1 Partial input

A generalizable model should base its decision on the *essential* parts of the input, which we define as those segments that, if removed, the model should not be able to reasonably produce the output. For instance, in the natural language inference task, where the goal is to determine the entailment relationship between a *premise* and a *hypothesis*, both parts are essential, such that the model should not be able to produce a sensible output without each of them. By *partial input* we mean the input with some of its essential parts removed. The high accuracy of a model trained on the partial inputs reveals the existence of some superficial cues, also known as dataset biases, which can be exploited by the model (Gururangan et al., 2018). This results in solving the dataset instead of the task, which is highlighted by high in-distribution accuracy and substantially lower out-of-distribution performance.

To evaluate the performance of transformers-based reasoners on partial inputs, we train a BERT (Devlin et al., 2019) and a RoBERTa (Liu et al., 2019) classifier just on the *question* part of the input, ignoring the *theory* which is obviously an essential part. As we show in §4, the classifier obtains an accuracy which is significantly higher than the chance level on this partial input data, which reveals the existence of exploitable biases in RuleTaker dataset.

3.2 Adversaries

To determine whether transformers' reasoning is limited to simple patterns, we introduce adversaries by altering the test set of the RuleTaker dataset. Consider the example in Figure 1 in which the the label is changed from true to unknown when one word is changed in the theory.

We define three adversary functions that modify the theory and answer in an example $(theory, question, answer, proof)$. To ensure that the functions generate valid examples, a human labeler evaluates some random generated instances (§4). We just target depth-1 examples. All theories in depth-1 contain a syllogism. Syllogistic reasoning has two conditions: 1) The middle-term should convey the same meaning in the premises.[1] 2) The middle-term should not come with modifying terms in one of the premises.

To check the sensitivity of the RuleTaker-trained model to the first condition, the ReplaceMid function replaces the middle-term with another term. To

[1] In a more general definition, the semantic meaning of the middle-term in one premise can be a synonym, hypernym, or hyponym of the other.

	RuleTaker train set	RuleTaker test set	Replace Mid	Add Mid	Change Polarity	Adversarial test set
All	69616	20210	1914	360	708	2982
True	18016	5214	-	-	708	708
False	18016	5214	-	-	-	-
Unknown	33584	9782	1914	360	-	2274

Table 1: RuleTaker dataset and adversarial data statistics.

check the sensitivity to the second condition, the AddMid function adds a new term to the middle-term. The added term is not included in the reference theory, but it is selected from the RuleTaker dataset. So, under the OWA, AddMid modifies the true labels to unknown. Refer to Figure 2 for an illustration of these steps. We also introduce ChangePolarity function to examine the sensitivity of the models to the negation of the premises. All three mentioned functions use a three-step procedure to generate sentences that are similar to the original premises in theory, but lead to different answers. We will provide more details in the following.

- **ReplaceMid**. To create an adversarial example for a question, we focus on facts and rules in the proof of the answer. The ReplaceMid function modifies an example with true label by replacing the middle-term in the second premise with a new selected word. After this change, the first and second premises do not match, and the label will not be true. On the other hand, as noted earlier, the added term is not included in the reference theory. So, there is no other sentence in theory that matches the adversary sentence, gives more information about it, or contradicts with the original answer. Therefore, under OWA, the label of the adversarial example would be unknown. Figure 1 (right) shows an example of this adversary.

- **AddMid**. This function changes the middle-term by adding extra items to it. In some true questions, the first premise includes two facts. Here again, the first and second premises do not match, and because there is no other sentence in theory to match the adversary sentence or give more information about it, we expect the label to be unknown. Figure 2 shows an example of this adversary.

- **ChangePolarity**. This function works on examples with false label by negating the second premise. Because the adversary sentence contradicts with the original sentence, we expect the false label to be changed to true.

Table 1 includes the statistics of the adversarial dataset. The aggregation of instances generated by ReplaceMid, AddMidd, and ChangePolarity functions constitutes the adversarial test set. Each dataset is split into 80/20% training and test sets, respectively. To make sure that after manipulating the data, the original answer cannot be obtained through an alternative proof, we only used the part of the original test set that has single proofs.

4 Results and Discussion

In this section, we present and discuss the results of evaluations introduced in §3. In all experiments, we utilized Google Colab platform which provides GPU Tesla K80.

Dataset Bias The *partial input* line in Table 2 shows the results of training BERT and RoBERTa models just on the *question* part of the input (detailed in §3.1). The obtained accuracies are obviously higher than chance-level baseline which is 33.3%. This reveals the existence of dataset biases which can be exploited by models to obtain high accuracies without solving the task. Note that this experiment does not prove that the models necessarily exploit these biases (Amirkhani and Pilehvar, 2021). If the models base their decision on these superficial cues, they cannot generalize to adversarial settings, which is examined in the following.

Adversarial Evaluation Table 2 shows the accuracy of the BERT and RoBERTa models against all three adversaries. According to this table, the transformers-based models trained on the RuleTaker dataset are significantly fragile against adversaries introduced in §3.2. For instance, AddMid adversary decreases the accuracy of RoBERTa model

Step 1: Extract the proof

Input: $answer$; $question$=Harry is blue?; $context$ =Dave is quiet. Fiona is blue. Harry is furry. Harry is green. Harry is kind. Harry is quiet. Harry is young. Green, young people are blue. If Dave is furry then Dave is cold. If someone is furry and cold then they are green.

Output: $answer$ = True; $proof$ = [triple4 triple7 → rule1]

Harry is green. Harry is young. Green, young people are blue.

Step 2: Generate adversary sentence

Green, young and **rough** people are blue.

Step 3: Replace in original theory

Input: $answer$; $question$ =Harry is blue? ; $context$ =Dave is quiet. Fiona is blue. Harry is furry. Harry is green. Harry is kind. Harry is quiet. Harry is young. Green, young and rough people are blue. If Dave is furry then Dave is cold. If someone is furry and cold then they are green.

Output: $answer$ = **Unknown**; $proof$ = [rule1 ← FAIL]

Figure 2: An example of the AddMid adversary procedure.

from 97.55% to 61.39%. ReplaceMid is even more effective, where it drops accuracy to 36.62%. Low accuracy of BERT under ChangePolarity function verifies its insensitivity to negation, which has also been reported in previous research (Ettinger, 2020).

Failure Analysis Investigating the adversarial cases that fool the model shows that for ReplaceMid adversary, in 99.42% of cases, the model outputs the same label produced for attack-free instance. This ratio is 97.10% for the AddMid adversary. This finding is in line with the known fact that NLP models are overly stable and do not realize that a small change can completely change the meaning of one sentence (Jia and Liang, 2017).

Human Evaluation To ensure that the generated adversarial instances are valid and trivial for humans, we randomly sampled and labeled 50 examples of AddMid, ReplaceMid, and ChangePolarity instances. We also added some attack-free records from the original dataset to make the datasets more natural. The results of this human evaluation presented in Table 3 show that the instances are completely trivial for humans, while BERT model has a considerable low accuracy on them.

	RoBERTa Acc.	BERT Acc.
RuleTaker train set	97.29	80.00
RuleTaker test set	97.55	79.25
Partial input	52.27	52.33
Adversarial test set	**67.10**	**58.71**
ReplaceMid	36.62	70.68
AddMid	61.39	69.44
ChangePolarity	99.01	21.23

Table 2: Accuracy of BERT-base and RoBERTa-base models in different settings (random baseline is 33.3%).

Sub Datasets	Records without attack	BERT Acc.	Human Acc.
ReplaceMid	17.53%	68.63	100
AddMid	13.46%	65.86	100
ChangePolarity	69.23%	41.06	100

Table 3: Human evaluation on random subsets of adversarial instances.

5 Conclusions

In this paper, we evaluated the recent success of transformers-based models in reasoning over natural language texts (Clark et al., 2020; Tafjord et al., 2020). We first showed that there are some dataset biases in the RuleTaker dataset which are likely to be exploited by the models to obtain fake high accuracies. Then, three adversarial functions were introduced to investigate true reasoning capabilities of the learned models. We observed that the learned models are significantly fragile against minimal adversarial changes. This illustrates that transformers partly use non-generalizable patterns to perform reasoning, highlighting the need for developing more robust real reasoners.

References

Zeinab Aghahadi and Alireza Talebpour. 2021. Language-based syllogistic reasoning using deep neural networks. *Cognitive Semantics*, 8(2).

Hossein Amirkhani and Mohammad Taher Pilehvar. 2021. Don't discard all the biased instances: Investigating a core assumption in dataset bias mitigation techniques. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Findings*.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2020. A survey on machine reading comprehension systems. *arXiv preprint arXiv:2001.01582*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*.

Randall Davis. 1979. Interactive transfer of expertise: Acquisition of new inference rules. *Artificial intelligence*, 12(2):121–157.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Sangeet S. Khemlani and Philip N. Johnson-Laird. 2012. Theories of the syllogism: A meta-analysis. *Psychological bulletin*, 138 3:427–57.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Varun Manjunatha, Nirat Saini, and Larry S Davis. 2019. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9562–9571.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2020. Proofwriter: Generating implications, proofs, and abductive statements over natural language. *arXiv preprint arXiv:2012.13048*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.