Fine-Tuning a QA Model Against Adversaries

Anonymous submission

Abstract

Machine learning models are often evaluated by measuring their performance on a dataset. These models can perform well by learning predictive patterns, known as dataset artifacts, instead of the intended behavior. We explore the impacts of dataset artifacts on an ELECTRA-small model trained on the SQuAD reading comprehension dataset by testing against an adversarial dataset. We then make our model robust to such adversaries through fine-tuning by inoculation, fine-tuning the model on a small sample of adversarial examples. We find that adversaries trick the original model by exploiting deficiencies in the original SQuAD dataset.

1 Introduction

Machine learning models are often evaluated by measuring the error on a held-out test set after training on an associated dataset. However, models may achieve high performance on these datasets by recognizing spurious yet predictive patterns ("dataset artifacts") while not learning the intended behavior. For example, models can perform well on test sets where the inputs are modified to be impossible to predict, such as hypothesis only baselines in natural language inference (Poliak et al., 2018), and poorly on adversarial sets with examples similar to the training data (Jia and Liang, 2017).

To investigate this behavior, we use a pre-trained model to predict answers to reading comprehension questions. Reading comprehension is a strong candidate task for investigating dataset artifacts, as models perform comparably to humans while likely not achieving similar levels of comprehension. We introduce adversaries to the input to test whether the model is learning as intended and observe the impact of dataset artifacts. We then attempt to address the model's shortcomings on the adversarial dataset by fine-tuning the model via "inoculation," exposing it to a sample of the adversaries.

2 Base Model Evaluation

2.1 Model

For our analysis we use the ELECTRA-small model (Clark et al., 2020), which uses an improved training method for the BERT architecture. We implement model training and evaluation using the HuggingFace `transformers` library, and also obtain the necessary datasets from the library.[1] For both the initial training and fine-tuning, we train the model over 3 epochs.

2.2 Original SQuAD

SQuAD is a reading comprehension dataset about Wikipedia articles (Rajpurkar et al., 2016). It contains 107,785 questions on specific paragraphs, which contain the answers. SQuAD is a particularly attractive dataset for exploring dataset artifacts as others have previously argued that many questions can be answered with heuristics such as type-matching (Weissenborn et al., 2017).

Article: Super Bowl 50
Paragraph: *In early 2012, NFL Commissioner Roger Goodell stated that the league planned to make the 50th Super Bowl "spectacular" and that it would be "an important game for us as a league".*
Question: *When did he make the quoted remarks about Super Bowl 50?*
Answer: *In early 2012*

Figure 1: An example from the SQuAD dataset.

Looking through the training data, we can find several supporting examples. Consider the example shown in Figure 1. Given the question type of "when," we logically see that "early 2012" is an easy prediction just by being a date, with "the 50th Super Bowl" being the only competing answer as

[1] https://huggingface.co/

an event. We can further test this by changing the question to just "When?", and find that the model still answers correctly.

Dataset	Model F1
Original SQuAD	86.1
SQuAD "When" questions	92.9
SQuAD "When?"	50.8

Figure 2: The performance of the model on SQuAD. We show the performance on the overall dataset, and for questions beginning with "When" for simplicity. We also show the performance if those questions are replaced with only "When?".

We apply this procedure to every question that begins with "When." The performance of the model on such questions is shown in Figure 2. After replacing each of these questions with just "When?", the model still achieves a high F1 score of 50.8%. This high performance even with no other context supports Weissenborn's claims, and suggests that the model can rely on heuristics to answer a significant number of questions without the need for actual comprehension.

2.3 Adversarial SQuAD

To further explore the impact of dataset artifacts, we test our original model against the adversarial SQuAD dataset created by Jia and Liang (2017), specifically the ADDSENT challenge set. The dataset is constructed by automatically generating and concatenating a distracting sentence to the input paragraphs of a subset of the original SQuAD dataset. These adversarial sentences are specifically designed to leave the original answer unchanged, while being grammatical sentences that look similar to their respective questions. For each of the sampled examples, the adversarial dataset contains the original question and up to three human-reviewed adversarial variations, for a total of $3{,}560$ questions. Because we are testing on this "full" dataset and not just the adversarial examples, our performance will be higher than in the original work.

Compared to the F1 of 86.1% on the original SQuAD, our model achieves a much lower 61.3% on ADDSENT. We find that the model performs comparatively worse on longer questions (see Figure 3). To explain this, we know that ADDSENT will change at least one word from the question to form the adversarial sentence. For longer questions, many of the words in the adversarial sentence will be shared with the questions, leaving more room to trick the model. With shorter questions each individual change can be significant, to the point where the model might completely ignore the adversarial sentence.

We can further see the impact of question length on performance by calculating the cosine similarity between the questions and adversarial sentences, which we graph in Figure 4. From the figure we see that most of the examples the model predicts incorrectly have low cosine similarity between the question and adversarial sentence. We can generally associate these low cosine similarities with shorter questions, because ADDSENT changes a higher percentage of their words. Thus for adversarial sets, we can reasonably conclude that examples that have high similarity (e.g. n-gram overlap) with spans in the paragraph that don't contain the answer are challenging for the model.

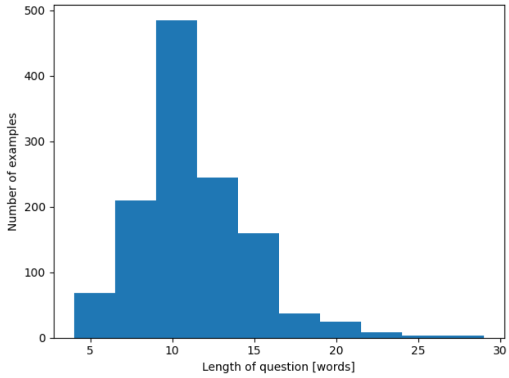

Figure 3: A histogram of the number of words in questions the model predicts incorrectly.

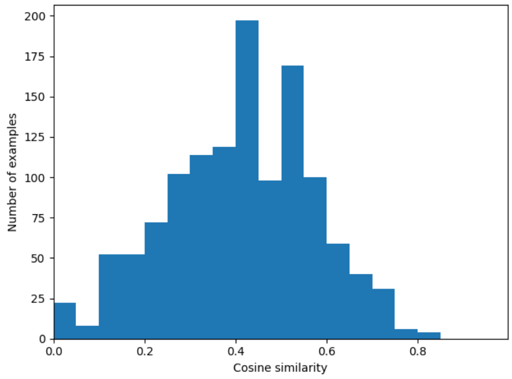

Figure 4: A histogram of the cosine similarity between the questions and adversarial sentences for examples the model predicts incorrectly. Stop words are omitted prior to calculation.

3 Inoculation by Fine-Tuning

To make our model robust to these adversarial sentences, we "inoculate" it by fine-tuning on a small sample of adversarial examples (Liu et al., 2019). Following the methodology of Liu et al. (2019), we test the performance using 10, 50, 100, 400, 500, 750, and 1,000 randomly chosen samples. Given that the adversarial challenge set has 3,560 total examples, these are roughly 0.3%, 1.4%, 2.8%, 11.2%, 14%, 21%, and 28% of the total dataset, respectively. For these samples, we move the adversarial sentence to a random sentence position, such that we don't break up any of the other sentences. While this method admittedly could potentially break co-reference links between sentences (and our assumption that the adversary doesn't change the original answer), it avoids the known issue of the model learning to ignore the last sentence when training on this data (Jia and Liang, 2017; Liu et al., 2019). We show our results in Figure 5.

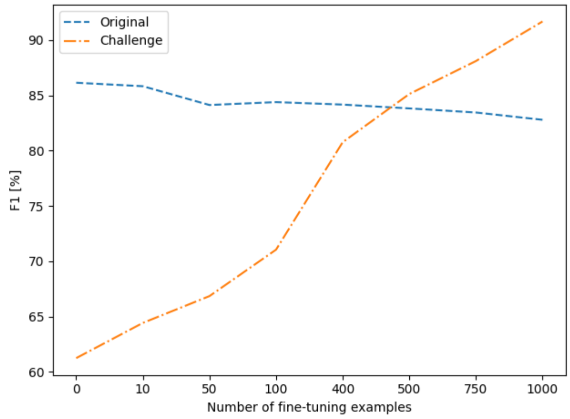

Figure 5: F1 scores of the fine-tuned models on the two datasets.

4 Discussion

The base model has F1 scores of 86.1% and 61.3% on the original SQuAD and adversarial datasets, respectively. After fine-tuning with 1,000 examples, the model has F1 scores of 82.8% and 91.7%, respectively. We find that the model quickly recovers the lost performance on the challenge set through this fine-tuning (getting 78.4% of the lost performance back fine-tuning on ∼10% of examples). Although the model loses some accuracy on the original SQuAD dataset (2.0% after 400 examples, 3.3% after 1000), it overall maintains a comparable high score.

These results most closely match outcome 1 from Liu et al. (2019) - the inoculated model retrains its high performance on the original test set and quickly performs well on the challenge dataset. The model actually performs better on the challenge set than the original after 500 examples, although at that point a significant portion of the test set (∼14%) is used for fine-tuning. This case suggests that the adversarial sentences reveal a lack of diversity in the original SQuAD examples. Considering the presence of examples like the one seen in Figure 1, this agrees with our analysis.

Surprisingly, these findings disagree with those of Liu et al. (2019). On their trials their BiDAF and QANet models lost more than twice the performance on the original dataset as our model. They speculate that their models take advantage of the adversarial sentence always occurring at the end of the input paragraphs, introducing a dataset artifact. Jia and Liang (2017) also found that when training on the adversarial data, their model just learned to ignore the last sentence for answering the question. By moving around the adversarial sentence in our fine-tuning, we can be reasonably confident that our model cannot make these same shortcuts. However, our model still similarly loses a bit of performance on the original dataset. Although it's difficult to distinguish the reasoning behind this, we can speculate that the presence of the adversary itself can be a dataset artifact. By design, the ADDSENT adversaries have a high degree of n-gram overlap with the original question, which was a heuristic commonly used by the base model. When fine-tuning on these adversaries we could be training the model to not use overlap as a heuristic as much, which could then lead to the decreased performance on the original set.

5 Conclusion

We trained a model on the SQuAD dataset and exposed the effect of dataset artifacts using an adversarial challenge dataset. Fine-tuning the model via inoculation, we make our model robust to such adversarial sentences, while retaining performance on the original dataset. Through our analysis, we speculate that the adversarial dataset breaks SQuAD-trained models by exploiting deficiencies in the original training data.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural QA as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 271–280, Vancouver, Canada. Association for Computational Linguistics.