A Comprehensive Analysis and Enhancement of Question Answering Task on Pre-trained Model

Doeun Lee, Kate Lee

Abstract

In pre-trained models, high performance can be achieved by unintended trends in the dataset, also known as dataset artifacts. To discover dataset artifacts in the SQuAD (Rajpurkar et al., 2016) dataset, we use adversarial attacks to confuse the model (Jia and Liang, 2017). From an in-depth error analysis on a randomly sampled set from the adversarial evaluation data, we propose training the model on various "challenging" datasets to mitigate the spurious correlations. After training the model on the modified training data that blends adversarial and original data, the same set of sample tasks are then revisited for a side-by-side before and after comparison. The overall performance of the model against adversarial attacks improved significantly without harming the performance of the original examples. Additionally, through analyzing question types and the error causations, our method is proven particularly effective in improving the named entity recognition error as well as accuracy on temporal questions.

1 Introduction

As the popularity of pre-trained models in machine learning grows, the efforts to solve real-world problems using machine learning have expanded its scope and variety exponentially. However, the increasing complexity of real-world data introduces a series of challenges. Most pre-trained models perform extremely well on train data and its test data. However, can the model perform well or generalize out of its normally encountered dataset?

Dataset artifacts refer to unintended biases, patterns, or irregularities present in a dataset that can impact the performance and generalization of machine learning models. Understanding and addressing dataset artifacts is crucial for developing robust and unbiased machine learning models. Failure to recognize and mitigate these artifacts can lead to models performing poorly in real-world scenarios.

In this project, we specifically aim to analyze and mitigate the dataset artifacts in SQuAD dataset (Rajpurkar et al., 2016) using ELECTRA-small model (Clark et al., 2020).

2 Analysis of Dataset Artifacts

In order to discover the spurious correlations in the dataset, we made various "versions" of the SQuAD dataset by applying contrast and adversarial methods. Both methods make minimal changes to the dataset, breaking any overly simple and non-generalizing patterns the model may have learned.

2.1 Contrast Set

Contrast sets (Gardner et al., 2020) are formed by making slight modifications to the example that change the golden value. We used a pre-constructed contrast set for SQuAD (Mlxen, 2022). This contrast set predominantly contained one-word substitutions such as replacing "last" with "first", "feature" with "lack", and "early" with "middle".

2.1.1 Contrast Set Error Analysis

Evaluation of the model on the contrast set resulted in a performance degradation from 78.35% accuracy to 76.18% compared to the original SQuAD dataset. We noticed that evaluation on the contrast set had a very minimal impact on the predictions due to the limited scope of the change in the chosen contrast set. Also, the question answering task in nature is not heavily affected by a few word modifications unlike NLI or sentiment analysis. Therefore, we decided to focus our analysis on the adversarial set instead.

2.2 Adversarial Set

Adversarial sets are formed by adding a sentence that is designed in a very similar context to the question to "confuse" the model. We specifically used the onesent adversarial examples published by Jia and Liang (Jia and Liang, 2017). Onesent adversarial examples choose one of the five randomly generated distractor sentences and add it to the end of the context.

2.2.1 Adversarial Set Error Analysis

The model's accuracy dropped from 78.35% to 47.5% when evaluated on the adversarial examples. This drop was expected because the sentence added at the end of the context is specifically tailored to each question to confuse the model. The distractor sentence often exhibits more similarity to the question than the actual context in which answer can be found and causes most models to drop below 60% in accuracy.

2.2.1.1 Analysis Based on Question Type

We first categorized the errors based on the type of question and investigated any correlation between accuracy and the question type. The analysis was concluded from randomly selected 50 adversarial examples.

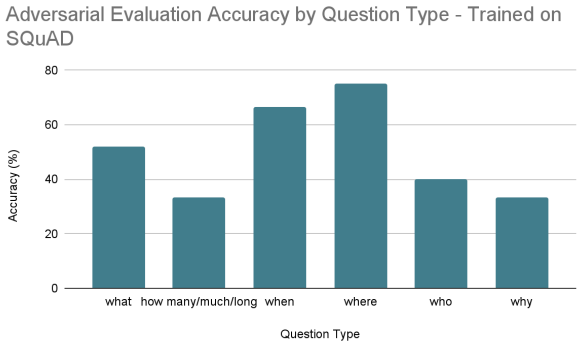

Figure 1: Adversarial evaluation accuracy based on question type (original model)

The average accuracy was around 50% across the categories. Performance of the model for "who", "why", and "how many/much/long" questions were below 40%. This indicates that questions involving numbers or names were particularly susceptible to adversarial attacks. Reasoning questions tend to have lengthy answers and require more in-depth understanding of context. For other types, the discrepancy in accuracy seemed to be more related to how "disruptive" or "confusing" the added sentence is than the difference in the question type itself.

2.2.1.2 Analysis of Sources of Error

From the analysis of adversarial set evaluation, the most common sources of error were word overlap, lack of contextual understanding, and named entity recognition.

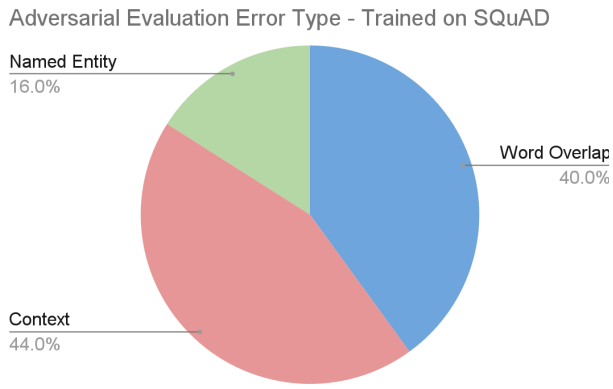

Figure 2: Distribution of error types in adversarial evaluation (original model)

[1] Word overlap between added sentence and question

If there is a sentence with greater word overlap than the answer sentence, the model may be confused into giving the answer according to the sentence with greater word overlap.

Example:

Question: What car is licensed by the FSO Car Factory and built in Egypt?
Added sentence: The American car is licensed by the VALCO Vehicle Factory and subsequently built in Arabia.
Predicted Answer: American car

In this example, the added sentence contains the majority of the exact words from the question whereas the context to the right answer “Polonez" is scattered across 3 sentences with different wordings.

[2] Lack of context understanding

The model may struggle to understand complex contexts or nuances in the questions and passages. This is especially challenging for questions that require reasoning over multiple sentences or paragraphs.

Example:

Question: How many hymns of Luther were included in the Achtliederbuch?
Added sentence: Vandross had 9 hymns included in the Achtliederbuch.
Predicted Answer: 9

The correct answer to this question is 4 according to this sentence in the context: “He supplied four of eight songs of the First Lutheran hymnal Achtliederbuch”. However, the model fails to take into account the complex multi-sentence context.

[3] Named Entity Recognition:
Mistakes may occur in recognizing named entities, such as people, locations, and organizations. Misidentifying entities can lead to incorrect answers.

Example:

Question: Where did Super Bowl 50 take place?
Added sentence: Champ Bowl 40 took place in Chicago.
Predicted Answer: Chicago

The correct answer to this question is inferred from “The game was played at Levi's Stadium in the San Francisco…”. The model fails to recognize the difference between the entity “Super Bowl” and “Champ Bowl”. However, this example may have a combination of other potential sources of error such as ambiguity of subject and word overlap.

3 Fix: Mitigating Dataset Artifacts

To mitigate the dataset artifacts described above, we explored different ways of training the model on adversarial data. By training on a more challenging dataset, the model can avoid making

spurious correlations and learn more complex decision boundaries.

3.1 Training Solely on Adversarial Data

We first tried training the model entirely on the adversarial training dataset. Although accuracy when evaluated on the adversarial data increased from 47.5% to 72%, the accuracy when evaluated on the original set dropped from 78.35% to 55.5%. This drop was significant and proved that training solely on adversarial data was not effective.

Due to the complicated and confusing nature of the adversarial examples, training solely on adversarial data seemed to have hindered the model from learning the necessary patterns for question answering tasks. Another contributing factor to the performance degradation could be the change in train and evaluation size. The adversarial training set had slightly fewer training examples due to data processing (from 87000 to 69000) which could limit the model's learning. Also, the original SQuAD evaluation dataset is bigger than the adversarial evaluation dataset, which could introduce more patterns and errors that weren't encountered during the training as well.

3.2 Training on Both Datasets

We considered further training the model on adversarial training set after training on the original dataset. However, we recognized the risk of overfitting to the training set due to the data duplication, which may lead to performance degradation on a different dataset. From evaluating the approaches from 3.1 and 3.2, we decided to proportionally mix two datasets while keeping the original training dataset size.

3.3 Solution: Training on Adversarial Augmented Set

We created an augmented training set by replacing 40% of the original training set with their adversarial pair. When trained on the adversarial augmented dataset, the model showed a significant improvement on adversarial evaluation set from 47.5% to 78% while maintaining the accuracy on the original evaluation dataset as 77.6%. This was a significant improvement compared to the two previous experiments on the original dataset and adversarial-only dataset.

4 Performance Analysis

In addition to the overall performance improvement described in 3.3, we revisited the 50 randomly sampled adversarial examples from the initial analysis to conduct in-depth quantitative and qualitative comparison between the original model and the model trained on the augmented dataset.

4.1 Analysis of Adversarial Examples Based on Question Type

Similar to last time, we categorized the questions into different types to scrutinize improvements according to the type of questions answered.

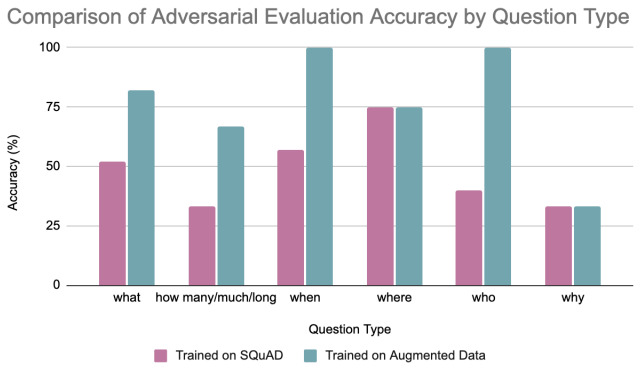

Figure 3: Adversarial evaluation accuracy before vs. after augmented data training

The average accuracy of this sample across all types improved from 50% to 80%, which shows that the model generally became more resistant to

adversarial attacks. The model trained on the augmented dataset had particularly high accuracy for answering “when” and “who” questions. Increase in accuracy was also observed for “what” and “how many/much/long” questions. The accuracy of “why” and “where” questions remained the same. However, the pie chart below shows that those two question types make up the least portion of the examples and thus improvement on “why” and “where” is difficult to conclude given the limited sample size.

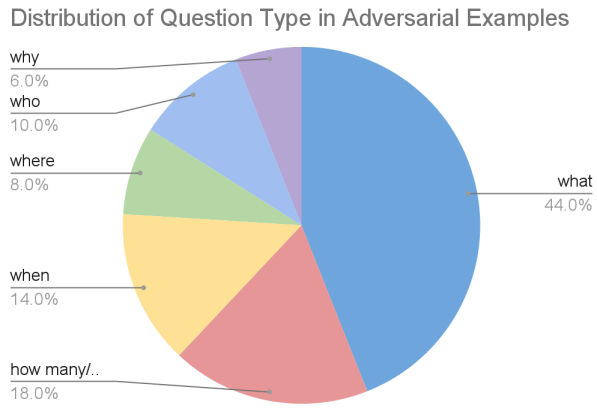

Figure 4: Distribution of question types in 50 randomly selected adversarial examples

4.2 Analysis of the New Source of Error

While the model trained on the augmented dataset exhibits better performance, we encountered a new source of error: partial answer.

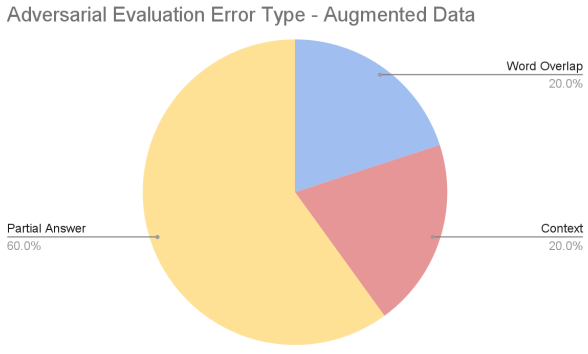

Figure 5: Distribution of error types in adversarial evaluation (augmented model)

There are two scenarios where this error can occur. First, if the intended answer is lengthy, the predicted answer tends to only partially match them. Second, the context gives details not reflected by the list of answers but by prediction.

Example 1:

Question: Why did OPEC block oil deliveries to the United States?
Answer: OAPEC proclaimed the embargo that curbed exports to various countries and blocked all oil deliveries to the US as a “principal hostile country
Predicted Answer: a “principal hostile country

In this example, the predicted answer includes only four words of the full answer.

Example 2:

Question: In the definition based off the mountain range, which region would the desert portions of north Los Angeles County be included in?
Answer: southern California
Predicted Answer: the southern California region

While the predicted answer essentially conveys the same meaning as the answer, it is classified as incorrect due to unnecessary elaboration on prediction.

Having a partially correct answer was the most prominent type of error in adversarial evaluation with the augmented model. All errors due to deficiency in named entity recognition were fixed, along with Word Overlap. Many context comprehension errors were converted to partial answer errors. This is an advancement in answer prediction, as what was once troubled by a fundamental understanding of the context over multiple sentences is now outputting predictions similar to the answers.

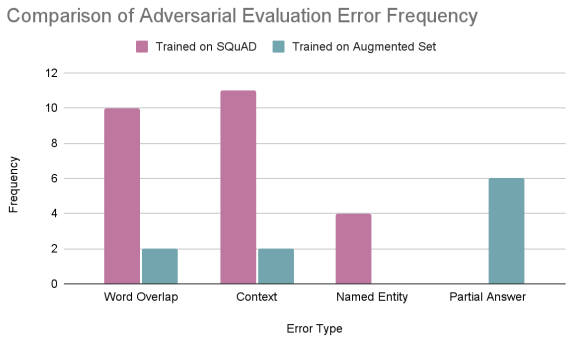

Figure 6: Adversarial evaluation error frequency before vs. after augmented data training

Overall, training on a mixture of original and adversarial data showed substantial performance enhancement when evaluated on adversarial examples while maintaining the accuracy on the original SQuAD evaluation set. Among the three commonly observed errors, named entity recognition error and word overlap errors showed significant improvement. A large portion of context comprehension errors were also mitigated to produce partially correct answers.

6 Conclusion

This study investigated dataset artifacts in the SQuAD (Rajpurkar et al., 2016) dataset using adversarial attacks to find vulnerabilities and suggest ways to mitigate them. Training on the SQuAD dataset and evaluating question answering tasks using adversarial examples revealed low performance against adversarial attacks primarily due to low comprehension of context, high word overlap between questions and adversarial sentences, and failure to recognize named entities. We proposed training the model on the augmented adversarial dataset to alleviate these vulnerabilities. This method challenged the model to learn more in-depth and complex correlations and thus led to a significant overall performance improvement when encountered with adversarial attacks.

References

[Clark et al., 2020] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pretraining Text Encoders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations (ICLR).

[Gardner et al., 2020] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets

[Jia and Liang, 2017] 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, Copenhagen, Denmark, September. Association for Computational Linguistics.

[Mlxen, 2023]. Mlxen, Hugging Face. Squad Contrasting Validation Dataset. Retrieved December 8, 2023, from https://huggingface.co/datasets/mlxen/squad_contrasting_validation_dataset

[Rajpurkar et al., 2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on

Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.