Investigating and Improving the Performance of ELECTRA-small on Adversarial QA

Varun Sridhar
Natural Language Processing
Fall 2021

Abstract

Our goal is to investigate the performance of ELECTRA-small on adversarial question-answering examples after training on the standard SQuAD dataset. We see that the model struggles to correctly answer the more challenging and complex questions in a dataset specifically comprised of adversarial questions. We examine why the model's performance is significantly worse on these questions compared to the questions in the SQuAD dataset by looking at specific examples from both datasets, and identifying general trends and patterns in the questions asked in both datasets. Finally, we attempt to improve the performance of ELECTRA-small on a combined dataset of SQuAD and adversarial examples with a variety of techniques centered around training on adversarial examples along with examples from SQuAD. Our best configuration results in an increase of more than 2% in exact match and F1 score in evaluation on a combined dataset.

1 Introduction

In natural language processing tasks, pretrained models can achieve high performance on a variety of tasks and benchmark datasets, but it can be hard to evaluate what exactly a model has learned in training. Even when a model performs well on a dataset, it could be exploiting certain patterns in the data and the task rather than actually comprehending the context and understanding the task it has been given. These spurious correlations that the model learns are known as dataset artifacts. In this paper, we explore the idea of dataset artifacts in the task of question answering.

One way to understand the effect of dataset artifacts on a model's performance is to construct adversarial examples and evaluate the model on these. Adversarial examples are purposely designed to be tricky or tough for a model. In the case of question answering, examples of adversarial data could include questions that are long and drawn-out, with confusing and extra information that may not totally relate to the true subject of the question, or questions that ask about obscure information in the context. In these datasets, the model cannot take advantage of the correlations from dataset artifacts, and thus tends to perform worse, calling into question how much is actually learned.

In this paper, we first train Clark et al.'s ELECTRA-small model on Rajpurkar et al.'s SQuAD dataset for three epochs, and then examine the performance of the trained model on a dataset comprised of adversarial examples (2020, 2016). We see that the trained model performs much worse on these examples. We then examine the structure of these adversarial examples to gain insight into the challenges of adversarial question answering. Finally, we use a variety of techniques to combine the SQuAD dataset with the adversarial examples into one training set, to see if we can improve performance of the model on the adversarial examples as well as general examples.

2 Analysis of SQuAD and Adversarial Dataset

For our initial analysis, we trained the ELECTRA-small model on the SQuAD dataset. The dataset had 87,599 examples of contexts and questions. We trained the model for three epochs on the entire dataset, and this resulted in the scores seen in Table 1 on an evaluation set. This evaluation set had 10,570 examples of contexts and questions (Rajpurkar et al., 2016).

We then ran our trained ELECTRA-small model on Bartolo et al.'s AdversarialQA dataset. AdversarialQA is made up of three datasets, constructed using three adversarial models: BiDAF,

Evaluation Set	Exact Match	F1 Score
SQuAD	78.86	86.50
AdversarialQA	18.23	28.27
Combined	65.45	73.62

Table 1: Exact match and F1 scores of our first ELECTRA-small model trained on SQuAD, evaluated on SQuAD, adversarialQA, and a dataset of combined examples.

BERTLarge, and RoBERTaLarge. These datasets consist of questions that current state-of-the-art models (including the ones used as adversaries) find challenging (2020). As seen in Table 1, the ELECTRA-small model performed significantly worse on the AdversarialQA evaluation set in comparison to SQuAD. This evaluation set had 3,000 examples of contexts and questions. We also evaluated our trained model on a combined dataset of SQuAD and AdversarialQA examples. To construct this dataset, we combined SQuAD and AdversarialQA and shuffled the order.

Why was performance so much worse on the adversarial examples? To gain insight into answering this question, we looked at some examples of contexts and questions in AdversarialQA. Consider the following example (text styling added for emphasis here) (Bartolo et al., 2020):

> Other shopping destinations **in Newcastle** include Grainger Street and the area around Grey's Monument, the relatively modern **Eldon Garden** and Monument Mall complexes, the Newgate Centre, Central Arcade and the traditional Grainger Market. Outside the city centre, the largest suburban shopping areas are Gosforth and Byker. The largest Tesco store in the United Kingdom is located in Kingston Park on the edge of Newcastle. **Close to Newcastle**, the largest indoor shopping centre in Europe, **the MetroCentre**, is located in Gateshead.
>
> Question: Which of the following is **inside Newcastle**, Eldon Garden or the MetroCentre?
> Answer: Eldon Garden
> Predicted Answer: MetroCentre

Here, the question asked tests how well the model comprehended the context. At the beginning of the context, we see a list of shopping destinations in Newcastle, one of which is Eldon Garden, which is the answer to the question. However later on, we see that the MetroCentre is located close to Newcastle. The model sees that Newcastle and the MetroCentre are located close to one another in the same sentence, and outputs MetroCentre as the answer. The sentence containing the correct answer has a lot of locations in it, and this may be confusing to the model. The question here tests if the model can understand the difference between "located in Newcastle" and "located close to Newcastle," and it ends up producing the wrong answer.

Consider another example from the AdversarialQA dataset (Bartolo et al., 2020):

> The development of fundamental theories for forces proceeded along the lines of unification of disparate ideas. For example, **Isaac Newton** unified the force responsible for objects falling at the surface of the Earth with the force responsible for the orbits of celestial mechanics in his universal theory of gravitation. **Michael Faraday and James Clerk Maxwell** demonstrated that electric and magnetic forces were unified through one consistent theory of electromagnetism. In the 20th century, the development of quantum mechanics led to a modern understanding that the first three fundamental forces (all except gravity) are manifestations of matter (fermions) interacting by exchanging virtual particles called gauge bosons. This standard model of particle physics posits a similarity between the forces and led scientists to predict the unification of the weak and electromagnetic forces in electroweak theory subsequently confirmed by observation. The complete formulation of the standard model predicts an as yet unobserved Higgs mechanism, but observations such as neutrino oscillations indicate that the standard model is incomplete. A Grand Unified Theory allowing for the combination of the electroweak interaction with the strong force is held out as a possibility with candidate theories such as supersymmetry proposed to accommodate some of the outstanding unsolved problems in physics. Physicists are still attempting to develop self-consistent unification models that would combine all four fundamental interactions into a theory of everything. **Einstein** tried and failed at this endeavor, but currently the most popular approach to answering this question is string theory.
>
> Question: What **physicist worked alone** other than **Einstein**?
> Answer: Isaac Newton
> Predicted Answer: James Clerk Maxwell

Here, the model just outputs the last name that it sees before Einstein in the context. We posit that the model doesn't understand the idea of "working alone" from the question. It has enough understanding and context to output a name, but doesn't comprehend the true meaning of the question.

We then compared the examples seen in the AdversarialQA dataset to examples from the SQuAD data. Consider the following example (Rajpurkar et al., 2016):

> Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The

American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24-10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Question: Which NFL team **represented the AFC** at Super Bowl 50?
Answer: Denver Broncos
Predicted Answer: Denver Broncos

Here, the question is fairly simple, involving an extraction task. The model can just search the context for an instance of "AFC", and see that the team mentioned near it is the "Denver Broncos," and return that as the answer. Consider another example from SQuAD (Rajpurkar et al., 2016):

In India, private schools are called independent schools, but since some private schools receive financial aid from the government, it can be an aided or an unaided school. So, in a strict sense, a private school is an unaided independent school. For the purpose of this definition, only receipt of financial aid is considered, not land purchased from the government at a subsidized rate. It is within the power of both the union government and the state governments to govern schools since Education appears in the Concurrent list of legislative subjects in the constitution. The practice has been for the **union government to provide the broad policy directions** while the states create their own rules and regulations for the administration of the sector. Among other things, this has also resulted in 30 different Examination Boards or academic authorities that conduct examinations for school leaving certificates. Prominent Examination Boards that are present in multiple states are the CBSE and the CISCE, NENBSE...

Question: What body in India **provides policy directions** to schools?
Answer: union government
Predicted Answer: union government

Again, the question here is relatively simple. The model can simply extract the statement given by the question from the context and quickly find that "union government" answers the question. So, through these specific examples, we see some of the differences in difficulty and complexity of questions in the AdversarialQA dataset as compared to the SQuAD dataset.

Examining these specific examples as well as others in both datasets allowed us to identify general trends. In adversarial examples, the questions generally contained a lot of information, some of which was irrelevant to the true subject matter of those questions. This could confuse the model about what in the context is relevant to the question. There were also examples where the question asked something that was completely irrelevant to the context. Generally, the model was not able to pick this up, and would answer with something from the context that was incorrect. In contrast, we noticed that questions in the SQuAD dataset tested the model on more immediate recall-style tasks, where the model can learn a nearest neighbors-like answering strategy. It could generally be successful searching for words from the question in the context, and once it finds overlap, examining the sentences around the overlap to find an answer. These questions in some cases did not require a deep understanding of the context or question prompt, especially compared to the AdversarialQA questions.

To quantify some of these comparisons between the two datasets, we first examined the average number of unique words in questions in both datasets. We hypothesized that the AdversarialQA dataset would contain complex questions with more unique words than SQuAD. This was actually not the case, as the AdversarialQA dataset contained an average of 10.53 unique words per question, whereas the SQuAD dataset contained an average of 11.02 unique words per question. In addition, we examined overlap between the questions and the corresponding contexts in both datasets, to confirm our hypothesis that the SQuAD questions generally had a higher degree of overlap with their corresponding contexts. We examined the percent of words in each question that were also in the context, and then computed an average over all the examples in each dataset. Here, our hypothesis was confirmed: on average, 65.20% of the words in a question from the SQuAD dataset were also in the corresponding context. In contrast, this was 53.41% for the AdversarialQA dataset. Since there is more overlap with the SQuAD dataset, this could indicate that our model is indeed using simple word-finding strategies as opposed to truly understanding the context and question.

3 Methods

After conducting analysis of the two datasets, we focused on improving performance of the

ELECTRA-small model the adversarial examples, and more generally, on the combined dataset as a whole. To improve performance, we focused on including adversarial data in the training process, to see if the model could learn the strategies necessary to answer the more difficult questions in this data. Our hope was that seeing these adversarial examples would push the model away from the strategies that it picked up due to artifacts in SQuAD. From these harder examples, the model could potentially learn to better interpret the context and question and therefore improve overall performance.

We tried three different approaches to incorporate the AdversarialQA dataset into the training process. First, we trained a model on a combination of the SQuAD and AdversarialQA datasets by simply concatenating both datasets. This training set had 117,599 examples total, 87,599 from SQuAD and 30,000 from AdversarialQA (Rajpurkar et al., 2016; Bartolo et al., 2020). Second, we trained the ELECTRA-small model on a dataset with an equal number of examples from both datasets. This training set had 60,000 examples, 30,000 from SQuAD and 30,000 from AdversarialQA. Lastly, we took the trained ELECTRA-small model from our initial analysis (trained just on SQuAD) and fine-tuned by further training on the AdversarialQA training set, consisting of 30,000 examples. We discuss and analyze the results of all of our experiments in the following section.

4 Results

To evaluate our results, we use a baseline ELECTRA-small model trained on the SQuAD training set, and evaluated on the following evaluation sets: SQuAD, AdversarialQA, and a combined evaluation set made up of the SQuAD and AdversarialQA evaluation sets. Our baseline results are summarized in Table 1. The model performs fairly strongly on the SQuAD evaluation set, but has much more difficulty with the adversarial examples, and therefore performs worse on the combined set.

Table 2 shows the evaluation results of our first attempt at including adversarial examples in the training set. Here, we simply concatenated the full SQuAD and full AdversarialQA datasets, and used this combined dataset to train an ELECTRA-small model. As seen in Table 2, we see significant improvement on the exact match and F1 scores in evaluation of adversarial examples compared to our baseline. As a result, we see an overall improvement in performance on the combined dataset as well, and a small increase in the F1 score on just the SQuAD evaluation set.

Evaluation Set	Exact Match	F1 Score
SQuAD	78.76	86.58
AdversarialQA	27.63	38.40
Combined	67.46	75.93

Table 2: Exact match and F1 scores of an ELECTRA-small model trained on a combined dataset with 87,599 SQuAD examples and 30,000 AdversarialQA examples.

Evaluation Set	Exact Match	F1 Score
SQuAD	74.11	82.45
AdversarialQA	25.00	35.56
Combined	63.26	72.08

Table 3: Exact match and F1 scores of an ELECTRA-small model trained on a dataset with an equal number of SQuAD and AdversarialQA examples, 30,000 from each.

Table 3 displays the evaluation results of our second attempt at training on adversarial examples. In this configuration, we created a dataset with an equal number of examples, 30,000, from the SQuAD and AdversarialQA training sets. The rationale for balancing the number of examples from both datasets was so that the model could potentially learn as much from the adversarial examples as it did from the SQuAD examples. As shown in Table 3, we see improvement in performance specifically on the AdversarialQA evaluation set, but performance on SQuAD and on the combined dataset actually declines. This could be because the model loses some of the knowledge from the SQuAD examples we did not train on.

Table 4 shows the evaluation results from our last attempt at incorporating the adversarial examples in training. Here, we started with the ELECTRA-small model referenced earlier in this paper that we trained on the SQuAD data. We then finetuned this trained model by further training it on the AdversarialQA training set, consisting of 30,000 training examples. This training configuration resulted in our best performance on the AdversarialQA dataset, but our performance on SQuAD and the combined evaluation set declined

Evaluation Set	Exact Match	F1 Score
SQuAD	68.38	77.44
AdversarialQA	28.93	39.07
Combined	65.45	73.62

Table 4: Exact match and F1 scores of our original ELECTRA-small model, finetuned on the AdversarialQA training set consisting of 30,000 examples.

somewhat significantly.

For the final part of our analysis, we investigated some examples from the AdversarialQA evaluation set and examined if our best model (ELECTRA-small trained on both full training sets, results shown in Table 2) showed any differences in understanding of these examples. Consider the following example (styling added here is ours) (Bartolo et al., 2020):

> **Newcastle** International Airport is located approximately 6 miles (9.7 km) from the city centre on the northern outskirts of the city near **Ponteland** and is the larger of the two main airports serving the North East. It is connected to the city via the Metro Light Rail system and a journey into Newcastle city centre takes approximately 20 minutes. The airport handles over five million passengers per year, and is the tenth largest, and the fastest growing regional airport in the UK, expecting to reach 10 million passengers by 2016, and 15 million by 2030. As of 2007[update], over 90 destinations are available worldwide.
>
> Question: What is the second city mentioned?
> Answer: Ponteland
> Predicted Answer: Ponteland

In this example, the question tests how well the model comprehended the context. This question is phrased in a way that doesn't necessarily overlap too much with the context. There is no notion of "first" city and "second" city in the context, so the model must have some deeper understanding of what the question is asking besides just being able to extract the contents of the question from the context. Training on these adversarial examples possibly gave the model a better sense of how to comprehend these more difficult questions.

Consider another example showing improvement of model comprehension after training on adversarial data (Bartolo et al., 2020):

> In a report, published in early February 2007 by the Ear Institute at the University College London, and Widex, a Danish hearing aid manufacturer, **Newcastle** was named as the noisiest city in the whole of the UK, with an average level of 80.4 decibels. The report claimed that these **noise levels would have a negative long-term impact on the health of the city's residents**. The report was criticized, however, for attaching too much weight to readings at arbitrarily selected locations, which in Newcastle's case included a motorway underpass without pedestrian access.
>
> Question: People living in **which area** would have their **hearing ability adversely affected**?
> Answer: Newcastle
> Predicted Answer: Newcastle

This example formulates the question in a way that could be challenging for the model to interpret. The question contains the term "hearing ability adversely affected," which is not how the hearing damage discussed in the context was exactly described. The simple extraction strategy with overlap from the question would not work here, so the model needs to have some understanding of the context and question to answer correctly. These examples and the improved performance we show in Tables 2, 3, and 4 show that performance on question answering on adversarial examples can be improved with adding these examples to the training set.

5 Conclusion

In this paper, we presented some of the issues that pretrained models have in question answering tasks on adversarial examples, which are constructed to be challenging for these models that generally perform well on standard datasets. We analyzed the differences between examples in a standard dataset, SQuAD, and an adversarial dataset, AdversarialQA, so see why the adversarial examples can be more difficult for pretrained models to predict on. We then incorporated these adversarial examples into training an ELECTRA-small model, and compared its performance to an ELECTRA-small model trained only on SQuAD. We saw that the model trained on a full combined dataset of SQuAD and AdversarialQA had the best performance, improving on our initial model by more than 2% on exact match and F1 scores on a combined set of SQuAD and AdversarialQA testing examples. These results show that this approach of incorporating adversarial examples in training is a promising way to improve a pretrained model's performance on adversarial examples.

Future work in this task could explore other methods to improve model performance on these

adversarial examples. Another related future interest in this area of research is to see if we can get a better sense of what a model understands about a context and question. We explored this idea briefly by looking at specific examples of contexts and questions to see what the model could answer correctly, but having a way to quantify model understanding beyond simple evaluation of correctness or exact match / F1 score would be quite useful in this task and other NLP tasks. This could be a very exciting and interesting area of research moving forward, with application in a variety of fields.

6 Acknowledgements

Thanks to Dr. Durrett and the TAs for a great semester!

References

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotations for reading comprehension. *arXiv preprint arXiv:2002.00293*.

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*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.