Enhancing ELECTRA-Small Model Robustness through Adversarial Training and Challenge Sets

Sandeep Singh
Fall 2023

Abstract

This study aims to enhance the robustness and performance of the ELECTRA-small model by incorporating adversarial training alongside standard dataset training SQUAD. It involves using adversarial challenge sets from various sources to target language model vulnerabilities, such as susceptibility to syntactic and semantic misinterpretations. The training begins with a baseline performance established on conventional datasets, followed by the integration of adversarial examples through direct inclusion and data augmentation, aimed at addressing the model's specific weaknesses. The model's effectiveness is evaluated against both standard and adversarial benchmarks, hypothesizing that this comprehensive approach will result in a more resilient and generalized model. The research expects to contribute significantly to the development of more robust NLP models, enhancing reliability in various applications and informing future strategies against adversarial tactics in NLP system development.

1 Introduction

This study focuses on augmenting the robustness and performance of the ELECTRA-small model (Clark et al., 2020) by training it on selected datasets complemented by various adversarial challenge sets. Recognizing the vulnerability of language models to adversarially crafted inputs, we aim to explore the efficacy of adversarial training in improving model resilience and understanding.

The training regimen integrates standard datasets with adversarial challenge sets from notable sources such as Rajpurkar et al.'s SQuAD dataset and adversarial data sets that uses three different model ss3w4 from BiDAF (Seo et al., 2016), BERTLarge (Devlin et al., 2018), and RoBERTaLarge (Liu et al., 2019) in the annotation loop and construct three datasets; D(BiDAF), D(BERT), and D(RoBERTa). These sets are carefully designed to expose and target specific weaknesses in NLP models, including but not limited to, susceptibility to syntactic tricks, logical fallacies, and semantic misunderstandings.

Initially, the ELECTRA-small model is trained on a conventional dataset to establish a baseline performance. Subsequently, we introduce adversarial examples. This process involves two key strategies: direct incorporation of challenge sets into the training data SQuAD and adversarial data AdversialQA augmentation, where new adversarial examples are generated based on the model's existing vulnerabilities.

Throughout the training process, the model's performance is evaluated not just on standard benchmarks but also on separate adversarial test sets to assess its robustness against complex and deceptive inputs. We hypothesize that this dual approach of direct training combined with adversarial data augmentation will yield a model that is not only more robust but also exhibits improved generalizability and understanding in real-world scenarios.

The outcomes of this study are expected to contribute valuable insights into the development of more resilient NLP models, paving the way for more reliable applications in various domains such as automated question answering, sentiment analysis, and beyond. The enhanced understanding of adversarial impacts in training also aims to inform future research directions in the development of NLP systems resistant to evolving adversarial techniques.

2 Approach

2.1 SQUAD training

In the initial phase of our research, we focused on evaluating the performance of the ELECTRA-small model, a variant of the ELECTRA model known for its efficiency and smaller size, making it suitable for datasets that demand less computational power. The training was conducted on the Stanford Question Answering Dataset (SQuAD), a widely recognized benchmark in the field of natural language processing for evaluating reading comprehension abilities of models.

The SQuAD dataset is known for its challenging nature, consisting of 87,599 training examples, each comprising a context paragraph and a corresponding question. These examples are derived from a diverse range of Wikipedia articles, ensuring a broad spectrum of topics and styles in the training data. This diversity is crucial in developing a model capable of generalizing well across various domains and text formats.

Our training regimen involved running the ELECTRA-small model over this extensive dataset for three epochs. An epoch in machine learning represents a full pass through the entire training dataset. By iterating over the dataset multiple times, the model is given the opportunity to learn and adjust its parameters more effectively, potentially leading to better performance.

Following the training, we evaluated the model's performance on a separate evaluation set, as is standard in machine learning to assess generalization capabilities. This evaluation set, as referenced in the original SQuAD paper by Rajpurkar et al. (2016), contained 10,570 examples, also consisting of contexts and questions. These examples are distinct from the training set to ensure that the model is evaluated on previously unseen data, providing a more accurate measure of its real-world applicability and understanding.

The results of this evaluation, as detailed in Table 1 of our report, offer insights into how well the ELECTRA-small model, trained on a significant volume of data over multiple epochs, can comprehend and respond to a range of questions based on varying contexts.

2.1 AdversialQA training

The AdversarialQA dataset and its model training using the Electra Small model for 3 epochs with a batch size of 40 can be quite an interesting case to analyze. Let's break down the details you provided and understand each aspect in a simpler context.

Understanding the Context and Data:

AdversarialQA Dataset: This dataset is designed for benchmarking question answering models. It typically contains questions formulated in a way to challenge the comprehension abilities of the model, often including tricky or misleading elements. This makes it a robust dataset for testing the limits of a model's understanding capabilities.

Electra Small Model: Electra is a type of transformer-based model, similar to BERT, but with some key differences in training methodology. The 'Small' variant refers to a smaller, more efficient version of the Electra model. It's designed to be faster and require less computational resources compared to its larger counterparts, making it suitable for experimentation or scenarios with limited resources.

Training Parameters:

Epochs (3): An epoch is a full pass through the entire training dataset. Training the model for 3 epochs means that the model gets to see and learn from the entire dataset three times. This is a relatively short training duration, indicating a quick training process or an experimental setup.

Batch Size (40): The batch size dictates how many samples from the dataset are processed together in one step of training. A batch size of 40 strikes a balance between computational efficiency and the model's ability to generalize from the data.

Training Metrics:

Train Runtime (1740.3106 seconds): This indicates the total time taken for the training process. In this case, it took approximately 29 minutes, which is relatively quick, demonstrating the efficiency of the Electra Small model.

Train Samples Per Second (51.801): This metric shows how many individual data samples are processed per second. A rate of 51.801 samples per second is quite efficient, especially considering the complexity of natural language processing tasks.

Train Steps Per Second (6.476): This refers to the number of training steps (or batches) processed each second. With 6.476 steps per second and a batch size of 40, it indicates a steady processing rate, balancing between speed and the learning quality of the model.

Train Loss (2.7254956925457896): The training loss is a numerical representation of how well the model is performing. A loss of 2.72,

while not exceptionally low, might be expected given the challenging nature of the AdversarialQA dataset and the relatively short training duration. It reflects the error between the model's predictions and the actual outcomes.

Epoch (3.0): This simply reconfirms that the training was conducted over 3 epochs.

Interpretation:

Given these metrics, it appears that the Electra Small model was trained quickly and efficiently on the AdversarialQA dataset. The training loss suggests that while the model is learning, there is still room for improvement. This could be achieved by increasing the number of epochs, adjusting the batch size, or fine-tuning other hyperparameters.

2.3 Combined training : Squad and AdversialQA

After the initial training on the SQuAD dataset, our next step involved enhancing the complexity and robustness of our model's training. We achieved this by incorporating the AdversarialQA dataset into our training regime. This dataset is known for its challenging nature, designed specifically to test the resilience of models against adversarially crafted questions and contexts. The combination of AdversarialQA with SQuAD provided a rich, diverse, and challenging training environment, ideally suited for pushing the limits of the model's comprehension and reasoning capabilities.

During this training phase, we configured the model to process the data in batches of 40. Batch processing is a standard approach in machine learning that allows for efficient handling of large datasets. By processing 40 examples at a time, the model could optimize its learning process, balancing between computational efficiency and the granularity of learning from individual examples.

The training runtime is noted as 6405.6715 seconds, which indicates the total time taken to complete the training process. This duration is significant as it reflects the computational effort required to train the model over a complex and extensive combined dataset.

Further, the metrics 'train_samples_per_second' and 'train_steps_per_second' are critical in understanding the efficiency of the training process. The model processed approximately 55.153 samples per second and executed around 1.379 training steps per second. These figures provide insights into the model's speed and efficiency in handling the training data, which is particularly relevant when dealing with large and complex datasets.

The 'train_loss' value, recorded at 1.6532853770404463, represents the model's average loss over the training period. In machine learning, the loss function quantifies the difference between the predicted and actual outcomes. A lower loss value typically indicates better model performance. However, interpreting this value requires context, such as comparison with baseline models or previous iterations of training.

Finally, the training was carried out over 3 epochs. This means the model was exposed to the entire combined dataset three times, allowing it to learn and adapt its parameters iteratively. Multiple epochs are often necessary for models to effectively capture and learn from the nuances present in complex datasets.

Table 1 : Training data run on Google Colab for datasets.

Data Sets	Batch	Epochs	Data Count	Train Loss
SQuAD	40	3	87599	1.227
AdversarialQA	40	3	30000	2.725
Combined	40	3	117599	1.65328
Combined equal	40	2	60000	2.129

2.4 Complexity and Overlap

Average Complexity:

- The average complexity of a dataset, quantified here for SQuAD as 11.015 and for AdversarialQA as 10.532, likely reflects the level of linguistic or conceptual difficulty inherent in the dataset. This could encompass factors such as sentence length, grammatical structure, vocabulary level, and the intricacy of the ideas presented.
- A higher complexity score (as seen with SQuAD) suggests that the dataset contains more challenging texts or questions, possibly requiring more sophisticated comprehension and reasoning capabilities from models. This can be beneficial for training robust models but may also necessitate more advanced or specialized NLP techniques.
- The slightly lower complexity score for AdversarialQA indicates that, on average, its texts or questions might be somewhat simpler or more straightforward compared to SQuAD. However, given that it's specifically designed to include adversarial examples, this simplicity might be deceptive, with the challenge instead lying in the dataset's ability to mislead or trick models.

Overlap Score:

- Overlap scores measure how much similarity exists between different parts of the dataset. For SQuAD, the overlap score is 65.197, and for AdversarialQA, it's 53.412.

- A higher overlap score, as in SQuAD, suggests that there are more commonalities between different data points. This could mean that certain phrases, sentence structures, or types of questions and answers recur throughout the dataset. While this might aid in learning recurring patterns, there's a risk that models might overfit to these specific patterns rather than learning to generalize.
- AdversarialQA's lower overlap score indicates that it has a greater diversity in its content. This variety can be crucial in training models to handle a wider range of questions and contexts, particularly valuable for adversarial robustness. It reduces the risk of overfitting to specific patterns but might increase the complexity of the training process, requiring a model to learn from a broader array of examples.

Dataset	**Complexity**	**Overlap**
SQuAD	11.015	65.197
AdversialQA	10.532	53.412

Table 2: Complexity and Overlap

In summary, these metrics shed light on the nature of the datasets: SQuAD, while more complex, shows higher overlap, suggesting intricate but possibly more repetitive content. AdversarialQA, slightly less complex, offers more diversity, aligning with its goal to challenge and strengthen model robustness against a wider range of adversarial inputs.

3. Process and Results

In this detailed study, the primary objective was to enhance the performance of the ELECTRA-small model, particularly in handling adversarial examples and improving its overall effectiveness across a combined dataset of standard and adversarial questions. Recognizing the potential for model biases and limitations due to artifacts inherent in the SQuAD dataset, the introduction of adversarial data was intended to guide the model towards more robust and contextually nuanced strategies.

Methods:

Three distinct approaches were employed to dataset into the training regimen:

Baseline with SQuAD data Training: The first approach is to create a baseline model with SQuAD dataset and then run different dataset combinations to see the metrics.

Data Set	Exact Match	F1 score
SQuAD	**77.34**	85.42
AdversialQA	**17.27**	**27.32**
Combined	**63.93**	**72.46**

Table 3: Evaluation of dataset run on SQuAD trained Electra small model

Combined Dataset Training: The second approach involved creating a merged training set consisting of both SQuAD and AdversarialQA data, totaling 117,599 examples. This set comprised 87,599 examples from SQuAD and 30,000 from AdversarialQA. By training the ELECTRA-small model on this combined dataset, the intention was to expose the model to a broader range of question-answering scenarios, encompassing both standard and adversarial structures.

Data Set	Exact Match	F1 score
SQuAD	**77.465**	85.41
AdversialQA	**24.73**	**35.13**
Combined	**65.81**	**74.21**

Table 4: Evaluation of dataset run on a combined SQuAD and AdversialQA trained Electra small model

Balanced Dataset Training: The third strategy sought to maintain an equal representation from both datasets, creating a training set with 60,000 examples — 30,000 from SQuAD and an equal number from AdversarialQA. This approach aimed to give equal weight to standard and adversarial examples, potentially mitigating any bias or overfitting towards the styles and patterns predominant in the larger SQuAD dataset.

Data Set	Exact Match	F1 score
SQuAD	72.03	81.23
AdversialQA	**22.1**	**33.15**
Combined	61	70.6

Table 5: Evaluation of dataset run on SQuAD- Adversarial with same data number of data set
trained on Electra small model

The subsequent section of the study presents a thorough analysis and discussion of the outcomes from these experiments. By examining the model's performance across these different training scenarios, insights were sought into the most effective strategies for enhancing model robustness and accuracy, particularly in the face of challenging, adversarial inputs.

4. Conclusions

This paper delves into the realm of natural language processing (NLP), specifically focusing on the performance of pretrained models like ELECTRA in question-answering tasks. The core of the study is centered around understanding how these models, which typically excel on standard datasets like SQuAD (Stanford Question Answering Dataset), fare against adversarial examples. Adversarial examples are specially designed to be challenging and can expose the weaknesses of AI models.

I conducted a comparative analysis between standard examples found in SQuAD and those in an adversarially crafted dataset, AdversarialQA. This comparison was crucial to understand why adversarial examples pose a greater challenge for pretrained models. It's hypothesized that adversarial examples differ significantly in structure or content, requiring a deeper understanding or different approach for successful prediction.

To test this, I trained an ELECTRA-small model (a compact version of the larger ELECTRA model known for efficiency in NLP tasks) on a combined dataset of both SQuAD and AdversarialQA. This model's performance was then compared to an ELECTRA-small model trained solely on SQuAD. The results were noteworthy - the model trained on the combined dataset outperformed the one trained only on SQuAD, showcasing *more than a 2% improvement in both exact match and F1 scores when tested on a mix of SQuAD and AdversarialQA examples.*

These findings are significant for several reasons. First, they demonstrate that including adversarial examples in the training process can substantially improve a model's ability to handle such complex cases. This suggests that the robustness of pretrained models to adversarial examples can be enhanced, which is crucial for applications where accuracy and reliability are paramount.

Secondly, the research highlights the potential limitations of current pretrained models when faced with adversarially designed inputs. This has broader implications for AI safety and reliability, especially in fields where AI decision-making is critical.

Lastly, the improvement in performance metrics like exact match and F1 scores is a strong indicator of the effectiveness of this training approach. It opens avenues for further research into model training methodologies, especially in the realm of NLP and machine learning, where adversarial robustness is becoming increasingly important.

5. Acknowledgements

I respectfully express my sincere gratitude to **Dr. Durrett** and all the teaching assistants for their invaluable guidance and support throughout this remarkable semester. Their insights and assistance have been integral to our learning and research progress.

6. References

My github code references:
https://github.com/sandeepshabd/fp-dataset-artifacts

Huggingface Adversial QA
https://huggingface.co/datasets/adversarial_qa

Huggingface SQUAD schema
https://huggingface.co/datasets/squad/blob/main/README.md

Dr. Durrett Github page
https://github.com/gregdurrett/fp-dataset-artifacts

BiDAF (Seo et al., 2016) Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). Bidirectional Attention Flow for Machine Comprehension. In Proceedings of the 5th International Conference on Learning Representations (ICLR 2016).

BERTLarge (Devlin et al., 2018) Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019).

RoBERTaLarge (Liu et al., 2019)Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

Bartolo, M., Roberts, A., Welbl, J., Riedel, S., & Stenetorp, P. (2020). Beat the AI: Investigating Adversarial Human Annotations for Reading Comprehension. arXiv preprint arXiv:2002.00293.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In Proceedings of the International Conference on Learning Representations (ICLR).

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv e-prints, arXiv:1606.05250.