Mitigating Positional Biases with Adversarial SQuAD

Anonymous submission

Abstract

Pre-trained language models achieve high accuracy on benchmark datasets like SQuAD, but their success often stems from exploiting dataset artifacts rather than genuinely solving the underlying tasks. Building on the inoculation by fine-tuning strategy (Liu et al., 2019) and the Adversarial SQuAD dataset (Jia and Liang, 2017), I introduce a method for coreference replacement and adversarial sentence shuffling, encouraging models to rely on more comprehensive reasoning, rather than heuristic positional cues. I generate permutations of the SQuAD and Adversarial SQuAD datasets by altering the context through coreference replacement and random insertion of adversarial sentences within the context. I then fine-tune the ELECTRA-small model on each of these modified dataset perturbations. My results demonstrate that coreference replacement and adversarial sentence shuffling serve as effective evaluation augmentations but offer limited benefits for fine-tuning.

1 Introduction

The SQuAD reading comprehension task (Rajpurkar et al., 2016) is a standard benchmark for evaluating question-answering systems. Despite high performance, current state-of-the-art transformer models still struggle with true language understanding and reasoning, often relying on superficial cues rather than genuine comprehension, which can result in brittle performance in real-world applications.

Prior work from Jia and Liang (2017) introduced the Adversarial SQuAD dataset, which concatenates adversarial sentences at the end of context paragraphs to expose weaknesses in model robustness. Further research by Liu et al. (2019) utilized inoculation by fine-tuning with Adversarial SQuAD, a strategy where models are exposed to adversarial challenges during training to assess their adaptability. While these methods expose key weaknesses in model robustness, they fail to address the reliance on sentence position, motivating the need for approaches that challenge positional biases more directly. A key limitation of Adversarial SQuAD is its positional bias due to the fixed placement of adversarial distractor sentences at the end of the context. Models can exploit this predictable structure during training, learning to ignore the final sentence to improve performance. This positional bias restricts the model's ability to generalize to adversarial conditions where distractor sentences are placed in unpredictable locations. To address this, a more comprehensive approach is needed to reduce reliance on adversarial sentence placement and encourage the model to engage in deeper reasoning.

I introduce a method for coreference replacement and adversarial sentence shuffling, encouraging models to rely on more comprehensive reasoning rather than heuristic positional cues. I generate permutations of the SQuAD and Adversarial SQuAD datasets by altering the context through coreference replacement and random insertion of adversarial sentences within the context. I then fine-tune the ELECTRA-small model on each of these modified dataset perturbations. To evaluate the model's adaptability, I report the F1 score of the fine-tuned model across each variation of the datasets. My analysis reveals that before fine-tuning, the model's F1 score improves when coreference links are replaced and adversarial sentences are shuffled, compared to the baseline Adversarial SQuAD. After fine-tuning, the highest performance across adversarial dataset permutations is achieved by the model fine-tuned on the original Adversarial SQuAD with no modifications, indicating that coreference replacement and adversarial sentence shuffling serve as effective evaluation augmentations but offer limited benefits for fine-tuning. This result highlights the challenge of balancing model robustness across different dataset perturbations

and suggests that coreference replacement and adversarial sentence shuffling are more valuable as evaluation tools than as strategies for data augmentation during fine-tuning.

2 Methodology

Current question-answering models trained on Adversarial SQuAD rely on positional cues to identify correct answers, often ignoring adversarial distractor sentences placed at the end of the context. This reliance on positional cues limits model generalization to adversarial conditions where distractors are placed unpredictably within the context. To address this, I aim to mitigate positional bias by introducing coreference replacement and shuffling adversarial distractors throughout the context, encouraging models to engage in deeper reasoning. By modifying the coreference structure, I ensure that adversarial distractor sentences can be randomly inserted at different positions without breaking the logical flow of the context. The model is then fine-tuned on multiple dataset permutations, and its performance is evaluated to assess its ability to generalize across varying adversarial conditions.

The SQuAD (Stanford Question Answering Dataset) is a widely used benchmark for evaluating the ability of models to extract answers from context paragraphs in response to natural language questions. Each context paragraph is paired with a question, and models must predict the exact span of text containing the answer. The Adversarial SQuAD extends this task by introducing adversarial distractor sentences, which test the model's robustness to misleading information within the context. Concatenative adversaries, introduced by Jia and Liang (2017), place these distractors at the end of the paragraph, allowing models to exploit positional cues to improve performance.

Inoculation by fine-tuning, as introduced by Liu et al. (2019), is a fine-tuning strategy where models are exposed to a small amount of adversarial data during training to improve robustness on unseen challenge datasets. By fine-tuning on challenge-specific data, the models learns to adapt to new adversarial conditions. In this work, I apply inoculation by fine-tuning on modified versions of SQuAD and Adversarial SQuAD to evaluate the model's ability to generalize multiple adversarial dataset variations.

3 Experiments

I use the ELECTRA-small model from the HuggingFace library as the base model for all training and fine-tuning experiments. Training is conducted on an NVIDIA Quadro RTX 3000 GPU using PyTorch and HuggingFace Transformers. For initial training, I use a batch size of 16, a learning rate of 5e-5, and train for 3 epochs. During fine-tuning, I freeze all layers of the model except for the final layer to limit parameter updates and focus adaptation on adversarial dataset variations, using a batch size of 8, a learning rate of 3e-6, and train for 50 epochs.

To challenge the model's reliance on positional cues, I introduce a random adversarial sentence shuffling strategy. For each context paragraph, an adversarial distractor sentence is randomly inserted at any position except the first sentence, breaking the fixed end-of-paragraph position used in prior work by Jia and Liang (2017). To ensure logical consistency, I first perform coreference replacement, modifying pronouns and references so that the adversarial sentence can be inserted at arbitrary positions without disrupting context coherence.

Model performance is primarily evaluated using the F1 score, which balances precision and recall to provide a comprehensive measure of prediction quality. Evaluation is conducted at the end of training on a held-out portion of the dataset, with no intermediate evaluations during training. Additionally, I track training loss over epochs to monitor convergence.

3.1 Datasets

I use the SQuAD v1.1 dataset from the HuggingFace library, which contains 87,599 training examples and 10,570 validation examples. I limit the size to 2,000 examples when fine-tuning on CSQuAD. I also use the Adversarial SQuAD dev set from Stanford NLP in the HuggingFace library. This dataset is based on the SQuAD development set with adversarial sentences added. I use the AddSent variant, which contains 3,560 examples. The dataset is split into an 80/20 training and evaluation split, resulting in 2,051 training examples and 509 evaluation examples after filtering out non-adversarial examples that match the original SQuAD format.

CSQuAD is created by applying coreference replacement to all context sentences in 2,000 training examples from the SQuAD v1.1 dataset for fine-

Model	SQuAD	CSQuAD	ASQuAD	ACSQuAD	SACSQuAD
ELECTRA-SD	85.91	84.38	46.24	46.34	48.69
ELECTRA-SD-CSQuAD	81.68	79.56	49.95	48.49	44.97
ELECTRA-SD-ASQuAD	82.99	81.12	79.46	76.85	67.73
ELECTRA-SD-ACSQuAD	80.73	78.37	73.72	71.58	62.11
ELECTRA-SD-SACSQuAD	79.28	77.20	68.56	68.04	62.44

Table 1: F1 scores for the ELECTRA-small model variants evaluated on SQuAD, CSQuAD, ASQuAD, ACSQuAD, and SACSQuAD.

tuning. For evaluation, coreference replacement is applied to the entire SQuAD v1.1 validation set, allowing for a more comprehensive assessment of model performance. ACSQuAD is created by applying coreference replacement to all context sentences, excluding the adversarial sentence at the end of the paragraph, in the Adversarial SQuAD AddSent variant. The dataset is reduced to 2,051 training examples and 509 evaluation examples by filtering out instances that do not contain an adversarial sentence. ACSQuAD is used for both fine-tuning and evaluation. SACSQuAD is created from the same Adversarial SQuAD AddSent variant by applying coreference replacement to all non-adversarial sentences in the context paragraph and then shuffling the adversarial sentence, inserting it at a random position in the context paragraph (excluding the first sentence). After filtering out instances that do not contain an adversarial sentence, the dataset is again reduced to 2,051 training examples and 509 evaluation examples. SACSQuAD is used for both fine-tuning and evaluation.

Context paragraphs are sentence-tokenized using NLTK, enabling precise extraction and removal of the adversarial sentence from the rest of the paragraph. Coreference resolution is performed on non-adversarial sentences using spaCy and neuralcoref, ensuring all references are consistently replaced before re-inserting the adversarial sentence. The adversarial sentence is then repositioned using Python's random module, inserted at a random index other than the first sentence if the paragraph contains multiple sentences, or appended after the single context sentence otherwise.

3.2 Results

Table 1 presents the F1 scores of the ELECTRA-small model variants evaluated on SQuAD, CSQuAD, Adversarial SQuAD, ACSQuAD, and SACSQuAD. Prior to fine-tuning on adversarial variations, the baseline ELECTRA-SD model achieves a notably higher F1 score on the ACSQuAD and SACSQuAD sets compared to the original ASQuAD evaluation. Specifically, while ELECTRA-SD attains an F1 score of 46.24 on ASQuAD, its performance marginally improves to 46.34 on ACSQuAD and 48.69 on SACSQuAD, indicating that coreference replacement and adversarial shuffling can increase evaluation difficulty for simple heuristic cues and thereby reduce the model's reliance on positional information.

However, after fine-tuning on these modified datasets, the anticipated improvements in robustness do not materialize. The ELECTRA-SD-ASQuAD model, fine-tuned on the original Adversarial SQuAD without modifications, achieves the highest F1 scores across the Adversarial SQuAD and adversarially perturbed datasets. For instance, while ELECTRA-SD-ASQuAD scores 79.46 on ASQuAD and 76.85 on ACSQuAD, the variants fine-tuned on ACSQuAD or SACSQuAD directly (ELECTRA-SD-ACSQuad, ELECTRA-SD-SACSQuAD) do not surpass these results. Notably, the model fine-tuned on SACSQuAD attains only a 68.56 F1 on ASQuAD and 68.04 on ACSQuAD, both lower than when fine-tuned on the unmodified Adversarial SQuAD.

These findings suggest that although coreference replacement and adversarial sentence shuffling can serve as effective evaluation techniques, highlighting a model's positional biases and prompting more careful reasoning at test time, these modifications confer limited benefits for fine-tuning. In other words, training on these perturbed contexts does not yield substantial gains in out-of-distribution adversarial robustness compared to training on the original adversarial dataset.

4 Conclusion

These findings demonstrate that altering the SQuAD and Adversarial SQuAD datasets through coreference replacement and adversarial sen-

tence shuffling can highlight positional biases in transformer-based QA models. While these modifications improve evaluation difficulty and prompt more careful reasoning at test time, they do not substantially enhance fine-tuning outcomes compared to training on the unmodified Adversarial SQuAD. As a result, the approach primarily serves as a valuable evaluation augmentation rather than a means of improving general robustness through fine-tuning. The results show that before fine-tuning, coreference replacement and adversarial sentence shuffling modestly boost F1 scores on adversarial evaluations, indicating a reduced reliance on simple positional cues. After fine-tuning, however, training on these modified datasets does not translate into improved performance on out-of-distribution adversarial conditions, underscoring the limited transferability of these perturbations for robustness gains. One limitation of the introduced perturbations is that, while effective at revealing positional dependencies, they do not lead to improved adversarial robustness during fine-tuning. Additionally, the scope of the experiments, restricted to SQuAD and its adversarial variants, may not capture the full complexity of other QA benchmarks or reasoning-intensive tasks.

I introduce a method for applying coreference replacement and adversarial sentence shuffling to SQuAD-based datasets. By generating multiple dataset permutations and evaluating a fine-tuned ELECTRA-small model, this approach provides insights into the conditions under which positional bias emerges and the effectiveness of these perturbations as evaluation tools. While coreference replacement and adversarial sentence shuffling increase the difficulty of evaluation tasks and reveal positional biases, they do not offer substantial fine-tuning advantages over the original adversarial training data. In other words, these dataset modifications function better as diagnostic tools than as methods for sustained performance improvements. These results suggest that efforts to reduce reliance on positional cues must be accompanied by additional strategies to foster genuine reasoning capabilities. The limited transfer of these benefits to fine-tuned models highlights the complexity of improving model robustness and the need for more diverse and conceptually challenging adversarial scenarios in QA research.

References

Liu, Nelson F., Roy Schwartz, and Noah A. Smith. 2019. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proc. of NAACL-HLT*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proc. of EMNLP*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP*.