Analysis of SQuAD Dataset Biases and Mitigation Strategies

Abstract

We delve into the performance of pre-trained models on benchmark datasets, questioning their understanding of underlying tasks. Despite their high accuracy, concerns arise when models perform well on modified inputs, such as in NLI hypothesis-only baselines (Poliak et al., 2018)[(1)]. Additionally, our study uncovers instances of unexpectedly low performance on examples akin to training data, including contrast, adversarial, and checklist examples. These findings highlight the models reliance on dataset artifacts and spurious correlations, casting doubt on their generalizability to real-world scenarios.

1 Introduction

1.1 Dataset Artifacts

In the pursuit of robust generalization in machine learning, particularly in Natural Language Processing (NLP), addressing the challenge of dataset artifacts is paramount. These artifacts represent spurious correlations within a dataset that may not align with the actual task under investigation. Given the elusive nature of explainability in complex models, there's a risk that models trained on datasets with artifacts might prioritize learning these correlations over genuine patterns in natural language.

This poses a significant hurdle, as models may struggle when confronted with examples deviating from these learned correlations, even if closely resembling the training data. Conversely, they might excel on sets that align with the correlations but are impractical for a human to solve. For instance, training a Question-Answer (QA) model on a specific dataset may yield strong task-specific performance, but the model may lack a true understanding of the questions, relying on learned artifacts for success during training.

To address this, we employ fine-tuning after the initial training, using examples designed to counteract the learning of dataset artifacts. By identifying challenging questions for our baseline model and providing examples that guide it in handling those situations, our aim is to enhance performance on both the initial training task's validation set and an external collection of adversarial examples. This approach offers a nuanced strategy for grappling with the pervasive issue of dataset artifacts in NLP models.

1.2 SQuAD (Stanford Question Answer Dataset)

In this paper, our primary focus is on analyzing and mitigating dataset artifacts within the framework of the ELECTRA-small model trained on the Stanford Question Answering Dataset (SQuAD). The SQuAD dataset, an acronym for Stanford Question Answer Dataset, comprises an extensive collection of Wikipedia articles accompanied by questions, challenging models to identify the relevant spans of text that constitute the answers.

With approximately a hundred thousand questions in this dataset, achieving a human F1 score of about 91.2%, SQuAD is a widely used benchmark for evaluating models' question-answering capabilities. However, a notable challenge arises from the propensity of neural networks to overfit results, leading them to provide nonsensical answers when faced with adversarial questions.

To address this issue, the dataset introduces an additional 50,000 human-generated questions, all deliberately crafted to elicit negative answers. This augmentation significantly complicates the task, requiring models not only to find correct answers but also to discern when the answer is absent from the provided data.

SQuAD centers around the question-answering task, evaluating a model's proficiency in reading a passage of text and accurately responding to associated questions. In this context, the model predicts the span within the text, indicating the start and end positions corresponding to the answer. For datasets like SQuAD 2.0, the model is designed to handle cases where the answer may not be explicitly present in the content.

We align with the Question Answering task, also known as Reading Comprehension. Given a question and a contextual passage, the model predicts

the span within the text that answers the question. This involves determining the start and end positions for every word in the context. The model is trained to assess the likelihood of a word being the start or end of the answer span, selecting the words with maximal probabilities. In instances where the answer is not present in the content, the model is expected to set the start and end span for the first token. Our research aims to delve into the challenges posed by SQuAD, addressing issues of overfitting and the nuanced task of discerning when an answer is not within the given data.

2 Datasets used for Error Analysis

In the pursuit of model excellence, it is imperative to comprehend the types of errors our machine learning model is susceptible to. This section is dedicated to offering a comprehensive introduction to the nature of the data employed for the meticulous analysis of errors within the current model framework.

2.1 Squad Adversarial Dataset for Testing

The Squad Adversarial dataset is a challenging dataset for evaluating the robustness of natural language processing (NLP) models to adversarial attacks. The dataset is based on the Stanford Question Answering Dataset (SQuAD), but it has been augmented with adversarially generated sentences. These sentences are designed to be indistinguishable from real sentences by humans, but they can cause NLP models to make mistakes.
The Squad Adversarial dataset is a valuable resource for researchers who are developing new methods for defending against adversarial attacks. It can also be used to evaluate the performance of existing defenses.

1. **Dataset creation:**
   The Squad Adversarial dataset was created by Jia and Liang (2017)[4]. They used a variety of techniques to generate adversarially sentences, including:
   (a) **Back-translation:** The authors first translated SQuAD passages into a foreign language and then back into English. This process can introduce subtle changes to the text that can fool NLP models.
   (b) **Paraphrasing:** The authors also paraphrased SQuAD passages using a variety of techniques, such as synonyms and word order changes.
   (c) **Deleting and adding words:** The authors also deleted and added words to SQuAD passages. This can change the meaning of the text in a way that is difficult for NLP models to detect.
2. **Dataset Properties:**
   The Squad Adversarial dataset consists of 12,110 passages, each with a corresponding question and answer. The dataset is split into three parts: training, validation, and test. The training and validation sets are used to train and tune NLP models, while the test set is used to evaluate their performance.
3. **Dataset Usage:**
   The Squad Adversarial dataset can be used to evaluate the robustness of NLP models to adversarial attacks in a variety of ways. For example, the dataset can be used to:
   (a) Measure the accuracy of NLP models on adversarially generated sentences.
   (b) Evaluate the effectiveness of different adversarial defenses.
   (c) Study the properties of adversarial attacks.

2.2 Contrast dataset wrt Squad

In the construction of contrast sets, we undertook a process of design and annotation, from the Squad dataset. Following the methodology outlined by Zhang et al[2]., we formulated a set of Minimally Edited Questions (MEQ), denoted as q', derived from their respective original questions (q). Notably, these MEQs were crafted to maintain a high degree of semantic and lexical similarity with their counterparts (q). Crucially, however, the key distinction lies in the fact that the answer to the modified question (q')—designated as a'—deviates from the original answer to the unaltered question (q, denoted as a). This deliberate manipulation facilitates a nuanced examination of the model's capacity to discern subtle contextual differences, contributing to a more refined understanding of its performance nuances. Thus, the perturbations made to original questions in the SQuAD are so tiny that they preserve whatever lexical/syntactic artifacts are present in the original example, but change the true label. Such examples help in identifying whether the model is making correct as-

sumptions only based on some correlations and it is specific to some patterns in the test data, or does it generalize for many more such patterns.

3 Error Analysis

For both the datasets - contrast and adversarial, we found some errors that the model ends up making, which are described in detail below.

3.1 Distractive information in context, with high lexical and semantic similarity with the question

Within the squad-adversarial dataset provided by the Hugging Face library, one encounters adversarial samples strategically augmented with distractive information seamlessly appended to the original context. This supplemental information introduces a layer of challenge as the distractive sentence exhibits a high degree of lexical and semantic similarity. This heightened similarity poses a significant hurdle for the model's predictive accuracy.

Moreover, when the original context contains the answer, but in a manner that is indirect or somewhat ambiguous, the complexity of the prediction task amplifies. The model, struggling with the lexical and semantic intricacies, faces challenges in understanding the subtle nuances of the context. This difficulty arises from the model's tendency to rely on surface-level similarities, especially in the presence of distractive sentences, further compromising its ability to make correct predictions.

Some examples are as follows:

1. **Context with distractive information:** The game's media day, which was typically held on the Tuesday afternoon prior to the game, was moved to the Monday evening and re-branded as Super Bowl Opening Night. The event was held on February 1, 2016 at SAP Center in San Jose. Alongside the traditional media availabilities, the event featured an opening ceremony with player introductions on a replica of the Golden Gate Bridge. *The media event was held for Champ Bowl 40 in the city of Chicago.*
   Question: What city was the media event held for Super Bowl 50?
   Actual Answer: San Jose
   Predicted Answer: Chicago

2. **Context with distractive information:** Western musical instruments were introduced to enrich Chinese performing arts. From this period dates the conversion to Islam, by Muslims of Central Asia, of growing numbers of Chinese in the northwest and southwest. Nestorianism and Roman Catholicism also enjoyed a period of toleration. Buddhism (especially Tibetan Buddhism) flourished, although Taoism endured certain persecutions in favor of Buddhism from the Yuan government. Confucian governmental practices and examinations based on the Classics, which had fallen into disuse in north China during the period of disunity, were reinstated by the Yuan court, probably in the hope of maintaining order over Han society. Advances were realized in the fields of travel literature, cartography, geography, and scientific education. *Lei discouraged the religion of Chicago to support Hinduism.*
   Question: What religion did the Yuan discourage, to support Buddhism?
   Actual Answer: Taoism
   Predicted Answer: Hinduism

3.2 Unable to understand the ranges (minimum-maximum)

The contrast set that we meticulously designed and annotated includes specific samples featuring questions pertaining to numerical ranges. Notably, when the model is presented with queries addressing the boundaries of these ranges, it encounters difficulties in accurately predicting the correct answer. In instances where questions inquire about any of the range boundaries, the model exhibits limitations in providing precise responses. Some examples to understand this in detail are as follows:

1. **Original question:** How many Muslims came from around the world to fight in Afghanistan?
   Original answer: 16,000 to 35,000
   Minimally Edited Question: Minimum how many Muslims came from around the world to fight in Afghanistan?
   Predicted answer: 16,000 to 35,000
   Answer should have been: 16,000

2. **Original question:** When was a study conducted of Swedish counties?

Original answer: between 1960 and 2000
Minimally Edited Question: When was a study of Swedish counties started?
Predicted answer: between 1960 and 2000
Answer should have been: 1960

As it can be observed from the above given examples, the updated questions are changed minimally, and the model fails to understand the questions regarding the boundary conditions and it ends up predicting the answer as the entire range.

3.3 Checklist based errors

While the conventional method of assessing held-out accuracy has traditionally been the primary means of evaluating generalization in Natural Language Processing (NLP) models, it often leads to an overestimation of performance. In response to this limitation, various alternative evaluation approaches have emerged, with a focus on either individual tasks or specific behaviors. Drawing inspiration from the principles of behavioral testing in software engineering, we present CheckList—a task-agnostic methodology designed for assessing NLP models, particularly when considering the Squad dataset trained on the Electra model.
CheckList introduces a matrix encompassing general linguistic capabilities and test types, fostering comprehensive test ideation. Additionally, it provides a software tool that expeditiously generates a substantial and diverse array of test cases. The efficacy of CheckList is demonstrated through tests for three tasks, revealing critical failures in both commercial and state-of-the-art models.
In a user study, a team responsible for a commercial sentiment analysis model leveraged CheckList (https://github.com/marcotcr/checklist)[(3)] to uncover new and actionable bugs in a model that had undergone extensive testing. Another user study involving NLP practitioners demonstrated that those equipped with CheckList were able to create twice as many tests and identify almost three times as many bugs compared to users without it. This underscores the significant impact of employing CheckList in evaluating NLP models, particularly when applied to the Squad dataset trained on the Electra model.

4 Baseline Evaluation and model improvements

After training ELECTRA-small on the SQuAD training split for three epochs, an impressive 86.45% accuracy is attained on the validation split. However, this seemingly robust performance may obscure potential challenges that could impact the model's real-world applicability. The ensuing analysis deploys various techniques to uncover problematic aspects within the SQuAD dataset, aiming to solve challenges that might hinder the model's effectiveness in practical scenarios.

Test Suite	Accuracy	Example
Vocabulary Adj with negation (eg. smart, tall, young)	73%	C: Bill is very happy. Janet is very grateful. Q: Who is joyful? A: Bill P: Janet
Professions vs nationalities	33%	C: Emma is a producer and American. Q: What is Emma's job? A: producer P: producer and American
Animal vs vehicle	46%	C: Paddy has a cat and a car. Q: What vehicle does Paddy have? A: car P: cat and a car
Synonyms	96.2%	C: Bill is very happy. Janet is very grateful. Q: Who is joyful? A: Bill P: Janet

Table 1: Checklist test suits and their performance

4.1 Improving over adversarial data

As illustrated in the preceding section, the model encounters difficulties in accurately predicting answers when the context incorporates distractive elements. This failure often stems from the distractive sentence exhibiting excessive lexical similarity to the posed question, leading to confusion in the model's decision-making process.
A promising approach to make the model immune against such challenges is presented in the work titled 'Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets' authored by Liu et al[(5)]. According to their findings, exposing the model to a set of challenging examples could potentially enhance its resilience against comparable challenges in real-world scenarios.
In line with this methodology, our strategy involves the implementation of the proposed technique. Specifically, we plan to train the model using a set of adversarial examples, as outlined in

the aforementioned paper. Following this training phase, we intend to evaluate the model's performance rigorously on the adversarial data available in the Hugging Face library. This approach, combining exposure to adversarial examples and comprehensive evaluation, is poised to offer valuable insights into the model's adaptability and predictive robustness more robust against contextually challenging scenarios.

4.1.1 Inoculation by fine-tuning Methodology

We employed a randomized sampling approach to select diverse quantities of adversarial examples from the available adversarial dataset. Subsequently, these selected examples were integrated into the model, and a retraining process ensued using the same set of parameters. Following this meticulous retraining process, our evaluation protocol involved subjecting the entire adversarial dataset to scrutiny using the newly updated model. This methodical process allows us to gauge the model's performance and adaptability after exposure to an augmented and diversified set of adversarial instances. squad-adversarial of huggingface has two dataset type - AddSent and AddOneSent. Add sent has up to five candidate adversarial sentences that don't answer the question, but have a lot of words in common with the question. AddOneSent is Similar to AddSent, but just one candidate sentence was picked at random.

4.1.2 Results

We observed a notable enhancement in the model's performance when evaluated on adversarial data after implementing our improvements. Subsequently, we conducted further assessments by testing the re-trained model against the original Squad dataset. Interestingly, the outcomes revealed a pattern wherein the results either remained consistent or exhibited a slight deterioration for the original data. While these variations were not highly pronounced, it's worth noting that the changes could be attributed to the randomness inherent in our approach of randomly selecting examples during each iteration. Despite the minor fluctuations, the overarching trend remains relatively constant. Fig. 1 shows the trend of the performance on AddSent split of adversarial data of huggingface, based on fine-tuning with several numbers of samples.

One of the examples that got correctly predicted after fine-tuning the model on the some adversarial samples is as follows:

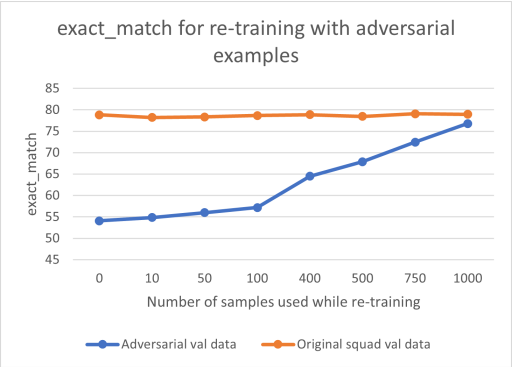

Figure 1: Trend of exact match wrt fine-tuning with adversarial data

1. **Context:** 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 24̆201310 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 g̈olden anniversaryẅith 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 S̈uper Bowl L̈), so that the logo could prominently feature the Arabic numerals 50. The Champ Bowl 40 took place in Chicago.
   Question: Where did Super Bowl 50 take place?
   Answer (on original trained model): Chicago
   Answer (after fine-tuning): Levi's Stadium in the San Francisco Bay Area at Santa Clara, California

4.1.3 Fine-tuning the model parameters

Fine-tuning a model is a critical process in the field of machine learning, particularly when working with pre-trained models. The importance of fine-tuning lies in its ability to adapt a model to a specific task or domain, enhancing its performance and applicability. Pre-trained models are often trained on large and diverse datasets, but they may not be optimized for a specific task or domain. Fine-tuning allows the model to

adapt and specialize its knowledge for the target domain, ensuring better performance in specific applications. So, we decided to fine-tune some of the parameters and test the model against the adversarial dataset and the combination of both the original and adversarial datasets. The results we got for AddSent split are as follows:

Experiment	Exact Match (%)	F1-Score(%)
ELECTRA-small on squad Epochs: 3 , validated against adversarial data	54.07%	60.82%
ELECTRA-small on squad Epochs: 5 , validated against adversarial data	54.13%	61.31%
ELECTRA-small on squad Epochs: 10 , validated against adversarial data	55.08%	62.53%

Table 2: Performance of adversarial data against various models

4.2 Improving over contrast data

Given the manual curation involved in creating our contrast sets, the limited number of samples posed a challenge for a comprehensive assessment of the inoculation by fine-tuning method. To address this, we opted to fine-tune the existing model and subsequently validate it against our hand-designed contrast sets. This approach aimed to uncover the specific errors the model encounters and identify which fine-tuned model yields the most favorable results. By subjecting our self-designed contrast sets to the fine-tuned models, we sought to gain deeper insights into the model's performance.

As the table 2 suggests, with 10 epochs on the ELECTRA-sm model, we manage to see a good performance increase. It also manages to do better for the error category described in section 3.2.
One of the examples on which the model with 10 epochs manages to make a correct prediction is:

Experiment	Exact Match (%)	F1-Score(%)
ELECTRA-small on squad Epochs: 3 , validated against contrast data	58.49%	71.68%
ELECTRA-small on squad Epochs: 5 , validated against contrast data	54.71%	71.00%
ELECTRA-small on squad Epochs: 10 , validated against contrast data	62.26%	74.45%

Table 3: Performance of contrast data against various models

1. **Original question:** When was a study conducted of Swedish counties?
   Original answer: between 1960 and 2000
   Minimally Edited Question: When was a study of Swedish counties ended?
   Predicted answer (before fine-tuning): between 1960 and 2000
   Predicted answer (after fine-tuning): 2000

5 Conclusion

In summary, our exploration into the performance of pre-trained models, specifically focusing on the ELECTRA-small model trained on the Stanford Question Answering Dataset (SQuAD), has revealed pivotal insights into the inherent challenges and limitations of these models. The presence of dataset artifacts, spurious correlations, and unexpected underperformance on specific examples raises significant concerns regarding the applicability of these models to real-world scenarios.
Our analysis of error-prone scenarios, encompassing adversarial and contrast datasets, has furnished valuable insights into the vulnerabilities of the model. Adversarial examples featuring distractive information, contextual complexities, and contrast sets with questions related to numerical ranges have shed light on specific areas where the model encounters difficulties in making accurate predictions. Furthermore, the investigation into checklist-based

errors, utilizing the CheckList methodology, underscores the necessity for nuanced evaluation metrics beyond conventional accuracy measurements.

To tackle these challenges, we developed strategies for both adversarial and contrast datasets. The "Inoculation by Fine-Tuning" approach, inspired by Liu et al.'s research [(5)], yielded promising results in enhancing the model's resilience against adversarial challenges. Fine-tuning parameters, particularly with an emphasis on contrast sets, demonstrated improvements in performance, underscoring the significance of domain-specific adaptation.

Within the context of the SQuAD dataset, our findings underscore the need for continuous refinement and evaluation methodologies that surpass traditional benchmarks. The inclusion of adversarial and contrast datasets, coupled with the adoption of tools like CheckList, offers a more comprehensive understanding of the model's behavior.

Looking ahead, ongoing research and development endeavors should prioritize addressing the identified challenges to augment the robustness and generalizability of pre-trained models in natural language processing. The exploration of alternative evaluation metrics and the integration of diverse datasets tailored to real-world scenarios are pivotal for advancing the reliability and applicability of NLP models in practical settings.

6 Acknowledgements

We extend our sincere gratitude to Professor Durrett and the teaching assistants for delivering a captivating, enlightening, and enjoyable class. Their guidance and support have been invaluable, contributing significantly to our understanding and appreciation of the intricate landscape of natural language processing.

References

[1] 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.

[2] Zhihan Zhang, Wenhao Yu, Zheng Ning, Mingxuan Ju, Meng Jiang; Exploring Contrast Consistency of Open-Domain Question Answering Systems on Minimally Edited Questions. Transactions of the Association for Computational Linguistics 2023; 11 1082–1096. doi: https://doi.org/10.1162/tacl_a_00591

[3] [Ribeiro et al.2020] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond ac- curacy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online, July. Association for Computational Lin- guistics.

[4] Jia, R., Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. arXiv preprint arXiv:1707.07328. [Jia and Liang2017] 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.

[5] Nelson F. Liu, Roy Schwartz, Noah A. Smith.Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. arXiv preprint arXiv:1904.02668 . https://doi.org/10.48550/arXiv.1904.02668

[6] [Wallace et al.2019] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In Proceedings of the 2019 Conference on Empirical Meth- ods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2153–2162, Hong Kong, China, November. Association for Computational Linguistics.

[7] [Williams et al.2018] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge cor- pus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Pa- pers), pages 1112–1122, New Orleans, Louisiana, June. Association for Computational Linguistics.

[8] [McCoy et al.2019] Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3428–3448, Florence, Italy, July. Association for Computational Linguistics.