# 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). 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 rk 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. Baseline Evaluation

## 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 reveal challenges that might hinder the model's effectiveness in practical scenarios.

## 2.1 Checklist

## 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)

## 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.

|  |  |  |
| --- | --- | --- |
| Test Suite | Acc | Example |
| Vocabulary Adj with negation (eg. smart, tall, young) | 73% | C: Andy is incredibly vocal about the project. John is somewhat vocal about the project  Q: Who is least vocal about the project?  A: John  P: Andy |
| 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 |

## 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.

## Dataset Creation

## The Squad Adversarial dataset was created by Jia and Liang (2017). They used a variety of techniques to generate adversarially sentences, including:

## 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.

## Paraphrasing: The authors also paraphrased SQuAD passages using a variety of techniques, such as synonyms and word order changes.

## 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.

## 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.

## 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:

## Measure the accuracy of NLP models on adversarially generated sentences.

## Evaluate the effectiveness of different adversarial defenses.

## Study the properties of adversarial attacks.

## We also investigated the impact of different model parameters on the effectiveness of adversarial training applied to the SQuAD dataset. Through a series of experiments, we evaluate model performance on the entire adversarial dataset, the high-confidence adversarial dataset, and explore the effects of retraining using a combination of SQuAD and high-confidence adversarial data. Additionally, we analyze the outcomes of experiments with varying training epochs, weight decay, and learning rates. The results provide insights into the optimal configuration for achieving improved performance in question-answering tasks

|  |  |  |
| --- | --- | --- |
| Experiment | Exact Match(%) | F1 Score(%) |
| Validation on entire Adv dataset | 63.79 | 70.77 |
| High-conf Adv dataset | 45.74 | 51.88 |
| Squad + High-conf Adv dataset | 79.14 | 86.86 |
| ELECTRA-small on squad  Epochs:5,  weight decay:0.01  lr: 3e-5 | 77.98 | 85.83 |
| ELECTRA-small on squad  Epochs:10,  weight decay:0.01  lr: 3e-5 | 75.85 | 84.88 |
| ELECTRA-small on squad  Epochs: 3 ,  weight decay:0.01  lr: 3e-5 | 79.23 | 86.63 |

**References**

[Poliak et al.2018] 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, June. Association for Computa- tional Linguistics.

[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.

[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 Lin- guistics.

[Clark et al.2020] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre- training 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, 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.

[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.

[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.

[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.