The effect of biased heuristics on BERT's performance

YangJunqing Qiao
Dept. of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
jqiao@umass.edu

Abstract—**Machine learning systems are opaque humans. We try to deobfuscate the current state of the art NLP model BERT by showing its reliance of unigram and bigram biases present in certain datasets.**

Index Terms—**adversarial, bert, bias, unigram, bigram, machine learning, understanding**

I. Introduction

Understanding how complex machine learning systems behave and make the predictions that they do is an ongoing research effort critical to the widespread adoption, acceptance, and use of these models in more generalized settings. Current state of the art models are often opaque at a low level to humans. It is difficult to tell what features the models are using to come to certain conclusions and what small changes to the input, if any, could create failures in the model's classification. In this paper we survey the works of several papers analyzing the effect of feature perturbations on the performance of BERT [1]. We then perform two new experiments combining the techniques of [2] with the task of [3] to provide further evidence that BERT is relying on spurious cues in data to make classification choices.

The paper is split into 3 major sections: background, experiments, and conclusion. In the background section we provide brief descriptions of all relevant information from the papers surveyed and outline the problem in more detail. The experiments section contains the results of running a BERT model on different types of adversarially perturbed data. Our results support the claims made in [2], [3], [6] that BERT is relying on statistical artifacts in data perhaps more strongly than it should. Finally the conclusion outlines futures directions for this work such as different biases that might exist in data and potential methods of detection.

II. Background

Bigram and unigram prediction features are basic NLP techniques used during teaching due to the ease by which humans can understand them. Machine learning models are capable of learning more complex mappings but come at the cost of increased opacity to humans. Determining whether these systems have learned meaningful information or if the learning process is relying on some bias or association in the dataset is still an active field of research.

This paper investigates the prevalance of bigram and unigram heuristics in two NLP tasks and determine that in these tasks, complicated systems like BERT rely partially on dataset biases. Specifically we provide some insight to the magnitude by which these dataset biases affect BERT's modeling capabilities.

A. BERT

BERT [1] is a transformer model which provides bidirectionally associative word embeddings and has achieved state of the art performance on a large number of benchmark NLP tasks. It is trained in an unsupervised fashion through two tasks. The first attempts to predict randomly masked words in a corpus and the second tries to predict the second sentence in a sequence. BERT's close to human level performance on complicated tasks has raised skepticism over its inner mechanisms. However, it is generally very difficult to distinguish the difference between a model which learns meaningful information about the task versus one which leverages some heuristic present in data.

III. Related Works

Recently, BERT's close to human level performance on certain 'hard' tasks has sparked concern over if it is learning heuristics on the training data causing a better than average performance simply due to biases present in the training set [3], [6]. In the argument reasoning comprehension task (ARCT), BERT was found to be heavily weighing the existence of specific words and bigrams such as 'not' and 'will not' towards its final classification determination [3].

A. Problem Formulation

Our main goal is to detect the presence and effect unigram and bigram biases have on machine learning models. To this end, we take inspiration from [3], and specifically focus on classification tasks with two distinct sets of features so that it is possible to count occurrences of unigrams and bigrams independently among these features. The first classification task is the same one analyzed in [3]. ARCT is the task of selecting the correct warrant which validates a claim given a reason.

1) ARCT: The ARCT [7] dataset consists of (reason, claim, warrant1, warrant2) tuples. Where the main goal is to choose the correct warrant that logically supports the claim given the reason. BERT's performance on this specific task drew skepticism as this task is hard for humans to perform with high accuracy.

ARCT Task:

Reason: Milk isn't a gateway drug even though most people drink it as children.
Claim: Marijuana is not a gateway drug.
Warrant 1: Milk is similar to marijuana
Warrant 2: Milk is not marijuana

Example taken from [7]. The correct warrant must be chosen which satisfies the claim given a reason.

2) NLI: The second task is a variation on the natural language inference task. In natural language inference [8], the goal is to find the correct judgement given a text and a hypothesis. For our experiments, we re-purpose this task and instead try to predict the correct hypothesis given a text and a judgement. The original data comes from the Stanford Natural Language Inference Corpus which is a 570k human written and labeled corpus consisting of text, judgement, and hypothesis. The data for our task is the NLI corpus with each member augmented with another randomly chosen hypothesis from the corpus.

Original NLI Task:

Example 1:
Text: A man inspects the uniform of a figure in some East Asian country.
Hypothesis: The man is sleeping
Judgement: Contradiction

Example 2:
Text: A soccer game with multiple males playing.
Hypothesis: Some men are playing a sport.
Judgement: Entailment

The **hypothesis** {**entails** | **contradicts** | is **neutral**} to the **text**

The modified task inputs a text, two hypothesis, and a judgement and requires selecting the correct hypothesis that matches the text and judgement.

Modified NLI Task:

Text: A man inspects the uniform of a figure in some East Asian country.
Hypothesis 1: The man is sleeping
Hypothesis 2: Some men are playing a sport. (adversarial)
Judgement: Contradiction

The {**Hypothesis 1** | **Hypothesis 2**} **contradicts** the **text**

This is another form of sentence selection problem very similar in structure to the ARCT task and was designed to provide a larger dataset on which we could test and verify the results from [3].

3) TextFooler: [2] Introduces a way to adversarially perturb inputs, replacing selected tokenized words with their approximate closest synonym using cosine similarity. This perturbation caused significant performance decreases in textual entailment from SOTA to worse than random chance. This method is similar in idea to [3], [6] in that it reduces the bias that might be inherent in a selected corpus of text. We provide results from applying this method to the ARCT and modified NLI task in the experiments section.

B. Prior results for tasks

BERT and other state of the art models on the ARCT task were shown to be following heuristics of the data set; specifically [3] found that in both the training and test dataset, the unigrams "is", "do", "are", and "not" had what they coined the highest *productivity*, *applicability*, and *coverage* over the data. This essentially means that these unigrams appear often in either the correct or alternative warrant at a relatively high frequency and do not appear often in the opposite warrant type. This makes it possible to use these distinguishing features between the correct and alternative warrant. These features are so powerful that BERT was able to achieve 71% accuracy just conditioning on the warrants, which is just 6% short of the SOTA having no information about the reason or claim at all due to the statistical artifacts present in the dataset.

IV. Experiments

We used the original codebases provided in [2], [3] to perform our experiments but were unable to produce exactly the same results; specifically, we were unable to achieve as high accuracies with the hugging face baseline BERT model. However, we did see similar trends as outlined in both papers.

A. ARCT Ablation

Since [3] provided results for the performance of a BERT model trained on unperturbed data on adversarial data but not the performance of a BERT model trained on perturbed data, we provide some results in this direction.

It is apparent that the model trained on unaugmented data performs better than all other models across all datasets until the proportion of adversarial data in the test set becomes the majority. This phenomenon supports the hypothesis that BERT can only learn spurious statistical biases in the dataset as a model trained on the original unaltered dataset can make better judgements on the remaining unaltered portion of the test dataset by relying on spurious cues giving it a higher accuracy for those features.

Additionally, if it is true that BERT relies on heuristics found in data, perturbing the data should remove these cues and so a model trained on adversarial data would not be able to learn these unigram and bigram heuristics for unperturbed data harming its performance for these datum.

Additionally, all of the models trained perform poorly on the test datasets with a high number of adversarially perturbed

% of Adversaral Data in Test Set	0%	25%	50%	75%	100%
0	61.7	60.1	60.5	61.3	59.9
25	60.5	53.2	56.8	56.3	56.7
50	61.5	52.2	58.7	55.5	53.5
75	49.2	53.2	56.2	53.3	58.9
100	51.6	55.9	52.7	57	52

Fig. 1. Accuracies of 5 different BERT models trained on test datasets with varying ratios of adversarial data.

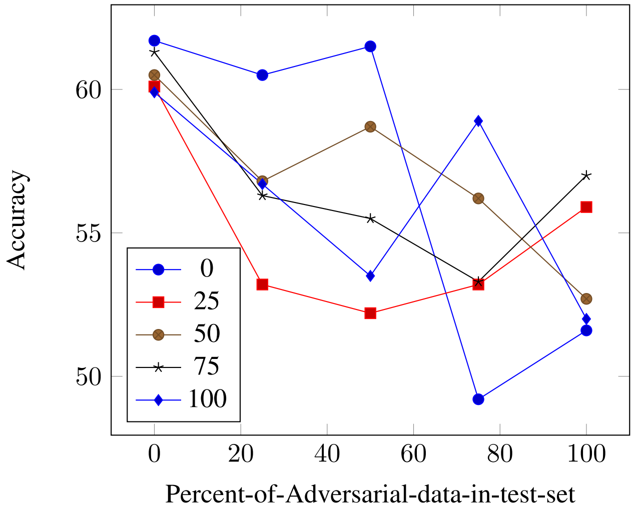

Fig. 2. Accuracies for 5 networks trained on test sets containing differing ratios of negated warrants. The original BERT model trained on no negated warrants outperforms other models up until a majority of the test set becomes adversarial.

datum. None doing significantly better than random chance. This is also consistent with the hypothesis that BERT might be overfitting on these heuristics as when these are removed BERT fails to perform significantly better than a baseline random chance.

[htbp]

B. Textfooler ARCT

TextFooler [2] is a system designed to replace tokens in an input with their closest cosine similarity synonyms. It was originally tested on IMDB, Yelp, and SNLI datasets for classification. It was able to reverse the classifications a >90% accuracy BERT model making it achieve <10% accuracy. We re-purposed the system to replace unigrams and bigrams identified in [3] as having high applicability, productivity, and coverage in the ARCT task with their closest cosine similarity match. The results are provided below:

Replacing "not" in inputs resulted in the sharpest decline in accuracy making BERT perform worse than random chance. This is not surprising as "not" was reported to be the strongest unigram bias for BERT. However we do not see the extreme poor performance outlined in [2]. This may be due to the fact that we only perturbed one word in each test point as opposed to a larger fraction of the input. Additionally, some of the datasets these words produced while understandable by a human reader were not completely grammatically correct.

Example of replaced unigram in ARCT data

Original Warrant: capitalism **is** widespread across the world

Adversarial Warrant: capitalism **be** widespread across the world

Replacing **is** results in a human understandable but not entirely grammatically correct sentence.

C. Modified NLI Ablation

The modified NLI problem we developed has no existing benchmarks by which we can compare against and we did not have time to do a baseline performance analysis using other models. Instead, we present just the results of BERT trained on this dataset and it's adversarial variants. For this problem we perform the same ablation test as we did for the ARCT dataset.

Here we define the unperturbed data to be the vanilla augmented dataset consisting of text, judgement, hypothesis 1, and hypothesis 2. The adversarial dataset is produced by passing both hypothesis through TextFooler.

The results are not too interesting as our data was generated with randomized hypothesis which likely have much different vector representations than the original text while the correct

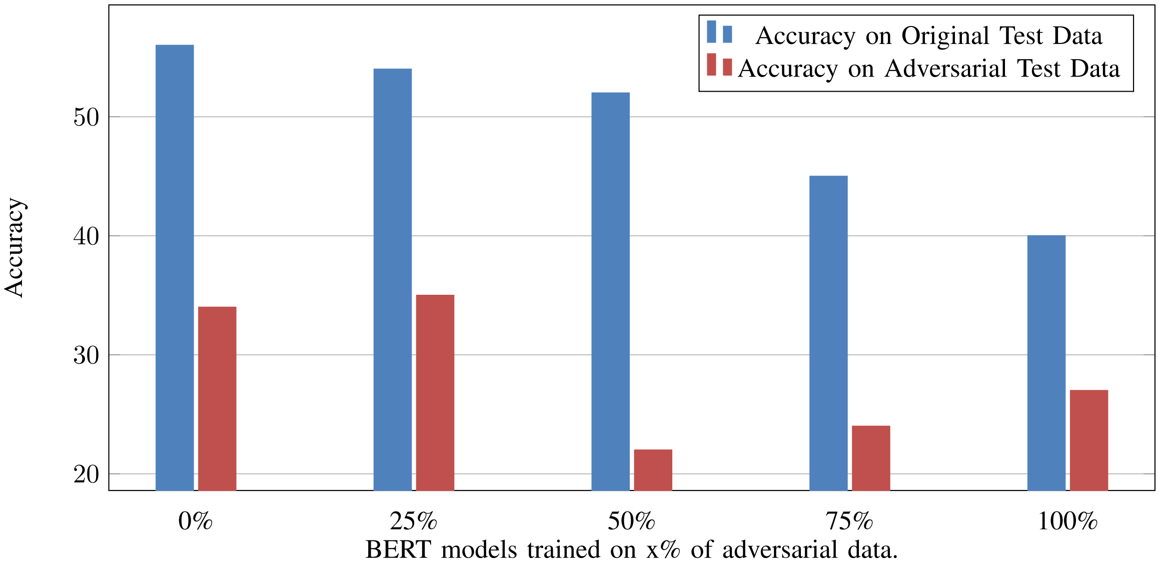

Fig. 3. Accuracies of the 5 models on the original and adversarial test data. It does not seem like training on adversarial test data improves network performance in classifying new adversarial test data. But this does show that the perturbations succeeded in confounding BERT.

Feature Replaced	Accuracy
is	46.5
do	43.7
are	44.2
not	40.8

Fig. 4. Accuracy of the baseline BERT model trained on no adversarial data on test sets with selected unigrams replaced

Adversarial-Training-Ratio	0	25	50	75	100
0	99.5	94.2	95.6	94.5	93.2
25	96.2	95.6	94.5	92.3	96.7
50	94.8	92.3	96.3	92.3	98.5
75	97.6	93.4	96.7	96.5	93.4
100	96.3	98.6	98.3	92.1	93.1

Fig. 5. Accuracies for 5 BERT models trained on test sets containing differing ratios of random adversarial hypotheses.

hypothesis likely shares similar words with the text allowing BERT to condition on similarity between words to make it's prediction.

V. Conclusion

We have validated the results seen in [2], [3] in this paper and shown that BERT is relying heavily on unigram and bigram cues in training data to make its predictions. Additionally we provided an additional task on which these biases could be shown in the form of the modified NLI task.

Future directions of work in this area involves creating better detection of more sophisticated biases in datasets and generating more reasonable adversarial hypothesis for the modified NLI task. We originally planned to have the adversarial hypothesis for modified NLI be a negated version of the original hypothesis but were not able to produce satisfiably coherent negations. This could be an avenue of investigation in the future as well.

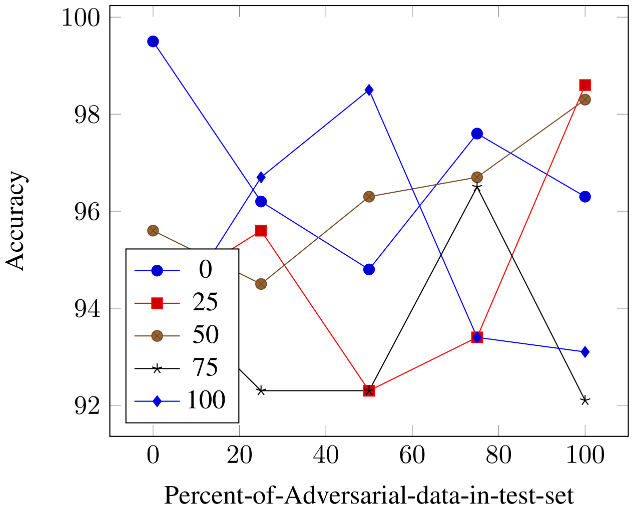

Fig. 6. Accuracies for 5 BERT models trained on test sets containing differing ratios of random adversarial hypotheses. All of the models achieve very high accuracy likely due to our problem formulation.

References

[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[2] Jin, Di, et al. "Is BERT Really Robust? Natural Language Attack on Text Classification and Entailment." arXiv preprint arXiv:1907.11932 (2019).

[3] Niven, Timothy, and Hung-Yu Kao. "Probing neural network comprehension of natural language arguments." arXiv preprint arXiv:1907.07355 (2019).

[4] Hsieh, Yu-Lun, et al. "On the robustness of self-attentive models." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[5] Goryachev, Sergey, et al. "Implementation and evaluation of four different methods of negation detection." Boston, MA: DSG (2006).

[6] McCoy, R. Thomas, Ellie Pavlick, and Tal Linzen. "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference." arXiv preprint arXiv:1902.01007 (2019).

[7] Habernal, Ivan, et al. "The argument reasoning comprehension task: Identification and reconstruction of implicit warrants." arXiv preprint arXiv:1708.01425 (2017).

[8] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).