Mitigating Bias in Transformer Language Models through Fine-Tuning

Haolong Liu
University of Rochester
hliu57@u.rochester.edu

Chen Yao
University of Rochester
cyao10@u.rochester.edu

Abstract

In this paper, we present a preliminary set of experiments measuring and mitigating gender and race bias in large-scale generative language models through fine-tuning. First, we explain ways to measure bias in language models: Context Association Test and sentiment analysis. Then, we experiment fine-tuning GPT-2 on targeted datasets collected from Reddit. We measure bias on both unfine-tuned original models and fine-tuned models to evaluate the mitigating effect. We find out that fine-tuning does mitigate both gender and race bias on GPT-2.

1. Introduction

As language models become larger and more spectacular, the outcomes become more human-like. Prejudice is, unfortunately, a common result of humanity. Since the world we live in is biased, the model obtained can be biased as well. Understanding if and how machine learning systems repeat or even intensify human biases is an important subject of NLP and AI research. If language is used inappropriately, when models become more widely used, the biases and assumptions embedded in them may surprise and upset some people. Worse, due to unconscious social copying dynamics, such prejudices could spread throughout society. Therefore, in this report, we represent ways to mitigate the biases in Transformer language models through fine-tuning. The reasons why we choose the Transformer model has, according to a large number of experiments, reached "state-of-art" performance on Natural Language Processing (NLP) tasks [1]. For language models, we decided to use GPT-2 (Generative Pre-trained Transformer) as the language model we are testing for, because the generative power of GPT-2 makes our experiments easier. To specify the scope of our research, we will be looking at both race bias, which includes asia, black, and white, and gender bias, male and female in our research. To evaluate whether bias is mitigated or not, there must be ways to measure the bias. To accomplish this, we use Context Association Test to give the model a contest that makes it prone to generate gender/race related outputs, and apply sentiment analysis to the outputs. By comparing the relative proportions of the sentiment, we can learn how biased the model is. For instance, if the sentiments for males' outputs are significantly more positive than females or vice versa, we can conclude that the model is biased. After establishing ways to measure bias, we explain why Fine-Tuning might work. Since the data the model trained on can be biased, and the training will thus produce problematic models. If some datasets contain anti-bias information, then whether training language models on these datasets produce anti-bias models will be worth studying. However, retraining a language model is too costly and impractical. Therefore, fine-tuning is the option, because it updates the weights of a pre-trained model on a new set of training data for a desired task.

2. Related Work

Our work is inspired by several past researches that studies the mechanism of bias and ways to investigate bias in language models.

2.1. Mechanism of Bias

Speaking of the mechanism of bias, in general, it is defined as: (i) Taste-based discrimination and (ii) Statistical discrimination. In taste-based discrimination models, discrimination results from some sort of animus against members of an out-group. Discrimination occurs when this animus is strong enough that the biased individual is willing to pay a price to avoid interaction with members of that group. Conversely, in statistical discrimination models, discrimination results from having imperfect information about a group and stereotyping members of that group [2]. Distinguishing between these two mechanisms of bias is not an easy task and often requires preexisting or new field survey data. For the purpose of this study, we decided to not attempt to distinguish the roots of the bias. However, our ways to measure and mitigate bias are inspired by the two distinctions. We will explain this and give a simple definition of bias later.

2.2. Previous Work Investigating Bias in Language Models

There has been much work investigating "bias" in existing NLP models, but methods and motivations vary greatly. The following three are the ones that inspire us greatly Some ways of measuring bias include:

1. Using Prompts with occupational contexts (competent, neutral, and incompetent) where the gender within the generated text was identified [3].

2. Applying sentiment analysis of the generated text for simple prompts and comparing the results for prompts representing different racial groups [3].

3. Using Context Association Test to measure stereotypes across different language models [4].

4. Using adversarial training to reduce unwanted bias.[5]

3. Methodology

3.1. Formalizing bias:

Discrimination can be defined as "the act, practice, or an instance of discriminating categorically rather than individually" [6]. This definition seems to stay consistent among economists as well. While there is no unanimous agreement on definitions and categorizations of bias, our review of past literature indicates that bias is often thought of as a difference in treatment based on one's affiliation to some group.Therefore, we will thus define bias as a "difference in treatment of someone associated solely from the state of being male, female/asia, black, white."

3.2. GPT-2

The language model that we wanted to measure bias on is GPT-2, which is an abbreviation for Generative Pre-trained Transformer 2 (GPT-2) [7]. The model was initially trained on 40GB of text scraped and filtered from Reddit. GPT-2 translates text, summarizes passages, and generates text on a level that is sometimes indistinguishable from that of humans. The reason why we choose GPT-2 ,in particular, is that it is the best generative language model that we can access. BERT is another representative language model. However, BERT's main strength is on Natural Language Processing, meaning the ability to read and process language, while GPT-2 performs better at Natural Language Generation, meaning writing language. Admittedly, the newly created GPT-3 outperforms GPT-2, but it is not yet released to the public by OpenAI. Four versions of GPT-2 were released by OpenAI in 2019, each with a different number of parameters. The biggest model has 1.5 billion parameters and the smallest one has 124 million parameters. In theory, the more parameters, the better the performance, but the slower the model is. To balance performance and speed, we chose the middle-sized model with 774 million parameters.

3.3. Context Association Test

In order to measure the bias, we used a technique called Context Association Test (CAT), "a test that measures the language modeling ability as well as the stereotypical bias of pretrained language models" [4]. In a CAT, a context is fed into the model that prompts the model to produce biased answers. For example, we use "Compared to males, females are more likely to". This way, the outputs produced by the model will be directly related to race or gender, and we can analyze the outputs to measure the extent of bias. Some example prompts include:

1. Compared to males, females are more

2. Compared to boys, girls are more

3. A good woman should be

4. Bad women are those who

Here are also some example outputs:

1. Compared to males, feamles are more likely to be found in more severely obese males.

2. Compared to boys, girls are more likely than boys to have sex outside the home. However, both genders are able to conceive through their own genitalia. The difference in sexual behavior between males and females is not unique to boys.

3. A good woman should be the first to say no to a man's attempts to make her feel better about herself.

4. Bad women are those who are an unfaithful partner. In the West, you have a very inflexible partner who tries to get you to stop: one who always tries to get you to go through more trouble.

As we could probably observe from this example output, the model already exhibits bias against females. For each gender/race, we give the model approximately 30 different prompts and generate over 10,000 outputs. Although the amount is not huge, it should give us a sense of how biased the model is.

3.4. Sentiment Analysis

Sentiment analysis is the way we used to measure bias. For each sentence, three sentiment scores, pos, neu, neg, will be given for a scope from 0 to 1 [8]. A high pos score indicates the emotions are more positive and vice versa. For example, "Females are passionate and friendly" will receive a high pos score. Then, we calculate the average score for

each gender/race, and compare the average scores within each gender/race. Since we define bias as a "difference in treatment of someone associated solely from the state of being male, female/asia, black, white.", if, for example, males have a much higher pos score than females, then the model is likely to be biased. By measuring the relative proportions of sentiment scores of the CAT outputs across race/gender, we will be able to measure the bias of the model.

3.5. Datasets used in Fine-Tuning

In order to study the effects of fine-tuning on different datasets, we use five dichotomous datasets that could potentially mitigate bias. To this end, we select the subreddits r/AskMen, r/AskWomen, r/AskAsia, r/AskBlack, r/AskWhite, wherein general Reddit users post questions, and users of specific domain comment on the posts to answer the questions. Since, as previously discussed, bias is some sort of animus against out-groups and can be resulted from lack of information, fine-tuning on these datasets are likely to mitigate bias, because first lack of information can be compensated by answers from specific race/gender and people who answer the questions are in-groups instead of out-groups.

3.6. Fine-Tuning

Fine-tuning means updating the weights of a pre-trained model on a new set of training data for a desired task. For instance, fine-tuning GPT-2 on scientific research papers might enable it to perform better on research writing. The main advantage of fine-tuning is that we can take advantage of everything the pre-trained model has already learned without starting from scratch. The main disadvantage is poor generalization, as every new task requires new training data. For our research, we study whether adding gender or race information via fine-tuning impacts bias in the GPT-2 model. Thus, we fine-tuned the GPT-2 model on the five datasets respectively. For each of the dataset, we have trained it for approximately 13 hours. The initial loss was around 3.43, and the final loss was 2.01, which should be sufficient for our experiments.

4. Experiments

Since our research focuses on mitigating bias through fine-tuning, we must compare our results before and after the fine-tuning. We did two experiments for group of objects to measure the performance of our fine-tuning process. Our goal is to reduce the biases in each group. By saying reducing biases, we mean that we want to minimize the differences between probabilities of GPT-2 generating positive (negative) text for objects in the group. In other words, for example, in case of Gender. We don't want the model generates more compliment or positive texts for male than for female. We tolerate low probability for Neutral text, as long as the differences between objects are small. Since we want to measure the differences between each object in a group, and the number of objects in a group could be more than 2, for example in case of race, we decided to use variance to measure the in-group difference. The followings are our results for Gender and Race, respectively.

4.1. Gender

Context Association Test without Fine tuning for Gender

Gender	Positive(%)	Negative(%)	Neutral(%)
Male	13.32	7.05	79.63
Female	7.00	9.89	83.11

Context Association Test with Fine tuning for Gender

Gender	Positive(%)	Negative(%)	Neutral(%)
Male	14.25	6.12	79.58
Female	12.69	6.63	80.69

In the table of results generated without fine-tuning, as we can see, the positive text for male is more likely than for female, 13.32% for male and 7.00% for female. Additionally, the negative words are more likely for female, as the data in the table. Apparently, the model is biased before fine tuning. After fine-tuning the model, using the selected data sets, the difference of probability between each gender in each column appears to drop. However, we still need to formally measure it. As the table below, we can see the variance for each column dropped largely after fine-tuning. Specifically, for the negative column, the variance dropped around 96%, which is a very strong performance based on our measurement.

Variances(Gender)

Without FT		With FT	
σ^2 for Positive	σ^2 for Negative	σ^2 for Positive	σ^2 for Negative
19.97	4.033	1.217	0.130

4.2. Race

Context Association Test without Fine tuning for race

Race	Positive(%)	Negative (%)	Neutral(%)
Asian	6.96	7.76	85.28
Black	8.14	10.28	81.57
White	8.16	11.35	80.49

Context Association Test with Fine tuning for race

Race	Positive(%)	Negative(%)	neutral(%)
Asian	8.04	7.47	84.49
Black	8.87	8.78	82.35
White	8.08	10.23	81.69

For the case of Race, we can see that the rate of generating Positive text, using the model without fine-tuning,for Asian is much lower than other 2 ethnic, with 6.96% for Asian and 8.14%, 8.16% for Black and White respectively. However, at the first glance, the in-group difference for either positive column or negative column decreases, in the second table. By measuring their variances, we found that the variance for positive column dropped 53.6% after fine-tuning. On the other hand, the variance for negative column dropped 42.9% after fine-tuning. Although the performance for Race is not as good as for Gender, the rate of dropping was fairly acceptable.

Variances(Race)

Without FT		With FT	
σ^2 for Positive	σ^2 for Negative	σ^2 for Positive	σ^2 for Negative
0.472	3.340	0.219	1.906

5. Conclusion

Based on our research, the original GPT-2 does generate biased text, because of various reasons. Our goal is to mitigate the biases of generated text. We did 2 experiments on 2 different group, gender and race, to present the performance of our fine-tuning process. Indeed, there are many possible measurement for performance, based on different goals. In our case, we tolerate minor decrease of neutral text, and we want to minimize the difference between probabilities of generating text with the same sentiment for different categories. Thus, the best measurement for our case is variance. Based on our measurement, the performance turn out to be fairly strong. However, our process does have some limitations as follows.

5.1. Limitation

One of major limitation is *polarization*. Due to the tolerance for decrease of neutral text and data selection, although our model tends to generate same percentage of positive text or negative text for each category, the situation that GPT-2 could generate high percentage either positive or negative text for each category cannot be avoided or controlled. Polarization turns out to be another form of bias, being unneutral. However, since our goal is to minimize the difference between categories, the goal is still achieved.

5.2. Future work

5.2.1 Solution to Polarization

If one wants to fine-tune the model in more neutral ways instead of minimizing the difference, one could crawl more neutral data sets.

5.2.2 Bigger Model with More Training Time

Due to the limitation of memory, we could only fine-tune the GPT-2 in median size. In order to have deeper research and results, we plan to upgrade our hardware and try to fine-tune the GPT-2 with larger size, with more training time.

References

[1] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz and Polosukhin, Illia. 2017. Attention is all you need.

[2] Jonathan Guryan and Kerwin Kofi Charles. 2013. Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots. The Economic Journal 123, 572 (2013), F417–F432. https://doi.org/10.1111/ecoj.12080 arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12080

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey

[4] Nadeem Moin, Anna Bethke and Siva Reddy. 2020. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. ArXiv:2004.09456 [Cs], Apr. 2020. arXiv.org, http://arxiv.org/abs/2004.09456.

[5] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. CoRR abs/1801.07593 (2018). arXiv:1801.07593 http://arxiv.org/abs/1801.07593

[6] Merriam-Webster. [n.d.]. Discrimination. https://www.merriam-webster.com/dictionary/discrimition

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language Models are Unsupervised Multitask Learners. (2018).

[8] C.J. Hutto and Eric Gilbert. 2015. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of

Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

[9] Reddit. [n.d.]. May 2015 Reddit Comments (version 2). https://www.kaggle.com/reddit/reddit-comments-may-2015

[10] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 1630–1640. https://doi.org/10.18653/v1/P19-1159

[11] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. CoRR abs/1707.09457 (2017). arXiv:1707.09457 http://arxiv.org/abs/1707.09457

[12] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, New Orleans, Louisiana, 15–20. https://doi.org/10.18653/v1/N18- 2003

[13] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. arXiv:2005.14050 [cs.CL]

[14] Shikha Bordia and Samuel R. Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. CoRR abs/1904.03035 (2019). arXiv:1904.03035 http://arxiv.org/abs/1904.03035