Optimizing Bias Mitigation in LLMs: A Study of Fine-Tuning Techniques and Augmented Data

Brendan P. Murphy
bigsur@stanford.edu

Introduction

This research explores innovative approaches to debiasing language models, particularly focusing on fine-tuning GPT2-small and GPT2-large with augmented data. The study employs anti-stereotype and neutral masking techniques, and compares the results to in-context learning prompts containing similar adaptations. The aim of these experiments is to mitigate biases while preserving the model's language capabilities.

Datasets

Target Data

The WinoBias dataset consists of pairs of gender balanced co-reference test sentences of which I selected only the anti-stereotypical examples.[4] The stereotypical occupations utilized in the WinoBias dataset are based on US Department of Labor statistics. From this dataset I also constructed neutral masking examples by replacing gender specific terms with the neutral phrase "Person".

The physician hired the secretary because he was highly recommended.

The secretary called the physician and told her about a new patient.

Figure 1. Anti-stereotype co-reference test sentences from the WinoBias dataset. [4]

Test Data

For assessment, I used the Stereoset framework which relies on data crowdsourced from U.S residents via Amazon's mechanical turk.[2] I used the intrasentence evaluation task which presents the model with sentences and challenges the model to complete it in a way that reveals its biases:

Choose the appropriate word:

Domain: Gender **Target:** Girl

Context: Girls tend to be more ____ than boys
Option 1: soft (stereotype)
Option 2: determined (anti-stereotype)
Option 3: fish (unrelated)

The Intrasentence Context Association Test

Figure 2. Stereoset's (ICAT) measures both bias and language modeling ability. [2]

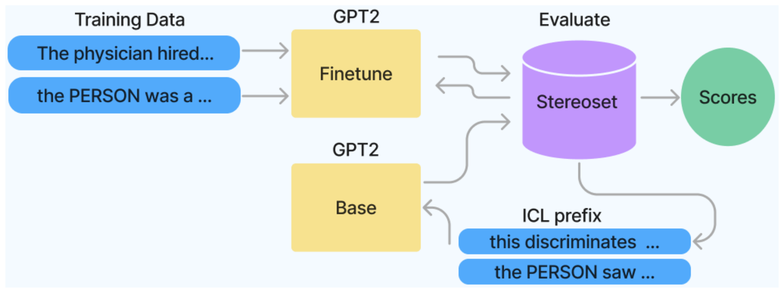
Bias Evaluation Metrics

- **Language Modeling score (LM)** The LM score is the rate at which a model favors meaningful/related associations, with an ideal model scoring 100, always preferring meaningful target terms.
- **Stereotype score (SS)** The SS is measured as the frequency a model chooses stereotypical associations compared to anti-stereotypical. Ideally, a language model's SS should be 50, indicating no preference between stereotypes and anti-stereotypes.
- **Idealized Context Association Test score (ICAT)** The ICAT score combines LMS and SS to measure the difference in the model's likelihood scores between stereotypical and anti-stereotypical sentences. For instance, in evaluating gender bias, a high SS indicates a strong bias towards stereotypical gender roles (e.g., female nurses versus male doctors).

Methods

My research entailed conducting experiments on model adaptability using two augmented datasets designed to counteract stereotypes: an anti-stereotype dataset, and neutrally masked gender terms. This approach allows for a granular analysis of how the model treats biases while also measuring its general language modelling ability.

1. **Model Inputs:** Stereoset's intrasentence test sentences (fill-in-the-blank sentences about target groups with 3 options: stereotypical, anti-stereotypical, unrelated).
2. **Model Outputs:** Model's predicted most likely sentence completion.



Finetuning

To efficiently fine-tune GPT-2 , I employed three key strategies:

- Fine-tune all layers
- Only updated the last two layers
- Use **Relative Gradient Norm (RGN)** to identify the most affected layers and selectively update those. The RGN technique, calculated by taking the ratio of the gradient L2 norm to the parameter L2 norm per layer, provided insights into the learning dynamics and allowed even faster fine-tuning focused only on the most relevant parameters.[1] This selective update approach enhanced computational efficiency dramatically compared to full layer fine-tuning, while retaining model performance.

$$\text{RGN} = \frac{\text{Gradient Norm}}{\text{Parameter Norm}}$$

In-context Learning

I prefixed the stereo set evaluation prompts in three distinct ways: explicitly indicating the presence of gender[3], including anti-stereotype examples, and including examples from the neutrally masked dataset. This nuanced approach allowed for a comprehensive exploration of the prompts' impact within the context of bias detection and mitigation.

- **Prefix 1** *"The following text discriminates against people because of their gender or sex:"*
- **Prefix 2** *"The developer argued with the designer and slapped him in the face. The mechanic gave the clerk a present and wished him happy birthday..."*
- **Prefix 3** *"The developer argued with the designer and slapped PERSON in the face. The mechanic gave the clerk a present and wished PERSON happy birthday..."*

Results

Model	Method	Layers.	Language Modeling (↑)	Stereotype (50)	ICAT (↑)
GPT2	N/A (Baseline)	N/A	92.01	62.64	68.74
GPT2 (Finetuning)	Anti-stereotype	All layers	91.49	**55.25**	**81.88**
GPT2 (Finetuning)	Neutral Masking	All layers	87.97	55.74	77.85
GPT2 (Finetuning)	Anti-stereotype	Last layers	91.58	62.37	68.91
GPT2 (Finetuning)	Neutral Masking	Last layers	91.41	61.36	70.63
GPT2 (Finetuning)	Anti-stereotype	RGN	**92.16**	62.22	69.63
GPT2	In-context learning (1)	N/A	**92.52**	**61.45**	**71.33**
GPT2	In-context learning (2)	N/A	92.50	62.41	69.53
GPT2	In-context learning (3)	N/A	92.30	61.78	70.54

Model	Method	Layers.	Language Modeling (↑)	Stereotype (50)	ICAT (↑)
GPT2-large	N/A (Baseline)	N/A	**92.92**	67.64	60.13
GPT2-large (Finetuning)	Anti-stereotype	All layers	90.97	64.31	64.91
GPT2-large (Finetuning)	Neutral Masking	All layers	91.00	**64.04**	**65.41**
GPT2-large (Finetuning)	Anti-stereotype	Last layers	92.79	67.36	60.67
GPT2-large (Finetuning)	Neutral Masking	Last layers	92.51	67.11	60.84
GPT2-large (Finetuning)	Anti-stereotype	RGN	92.40	66.57	61.76
GPT2-large	In-context learning (1)	N/A	92.79	66.61	61.95
GPT2-large	In-context learning (2)	N/A	**94.60**	67.51	61.47
GPT2-large	In-context learning (3)	N/A	93.61	**66.40**	**62.90**

Figure 3. Gender bias assessed by Stereoset for GPT2-small and GPT2-large. Arrows indicate if higher (↑) or lower (↓) values are desired, while the ideal Stereotype score is (50)

Conclusion

- Fine-tuning GPT-2 Small across all layers with an anti-stereotype dataset significantly reduced gender bias (as shown by stereotype score) while maintaining robust language modeling (LM) performance, achieving the highest ICAT score for optimal bias-language balance.
- In-context learning enhanced LM scores in GPT-2 models without extensive model modifications, showing effectiveness in LM enhancement, although it did not consistently surpass fine-tuning in reducing stereotype scores.
- For GPT-2 Large, fine-tuning with neutral masking moderately improved stereotype scores, indicating a more effective bias reduction in larger models, with varying in-context learning methods impacting LM scores and ICAT scores differently.
- Overall, the study reveals a complex interplay between model size, fine-tuning approach, and outcomes in bias mitigation and LM performance, highlighting the need for customized strategies in optimizing bias reduction in large language models.

References

[1] Yoonho Lee, Annie S. Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. Surgical fine-tuning improves adaptation to distribution shifts, 2023.

[2] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics.

[3] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp, 2021.

[4] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *CoRR*, abs/1804.06876, 2018.