Biases Mitigation and Expressiveness Preservation in Language Models: A Comprehensive Pipeline (Student Abstract)

Liu Yu, Ludie Guo, Ping Kuang [*], Fan Zhou

University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China
liu.yu@std.uestc.edu.cn, 202222090414@std.uestc.edu.cn, kuangping@uestc.edu.cn, fan.zhou@uestc.edu.cn

Abstract

Pre-trained language models (PLMs) have greatly transformed various downstream tasks, yet frequently display social biases from training data, raising fairness concerns. Recent efforts to debias PLMs come with limitations: they either fine-tune the *entire* parameters in PLMs, which is time-consuming and disregards the expressiveness of PLMs, or ignore the reintroducing biases from downstream tasks when applying debiased models to them. Hence, we propose a two-stage pipeline to mitigate biases from both internal and downstream contexts while preserving expressiveness in language models. Specifically, for the debiasing procedure, we resort to continuous prefix-tuning, not fully fine-tuning the PLM, in which we design a debiasing term for optimization and an alignment term to keep words' relative distances and ensure the model's expressiveness. For downstream tasks, we perform causal intervention across different demographic groups for invariant predictions. Results on three GLUE tasks show our method alleviates biases from internal and downstream contexts, while keeping PLM expressiveness intact.

Introduction

Pre-trained Language Models (PLMs) excel in diverse natural language tasks due to their training on extensive data. However, prior studies have revealed that PLMs inadvertently encode and propagate social biases from their unfiltered pre-training data. Take gender bias as an example: the PLM is more inclined towards associating *male* (*female*) attributes with *programmers* (*nurses*). Several solutions for mitigating the social biases have been proposed, including: (1) *Post-hoc*-based method add a post-training step to these sentence representations before applied to downstream tasks, including removing the estimated gender-direction subspace from sentence representation (Liang et al. 2020), or use pre-defined word tuples combine specific techniques to debias text encoder for a fair sentence representation (Cheng et al. 2021). (2) *Fine-tuning*-based models use specific loss terms to guide a PLM to remove biases, including distribution alignment loss for debiasing embedding space (Guo, Yang, and Abbasi 2022); orthogonal loss aims to promote irrelevance between stereotyped words and gender-specific words (Kaneko and Bollegala 2021), etc.

[*]Corresponding author.

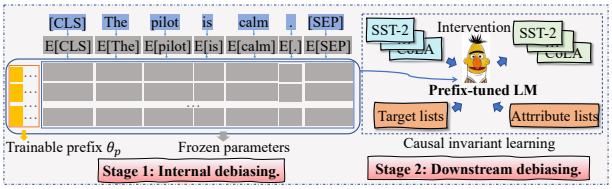


Figure 1: Our comprehensive debiasing pipeline.

Current debiasing methods for PLMs have shown promise but grapple with notable challenges: (1) demanding time-consuming to fine-tune *entire* parameters in PLMs; (2) disregarding the expressiveness of PLMs, which could potentially disrupt PLM's computational structure and undermine the benefits of pre-training; (3) reintroducing biases from downstream tasks into PLMs when applying debiased models to those tasks. Hence, we present a new two-stage pipeline that aims to simultaneously preserve the PLMs' expressiveness and mitigate biases from both internal and downstream contexts. As shown in Figure 1, in first stage, we keep PLM's parameter frozen, and only train the continuous prefix to reduce the magnitude of trainable parameters, towards mitigating internal bias and meanwhile preserving expressiveness. In second stage, we perform causal interventions on different demographic groups to eliminate the biases from downstream contexts.

Method

Let $\mathcal{W}_n$ and $\mathcal{W}_{a_i}$ denote the pre-defined *neutral* and *attribute* words tuples, where $i = 1, \ldots, d$ denotes the index of attributes types (e.g., $d = 2$ in binary gender case); we scrape natural sentences (i.e. $\mathcal{S}_n$ and $\mathcal{S}_{a_i}$) from News-Commentary v15 corpora containing at least one word in $\mathcal{W}_n$ or $\mathcal{W}_{a_i}$ for covering the diversity of demographic groups that reflects better with the real world. A fair PLM should offer equal attention to all groups without discrimination. To assess a model's perspective on various groups, we first extract embeddings for neutral words e^{n} and each stereotypical group e_{a_i}:

$$e_n = \mathcal{M}_\Theta\left(\mathcal{S}_n\right), e_{a_i} = \mathcal{M}_\Theta\left(\mathcal{S}_{a_i}\right)$$

where $e_n = [e_n^1, e_n^2, \ldots]$, $e_{a_i} = [e_{a_i}^1, e_{a_i}^2, \ldots]$ denote the neutral and attribute words embedding matrix extracted

from the associated neutral and attribute sentences, respectively, and Θ denotes the original parameters from a PLM $\mathcal{M}$. Then, we proceed with our two-stage debiasing pipeline.

• **Debiasing the PLM**: We seek to mitigate *internal* biases by minimizing the Wasserstein distance between pairwise attribute words and neutral words, i.e., effectively pushing pairwise attribute words closer to neutral word cluster:

$$\ell_{in\_bias} = \sum_{i,j \in \{1,\ldots,d\}, i<j} \left\{ D_{wass}\left(P_{a_i} \| P_{a_j}\right) \right\}$$

where P_{a_i} represents the distance from $E_{a_i} = \text{Aver}\left(e_{a_i}\right)$, which the average of attribute e_{a_i} to all neutral words e_n.

Prior studies have shown that debiasing can potentially damage the model's expressive ability. To mitigate this impact on PLM and preserve its benefits obtained from pretraining, we devise a KL divergence term to keep PLM's parameters unchanged before & after the debiasing procedure:

$$D_{KL}(\mathcal{M}_\Theta(\mathcal{S}) \| \mathcal{M}'_\Theta(\mathcal{S})) = \sum_{i=1}^{\|V\|} \sum_{j=1}^{\|V\|} \mathcal{M}_\Theta(\mathcal{S})_{ij} \log_2 \left(\frac{\mathcal{M}_\Theta(\mathcal{S})_{ij}}{\mathcal{M}'_\Theta(\mathcal{S})_{ij}} \right)$$

where V is the vocabulary size, and $\mathcal{M}_\Theta(\mathcal{S})_{ij}$ is a probability distribution matrix obtained from $\mathcal{M}$ that quantifies the degree to which the word w_i's information can be restored from the word w_j. $\mathcal{M}'_\Theta$ is short for $\mathcal{M}_{\theta_p \cup \Theta}$ as the debiased model. ℓ_{re} measures the differential between the original model's and the debiased model's hidden states.

In practice, instead of fine-tuning the entire PLM, we provide a set of continuous trainable prefixes θ_p before the language model's parameters as extra hints for optimization. The whole prefix-tuning loss for debiasing is as follows:

$$\min \mathcal{L}_p = \ell_{in\_bias} + D_{KL}$$

• **Fine-tuning downstream tasks**: When applied to downstream tasks, existing methods ignore new bias reintroduced into PLM, which neutralizes the impact of above debiasing. Hence, we propose a casual-inspired d-intervention on original sentence X_o from downstream stereotype groups, so the augmented datasets can be obtained:

$$X_a = X_o \cup X_c,$$

where X_c denotes counterfactual sentences via performing attribute word counterfactual augmentation. The risk under the n-interventional distribution is:

$$\mathcal{R}(\mathcal{M}(X_a), Y \mid do(N = n)) = \mathbb{E}_{C=m_C(x), N=n} l(\tilde{y}, y),$$

except for task prediction loss (i.e., $\mathcal{L}_\text{t}$), the PLM is required to predict the same results on X_o and X_c, which have equivalent semantics but different attribute words:

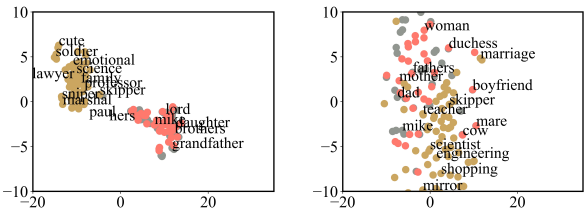
$$\min \mathcal{L} = \mathcal{L}_\text{t} + \mathbb{E}_n(\mathcal{R}) + \text{Var}_n(\mathcal{R}),$$

Experiment

For the gender bias evaluation, we report three stereotype scores on SEAT (6, 7, 8), Stereotype Score (SS), and CrowS-Pairs. For the expressiveness of PLM, we evaluate it by Language Modeling Score (LMS), and also visualize t-SNE in Figure 1 to explore the PLM's expressiveness on different methods. We evaluate our pipeline on three GLUE tasks, including **SST-2**, **RTE** and **CoLA**, based on `BERT-LARGE-UNCASED` (denoted as Orig.).

	Orig.	Context-Debias	Ours
C6	0.121	0.378	**0.023**
C7	0.253	**-0.091**	0.166
C8	-0.331	-0.038	**0.007**
LMS	**90.441**	84.420	90.019
SS	64.300	**59.657**	61.028
CrowS-Pairs	60.34	43.57	**53.32**
Acc. (SST-2)	0.924	0.927	**0.933**
Acc. (RTE)	0.527	0.487	**0.560**
Mcc. (CoLA)	0.588	0	**0.633**

Table 1: Evaluation results of debiasing.



(a) Context-Debias (b) Ours

Figure 2: Visualization of t-SNE plots.

As shown in Table 1, our pipeline performs better in bias mitigation and downstream tasks than baselines. Moreover, compared to Context-Debias (84.4), the LMS of our pipeline (90.0) remains merely unchanged to Orig. (90.4), and from Figure 1, it maintains words' relative distances, while simultaneously pulling pairwise attribute words closer, indicating its excellent expressiveness ability.

Acknowledgments

This work was supported in part by Key R&D Projects of Sichuan Provincial Science and Technology Plan (Grant No. 2023YFG0114 and 2023YFG0022), and Chengdu Key R&D Support Plan (Grant No. 2021YF0800019GX).

References

Cheng, P.; Hao, W.; Yuan, S.; Si, S.; and Carin, L. 2021. FairFil: Contrastive Neural Debiasing Method for Pretrained Text Encoders. In *ICLR*.

Guo, Y.; Yang, Y.; and Abbasi, A. 2022. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. In *ACL*, 1012–1023.

Kaneko, M.; and Bollegala, D. 2021. Debiasing pre-trained contextualised embeddings. In *EACL*.

Liang, P. P.; Li, I. M.; Zheng, E.; Lim, Y. C.; Salakhutdinov, R.; and Morency, L.-P. 2020. Towards Debiasing Sentence Representations. In *ACL*, 5502–5515.