Debiasing Intrinsic Bias and Application Bias Jointly via Invariant Risk Minimization (Student Abstract)

Yuzhou Mao[1*], Liu Yu[1*], Yi Yang[2], Fan Zhou[1†], Ting Zhong[1]

[1]University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China
[2] Hong Kong University of Science and Technology
yuzhou.mao@outlook.com, liu.yu@std.uestc.edu.cn, imyiyang@ust.hk, {fan.zhou, zhongting}@uestc.edu.cn

Abstract

Demographic biases and social stereotypes are common in pretrained language models (PLMs), while the fine-tuning in downstream applications can also produce new biases or amplify the impact of the original biases. Existing works separate the debiasing from the fine-tuning procedure, which results in a gap between *intrinsic bias* and *application bias*. In this work, we propose a debiasing framework CauDebias to eliminate both biases, which directly combines debiasing with fine-tuning and can be applied for any PLMs in downstream tasks. We distinguish the bias-relevant (non-causal factors) and label-relevant (causal factors) parts in sentences from a causal invariant perspective. Specifically, we perform intervention on non-causal factors in different demographic groups, and then devise an invariant risk minimization loss to trade-off performance between bias mitigation and task accuracy. Experimental results on three downstream tasks show that our CauDebias can remarkably reduce biases in PLMs while minimizing the impact on downstream tasks.

Introduction

With remarkable success of pretrained language models (PLMs) in many natural language processing (NLP) tasks, undesired stereotypes have been taken seriously. Take the cloze-style task as an example, as a result of gender occupational discrimination, the PLM fills in `[MASK]` of the sentence "The *boy*/*girl* got a job as `[MASK]`" with "*doctor*/*nurse*" respectively. Such demographic biases and social stereotypes would be inherited or amplified in downstream NLP tasks. Meanwhile, the fine-tuning procedure in downstream task even produces new biases, which two kinds of biases may lead to undesirable results.

The static word embedding debiasing method was firstly proposed in (Bolukbasi et al. 2016), and many debiasing methods have emerged in recent years. According to the stage of debiasing technique applied, existing methods fall into three main categories: (1) *Pretraining*: counterfactual data augmentation and increasing dropout parameters techniques are widely used in pretraining stage. (2) *Post-hoc*: Sent-debias (Liang et al. 2020) mitigates biases by removing the estimated gender-direction subspace from sentence representation. Fairfil (Cheng et al. 2021) uses contrastive learning to correct text encoder for a fair sentence representation. (3) *Fine-tuning*: Auto-Debias (Guo, Yang, and Abbasi 2022) automatically searches biased prompts and devises a distribution alignment loss for bias mitigation. Context-Debias (Kaneko and Bollegala 2021) designs a loss function to encourage the stereotype words and gender-specific words to be orthogonal, which, however, degrades debiasing performance due to over-reliance on orthogonality assumptions. Except for Context-Debias, the above debiasing works consider that the debiasing process is independent of downstream fine-tuning tasks, leading to a gap between *intrinsic bias* and *application bias*.

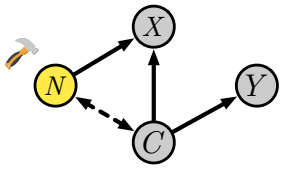

Figure 1: Structure Causal Model (SCM) of CauDebias. Each raw sentence of X is constructed by a mix of causal factors C and non-causal factors N. Note that only the causal factors effect the ground truth label Y, while the hammer indicates the intervention on non-causal factors.

Instead of separating debiasing and fine-tuning in prior studies, we present an approach to simultaneously mitigate the *intrinsic biases* in PLMs and the *application bias* in downstream tasks. Specifically, we start with Invariant Risk Minimization (IRM) theory (Arjovsky et al. 2019) to consider both biases. As shown in Figure 1, CauDebias reduces the impact of non-causal factors N (label-irrelevant factors) on PLMs by minimizing the distributional disparity between original bias-related sentences and augmented sentences.

Method

To alleviate the original biases (i.e. *intrinsic bias*) present in a PLM and newly incurred biases (i.e. *application bias*) when fine-tuning the PLM on downstream tasks, we propose a debiasing framework CauDebias from a causal invariant view. Due to the space limitation, we take the binary gender bias as an example-setting; however, we note that our ap-

*These authors contributed equally.
†Corresponding author.

proach is applicable to other types of biases, e.g., multi-class religion bias, racial bias, etc. CauDebias contains two steps: (1) Intervention: we extract bias-related sentences from external corpora to cooperate with the biased sentence in the downstream dataset to strengthen intervention on the non-causal factors; (2) Debiasing: after obtaining all bias-related sentences, we debias the PLM through fine-tuning to capture the actual label-related causal factors of the sentences with the assumption that the intervention on non-causal factors (related to biases) should be independent with ground truth.

Intervention. Let $\mathcal{W}_a$ and $\mathcal{W}_t$ denote two types of words: *attribute* words and *target* words, respectively. In the case of gender bias, attribute words are composed of feminine (e.g. *she, woman, mother*) and masculine (e.g. *he, man, father*) words, and target words consist of gender-neutral words (e.g. *nurse, engineer, professor*). We first obtain the original bias-related sentences S_o by extracting sentences containing any attribute or target word from downstream dataset. Then, we generate counterfactual sentences S_c by performing attribute word counterfactual augmentation on S_o – by merging S_c and S_o as S_d. Finally, we let S_o and S_c do semantic similarity matching with the external corpora respectively:

$$sim(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$$

, and then select two sets of sentences with top-k semantic similarity from external corpora (denoted by S_{ex}). Thus, we expand downstream dataset with S_{ex} as causal intervention.

Debiasing. For debiasing the language model and minimizing the impact of intervention on predictions, we assume that the PLM will predict the same results on S_d and S_{ex} inspired by the Invariant Risk Minimization theory. We choose Wasserstein distance as the optimization objective of causal invariant minimization:

$$\mathrm{P}_d = \mathrm{PLM}(S_d), \quad \mathrm{P}_{ex} = \mathrm{PLM}(S_{ex}),$$

$$\ell_{\text{invariant}} = D_{\text{Wasserstein}}(\mathrm{P}_d, \mathrm{P}_{ex}),$$

where P_d and P_{ex} are probability distributions predicted by the language model utilizing the sentences before and after counterfactual data augmentation S_d and the intervention sentence S_{ex}. Wasserstein distance has the property of being a very natural extension to consider the absolute difference between two random variables, formalized as below:

$$D_{\text{Wasserstein}}(x, y) = \inf_{\gamma(x,y)\in\prod} \mathcal{E}_{(x,y)\sim\gamma}||x - y||,$$

The overall optimization objective of CauDebias is a trade-off between target task performance and bias mitigation, which is summarized as follows:

$$\mathcal{L} = \ell_{\text{application}} + \tau \cdot \ell_{\text{invariant}},$$

where τ is the trade-off coefficiency, and $\ell_{\text{application}}$ is the loss function of a specific downstream task.

Experiment

We evaluate our CauDebias on three downstream tasks, including **CoLA**, **QNLI** and **SST-2**, based on three widely used PLMs: **BERT**, **ALBERT** and **RoBERTa**. We use Sentences Embedding Association Test (Caliskan, Bryson, and Narayanan 2017) to evaluate the biases by leveraging simple templates such as "This is a[n] $\langle word \rangle$". We report the average result of SEAT- 6, 6b, 7, 7b, 8, and 8b tests for measuring the gender bias. As shown in Table 1, our debiasing framework achieves superior performance in bias mitigation than baseline models and effective balance between task accuracy and bias mitigation.

Methods	SEAT avg.	Acc. avg.
BERT	0.35	0.79
+Sent-Debias	0.26	0.78
+Context-Debias	0.53	-
+FairFil	0.15	0.79
+Auto-Debias	0.14	0.78
+CauDebias†	**0.13**	0.80
ALBERT	0.28	0.81
+Context-Debias	0.33	-
+Auto-Debias	0.18	0.81
+CauDebias†	**0.17**	0.81
RoBERTa	0.67	0.79
+Context-Debias	1.09	-
+Auto-Debias	0.20	0.77
+CauDebias†	**0.15**	0.81

Table 1: Gender debiasing results of SEAT on BERT, ALBERT and RoBERTa. Absolute values closer to 0 are better. "†" denotes that average results of three tasks and "-" denotes that the original work does not be applied to all tasks.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (Grant No. 62072077 and No. 62176043), Natural Science Foundation of Sichuan Province (Grant No. 2022NSFSC0505).

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv :1907.02893*.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, volume 29.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

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.