BiasEdit: Debiasing Stereotyped Language Models via Model Editing

Anonymous Author(s)
Affiliation
Address
email

Abstract

Warning: This abstract explicitly contains offensive stereotypes.
Existing debiasing strategies, such as retraining a model with counterfactual data, representation projection, and prompting, often fail to efficiently eliminate bias or directly alter the models' biased internal representations. To address these issues, we propose **BiasEdit** (Figure 1), an efficient debiasing technique via model editing. BiasEdit employs a *debiasing loss* $\mathcal{L}_d = \mathrm{KL}(P_{\theta_{\tilde{\mathcal{W}}}}(x_{\text{stereo}}) \| P_{\theta_{\tilde{\mathcal{W}}}}(x_{\text{anti}})) + \mathrm{KL}(P_{\theta_{\tilde{\mathcal{W}}}}(x_{\text{anti}}) \| P_{\theta_{\tilde{\mathcal{W}}}}(x_{\text{stereo}}))$ guiding editor networks to conduct local edits on partial parameters of a language model for debiasing while preserving the language modeling abilities during editing through a *retention loss* $\mathcal{L}_r = \mathrm{KL}(P_{\theta_{\mathcal{W}}}(x_{\text{mless}}) \| P_{\theta_{\tilde{\mathcal{W}}}}(x_{\text{mless}}))$. Experiments on StereoSet and Crows-Pairs demonstrate the effectiveness, efficiency, and robustness of BiasEdit in eliminating bias compared to tangential debiasing baselines, and little to no impact on the language models' general capabilities.

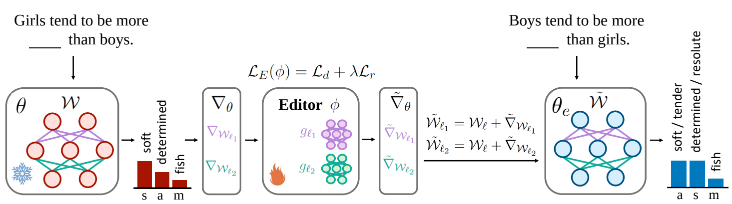

Figure 1: Editor networks ϕ are trained to produce edit shifts on partial parameters $\mathcal{W}$ of a language model while its parameters θ are frozen. After editing, an unbiased LM is obtained with the robustness of gender reversal and semantic generality. s: stereotyped. a: anti-s. m: meaningless.

Method	GPT2-medium						Gemma-2b					
	SS (%) $\to$ 50%			ΔLMS (%) $\to$ 0			SS (%) $\to$ 50%			ΔLMS (%) $\to$ 0		
	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion
Pre-edit	65.58	61.63	62.57	93.39	92.30	90.46	69.25	64.21	62.39	94.57	94.26	93.43
CDA	63.29	61.36	61.79	**-0.21**	-3.02	**0.00**				-		
SentenceDebias	67.99	58.97	56.64	+0.29	+1.52	+0.34	68.86	63.87	60.09	**-2.65**	-0.31	**-0.58**
Self-Debias	60.28	57.29	57.61	-3.47	-4.12	-1.35	65.70	58.29	58.02	-35.93	-30.39	-21.69
INLP	63.17	60.00	58.57	-5.15	**-1.49**	-2.48	52.17	62.96	58.57	-12.50	**-0.30**	-2.01
BiasEdit	**49.42**	**56.34**	**53.55**	-8.82	-5.12	-1.92	**48.59**	**55.86**	**47.36**	-4.78	-4.35	-5.44
Method	**Mistral-7B-v0.3**						**Llama3-8B**					
	SS (%) $\to$ 50%			ΔLMS (%) $\to$ 0			SS (%) $\to$ 50%			ΔLMS (%) $\to$ 0		
	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion	Gender	Race	Religion
Pre-edit	70.19	64.97	56.09	93.60	89.77	88.85	72.25	65.01	60.87	95.81	92.47	91.33
CDA				-						-		
SentenceDebias	68.36	64.54	54.94	-0.61	0.62	+0.09	68.55	64.97	59.91	**-0.22**	-1.14	-0.66
Self-Debias	61.79	**50.54**	60.68	-39.28	-29.17	-32.37	65.46	60.88	58.57	-40.04	-2.54	-28.64
INLP	69.22	65.23	55.90	**+0.35**	**-0.15**	-0.58	68.17	65.22	62.21	-1.43	**-0.09**	**0.00**
BiasEdit	**46.24**	51.46	**50.42**	-8.81	-8.59	**-0.03**	**49.18**	**53.51**	**51.13**	-13.42	-11.77	-10.02

Table 1: Performance of BiasEdit compared to previous debiasing baselines.