Multi-Feature Quantized Self-Attention for Fair Large Language Models

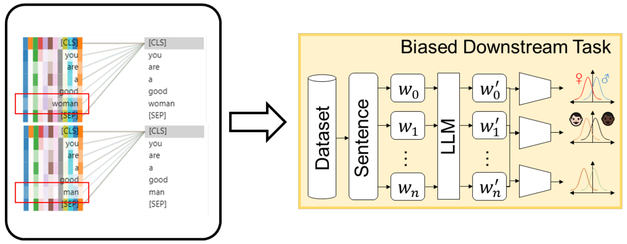
Jaeil Park and Sung-Bae Cho

Paper

Motivation

- LLMs deliver strong NLP performance
- Still inherit and amplify social bias!
- Bias in LLMs: often encoded in self-attention representations, not only in training data
- Existing methods: retraining, fine-tuning, or single-attribute assumptions

→ Real-world deployment demands a lightweight, multi-attribute, model-agnostic debiasing approach

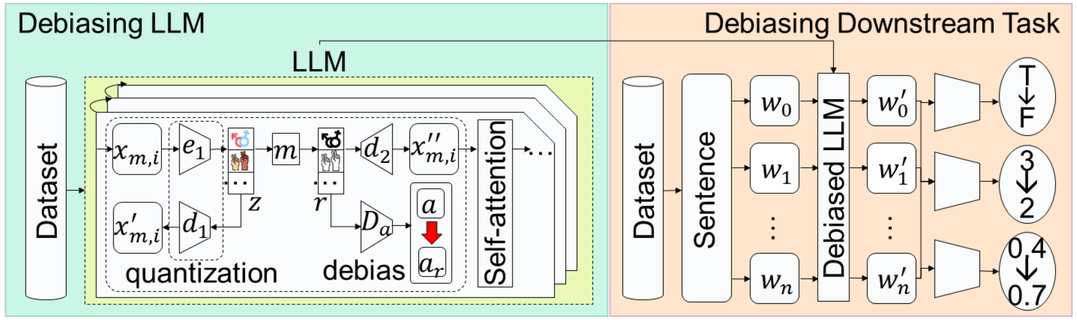


Contribution

- Model-agnostic & fine-tuning-free
  - A lightweight representation-level debiasing method for frozen LLM self-attention layers
  - Deployment across BERT, T5, GPT-Neo, Mixtral, and LLaMA 3.2
- Mitigating multi-attribute bias
  - Quantized autoencoder to addressing multi-feature regularization
  - Joint debiasing of multiple protected attributes and intersections within a single module
- Maintaining semantics with adversarial regularization
  - Removing biased information while preserving semantics

Multi-feature Quantized Attention Regularization

Proposed Method

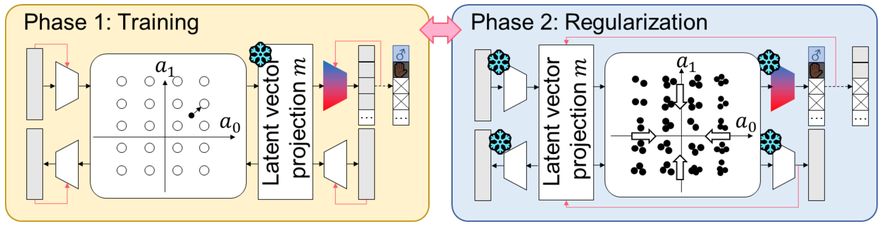


Phase 1. training the quantized autoencoder and discriminator

- Quantized autoencoder for multi-feature latent representation
- Discriminator for sensitive feature classification
- Decoder for preserving original vector information

Phase 2. regularization by discriminator

- Randomizing labels for the discriminator → removing sensitive feature information
- Adversarial learning with a discriminator and decoder



Strong duality holds for optimization

Theorem 4. *If $\epsilon_1, \epsilon_2, \epsilon_3 > 0$, then* ***strong duality holds*** *for the following optimization problem overdistributions p,e1,m:*

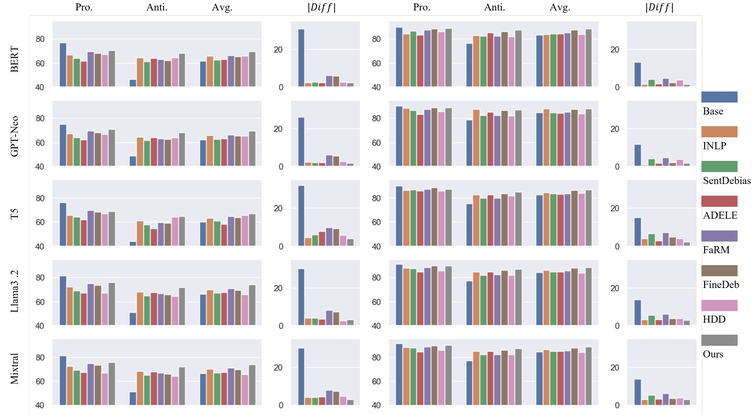
$$\min_{p,e_1,m} L \quad s.t. C_1 < \epsilon_1, C_2 < \epsilon_2$$

Lemma 2 and Corollary 3. *For any conditional distribution m,*

$$I(R;A) \le \int m(r|z)p(z)log\frac{m(r|z)}{s(r)} \coloneqq C_1$$

$$I(R;A) \le \mathbb{E}_{\mathbb{p}(\mathbb{r},\mathbb{a})}[\log p(\,a \mid r\,) - \log p(a)] \coloneqq C_2$$

Experiment on WinoBias Dataset



Experiment on Downstream Tasks

Task	Type	Metric	Model				
			Baseline	OSCaR	SentDebais	INLP	Ours
Abusive language detection (Founta)	Original Data	AUC	**93.8**	93.5	93.6	93.7	93.7
		FPED	2.32	**1.20**	2.53	1.92	1.87
		FNED	3.71	6.21	3.46	6.34	**3.44**
	Generated Data	AUC	92.3	91.9	92.5	91.5	**92.8**
		FPED	0.262	0.654	0.131	0.314	**0.0654**
		FNED	0.251	**0.036**	0.0835	0.332	0.167
Hate speech detection (CMSB)	Original Data	AUC	**96.5**	94.3	96.3	88.2	95.1
		FPED	0.121	0.443	**0.0117**	0.502	0.060
		FNED	9.54	3.61	4.43	12.4	**3.21**
	Generated Data	AUC	94.7	89.2	**94.9**	84.8	94.3
		FPED	0.0584	0.0562	**0.0137**	0.0192	0.0188
		FNED	3.01	1.01	0.0442	0.0257	**0.0218**
Sentiment analysis (EEC)	Anger Emotion	$\Delta_{F\uparrow-M\downarrow}$	0.0074	0.0092	0.0121	**0.0052**	**0.0052**
		$\Delta_{F\downarrow-M\uparrow}$	0.0316	0.0217	**0.0149**	0.0175	0.0163
	Anger Valence	$\Delta_{F\uparrow-M\downarrow}$	0.0219	0.0159	0.0130	**0.0121**	0.0133
		$\Delta_{F\downarrow-M\uparrow}$	0.0198	0.0137	0.0119	0.0130	**0.0105**
Text generation	GPTScore BLUE	Bias	8.73	4.36	5.82	4.31	**4.21**
		Bias	0.1	**0.07**	0.11	0.08	0.08
Question answering		Bias	36.8	33.5	35.3	34.6	35.5

Efficiency and Overhead, Ablation Study

Method	Param	FLOPs	Latency
Backbone	0.0	0.0	97.8
FineDeb	+107.3	+191.8	189.8
MQAR	+48.1	+87.0	132.7

Dataset	Metric	Original	(a)	(b)
Original	AUC	93.9	50.3	93.9
	FPED	1.84	20.2	2.50
	FNED	3.46	18.3	3.46
Generated	AUC	92.8	48.9	92.4
	FPED	0.0654	22.1	0.392
	FNED	0.167	17.5	0.250

This work was supported by the Yonsei Fellow Program funded by Lee Youn Jae, IITP grant funded by the Korea government (MSIT) (RS-2022-II220113 , Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework), and Air Force Defense Research Sciences Pro-gram funded by Air Force Office of Scientific Research.