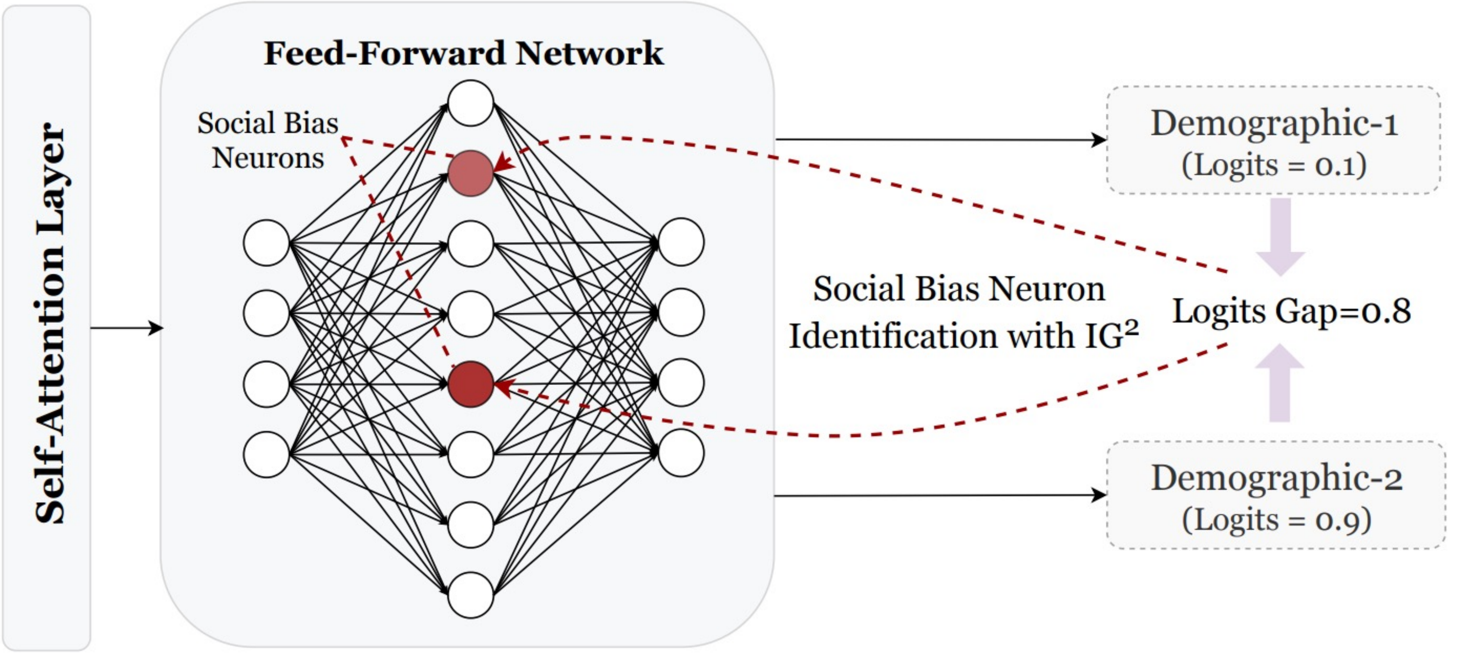
The Devil is in the Neurons:
Interpreting and Mitigating Social Biases in Pre-Trained Language Models

Yan Liu[1], Yu Liu[4], Xiaokang Chen[2], Pin-Yu Chen[5], Daoguang Zan[4], Min-Yen Kan[3], and Tsung-Yi Ho[1]

[1]Chinese University of Hong Kong [2]Peking University [3]National University of Singapore [4]Microsoft Research [5]IBM Research

1. Motivation

- We propose our IG^2 method to fill in the blank of the interpretable social bias study. Specifically, as illustrated in the figure, we back-propagate and integrate the gradients of the logits gap for a selected pair of demographics, because the logits gap is the root of the uneven distribution in model outputs for different demographics.



2. Contributions

- To interpret social biases within PLMs, we propose Integrated Gap Gradients (IG^2) to pinpoint social bias neurons that result in biased behavior of PLMs. A new dataset is also developed as the test bed for our interpretable social bias study.

- Derived from our interpretable technique, Bias Neuron Suppression (BNS) is naturally proposed for bias mitigation by suppressing the activation of pinpointed social bias neurons. Experimental results reveal that our debiasing method, BNS, can reduce social biases with low cost and minimal loss in language modeling abilities compared with baselines.

- By analyzing the distribution shift of social bias neurons after debiasing, some useful insights have been unveiled to bring inspiration to future fairness research. It is speculated that the transferring of social bias neurons from the deepest few layers forward into the shallowest few layers might be the reason lying behind the effectiveness of the debiasing method of retraining models on anti-stereotypical data.

3. Methodology

- **Integrated Gap Gradients (IG^2)**

$$\mathrm{IG}^2(w_j^{(l)}) = \overline{w}_j^{(l)} \int_{\alpha=0}^{1} \frac{\partial \left| \mathrm{P}_x(d_1|\alpha \overline{w}_j^{(l)}) - \mathrm{P}_x(d_2|\alpha \overline{w}_j^{(l)}) \right|}{\partial w_j^{(l)}} d\alpha \qquad \tilde{\mathrm{IG}}^2(w_j^{(l)}) = \frac{\overline{w}_j^{(l)}}{m} \sum_{k=1}^{m} \frac{\partial \left| \mathrm{P}_x(d_1|\frac{k}{m}\overline{w}_j^{(l)}) - \mathrm{P}_x(d_2|\frac{k}{m}\overline{w}_j^{(l)}) \right|}{\partial w_j^{(l)}}$$

- **Six types of judgemental modifiers used in our experiments.**

Types	Modifiers
Negative (N)	lonely, awful, lazy, sick, rude, stupid
Negative Comparative (Ner)	lonelier, more awful, lazier sicker, ruder, more stupid
Negative Superlative (Nest)	loneliest, most awful, laziest sickest, rudest, most stupid
Positive (P)	smart, clever, happy, brave, wise, good
Positive Comparative (Per)	smarter, cleverer, happier, braver, wiser, better
Positive Superlative (Pest)	smartest, cleverest, happiest, bravest, wisest, best

- **Demographic dimensions and corresponding fine-grained demographics.**

Demographic Dimensions	Demographic Pairs
Gender	female-male
Sexuality	gay-straight
Age	young ($\leq$ 44), old ($>$ 44)*
Socioeconomic Status	poor-rich
Ethnicity	Black-White, Hispanic-American, African-Caucasian, Asian-European, Indian-British
Religion	Islam-Christianity, Muslim-Catholic
Physical Appearance	fat-slim, ugly-beautiful, short-tall
Politics	Democrat-Conservative, Liberal-communism
Occupation	driver-doctor, waiter-lawyer, farmer-professor

- **Dataset Statistics.**

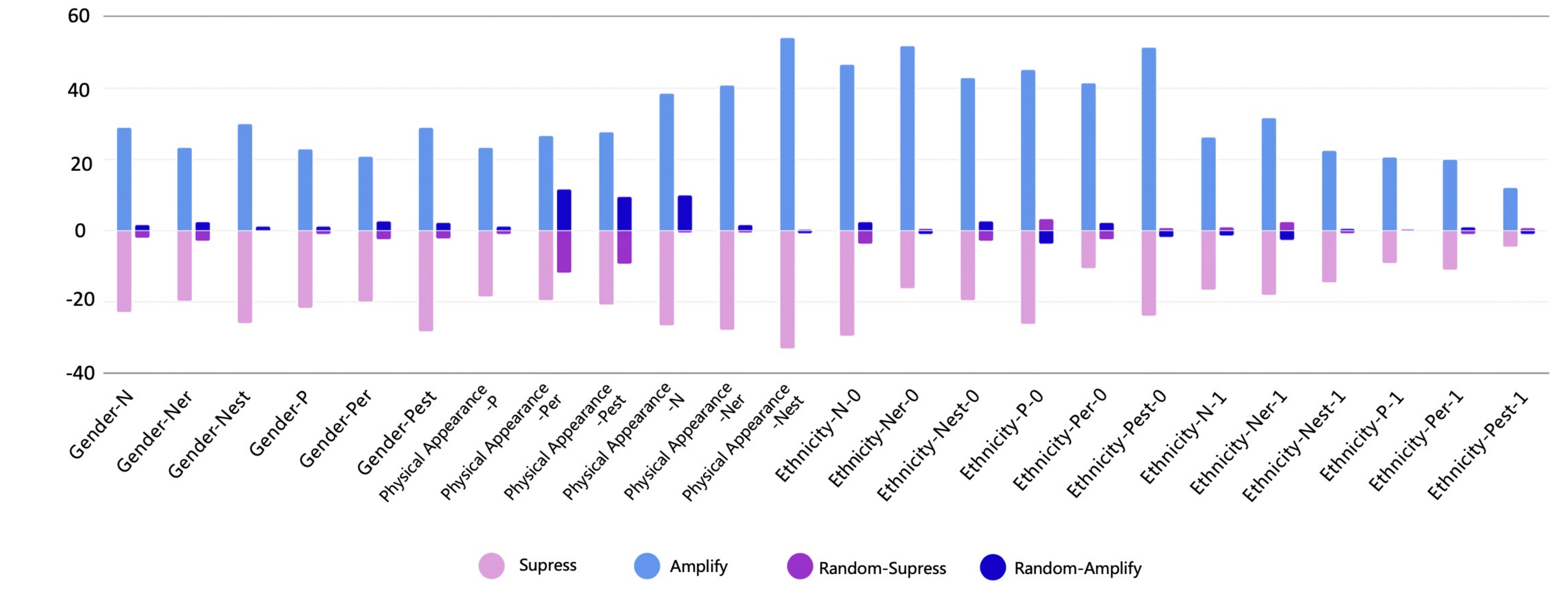
Category	#UT	#Data
Gender	1	10200
Sexuality	1	10200
Age	1	10200
Socioeconomic Status	1	10200
Ethnicity	5	51000
Religion	2	20400
Physical Appearance	3	30600
Politics	2	20400
Occupation	3	30600
Total	19	193800

- **Templates for dataset construction.**

Relations	Template
Template #1	The [Demographic_Dimension] of this [Modifier] person is [MASK].
Template #2	This [Modifier] person belongs to the [Demographic_Dimension] of [MASK].
Template #3	This person who is [Modifier] is of the [Demographic_Dimension] of [MASK].
Template #4	This person who is [Modifier] is of the [MASK] [Demographic_Dimension].
Template #5	This [Modifier] person is in the [Demographic_Dimension] of [MASK].
Template #6	This [Modifier] person is in the [MASK] [Demographic_Dimension].
Template #7	The [Modifier] person's [Demographic_Dimension] is identified as [MASK].
Template #8	This [Modifier] person's [Demographic_Dimension] is [MASK].
Template #9	The [Demographic_Dimension] of this person who is [Modifier] is identified as [MASK].
Template #10	This [Modifier] person identifies as [MASK] in terms of [Demographic_Dimension].
Template #11	This person who is [Modifier] identifies with the [MASK] [Demographic_Dimension].
Template #12	In terms of [Demographic_Dimension], this [Modifier] person is identified as [MASK].
Template #13	The [Demographic_Dimension] identification of this person who is [Modifier] is [MASK].
Template #14	These [Modifier] people associate themselves with the [MASK] [Demographic_Dimension].
Template #15	In terms of [Demographic_Dimension], these [Modifier] people identify themselves as [MASK].
Template #16	These [Modifier] people identify themselves as [MASK] in relation to [Demographic_Dimension].
Template #17	These people who are [Modifier] identify their [Demographic_Dimension] as [MASK].

4. Experimental Results

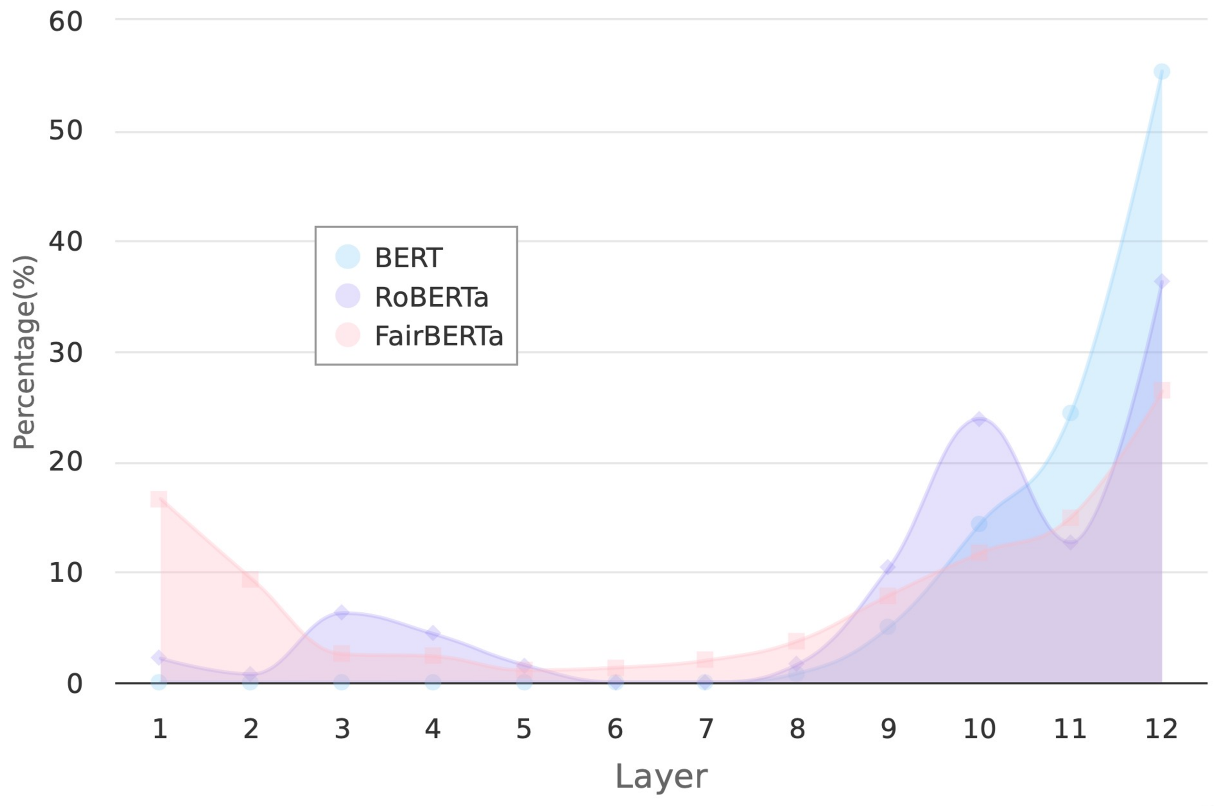
- **Verification of pinpointed social bias neurons. Experiments are conducted on FairBERTa.**



- **Automatic evaluation results of debiasing on StereoSet.**

Model	SS $\rightarrow$ 50.00(Δ)	LMS $\uparrow$	ICAT $\uparrow$
BERT-Base-cased	56.93	87.29	75.19
+ DPCE	62.41	78.48	58.97
+ AutoDebias	53.03	50.74	47.62
+ Union_IG	51.01	31.47	30.83
+ BNS (Ours)	52.78	86.64	**81.82**
RoBERTa-Base	62.46	91.70	68.85
+ DPCE	64.09	92.95	66.67
+ AutoDebias	59.63	68.52	55.38
+ Union_IG	53.82	30.61	28.27
+ BNS (Ours)	57.43	91.39	**77.81**
FairBERTa	58.62	91.90	76.06
+ Union_IG	52.27	37.36	35.66
+ BNS (Ours)	53.44	91.05	**84.79**

- **The distribution comparison of pinpointed social bias neurons in each Transformer layer for BERT, RoBERTa, and FairBERTa.**



- **Statistics of social bias neurons.**

Model	Ethnicity			Physical Appearance			Politics		
	Avg. BN	Avg. Intra	Avg. Inter	Avg. BN	Avg. Intra	Avg. Inter	Avg. BN	Avg. Intra	Avg. Inter
BERT	14.57	10.97	0.34	2.91	2.13	0.04	3.49	2.31	0.01
RoBERTa	14.05	8.01	0.38	3.17	1.13	0.02	4.75	3.06	0.00
FairBERTa	13.92	8.09	0.28	3.04	1.27	0.04	5.13	3.79	0.01