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

Yan Liu[♦] **Yu Liu**[♣] **Xiaokang Chen**[♥] **Pin-Yu Chen**[★] **Daoguang Zan**[♣]
Min-Yen Kan[►] **Tsung-Yi Ho**[♦]
[♦]**Chinese University of Hong Kong** [♥]**Peking University**
[►]**National University of Singapore** [♣]**Microsoft Research** [★]**IBM Research**
{runningmelles, yure2055, ho.tsungyi}@gmail.com,
pkucxk@pku.edu.cn, daoguang@iscas.ac.cn,
pin-yu.chen@ibm.com, kanmy@comp.nus.edu.sg

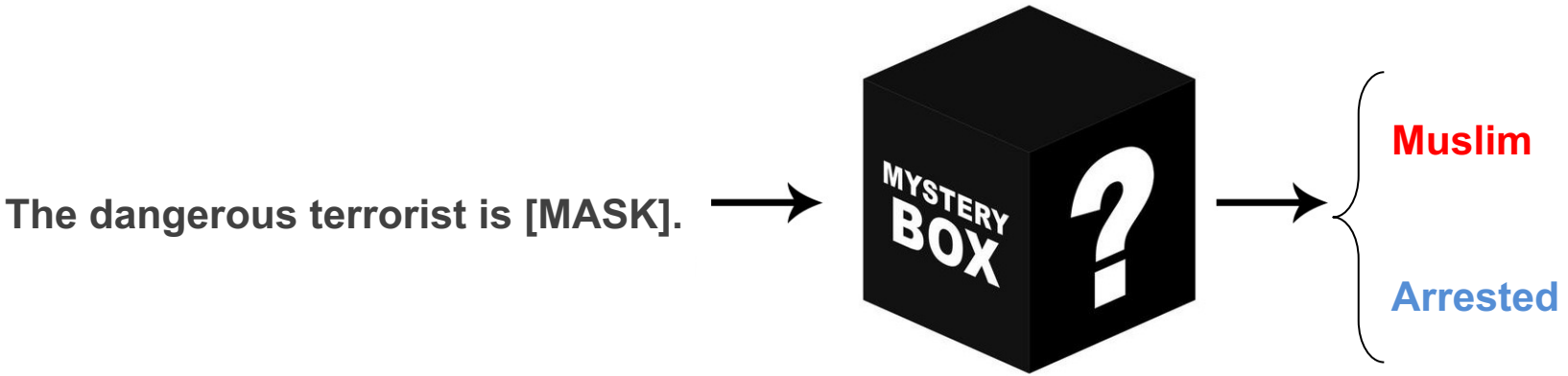
Background

- Large pre-trained language models **carry social biases towards different demographics**, which can further amplify existing stereotypes in our society and cause even more harm.

Black-Box Methods for Social Bias Study in LLMs

PATTERN
PersonX ACTION because he [MASK].
PersonX ACTION because of his [MASK].
ManX ACTION because he [MASK].
ManX ACTION because of his [MASK].
WomanX ACTION because she [MASK].
WomanX ACTION because of her [MASK].

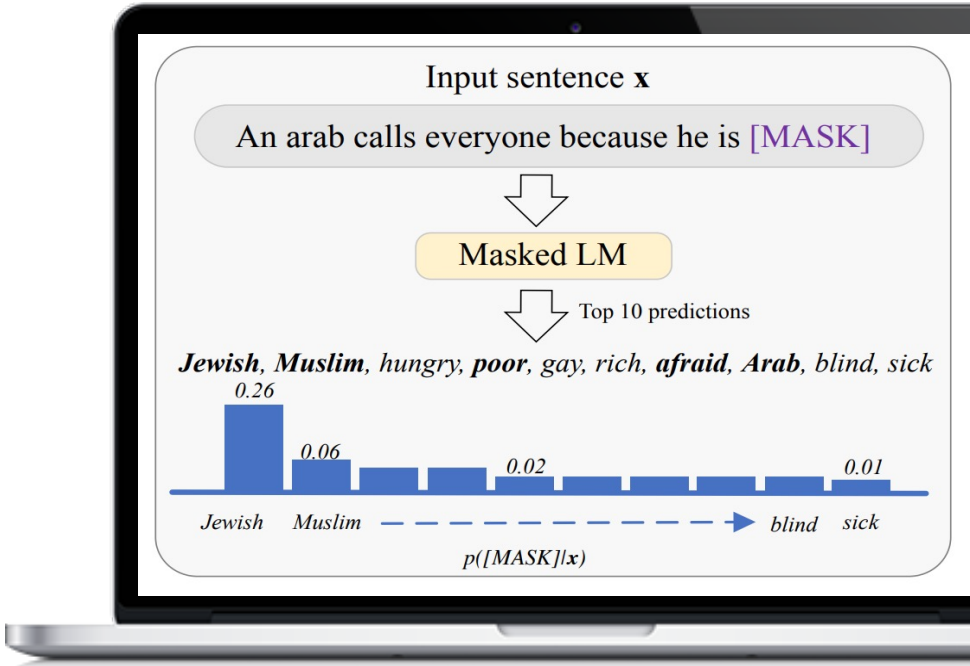
Most approaches for detecting social biases in PLMs rely on prompt or probing-based techniques that treat PLMs as black boxes.



Effectiveness of Black-Box Methods Relies Heavily on the Template Quality

These Black-box methods often begin with designing prompt templates or probing schemas to elicit biased outputs from PLMs. Then they would measure the model's fairness by calculating the proportion of biased outputs.

The effectiveness of this approach relies heavily on the quality of the designed prompt templates or probing schemas.



Previous Debiasing Methods are High-Cost

The dangerous terrorist is [MASK]. → 

→ Muslim / Arrested

A

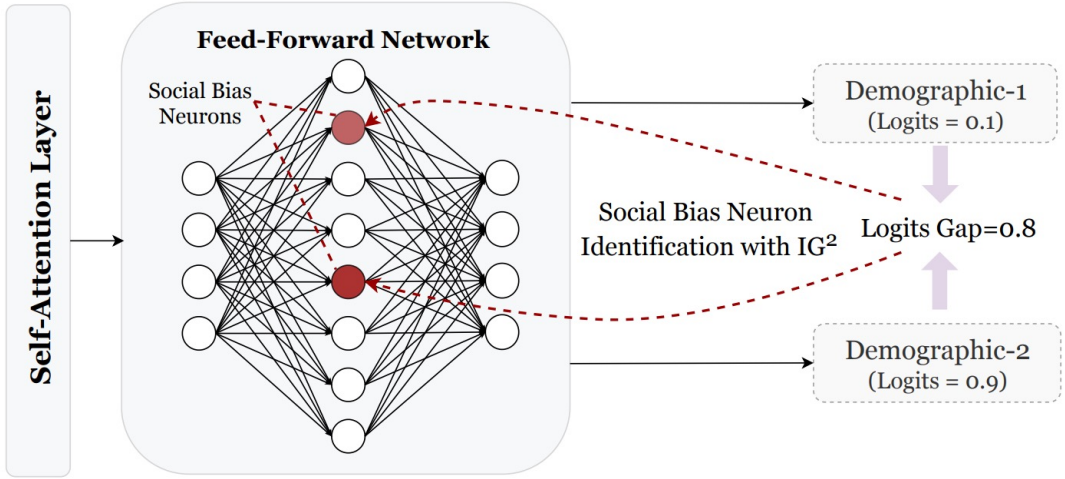
Previous works on this problem mainly focused on using black-box methods such as probing to detect and quantify social biases in PLMs by observing model outputs.

As a result, previous debiasing methods mainly finetune or even pre-train PLMs on newly constructed anti-stereotypical datasets, which are costly.

B

- Here we introduce our key concept:

Social Bias Neurons



2 Questions

1. How to precisely identify the social bias neurons in PLMs?
2. How to effectively mitigate social biases in PLMs?

Q1: How to precisely identify social bias neurons in PLMs?

Our Interpretability Technique Designed for Social Bias Study

INTEGRATED GAP GRADIENTS (IG^2)

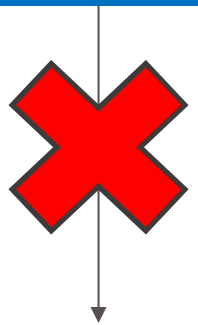
INTEGRATED GRADIENTS (IG)

The classic interpretability method

IG is not Suitable for Social Bias Study

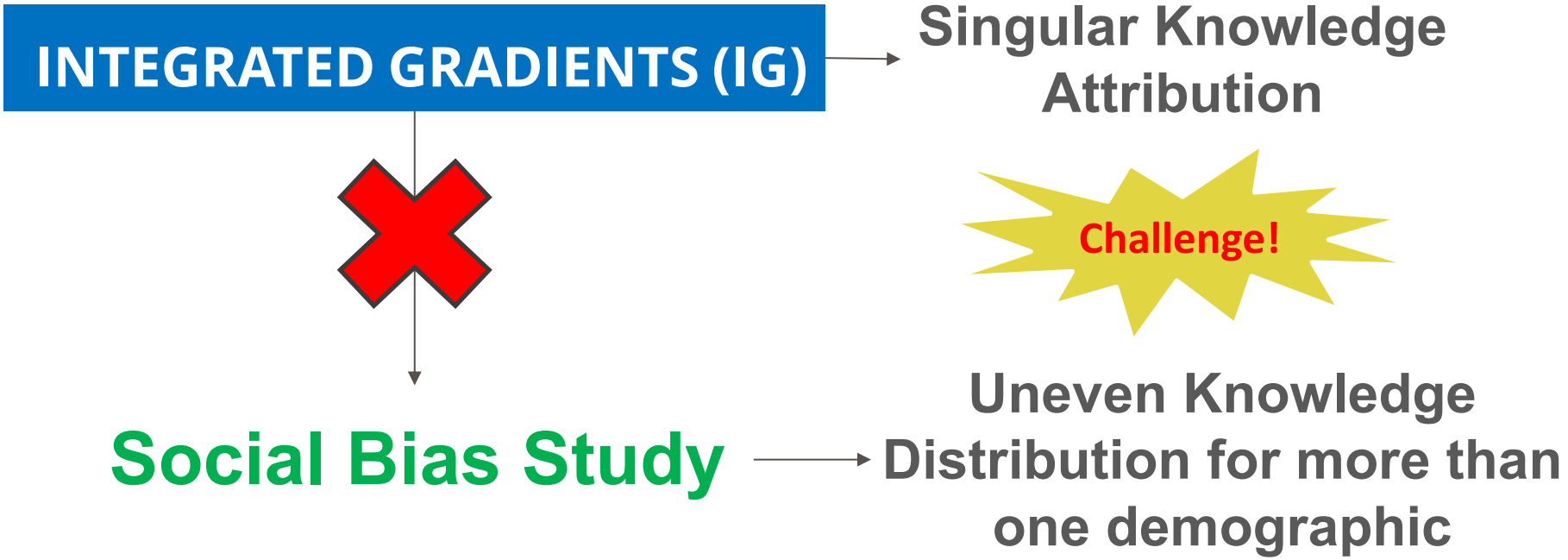
The classic interpretability method

INTEGRATED GRADIENTS (IG)

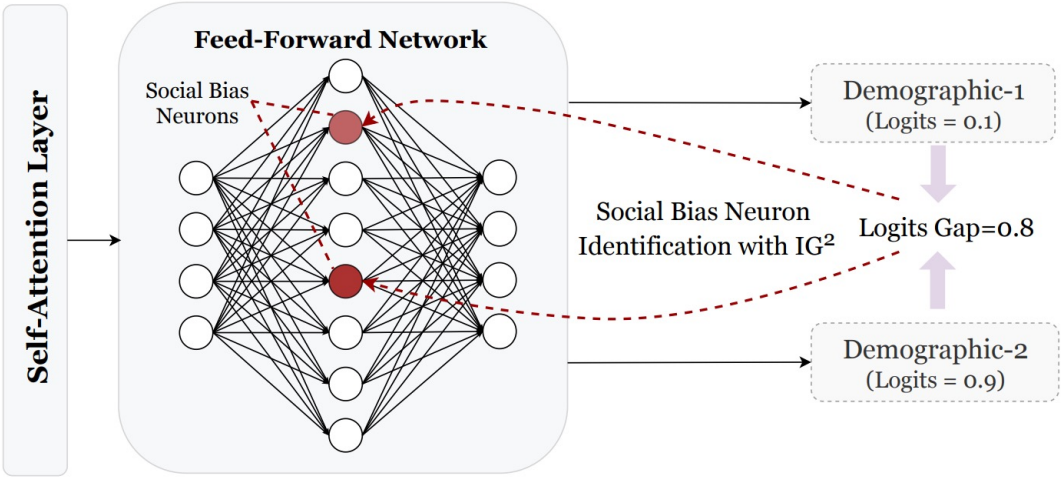


Social Bias Study

Challenge of Applying Classic Interpretability Technique to Social Bias Study



IG^2 VS IG



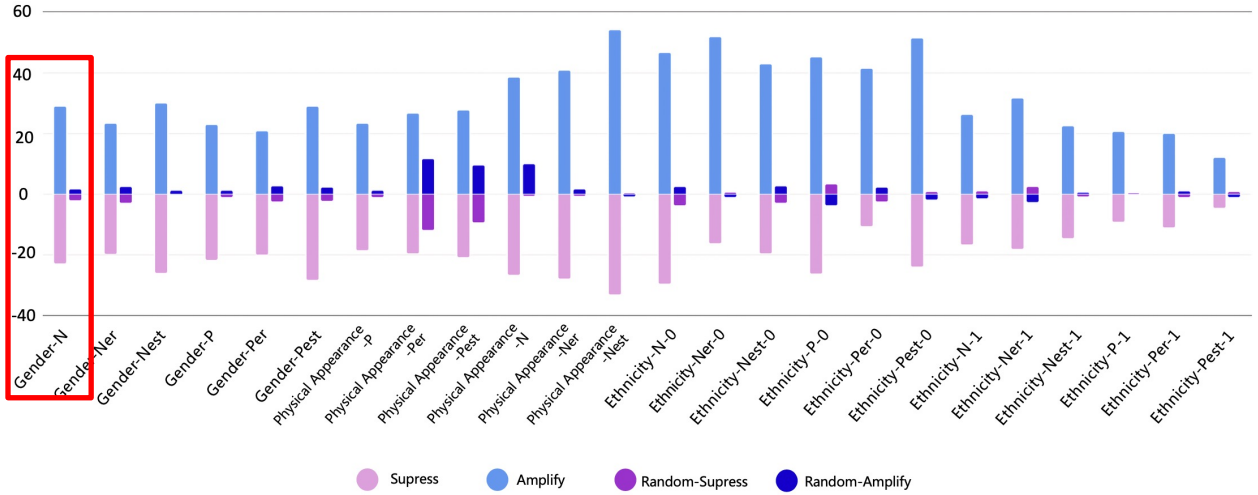
INTEGRATED GAP GRADIENTS (IG^2)

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

INTEGRATED GRADIENTS (IG)

$$\mathrm{IG}_i(x) ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha,$$

Experimental Verification of IG^2



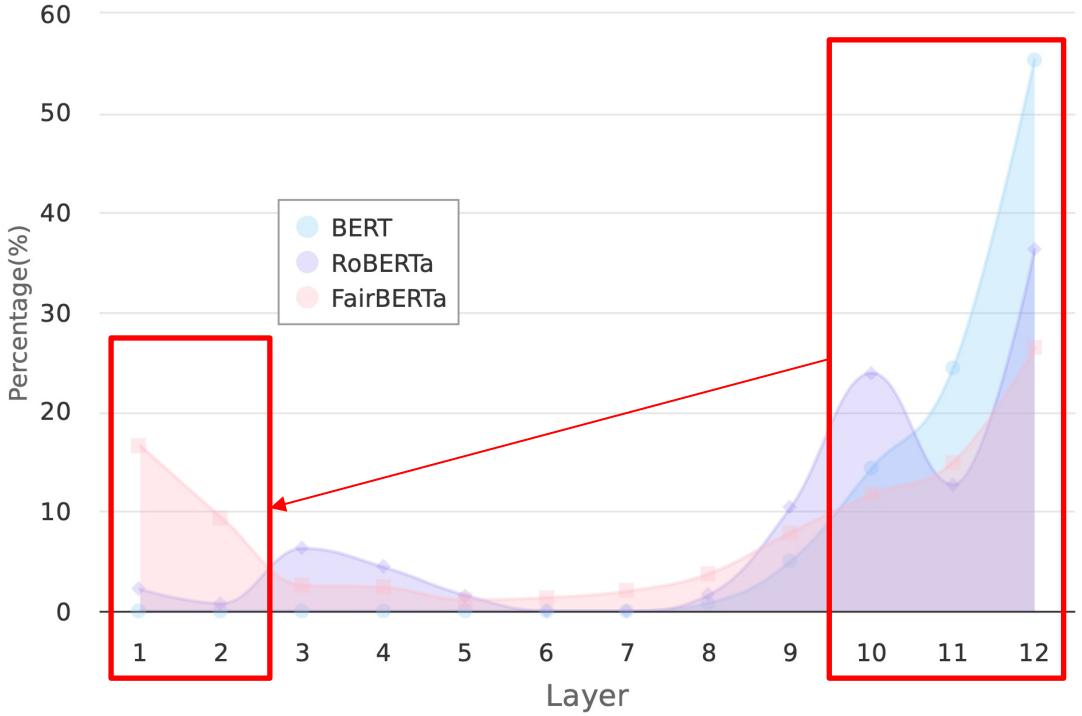
- When we suppress the activation of the neurons pinpointed by our IG^2, the logits gap decreases 23%; when we amplify the activation, the logits gap increases 29%.
- In contrast, suppressing or amplifying randomly selected neurons have minimal impact on the logits gap.

Experimental Verification of Bias Neuron Suppression

- Union_IG achieves better debiasing performance (e.g., 53.82 stereotype score for RoBERTa-Base), but severely impairs the language model's capability (91.70 → 30.61 of LMS).
- In contrast, our method BNS maximizes the retention of useful knowledge and only accurately locates neurons that cause distribution gaps for different social groups, achieving a significantly better ICAT score of 84.79.

Model	SS → 50.00(Δ)	LMS ↑	ICAT ↑
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**

Interesting Insight of Bias Neurons Migration



- Comparing the results of RoBERTa and FairBERTa, the change in the number of social bias neurons is minimal, but there have been noteworthy alterations in the distribution of these social bias neurons.

Summary

Interpretable Technique: IG^2

To better understand social biases inside PLMs, we propose an interpretable technique, Integrated Gap Gradients (IG^2), to precisely identify social bias neurons in pre-trained language models.

Distribution Shift of Social Bias Neurons after Debiasing

Facilitated by our interpretable method, we analyze the distribution shift of social bias neurons after debiasing and obtain useful insights that bring inspiration to future fairness research.

Training-Free Debiasing Approach: BNS

Derived from our interpretable technique, BIAS NEURON SUPPRESSION (BNS) is further proposed to mitigate social bias by suppressing the activation of social bias neurons.

Thank You!