CS689 PROJECT REPORT: BIAS IN LANGUAGE MODELS USING CAUSAL MEDIATION ANALYSIS

Yash Gupta
IIT Bombay
yashgupta@cse.iitb.ac.in

November 30, 2021

ABSTRACT

Regarding information inside hidden representations, most of the methods can only measure whether the information exists, not whether it is actually used by the model. [1] proposes a methodology of causal mediation analysis for interpreting which parts of a model are causally implicated in its behavior. The approach enables us to analyze the mechanisms that facilitate the flow of information from input to output through various model components, known as mediators. The authors apply this for the case of analyzing gender bias in pre trained LMs. We study neurons and attention heads in two datasets. The paper concludes that gender bias effects are concentrated in specific components of the model that may exhibit highly specialized behavior.

1 Introduction

1.1 Causal Mediation Analysis

Causal mediation analysis (CMA) is a method to dissect total effect of a process into direct and indirect effect. We consider each neuron in a neural network – the neuron is influenced by the input and, in turn, affects the model output. There also exist direct pathways from the input to the output that do not pass through the neuron. We can thus decouple model components from the rest of the model by framing them as intermediaries in the causal path from inputs to outputs. We propose the following as the gender bias of the model:

$$y(u) = \frac{p_\theta(anti\ stereotypical \mid u)}{p_\theta(stereotypical \mid u)} \tag{1}$$

For causal analysis, we define the following do-operations: (a) set-gender: replace the ambiguous profession with an anti-stereotypical gender-specific word (that is, replace nurse with man, doctor with woman, etc.); (b) null: leave the sentence as is. The population of units for this analysis is a set of example sentences such as the above prompt. We define $y_x(u)$ as the value that y attains in unit u under the intervention do(x).
The unit-level total effect (TE) of x on y in unit u is the proportional difference between the amount of bias under a gendered reading and under an ambiguous reading.

$$TE(set - gender, null; y, u) = \frac{y_{set-gender}(u) - y_{null}(u)}{y_{null}(u)} \tag{2}$$

The average total effect of x = x on y is calculated by taking the expectation over the population u. We then analyze the causal role of specific mediators, or intermediary variables, which lie between x and y. The mediator, denoted as z, might be a particular neuron, a full layer, an attention head, or a certain attention weight.

2 ABOUT THE DATASETS

2.1 Professions

For neuron intervention experiments, we have a list of templates instantiated with professions. The templates have the form "The [occupation] [verb] because". The professions are accompanied by crowdsourced ratings between -1 and 1 for definitionality and stereotypicality. Actress is definitionally female, while nurse is stereotypically female. None of the professions are stereotypically or definitionally gender-neutral in the sense that those people working in the profession are referred to in singular they. To simplify processing by GPT2 and focus on common professions, we only use examples that are not split into sub-word units, resulting in 17 templates and 169 professions, 2,873 examples in total. The dataset we use consists of a list of templates that are instantiated by profession terms, resulting in examples such as The nurse said that.

2.2 Winobias

This dataset is meant for understanding the gender bias issues lying in such systems. Providing the same sentence to the system but only changing the gender of the pronoun in the sentence, the performance of the systems varies. WinoBias contains 3,160 sentences (only 160/130 fir our formulation), split equally for development and test. Sentences were created to follow two prototypical templates but annotators were encouraged to come up with scenarios where entities could be interacting in plausible ways. Templates were selected to be challenging and designed to cover cases requiring semantics and syntax separately.

3 Methodology

3.1 Neuron Intervention

We study neuron intervention on the Professions Dataset. To study the role of individual neurons in mediating gender bias, we assign z to each neuron $h_{l,,k}$ in the LM. For each example, we define the set-gender operation to move in the anti-stereotypical direction, changing female-stereotypical professions like nurse to man and male-stereotypical professions like doctor to woman. In all cases, the mediator is in the representation corresponding to the profession word, such as nurse in the example.

3.2 Attention Intervention

The mediators z, in this case, are the attention heads $\alpha_{l,h}$, each of which defines a distinct attention mechanism.

Prompt u: The nurse examined the farmer for injuries because she
Stereotypical candidate: was caring
Anti-stereotypical candidate: was screaming

According to the stereotypical reading, the pronoun she refers to the nurse, implying the continuation was caring. The anti-stereotypical reading links she to the farmer, this time implying the continuation was screaming. The bias measure is

$$y(u) = \frac{p_\theta(was\ screaming \mid u)}{p_\theta(was\ caring \mid u)} \tag{3}$$

In this case, we define the swap-gender operation, which changes she to he. The total effect is

$$TE(swap - gender, null; y, u) = \frac{y_{swap-gender}(u)}{y_{null}(u)} - 1 \tag{4}$$

In the experiments, we study the effect of the attention from the last word (she or he) to the rest of the sentence. Intuitively, in the above example, if the word she attends more to nurse than to farmer, then the more likely continuation might be was caring. We compute the NDE and NIE for each head individually by intervening on the attention weights.

4 Results

Our project link is :

`https://github.com/yashgupta-7/689_project`

Note: The code is adapted from `https://github.com/sebastianGehrmann/CausalMediationAnalysis`

4.1 Neuron Intervention

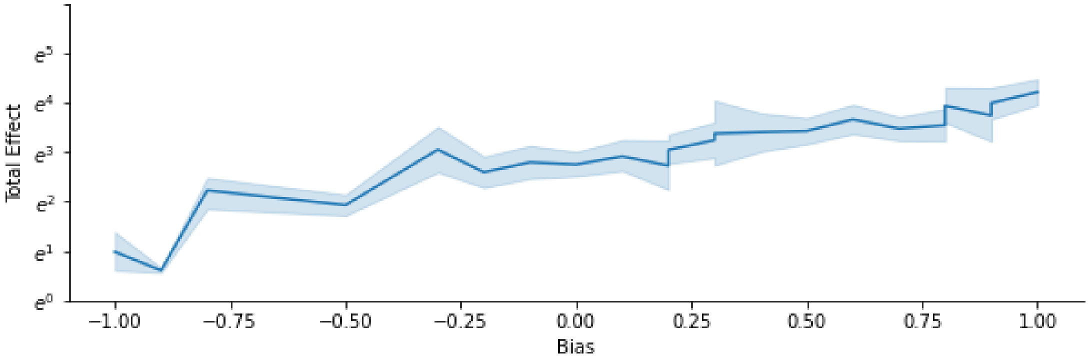

Figure 1: Model is distil gpt2. We can see that total effect roughly increases with bias. The total effect of this model is 2.237. The total (male) effect of this model is 3.080. The total (female) effect of this model is 48.186. The correlation between bias value and (log) effect is 0.60 (p=0.000).

Table 1: Total Effect on GPT2 variants

Dataset/Model	random GPT2	distilgpt2	small gpt2
Professions	0.120	142.45	130.28

4.2 Attention Intervention

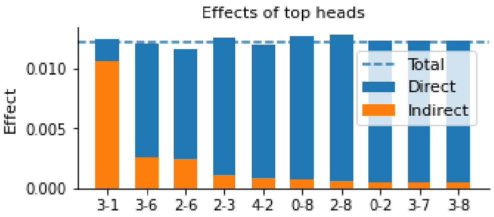

Figure 2: Top-10 heads effect by distilGPT2. The effects are additive.

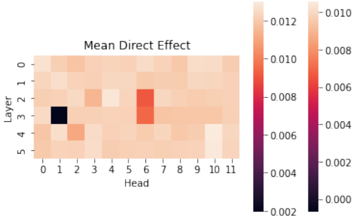

Figure 3: Direct effects in distilGPT2 on Winobias for heads and layers.

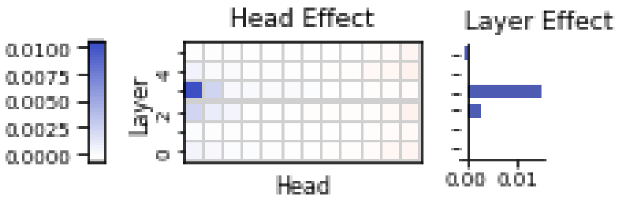

Figure 4: Indirect effects in distilGPT2 on Winobias for heads and layers.

5 Discussion

We do all our experiments with distilgpt2 due to compute limitations. From Table 1, we see that larger models are more sensitive to gender bias. This is because larger models can remeber more data (which is gender biased) and thus show more bias. Figure 1 In Professions: positive correlations between external gender bias and the log-total effect 0.35-0.45.
Figure 4 show that a small number of heads, concentrated in the middle layers of the model, have much higher indirect effects than others. We do not find such effects in randomised model so they do not occur by chance. Moreover, the total effect roughly equals the sum of the direct and indirect effects even though the models themselves are non linear (Figure 2).

6 Conclusion

- A framework for interpreting neural NLP models based on causal mediation analysis.
- Larger models have a greater capacity to absorb gender bias, though this bias manifests in a relatively small proportion of neurons and attention heads.
- Model components may take on specialized roles in propagating gender bias.

7 Acknowledgement

I would like to thank Prof. Harish, who guided me on this project and throughout this course.

References

[1] Vig, Jesse et al. “Investigating Gender Bias in Language Models Using Causal Mediation Analysis.” NeurIPS (2020).

[2] Zhao, Jieyu et al. “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods.” NAACL (2018).