Uncovering Stereotypical Bias in Multimodal LLMs - Report for LT2318 -

Dylan Massey
University of Gothenburg
gusmasdy@student.gu.se

Abstract

The advent of performant and readily available large-language models (LLMs) motivates a renewed discussion about the potential harms such models can bring cause. Numerous research has shown that LLMs exhibit stereotypical biases similar to those already found in other systems relying on distributional representations and, as such, if deployed in a careless manner, pose a risk. While most research in the realm of stereotypical bias has focussed so on unimodal models, i.e. text or vision, little research has been performed on bias existant in multi-modal models deployed for cross-modal tasks. WE FIND XXX, WE SHOW ...

1 Introduction

Research on stereotypes propagated by LLMs at the interface between text and vision appears to be scarce. Given the recent advances in the field of multimodal LLMs (Ruggeri and Nozza, 2023) and the subsequent appearance of multimodal instruction-tuned models such as chatGPT, Gemini, LLaMA3.3-vision, an investigation into the harms they might bring is warranted. The present paper aims to elicit stereotypical biases in multimodal LLMs by *means of probing*. Probing does not require access to model internals (i.e., model parameters) and is therefore suited in black-box settings, where interactions with the model are limited of an API with constraints.

Stereotypical bias A stereotype can be understood as an "over-generalised belief about a particular group of people" (Nadeem et al., 2021, 1). Since LLMs are trained on large corpora of real-data often scraped from the internet, the stereotypical associations found in the "real-world" are also present in LLMs. For a seminal review on bias the reader is referred to Blodgett et al. (2020) and for a more recent discussion to Navigli et al. (2023).

Probing Dataset To probe an LLM some dataset and constructed for an accompanying tasks are required. One such task is presented by Nadeem et al. (2021), who elicit stereotypical bias in pure-text LMs with the **StereoSet**. An extension of StereoSet for multimodal settings ewas introduced by Zhou et al. (2022), called **VLStereoSet**, which consists of stereotypical and anti-stereotypical images, along with captions.

In the present paper we aim to elicit such biases by means of probing. Our main contributions are as follows:

- We aim to elicit the bias present in two mid-sized open-source multi-model LLMs: LLaVA & LLaMA3.2-vision.
- We investigate the bias robustness on these two LLMs with the help of paraphrasing.
- We extend two metrics aiming to capture bias to the case of multiple perturbations of captions in a single data point.

2 Materials and methods

To elicit bias in multimodal instruction-tuned LLMs, we choose the cross-modal task of image-caption matching. Formally, given an image $\mathcal{I}$ along with a set of possible captions $\mathcal{O} = \{S_1, \dots, S_n\}$, the model is tasked to choose the textual caption that most appropriately matches given image.

Dataset We use the VLStereoSet (Zhou et al., 2022) mentioned in the introduction and transform it to fit an image caption matching task in a dialogue-based instruction setting. The intent is to investigate how often a stereotypical caption is chosen under an anti-stereotypical image. The model therefore is presented either with an stereotypical image $\mathcal{I}_{\int}$, or an anti-stereotypical image $\mathcal{I}_{\dashv}$. Each of the images is accompanied with a set

of options $\mathcal{O}$ corresponding to a stereotypical (S_s, an anti-stereotypical (S_a) and an nonsensical (S_n) caption. An example is displayed in ...

Robustness Datasets such as the one used the present study are generated with the help of templates. Using templates for bias elicitation does not account for the richness of how such content can be phrased (Dev et al., 2022). To circumvent this limitation we experiment with paraphrasing methods. That is, the parrot paraphraser Damodaran (2021) and `LLaMA3.3-70b` [1]. After an *ad-hoc* evaluation, we noticed that LLaMA offers more fine-grained control over how a paraphrase is generated (through prompting) and appears to generate more diverse outputs than the parrot paraphraser, which is based on T5 generated outputs which are ranked and "accepted" as viable solutions by further models, such as RoBERTa. With LLaMA, we experiment with multiple prompts and opt for a two-step solution. First we instruct the model to generate three variants paraphrasing only the most "pertinent" noun-phrase in the caption provided. A brief qualitative manual evaluation shows that the generated variations appear viable both syntactically and semantically without altering the content too much.

Evaluation metrics When evaluating the stereotypical bias of language models, papers usually try to capture the association strength between a target (the individuals / group) which in combination with an attribute-term evokes a certain stereotype. Historically in the case of distributional representation, the undertaking of quantifying such associations has been measured throuhg so-called "association tests". Caliskan et al. (2017) for example introduce the Word-embedding association test (WEAT) to capture associative biases in static word embeddings such as Word2vec and GloVe.

3 Results

LLaVA

LLaMA

Paraphrasing Since we instructed the paraphrasing-LLM to replace the *most pertinent* noun-phrase with an appropriate paraphrase, there is a certain amount of arbitrarity in the selection of the noun phrase (unless there is only a single-one, in which case it is mostly the stereotypical target), such as e.g. in *My sister is caring*. In a caption such as *The chess player was a genious person* however there are two possible phrases that can be replaced. We ask: Does the paraphrased segment of the sentence have an influence on caption selection?

4 Discussion

5 Conclusions and further work

6 Limitations

[1] Information available at: `https://www.llama.com/`

References

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Prithiviraj Damodaran. 2021. Parrot: Paraphrase generation for NLU. Version Number: v1.0.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On Measures of Biases and Harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality*, 15(2):1–21.

Gabriele Ruggeri and Debora Nozza. 2023. A Multidimensional study on Bias in Vision-Language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6445–6455, Toronto, Canada. Association for Computational Linguistics.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. VLStereoSet: A Study of Stereotypical Bias in Pretrained Vision-Language Models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and*

the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 527–538, Online only. Association for Computational Linguistics.

A Appendix: Prompts