All Humans are Human, but Some Ethnicities are More Ethnic than Others - Measuring Model Reporting Bias with Regards to Marginalized Ethnicities

Tom Södahl Bladsjö
Course: LT2318

Abstract

Human reporting bias is the tendency of humans to only mention information that they consider relevant or surprising, while omitting things that are considered obvious. While some work has been done on the effects of human reporting bias on the common sense knowledge models can obtain from training on text, no work has (to my knowledge) been done on the effects of human reporting bias on how models talk about marginalized social groups. This study aims to begin to fill this gap by investigating how two image captioning models (BLIP for Conditional Generation and ViT-GPT2) perform on the MMERB dataset (Södahl Bladsjö, 2024). It is demonstrated that both models are significantly more surprised to see ethnicity mentioned in captions for images depicting white people than for images depicting non-white people. This supports the hypothesis that the models encode a perspective where whiteness is considered the default for humans.

1 Introduction

It is sometimes said that a picture is worth a thousand words; a single image can contain more information than can easily be expressed in a single sentence, or even in multiple sentences. When describing an image, one needs to choose what information is relevant enough to be included and what can be omitted. There are often countless different, but equally valid, ways of describing a single image, both in terms of what information to include and how to present it. This makes image captioning a task that falls on the far subjective side of the *intersubjectivity spectrum* (Basile et al., 2023); inter-annotator agreement is likely to be very low, not because of annotator mistakes but because of the inherently subjective nature of the task. Different annotators are likely make different judgements about what information is relevant to include, and this will to some extent depend on their own experiences; what they consider important or surprising

(a) (b)

Figure 1: Two images of bananas.

in the context, and what they think can be assumed without explicit mention. This is an example of human reporting bias, described by some as "the tendency of people to not state the obvious" (Paik et al., 2021). For example, when asked to describe the image in Figure 1b, many people would likely mention that the bananas are brown. On the other hand, few people would deem it necessary to mention that the bananas in Figure 1a are yellow, since that is the color we expect and prefer bananas to be. When it comes to images of people, however, this subconscious filtering of information can manifest in more problematic ways. Consider the images and corresponding captions in Figure 2, taken from the Flickr8k dataset (Hodosh et al., 2013). Similar to the example with the bananas, the annotator who described the girl in Figure 2a considered it unnecessary to mention her skin color or perceived ethnicity, while the annotator who described the girl in Figure 2b considered her ethnicity to be relevant or unexpected enough to be worth mentioning. Just like yellow is the expected color for bananas, white seems to be the expected color for people in available image captioning datasets like Flickr8k and MS COCO (Södahl Bladsjö, 2024, (my course paper for LT2314)). The aim of this project is to investigate how this affects the representations learned by models trained on these and similar datasets. Specifically, I want to find out:

1. Regarding race or ethnicity, are there between-group differences when it comes to the

(a) A little girl in a pink dress going into a wooden cabin.

(b) An Asian girl in a pink dress is smiling whilst out in the countryside.

Figure 2: Two examples from the Flickr8k dataset.

amount and type of information a model includes in descriptions of humans in images?

2. Based on these differences, which group characteristic do models tend to consider "default" (that is, obvious and therefore unnecessary to mention)?

2 Background and Related Work

Human reporting bias as a phenomenon is not unique to descriptions of images, but pertains to most human language use. In Gricean pragmatics it can be explained in terms of the Maxims of Quantity (Grice, 1975):

1. Make your contribution as informative as is required (for the current purposes of the exchange).

2. Do not make your contribution more informative than is required.

In Neo-Gricean frameworks, the same phenomenon can be described in terms of the Q- and R-principles (Horn, 1984). In short, humans, when communicating with other humans, try to make their communication as efficient as possible. They achieve this by including only the information that is relevant, and excluding information that can be implicitly assumed or deduced from context. When training models on language data produced by humans, this phenomenon can cause problems. In the words of Gordon and Durme (2013),

> Much work in knowledge extraction from text tacitly assumes that the frequency with which people write about actions, outcomes, or properties is a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals.
> (Gordon and Durme, 2013)

In reality, models trained on text are likely to end up with a distorted representation of the world. For example, Paik et al. (2021) find that the distribution that language models learn of the colors of objects correlates more strongly with the distribution reported in text than with the actual distribution in the world. Other work on human reporting bias includes Misra et al. (2016), who note that "what's in the image" is not necessarily the same as "what's worth saying", and propose a method from decoupling the human reporting bias from what is actually present in the image. Hagström and Johansson (2022) investigate whether multimodal learning can help mitigate the effects of reporting bias in text, but find no clear evidence that it does.

Much existing work on bias related to social inequalities uses measures based on the Implicit Association Test (IAT) (Greenwald et al., 1998) used in psychology (see e.g. Caliskan et al., 2017; May et al., 2019). IAT is used to measure subconscious associations between certain concepts. For example, a recent study by Morehouse et al. (2023) found that white participants consistently associated *Human* (as opposed to *Animal*) more closely with their own group (white) than with other groups, while Black, Latinx, and Asian participants showed no such own-group=Human bias. IAT-based tests in NLP usually focus on associations between different social groups and pleasantness/unpleasantness, or associations between social groups and stereotypes. In contrast, this work focuses on a different type of bias – not what stereotypes of different ethnic groups are encoded in a model, but rather what the model has learned

to consider "the default human". I believe that insights into this will be useful for the general discussion of bias in the NLP community, by drawing attention to whose perspective is being represented in language models.

3 Experiment

3.1 Data

The MMERB dataset (created by me for a project in the course LT2314) consists of 1313 images, of which 633 depict people interpreted as white and 680 depict people interpreted as non-white. Each image comes with two contrasting captions; one where the ethnicity of the person (or persons) in the image is mentioned, and one where it is not. Apart from the presence or absence of the ethnicity word, the two captions in each pair are identical. When testing a model on this dataset, each image will be viewed twice (once with each version of the caption), resulting in a total of 2626 examples. More details, as well as download- and usage instructions can be found on Github[1].

3.2 Models

For this project I tested two publicly available image captioning models: BLIP for conditional generation[2], and ViT-GPT2[3]. Both are generative transformer models that combine an image encoder module with a text decoder module.

3.3 Method

In order to test if there are between-group differences regarding how likely a model is to mention ethnicity in a caption, each model was fed the examples from each of the four subsets of MMERB: test_mention (680 instances), test_no_mention (680 instances), norm_mention (633 instances) and norm_no_mention (633 instances). The normalized sequence perplexity was calculated as the exponentiated average negative log likelihood, or

$$ppl = exp\left\{-\frac{1}{t}\sum_{i}^{t}\log p(x_i|x_0, x_1, ..., x_{i-1})\right\}$$

where t is the length of the sequence and $\log p(x_i|x_0, x_1, ..., x_{i-1})$ is the model's predicted log probability of word i given the preceding words in the sequence. The difference in perplexity between the two captions for each image was then calculated by subtracting the perplexity of the caption then mentions ethnicity from the one that does not. This resulted in two sets of pairwise differences; one for the norm group (images of white people) and one for the test group (images of non-white people). Welch's t-test was performed on the two groups to determine if their means differed significantly – that is, if the model was more surprised to see a white person described as white than to see a non-white person described with ethnicity. A positive t-statistic means that the model was on average more surprised to see white people described as white than to see non-white people described with ethnicity, while a negative t-statistic would mean the opposite. A t-statistic of 0 would mean that there was no difference in group means.

3.4 Results

Model	t-statistic	p-value	DF
BLIP	11.4	1e-28	1226.4
ViT-GPT2	6.3	3e-10	1030.6

Table 1: T-test results for both models.

As can be seen in Table 1, the results are statistically significant ($p < 0.01$) for both models. The t-statistic is somewhat lower for the ViT-GPT2 model, suggesting that it is slightly less biased than BLIP. However, since the two models use different tokenization schemes, the resulting perplexities are not directly comparable. The distribution of perplexities for both models shown in Figure 3. As can be seen in the plots, the models are generally more surprised to see descriptions of people mentioning ethnicity, regardless of group. But the difference is greater for the norm group than for the test group, that is, models are more surprised to see white people described with ethnicity than to see non-white people described with ethnicity.

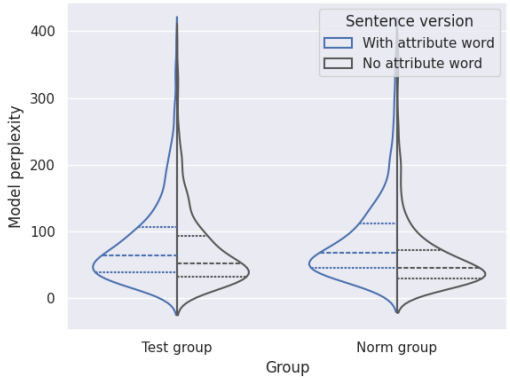
4 Discussion

Clearly, both models seem to consider white to be the default for humans. This in itself is not very surprising, but it draws attention to some limitations on the "objectivity" or "generality" that can be achieved in language modelling. Language does not exist in a vacuum. It is produced in specific

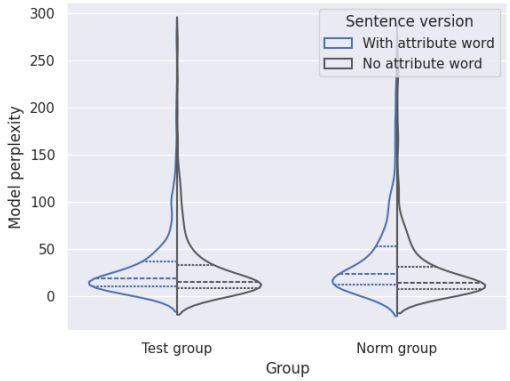
[1] https://github.com/TomBladsjo/LT-Resources-project

[2] https://huggingface.co/docs/transformers/model_doc/blip

[3] https://huggingface.co/nlpconnect/vit-gpt2-image-captioning



(a) BLIP perplexities.



(b) ViT-GPT2 perplexities.

Figure 3

situations, by people who are part of a society, and usually with some hearer/reader in mind. As such, language is situated not only in a material context, but also in a social context, and pieces of language will, unavoidably, reflect the situations in which they were produced and the people who produced them. In this respect, a language model differs from a human in that it is not expected to be socially situated – how could it be, when it does not have a "self", or any type of lived experience? However, the language that models have been trained on was all produced in specific situations, by specific people. The resulting model is thus not independent from any social perspective, but rather an amalgamation of multiple perspectives, where the perspective that is most common in the training data wins out over less commonly represented perspectives at inference time. The results of this experiment demonstrate that the tested models are in fact socially situated in the sense that they describe images from a specific perspective, namely that of a white person who interacts mostly with other white people.

Many papers about bias in language models follow a structure where they first demonstrate that the model is biased, and then perform some kind of debiasing and show that the bias was reduced. I have chosen to follow a different structure in this project. My main reason is that I believe that the kind of over-reporting of group membership that this method measures can be one of several symptoms of a more general skewness in the data the model has been trained on. As such, it can be a useful indication of other, possibly more harmful, biases in a model. Gonen and Goldberg (2019) showed that popular debiasing methods cover up systematic biases but actually fail to remove them. Similarly, I believe that trying to remove the type of reporting bias investigated in this study without properly understanding its origins and relationship to other potentially harmful model behaviors would be rendering useless a possible diagnostic tool while failing to treat the underlying problems. Furthermore, Goldfarb-Tarrant et al. (2021) show that measures of bias within a model do not necessarily correlate with biased behavior in downstream tasks. It is clear that the issue of bias in models is more complicated than simply debiasing and considering the problem solved. Therefore, I believe that the phenomenon itself and its relationship to other aspects of model behavior should be thoroughly investigated before any attempts are made to remove it. I hope that work in this area can serve as a starting point for a more thorough discussion about whose language we are actually aiming to model when we model language, and whose language we end up modelling.

5 Conclusion and Future Work

This work has shown that there are clear and statistically significant between-group differences in how likely the tested models are to mention the ethnicity of a person in an image, and that both models clearly consider white to be the default color for humans. Future work in this area should aim to test a wider variety of models, and expand the investigation to include other marginalized attributes apart from ethnicity (such as gender, queerness, disability and religion, to name a few). It would also be interesting to look into the effects of intersectionality on this phenomenon; for example, will the likelihood of the model mentioning ethnicity differ depending on the gender of the person depicted, or

the presence of some other marginalized attribute? Last but not least, future work should investigate the relationship between this and other types of bias in a model, and to what extent they each correlate with problematic behavior in downstream tasks.

Limitations

This study has a number of limitations. One is that perplexity is not fully comparable across models with different tokenization schemes and vocabulary sizes. This means that while the results are consistent and valid within a single model (a t-statistic above 0 is a clear indication of bias in a model), the magnitude of the bias is not necessarily comparable between different models. In the future I would like to look into methods and metrics that are more robust when it comes to comparisons between different models. Another limitation is that only two models were tested; if one wants a more comprehensive indication of the state of available image captioning models, more should be included in future experiments.

Ethics Statement

This study deals with sensitive data, in that the dataset contains images of people grouped by ethnicity (in this case, white or non-white). Furthermore, the ethnicity labels are attributed to the people in the images by annotators based on their own visual judgement and with no access to information about how the people depicted in the images identify. This means that there is a considerable risk that this dataset contains incorrect labels. Another ethical concern is that the concept of ethnicity used in this study is insufficiently defined and theorized. If this work, or some variation of it, were ever to be published, the concept of ethnicity would need to be clearly grounded in existing theory, as well as further problematized and nuanced.

References

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2023. Toward a perspectivist turn in ground truthing for predictive computing.

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

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Durme. 2013. Reporting bias and knowledge acquisition. *AKBC 2013 - Proceedings of the 2013 Workshop on Automated Knowledge Base Construction, Co-located with CIKM 2013*, pages 25–30.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74 6:1464–80.

H. P. Grice. 1975. *Logic and Conversation*, pages 41 – 58. Brill, Leiden, The Netherlands.

Lovisa Hagström and Richard Johansson. 2022. What do models learn from training on more than text? measuring visual commonsense knowledge. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 252–261, Dublin, Ireland. Association for Computational Linguistics.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899.

Laurence R. Horn. 1984. Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. In *Cognitive Linguistics Bibliography (CogBib)*, Washington,.D.C. Georgetown University Press. 2010.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels.

Kirsten N. Morehouse, Keith Maddox, and Mahzarin R. Banaji. 2023. All human social groups are human, but some are more human than others: A

comprehensive investigation of the implicit association of "human" to us racial/ethnic groups. *Proceedings of the National Academy of Sciences*, 120(22):e2300995120.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 823–835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Södahl Bladsjö. 2024. MMERB: A multimodal dataset for measuring marginalized ethnicity reporting bias. Course paper for LT2314.