RACIAL AND GENDER STEREOTYPES ENCODED INTO CLIP REPRESENTATIONS

Vatsal Baherwani & Joseph Vincent
University of Maryland, College Park
College Park, MD 20742, USA
{vatsalb, jvincen3}@umd.edu

ABSTRACT

OpenAI's CLIP (Radford et al., 2021) is a vision language model widely used in current state-of-the-art architectures. This paper analyzes racial and gender biases present in CLIP's representation of human images. We evaluate images from the FairFace dataset (Kärkkäinen & Joo, 2019) grouped by race and gender on a series of traits describing demeanor, intelligence, and character. We find CLIP's understanding of these traits to be heavily influenced by race and gender, suggesting that this social bias propagates into many other architectures.

1 INTRODUCTION

The development of multimodal vision language models such as CLIP has sparked concern over fairness and the social bias present in these models (Lee et al., 2023). CLIP is used by generative models like Stable Diffusion (Rombach et al., 2021) and DALL-E (Ramesh et al., 2021), and the widespread availability of these models to millions of users has led to the reinforcement of Western stereotypes in generated content (Bianchi et al., 2023). Specifically, attributes such as race and gender have been shown to heavily influence the generated images from Stable Diffusion to reflect social biases (Howard et al., 2023). We attempt to identify the source of these biases by examining CLIP's inherent understanding of gender and race through its text embeddings, and we find that CLIP itself exhibits the biases which propagate into downstream models.

2 METHODS

We use the FairFace dataset, which labels each image with the person's age, race, and gender. The race is labeled as one of the following: East Asian, Latino/Hispanic, Southeast Asian, Indian, Black, White, and Middle Eastern, and the gender is labeled as male or female. We focus our experiments on the subset of adults aged 20-60 years old. The dataset is then grouped by all combinations of gender and race (e.g. Black males). We draw a random sample of size $n = 2000$ images for each of these groups and retrieve each image's CLIP embedding (using the CLIP ViT-B/32 architecture).

Each image is evaluated on six pairs of opposing traits: smart vs. dumb, happy vs. sad, hardworking vs. lazy, nice vs. mean, dominant vs. agentic, and honest vs. dishonest. We create two caption embeddings for each pair, specifying one trait from the pair (e.g. "a nice person" and "a mean person"). Using these embeddings, CLIP gives us a confidence level in predicting the positive trait versus the negative one. This confidence relies on the similarity between the embeddings (see A.1 for details). For each gender-race sample, we collect the average confidence in each of the four positive traits over all images to get four mean confidence levels. The mean confidence across each race is shown in Figure 1, with a bar plot for both male and female samples. We justify the statistical significance of the disparities between the mean confidence in each trait across racial groups using F-tests (see A.2).

We use the GradCAM library (Gildenblat, 2021) to generate class activation maps ("CAMs") given an input image and a trait pair. The CAM highlights portions of an image whose corresponding activations contributed the most to the classification. As seen in Figure 1, Indian men are disproportionately more likely to be predicted as smart. We analyze attributes that contribute to this disparity

by choosing five high-confidence inputs from the sample, and the results suggest that this distinction is not spurious (see A.3).

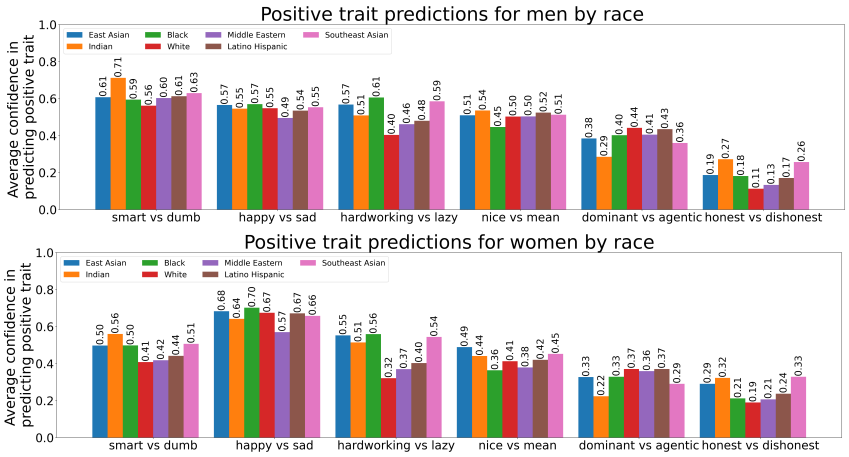

Figure 1: Average confidence splits for predicting each trait across the $n = 2000$ sample for each race category. Each bar value is the confidence of predicting the positive trait over the negative trait (based on the similarities between the CLIP embedding of the image and of the captions for each trait), averaged over all images in the sample.

3 Results & Analysis

In our experiments, we observe that for each chosen trait there are significant differences across race and gender. In Figure 1, men are more likely to be classified as smart than women across all racial groups. This reflects the "gender-brilliance stereotype" which contributes to the underrepresentation of women in science and technology (Storage et al., 2020). For all races, men are also perceived as more dominant than women, purporting the long-lasting belief that men should be assertive and risk-taking while women should be nurturing and caring as seen in Weisberg et al. (2011).

We also see that Indian men are more likely to be classified as smart than White men. Likewise, Indian and East Asian women are more likely to be classified as smart than women of other races. Both Asian and Indian men and women have stronger predictions for being hardworking than their White counterparts. These results align with the model minority stereotype that sets unrealistic expectations for Asian-Americans (Thompson et al., 2016), suggesting that these biases are incorporated into CLIP's understanding of race as well. Moreover, the CAM results in A.3 suggest that CLIP resorts to racial features for predicting this trait when there are no other relevant attributes. Black men are classified as meaner than all other races and this trend follows for women, reflecting a sentiment found often in news media in which Black men and women are perceived as dangerous or violent (Oliver, 2003). These results and the statistical significance of the differences shown in A.2 indicate that there is indeed gender and racial bias in CLIP's embeddings.

4 Conclusion

In an unbiased vision-language model, we expect the average confidence of predicting each trait to be similar across all races and genders. However, we determine that there are statistically significant differences in CLIP's perceptions of traits across these groups; these predictions propagate Western stereotypes, such as those of the "model minority", black aggression, and "gender brilliance", to generative models like Stable Diffusion. Thus, we demonstrate that one source of the bias observed in generative models is within the text embeddings from CLIP. Many other factors could contribute to this, including dataset bias (e.g. more Indians wearing glasses in FairFace, causing them to be perceived as smarter), or bias specific to CLIP's image and text encoder. Nonetheless, there is more work to be done in finding the cause of social bias in CLIP, and future research should look to mitigate these disparities to promote equity and fairness in vision language models.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504. Association for Computing Machinery, 2023. doi: 10.1145/3593013.3594095.

Jacob Gildenblat. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, and Vasudev Lal. Probing intersectional biases in vision-language models with counterfactual examples, 2023.

Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age, 2019.

Nayeon Lee, Yejin Bang, Holy Lovenia, Samuel Cahyawijaya, Wenliang Dai, and Pascale Fung. Survey of social bias in vision-language models, 2023.

Mary Beth Oliver. African american men as "criminal and dangerous": Implications of media portrayals of crime on the "criminalization" of african american men. *Journal of African American Studies*, 7(2):3–18, 2003. doi: 10.1007/s12111-003-1006-5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Daniel Storage, Tessa E.S. Charlesworth, Mahzarin R. Banaji, and Andrei Cimpian. Adults and children implicitly associate brilliance with men more than women. *Journal of Experimental Social Psychology*, 90:104020, 2020. doi: 10.1016/j.jesp.2020.104020.

Taylor L. Thompson, Lisa Kiang, and Melissa R. Witkow. You're asian; you're supposed to be smart: Adolescents' experiences with the model minority stereotype and longitudinal links with identity. *Asian American Journal of Psychology*, 7(2):108–119, 2016. doi: 10.1037/aap0000038.

Yanna J. Weisberg, Colin G. DeYoung, and Jacob B. Hirsh. Gender differences in personality across the ten aspects of the big five. *Frontiers in Psychology*, 2, 2011. doi: 10.3389/fpsyg.2011.00178.

A APPENDIX

A.1 CLIP CONFIDENCE SCORES

CLIP consists of an image and text encoder that each produce embeddings given inputs of the respective data type. Given an image and text caption, CLIP computes embedding vectors $\vec{v}_{image}$ and $\vec{v}_{text}$. The CLIP similarity between the text and image is simply the cosine similarity:

$$s = \frac{\vec{v}_{image} \cdot \vec{v}_{text}}{\|\vec{v}_{image}\|\|\vec{v}_{text}\|}$$

These CLIP similarities can be used as the softmax logits for classification. Given an image and two text captions, we can calculate the similarities between the image and each of the captions s_0, s_1 using the cosine similarity shown above. Then, the probability of predicting the image to correspond to caption 0 (versus caption 1) is the softmax probability $\frac{e^{s_0}}{e^{s_0}+e^{s_1}}$. In the context of our experiments, we would choose a trait pair (e.g. happy vs. sad), encode these into captions such as "a happy person" and "a sad person", and generate CLIP embeddings for both captions along with the image. Then, we calculate similarity between each caption and the image as shown above, and use that to compute the probability of CLIP predicting an image to depict a happy person (as oppposed to a sad person). The averages of these probabilities across race and gender groups are reflected in Figure 1.

A.2 Statistical Significance

We use F-tests to judge whether the mean confidence in predicting each of the six traits is consistent across races and genders. Given a sample divided into I groups, the F-test is used to gauge whether the true population means of a variable across all I groups is the same. In our case, the variable is CLIP confidence in predicting a trait, and we are testing whether that is consistent across groups split by either gender or race. Let I be the number of groups, J the sample size for each group, $\bar{X}_i$ the variable's mean value for group i, $\bar{X}$ the variable's mean value across all groups, and S_i the sample variance for group i. The F-statistic, following an F-distribution is given by $f = MST/MSE$ where $MST = \frac{J}{I-1}\sum_{i=1}^{I}(\bar{X}_i - \bar{X})^2$ and $MSE = \frac{1}{I}\sum_{i=1}^{I} S_i^2$.

Intuitively, MST represents the variance explained by deviations in the data, while MSE represents unexplainable noise/variance. A relatively high explainable variance then suggests that the true means are not in fact equal across all groups. The p-value for this F-test is then $P(F \geq f)$ where F is the F-distribution. A lower p-value suggests that disparities in the data are less likely to have occured by chance, and there is a statistically significant difference across groups. We use the Python scipy library to perform the calculations necessary for F-tests in this paper.

Table 1: F-statistics of difference between races in mean confidence for each trait (by gender)

Trait	Male	Female
smart vs dumb	141.48	194.05
happy vs sad	14.40	40.23
hardworking vs lazy	325.40	522.60
nice vs mean	33.20	77.26
dominant vs agentic	98.75	163.20
honest vs dishonest	415.36	286.50

Table 2: P-values corresponding to difference in mean confidence for each trait (by gender)

Trait	Male	Female
smart vs dumb	8.88e-175	2.00e-238
happy vs sad	1.92e-16	7.80e-49
hardworking vs lazy	0.00e+00	0.00e+00
nice vs mean	5.56e-40	2.40e-95
dominant vs agentic	3.92e-122	3.20e-201
honest vs dishonest	0.00e+00	0.00e+00

Table 3: F-statistics of difference between genders in mean confidence for each trait (by race)

Trait	East Asian	Indian	Black	White	Middle Eastern	Latino Hispanic	Southeast Asian
smart vs dumb	336.64	749.31	396.48	742.69	1181.74	911.44	446.16
happy vs sad	161.90	105.20	200.66	182.49	64.16	207.08	122.58
hardworking vs lazy	6.43	0.72	59.91	217.92	252.50	172.34	45.92
nice vs mean	8.83	189.88	136.65	170.90	355.08	220.08	72.70
dominant vs agentic	70.04	107.66	112.28	97.55	41.84	83.90	120.35
honest vs dishonest	515.47	82.84	60.56	536.31	379.57	239.56	176.19

Table 4: P-values corresponding to difference in mean confidence for each trait (by race)

Trait	East Asian	Indian	Black	White	Middle Eastern	Latino Hispanic	Southeast Asian
smart vs dumb	2.9e-72	2.3e-151	3.4e-84	3.8e-150	4.0e-227	1.5e-180	5.6e-94
happy vs sad	2.2e-36	2.2e-24	1.8e-44	1.1e-40	1.5e-15	8.1e-46	4.4e-28
hardworking vs lazy	1.1e-02	4.0e-01	1.2e-14	4.6e-48	3.5e-55	1.4e-38	1.4e-11
nice vs mean	3.0e-03	3.1e-42	4.6e-31	2.8e-38	5.9e-76	1.7e-48	2.1e-17
dominant vs agentic	7.9e-17	6.6e-25	6.8e-26	9.5e-23	1.1e-10	8.1e-20	1.3e-27
honest vs dishonest	1.9e-107	1.4e-19	9.0e-15	1.9e-111	7.8e-81	1.6e-52	2.2e-39

Table 1 displays the F-statistics when we conduct a test to judge whether the true mean confidence in classifying each trait varies across different race samples. These tests are done for each gender and each trait. The corresponding p-values are in Table 2. For example, the third row in the male column shows a very large F-statistic and a p-value computationally equivalent to 0, suggesting there is a statistically significant disparity in confidence when classifying men of different races as hardworking or lazy.

Table 3 displays F-statistics for two-sample F-tests comparing the mean confidence of classifying a trait with one sample of men and one sample of women of the same race. The corresponding p-values are in Table 4. Some values suggest a low statistical significance; for example, we cannot claim that Indian men are more or less likely than Indian women to be classified as hardworking vs. lazy based on the low f-statistic of 0.72 and high p-value of 0.4. However, many statistically significant gender disparities are present across all races for certain traits, e.g. smart vs. dumb and dominant vs. agentic. The wide range of values presented in this table suggests that some traits carry more significant bias than others and thus lead to stronger disparities in the average CLIP confidence in predicting them.

A.3 GradCAM Class Activation Maps

Figure 2: Random sample of Indian men from the FairFace dataset with a high predicted confidence ($> 66\%$) of being smart (as opposed to silly), along with their respective class activation maps. The corresponding confidence levels from left to right are $68\%, 69\%, 73\%, 72\%, 74\%$. While the second, third, and fifth images highlight attributes like glasses and suits used for classification, the first and fourth use skin and hair in the absence of those predictors.

We analyze class activation maps through the GradCAM library (Gildenblat, 2021) to probe for features that heavily influence classification into certain traits. In Figure 2 we sample high confidence inputs of Indian men that were classified as smart. The GradCAM images allow us to notice both racial and non racial factors that contribute to high confidence predictions. Specifically, they suggest that the differences in CLIP confidence for classifying traits across different races is not spurious and indeed influenced by racial features. We see this in the inputs whose class activation maps highlight the skin and hair of the person. However, we also see attributes like glasses and formal shirts influencing the classification. This makes CLIP potentially subject to dataset bias, as a higher volume of Indian men with glasses in the training dataset could lead CLIP to spuriously associate Indian men as being smarter.