

Investigation of Visual Bias in Generative AI

Jerome Agius

Supervisor: Dylan Seychell,

Co-Supervisor: John Abela

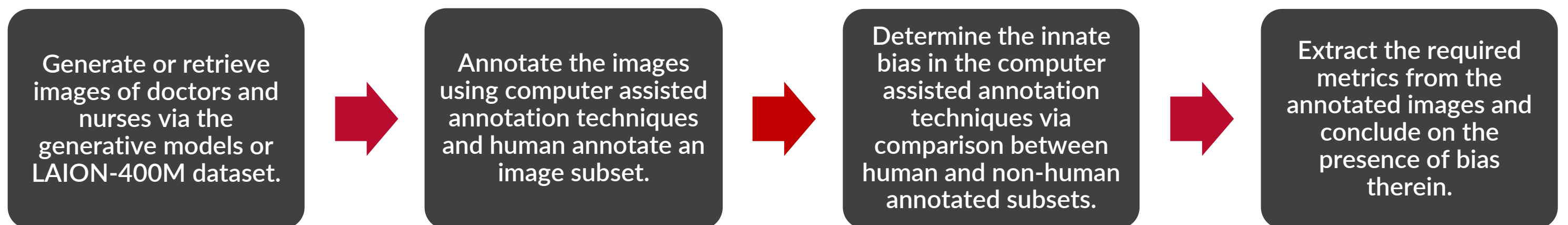
INTRODUCTION

Recent advancements in Generative AI have revolutionised visual content generation, particularly with regard to images. Models such as Stable Diffusion, Dall-E, and Midjourney have been at the forefront of this progress, facilitating the generation of high-quality, diverse images through simple text prompts. However, this progress has brought to light critical issues such as a lack of control over generated outputs, overfitting, privacy, ethical concerns, and bias [1, 2]. The latter serves as the focus of this research, particularly how it manifests within these generative models and its severity. Bias, particularly gender and racial, has led to detrimental consequences across various domains, from recidivism and credit scoring to online advertisement [3-5]. This study delved into the pervasive issue of bias within generative AI systems, aiming to identify biases present in the models and the LAION-400M training dataset, while outlining any bias mitigation techniques employed by such models.

AIM

In accordance with the introduction, the main aim of this research paper was to identify the types of bias present within the models and training dataset, with a particular focus on gender, racial, age, and prominence biases. This was achieved through the generation of innately biased images, particularly those of doctors and nurses, in conjunction with qualitative analysis of the appropriate metrics, leading to the identification of prominent bias forms, mitigation measures implemented, and the creation of a simple Python pipeline by which this can be replicated.

METHODOLOGY



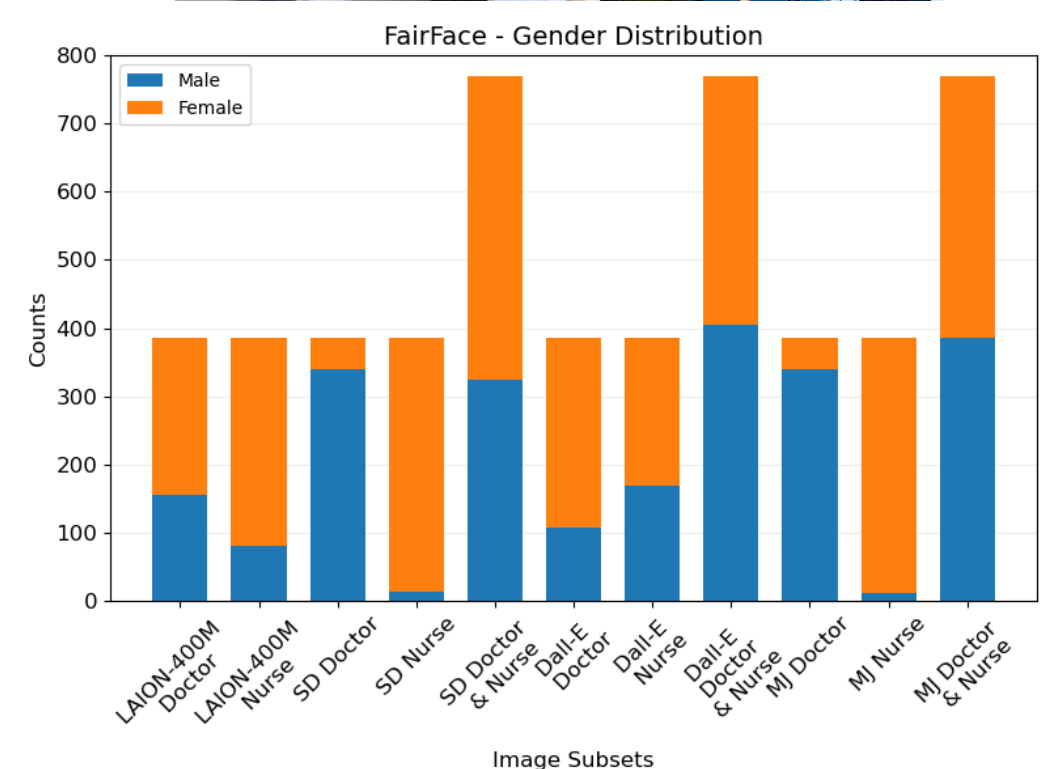
RESULTS

After extracting and interpreting the metrics according to the methodology pipeline, biases within both the dataset and models became evident. Firstly, in real-world data, doctors were predominantly young white males, while nurses were young white females – a crucial observation considering these models are trained on such data. Regarding the LAION-400M dataset, sourced from the common crawl, its inherent biases were evident as it depicted a far greater proportion of doctors younger than 55, with the dominant gender being female, although having similar racial demographics to real-world metrics. Similarly, nurse depictions consisted of individuals younger than 55; however, there was a lack of gender and race bias compared to real-world metrics. Observing the generative models, Stable Diffusion depicted severe gender, racial, and age bias, aligning with and exceeding real-world metrics. The prominence metrics denoted equal prominence among genders, whereas racially, Asians appeared to be more prominent overall; however, these results are not conclusive given minimal Asian depictions. Similarly, Midjourney depicted severe gender and age bias, exceeding real-world bias. Conversely, racial bias was reduced, with white remaining as the dominant race but with drastically reduced representation. Furthermore, prominence metrics depict an even prominence among genders and races. Finally, Dall-E produced the most promising results, having mostly balanced gender and racial depictions, with the latter being slightly biased towards Asian and Indian. Conversely, all depictions were younger than 55, denoting severe age bias. The prominence metrics depict an even distribution among gender and race, with some races being marginally more prominent. In relation to this, images depicting both doctors and nurses simultaneously were processed; however, they provided no additional insight.

CONCLUSIONS AND FUTURE WORK

In line with the achieved results, it was evident that Dall-E was the best model given its diverse representation and minimal bias. This lack of bias led to the discovery of its anti-bias measure, this being a prompt enhancing model which rewrites the initial prompt for safety reasons and to achieve higher quality images. Furthermore, the results indicated that amongst some of models steps are being taken to reduce gender and race bias, however this does not appear to be the case for age bias seeing as it was abundant across all three models. This research can be expanded upon by delving deeper into different forms of bias, studying alternate bias reduction measures similar to that applied in Dall-E and the creation of a small scale non-biased generative model by which bias reduction techniques can be further studied.

EXAMPLE IMAGES & RESULTS



REFERENCES

1. M. Lee and J. Seok, "Controllable generative adversarial network," IEEE Access, vol. 7, pp. 28 158–28 169, 2019. doi: 10.1109/ACCESS.2019.2899108.
2. A. Sudhir Bale et al., "The impact of generative content on individuals privacy and ethical concerns," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 1s, pp. 697–703, Sep. 2023. [Online]. Available: <https://www.ijisae.org/index.php/IJISAE/article/view/3503>
3. J. Angwin, J. Larson, S. Mattu, and L. Kirchner. "Machine bias." Accessed on 29 October 2023, ProPublica. (2016), [Online]. Available: <https://www.propublica.org/article/machine-bias-risk-assessmentsin-criminal-sentencing>.
4. R. Bartlett, A. Morse, R. Stanton, and N. Wallace, "Consumer-lending discrimination in the fintech era," National Bureau of Economic Research, Working Paper 25943, 2019. doi: 10.3386/w25943. [Online]. Available: <http://www.nber.org/papers/w25943>.
5. L. Sweeney, "Discrimination in online ad delivery," Commun. ACM, vol. 56, no. 5, pp. 44–54, May 2013, issn: 0001-0782. doi: 10.1145/2447976.2447990. [Online]. Available: <https://doi.org/10.1145/2447976.2447990>.