L-Università ta' Malta
Faculty of Information &
Communication Technology

Department
of Artificial
Intelligence

**B.Sc. IT (Hons.)
Artificial Intelligence**

Investigation of Visual Bias in Generative AI

Jerome Agius
Supervisor: Dylan Seychell,
Co-Supervisor: John Abela

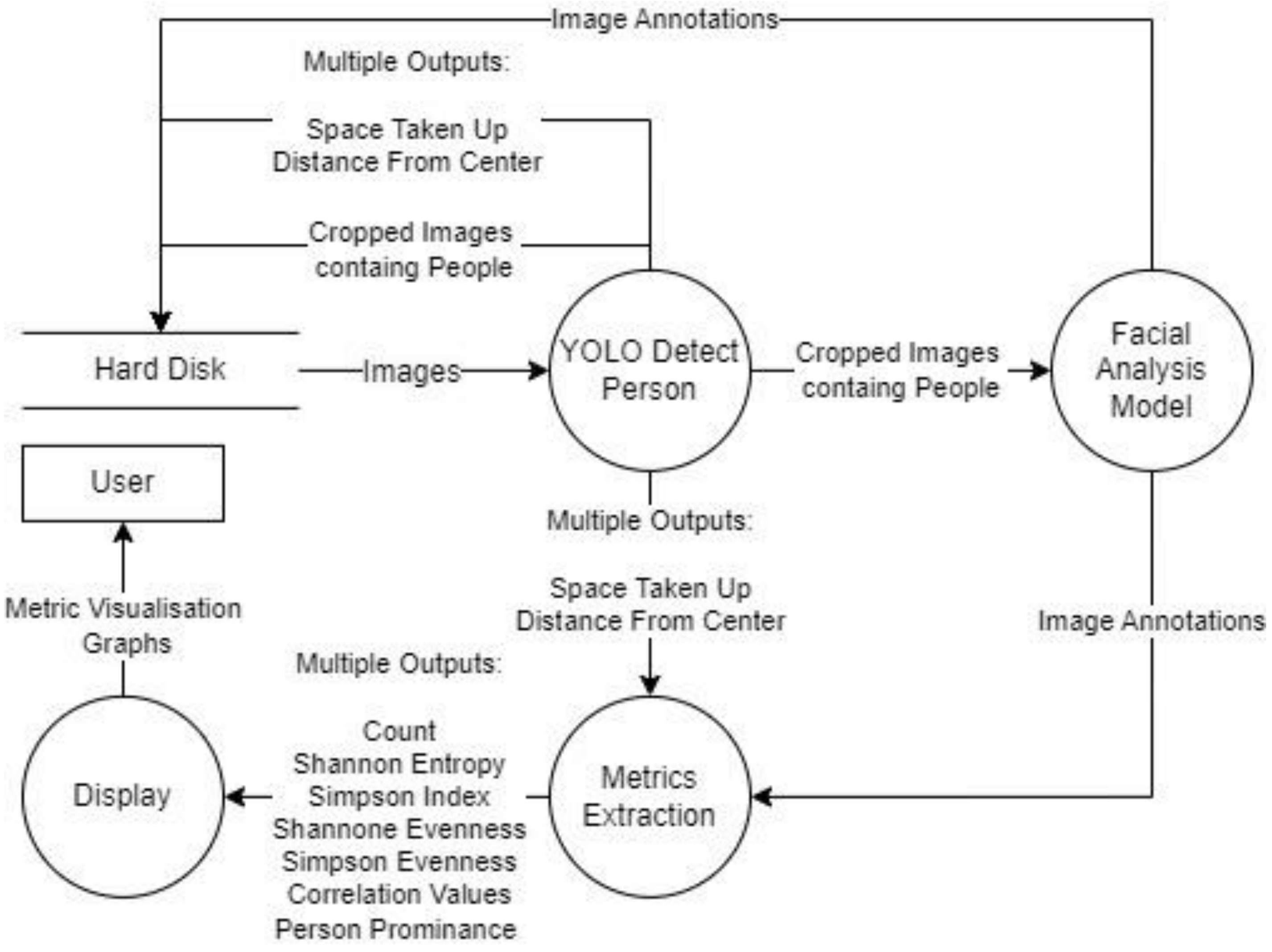
INTRODUCTION

Recent advancements in Generative AI have revolutionized visual content generation, particularly when it comes 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 serving as the focus of this research particularly how it presents itself within these generative models and the severity therein. 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 whilst outlining any bias mitigation techniques employed by such models.

AIM

In accordance with the introduction the main aim for 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. Through the generation of innately biased images in particular 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.

ARCHITECTURE DESIGN



METHODOLOGY

Generate or retrieve images of doctors and nurses via the generative models or LAION-400M dataset.

Annotate the images using computer assisted annotation techniques and human annotate an image subset.

Determine the innate bias in the computer assisted annotation techniques via comparison between human and non-human annotated subsets.

Extract the required metrics from the annotated images and conclude on the presence of bias therein.

RESULTS

Following the extraction of the metrics and their interpretation in line with the methodology pipeline the biases within the dataset and models were made clear. Firstly, it was noted that real-world data is not severely biased with doctors being predominately young white males whereas nurses' young white females, this being crucial information seeing as these models are trained on real-world data. In relation to the LAION-400M dataset which consisted of images from the common crawl it was evident that it contained innate bias seeing as it depicted a far greater degree of doctors younger than 55, whilst having majority female representation and similar racial demographics to real-world metrics. Furthermore, nurses saw most depictions younger than 55 however a reduction in bias when it comes to gender and race when compared to real-world metrics. Observing the generative models, Stable Diffusion depicted a sever gender, racial and age bias aligning and exceeding real-world metrics. The prominence metrics denoted equal prominence amongst genders, whereas racially, Asian appeared to be more prominent overall however these results are not conclusive given minimal Asian depictions. The Midjourney results depict a similar picture with sever gender and age bias, doctors as predominately young males and nurse's young females exceeding real-world metrics. Contrarily racial bias was reduced with white remaining as the dominant race however with a severely reduced representation. Furthermore, prominence metrics depict an even prominence amongst 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. Contrarily all depictions were younger than 55 denoting sever age bias. The prominence metrics depict even distribution amongst 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 results achieved it was clear that Dall-E was the most in line with the requirements of a non-biased generative model, that being that it had minimal to no 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 models steps are being taken to reduce gender and race bias, however this doesn't appear to be the case for age bias seeing as 20-29 was a dominant age amongst all results. 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.

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.