The Impact of Biases in Facial Recognition Artificial Neural Networks

Ezra Wingard
SUNY Oswego JKLab, Cognitive Dynamics Lab, Oswego, New York

Background

- Artificial Neural Networks (NNs) are a product of the fields of Artificial Intelligence (AI) and Machine Learning (ML). The general structure and behavior of these algorithms are based off of the biological processes and structures of human neural networks.
- Research and development on NNs has been growing steadily for years now, however the conversations surrounding ethics of using such technology has been lagging behind.
- It has been demonstrated through many research projects on facial recognition technology that classifications of gender and race on Black women and transgender people are typically much worse than on cisgender white men[2,3,4]. Many studies that have been conducted on the ethics within facial recognition algorithms have often left out transgender people, while including other marginalized groups such as non-white people and people of diverse age groups.

Current Study

- In this study, I aimed to examine bias/prejudice against transgender people and non-white people in CNN facial recognition networks. Specific studies have hypothesized that datasets that heavily include cisgender white men may contribute to biases learned during the training phase.
- A binary gender scale was used within this study. There is a difficulty in quantifying gender & gender expression (masculine to feminine) as a spectrum within a classification algorithm. Thus, binary transgender individuals were used in order to not intentionally misgender non-binary identifying individuals.
- I hypothesized whether a NN trained on cisgender people would be as accurate on binary transgender individuals as the cisgender population.

Methods and Materials

Programming Resources:
- Instaloader Python API[5]
- RStudio + R (programming language) – packages: tidyverse, rstatix, effectsize, ggplot

Facial Recognition CNN Models:
- FairFace[1] pre-trained neural network model trained on an image dataset

Image Datasets:
- Binary transgender dataset compiled from images scraped from Instagram
- Cisgender dataset compiled from images scraped from Instagram

Demographics:
- Number of images in the novel transgender dataset (n = 414) and cisgender dataset (n = 550).
- Transgender men (n = 192, ≈46.38% of dataset, 80 unique faces).
- Transgender women (n = 222, ≈ 53.62% of dataset, 102 unique faces).
- Cisgender men (n = 270, ≈ 49.09% of cis dataset, 170 unique faces).
- Cisgender women (n = 280, ≈50.9 of dataset, 162 unique faces).

Procedure:
- The Instaloader Python API was used to scrape images off of Instagram through searches of hashtags including #FTMtransgender , #MTFtransgender , and other buzzwords typically used within the respective communities to denote their self-identified gender.
- Any image that did not contain a face belonging to a person was removed, along with image with profiles that did not explicitly identify as a binary transgender individual (man vs woman) were removed from the transgender dataset.
- The FairFace NN model was tested on the two above datasets and classified gender from the facial images.
- Post-hoc analyses were conducted in R and google sheets, and are reported below.

Figure 1. Top: 2 images of the author that were used to test transgender classification.
Bottom: Same 2 images that were misclassified on race (East Asian) and age (30-39) and gender (Woman).

Results

$$\text{Accuracy} = \frac{\text{True Positives (Correctly Gendered)+ False Positives (Misgendered)}}{\text{N (Total images in binary transgender test dataset)}}$$

Figure 2. Preliminary accuracy rates were determined from the classification accuracy formula above

Using the accuracy formula above, general accuracy rates were determined for both transgender men and women, as well as an overall accuracy rate. The total gender accuracy rate of the FairFace model on the transgender dataset was 66.67% or $^{2}/_{3}$ and on the cisgender dataset, gender classification accuracy was 95.3%.

A logistic regression was conducted using R to investigate the gender classification accuracy on binary transgender men and women. It was found that gender classification accuracy on transgender men significantly predicted the gender classification accuracy of transgender women. It was found that the odds of a transgender woman being gendered correctly was 3.05 times more likely (95% CI [2.00, 4.69]) than a transgender man being gendered correctly by the neural network.

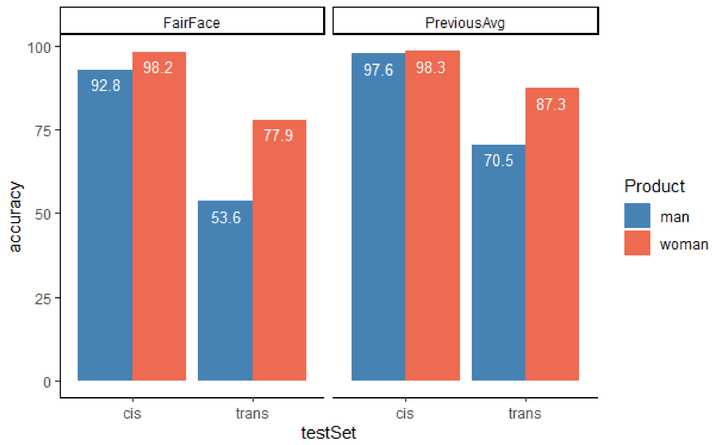

Figure 3. Bar Graph denoting the approximate accuracy rates between the transgender and cisgender datasets from the FairFace model and the previous averages located from the papers by Scheuerman, et al. on similar demographics (transgender vs cisgender men and women)[3, 4.]

Discussion

- This study found that despite these inclusions of race and age, there are still many biases that come about with respect to people whose presentation may not align with the strict, traditional rules of the gender binary.
- When compared to industry model averages for transgender data sets, including those from Amazon, IBM, Microsoft, and Clarifai, FairFace's performance is considered to be substandard[3,4].
- In the future, when considering minority populations that may be most detrimentally affected by biased technology such as neural networks, we need to make a conscious effort to include transgender people and other diverse identities. These results show that when identities such as transgender / gender diversity are not considered, they majorly skew results toward being unethical and unequitable, even when race and age as variables are accounted for.
- The FairFace model has demonstrated comparable performance to previous iterations on cisgender individuals, such as achieving a gender classification accuracy of 95.3% compared to 94.2% reported in the original FairFace publication.[1] However, when it comes to the transgender dataset, the model's performance is notably inadequate. These findings suggest that while FairFace may perform well on cisgender individuals, it is not currently an effective solution for facial recognition technology in the transgender community.

Contact Information

Ezra Wingard
SUNY Oswego
Email: ewingard@oswego.edu
Website: www.ewingard.xyz
Phone: (803) 605-4795

References

1. Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548-1558).
2. Buolamwini, J., Gebru, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency
3. Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. Proc. ACM Hum.-Comput. Interact. 4, CSCW1: Article 58.
4. Morgan Klaus Scheuerman, Jacob M Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services. Proc. ACM Hum.-Comput. Interact. 3, CSCW: Article 144.
5. Instaloader. (2023). GitHub - instaloader/instaloader: Download pictures (or videos) along with their captions and other metadata from Instagram. Retrieved 20 March 2023, from https://github.com/instaloader/instaload

Acknowledgements

I would like to thank first and foremost my two thesis advisors, Dr. Lindstedt and Dr. Rhodes for providing the sage advice to help me navigate this project. I would also like to thank the SUNY Oswego Honors Program for giving me many opportunities to grow my skills.