Defining and measuring racial bias in facial recognition models

Makena Robison
Mentor: Dr. Mandy Korpusik
Computer Science, Frank R. Seaver College of Science and Engineering
Loyola Marymount University

https://github.com/makenakalei/measuring-racial-bias

Abstract

Many facial recognition models possess embedded racial bias due to the imbalanced datasets they were trained upon. They often struggle to identify minority faces at the same rates as White ones. FairFace is a model built to combat racial bias by prioritizing balanced training data. When tested and analyzed it is found that while FairFace can recognize minority faces at a higher rate than White ones, its ability to accurately label them still presents a significant disparity when identifying those with darker skin tones.

Keywords

Facial Recognition, Implicit Bias, machine learning models, DEI, FairFace

Background

In day-to-day usage, bias can be defined as “prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.” Statistically, it can be defined as “a systematic distortion of a statistical result due to a factor not allowed for in its derivation.” Using both of these definitions, it can be shown that racial bias is present in AI models, especially facial recognition models. In the paper “A Tool for Detection and Analysis of a Human Face for Aestheical Quality Using Mobile Devices” by Shuyi Zhao, Scott King and Dulal Ka, an AI tool that was used to analyze aesthetic qualities of human faces was found to bias against black faces when categorizing image input as attractive or not. The tool used an active shape model algorithm to recognize facial features and then process those features through a golden ratio analysis. Based on that output, the data was labeled as one of four categories: animal, unattractive, ordinary, or beautiful (Zhao, King, Ka, 2018). To test the accuracy of the algorithm, the tool was tested on faces that were considered by humans to fit into those categories. When tested on animals, the only face that was recognized as a face at all was the monkey because of its similarity in composition to a human face (Zhao, King, Ka, 2018). In all other categories, the tool was tested on a set of human faces that were Caucasian, Indian, Asian, Latino, and African. In each human category other that unattractive, there were no significant differences between the residual average. However, when tested on the unattractive set of faces, the tool did not even identify the African face as a face to analyze. It produced the same result as when tested on a dog’s face, null (Zhao, King, Ka, 2018). Even through a ratio based system, the tool found a darker-skinned face impossible to identify which leads to the question of colorism within facial recognition models.

FairFace is a facial recognition model developed at UCLA in an effort to find ways to mitigate racial bias in facial recognition models. In their training data, they prioritized an evenly dispersed ratio of White, Black, Latino, East Asian, South East Asian, Indian, and Middle Eastern faces, unlike heavily disproportionate models that only prioritized White faces before (Karkkainen, Joo, 2021). This balanced training data was intended to make the model just as familiar with identifying minority faces as it was with White faces so that its overall performance rate would be increased.

Hypothesis

By analyzing the FairFace model based on its performance on a diverse face-image dataset, I expect to find that despite efforts to mitigate bias within the program, the model will more accurately predict 'White' labeled faces than any other ethnicity label and potentially also struggle to recognize those images as faces in general before predicting their ethnicity.

Methodology

To discover the presence of racial bias in the FairFace model, I analyzed its accuracy, precision, and error rate on the extensive Age, Gender, and Ethnicity (Face Data) CSV dataset found on Kaggle. This dataset had 27305 rows each containing the image name, age, ethnicity, gender, and array of the image pixels. To prepare the data, each image had to be converted from its respective pixel array to a .jpg file. Using the Python Imaging Library, I was able to convert each array to a 48*48 image ready to be processed. After converting each image, the data was split up into batches of various sizes to be processed, each documented with an individual CSV file containing their image paths, ready to be passed into the FairFace model.

The model outputs two forms of feedback I used to analyze the potential presence of bias in the system. The first output was whether FairFace detected the face within the image at all. If it did not detect a face within the image, it would return a message indicating so, along with the file name of the image. If the model was able to detect a face, it would then predict the gender, race, and age of the image. The race label outputted was split into two different predictions, one was a prediction out of 6 possible ethnicities and the other was made out of only 4. Each ethnicity within those two categories was assigned a score for each image and the max score was then determined to be the label.

To analyze the bias present within the predictions of the model, I created a Jupyter Notebook to find their precision, accuracy, and error rates. In order to perform the necessary computations, both the output CSV's race column and the initial dataset CSV's ethnicity column had to be mapped to corresponding values so that they could be accurately compared. After the columns were individually mapped and merged into one data frame, they were then input into a script to find their accuracy, precision, and error for each ethnicity label separately.

Results

Race	**Number of Undetected Faces**	**Percentage**
White	88	.44
Black	22	.11
Asian	73	.37
Indian	9	.05
Others (Hispanic, Latino, Middle Eastern)	7	.03

Table 1: The number of faces that the model could not recognize separated by the images' true ethnicity labels

Race	Total Predicted	Accuracy	Error Rate
East Asian	3191	.81	.19
South East Asian	3511	.009	.991
White	6975	.91	.09
Black	3501	.67	.33
Hispanic, Latino, Middle Eastern	1288	.27	.73

Table 2: The total amount of predicted race labels and the accuracy and error percentage for each

Analysis

Table 1 shows that a significant portion of unrecognized faces are labeled as White or Asian with Black, Indian, and Others (Hispanic, Latino, Middle Eastern) in the Age, Gender, and Ethnicity (Face Data) CSV. This is evidence that FairFace was successfully able to combat one major concern addressed by Zhao, King, and Ka: how to mitigate colorism when detecting faces in general. Minorities with historically darker skin tones were recognized at a higher rate than those with lighter skin tones indicating that the model is more attuned or familiar with those faces. However, the resulting error and accuracy percentages show an extreme discrepancy in FairFace's ability to accurately label minority faces in comparison to White faces. South East Asian faces were correctly labeled less than 1% of the time and Hispanic, Latino, and Middle Eastern faces were correctly labeled at a rate of only 27%. In contrast, White faces were accurately labeled 91% of the time, and the second highest accuracy rate was East Asian at 81%. These results show that while colorism may have been combated in the first task of facial recognition, the ability of the model to discern between those with darker skin tones is still lacking greatly.

Conclusion

Despite FairFace's efforts to mitigate racial bias within their facial recognition model, the model struggled to accurately differentiate between faces that typically have darker skin tones, performing at nearly unusable rates for all categories other than White and East Asian, two racial groups with typically lighter skin tones. It is imperative to recognize the implications of this performance disparity within larger facial recognition model usages. When used for cases like law enforcement or medical identification, accuracy within facial recognition models is of the utmost importance, and disproportional performance in machine learning models reinforces already existing systemic racism rather than combatting it. While training on more balanced datasets began to address colorism and embedded bias within facial recognition systems, there must be more steps taken to finetune models to ensure a continued ethical development of these tools.

Works Cited

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).

Narora, N. 2020. Age, Gender, and Ethnicity (Face Data) [Data set]. Kaggle. https://www.kaggle.com/datasets/nipunarora8/age-gender-and-ethnicity-face-data-csv

Zhao, S., King, S., & Kar, D., A Tool for Detection and Analysis of a Human Face for Aesthetical Quality Using Mobile Devices. Arabnia, H. R., Deligiannidis, L., & Tinetti, F. G. (2018). Image Processing, Computer Vision, and Pattern Recognition. CSREA.