Andrew Hu

COSI-159 Assignment 4

1. Large public face datasets are heavily biased towards Caucasian faces and underrepresent other races, particularly darker faces. Most models are based on these datasets, so any systems involving these models will perform worse with non-Caucasian faces. This has negative implications for any field involving computer vision, but is particularly harmful in security and medicine, where recognizing a face or a feature has vital importance.
2. Model and cross-dataset performance were measured on a variety of datasets, testing the new FairFace against UTKFace, LFWA+, and CelebA. CelebA was only used for gender classification since it does not have a parameter for race. Each dataset was trained on a ResNet-34 model with ADAM optimization and learning rate of 0.0001. After training, the model accuracy was obtained on the testing sets, and compared to the cross-dataset performance by running each model on the testing set from each other dataset. To test generalization performance, the models were tested with novel data from geo-tagged tweets, media photographs, and protest datasets, which were all manually annotated.

The researchers obtained several metrics from these tests. Accuracy is simply the proportion of correct classifications over total, measuring the model against the ground truth. The researchers also used standard deviation among different races and balanced accuracy, which weighs by class, to show how the models perform differently by race. Maximum accuracy disparity measures the greatest difference between classes on a logarithmic scale and shows how classes differ. Finally, pairwise distance analysis measures the distances between pairs of data points to show how well the model can separate faces.

1. The paper defines bias in face recognition as unequal outcomes among racial groups, with a less biased model having lower disparity. In this case, the researchers focus on how a biased dataset can cause a theoretically unbiased model to have worse performance for minority racial groups. They use the metrics listed above like maximum accuracy disparity to quantify the bias in each dataset and find that FairFace leads to better accuracy in all categories.