CS 109B / Stat 121B
Spring 2018

Scope of Work for Data Science Final Project

Topic: Algorithmic Biases in Facial Recognition Systems
Group: #17

Emma Dwight - edwight@college.harvard.edu
Agasthya Pradhan Shenoy - agasthya_pradhanshenoy@gse.harvard.edu
Mehul Smriti Raje - mraje@g.harvard.edu

Project Statement and Background

Face recognition software is now built into most smart phones and several companies have released commercial software that perform automated facial analysis. Some such products include Face++, Google Image search and even automatic recognition of people in Facebook and Apple photo libraries. However, much of this technology is plagued by shortcomings, especially with respect to women or people of colour. We have seen news articles about iPhones not recognising people of colour or a simple Google search of a black man's name returning more references to criminal activity or crime reporting. The latest gender classification report from NIST also shows that the algorithms they evaluated performed worse for female-labeled faces than male-labeled faces (Ngan et al., 2015).

Most large scale face collection depends on face recognition algorithms. This means that any systematic error found in face detectors will inevitably affect the composition of the benchmark. As an example, the LFW dataset composed of celebrity faces is 77.5% male and 83.5% White. In response, the IARPA has released the IJB-A dataset which does not use face detectors to select images. An Algorithmic Justice League (AJL) has also been set up at MIT to combat 'exclusionary experiences and discriminatory practices' caused by algorithms.

A 2012 study (Klare et. al.) on mug shots found that a facial recognition algorithm trained exclusively on either African American or Caucasian faces recognized members of the race in its training set better than those of other races. The effect of the composition of the training set used surely matters.

Literature Review

Age and Gender Classification using Convolutional Neural Networks

This paper demonstrates that a convolutional neural network with three convolutional layers and two fully-connected layers and a small number of neurons can dramatically increase performance in age and gender classification of unfiltered photographs. Although there is a relatively small existing dataset of faces with labeled age and gender, researchers show that a “shallow” CNN can still improve classification accuracy. This suggests that a more elaborate model with a larger training set can increase accuracy even more - exactly the avenue we intend to pursue.

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification
This paper uses the Fitzpatrick Skin Type classification system to characterize the gender and skin type distribution of IJB-A and Adience facial analysis benchmarks. It is demonstrated that these datasets are largely composed of lighter-skinned people. This paper goes on to introduce a new dataset of 1270 faces in which skin types are more phenotypically balanced than existing benchmarks. It also introduces the first intersectional demographic and phenotypic evaluation of face-based gender classification accuracy.

Algorithmic decision making and the cost of fairness
This paper discusses specifically the COMPAS system which is used to classify the risk of reoffence of criminals, and used in many states for parole decisions, and even for sentencing. It takes into account many factors and predictors (not explicitly including race) but was shown by some researchers to have higher rates of errors for people of color. Specifically, black defendants are substantially more likely than white defendants to be incorrectly classified as high risk, and among defendants who ultimately did not reoffend, black people were more than twice as likely as white people to be labeled as high-risk. This takes place in a larger conversation: What does algorithmic fairness mean?

In this context, the authors show that formal fairness constraints would require race-specific thresholds for risk, but applying an equal threshold for risk across all people leads to different rates of risk-classification by racial group. In our own context, achieving equal levels of gender classification could be achieved by artificially worsening predictions on white faces, for example. We will consider several different formulations of “algorithmic fairness” in our work, and consider their statistical implications, drawing from papers like this one.

Goals:

Current gender classification methods might perform well, but they have great disparities on performance in classifying gender and faces of different racial/ethnic groups. We hope to raise the accuracy of the model on these subgroups thereby improving measures of statistical fairness (for example, statistical parity and predictive equality).

In our project, we plan to merge some existing image datasets. We hope to explore the use of data augmentation techniques here too. We will then apply three different transfer learning techniques to a pre-trained CNN to improve gender classification accuracy of POC.

Data sources:

- SCUT-FBP5500 (diverse benchmark dataset for multi paradigm facial beauty, 5500 labeled diverse faces)
- pubfig (58797 images of 200 public figures, labeled by name not gender)
- adience: 26,580 photos, 2,284 subjects, labeled
- Color Feret Database: (2,413 still facial images representing 856 individuals)

Preliminary EDA

The below is excerpted from NYT article: “Facial Recognition Is Accurate, if You’re a White Guy”, which itself draws this section from the work of Joy Buolamwini of the M.I.T. Media Lab.

Color Matters in Computer Vision

Facial recognition algorithms made by Microsoft, IBM and Face++ were more likely to misidentify the gender of black women than white men.

*Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.*

Gender was misidentified in ***up to 7 percent of lighter-skinned females*** *in a set of 296 photos.*

Gender was misidentified in ***up to 12 percent of darker-skinned males*** *in a set of 318 photos.*

*Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.*