Gender Classification of Image Data

Godwin Anguzu | Haoliang Jiang | Emma Mavis | Nansu Wang

Machine Learning, Spring 2022

1. Abstract

We present an analysis of experimentation across different age, gender, and racially imbalanced groups of image data. We first generate a model on the data as a whole, which is inherently imbalanced to see how well it performs on a balanced data set. Additionally, we train logistic regression and neural network models on imbalanced data (leaving out one age and race subgroup, skewed gender distribution) and evaluate the performance on a balanced data set.

1. Introduction

The world of image classification is filled with many experiments and justifications of how important facial recognition technology is in a progressing world. The specific classifications of gender, age, and race aid in this branch of machine learning applications, opening new frontiers of research and data ethics.

This field began as early as the 1960s from the work of Woody Bledsoe, Helen Chan Wolf, and Charles Bisson, in which they created an important first step of facial recognition. It was not until the 1980s when Sirovich and Kirby began to apply linear algebra to this problem, allowing great advancements to occur moving into the new century[[1]](#endnote-1).

It may not come as a surprise that one of the key turning points in the history of facial recognition created a controversy: in 2001 law enforcement used this technology on crowds at the Superbowl that year, invoking critics to call attention to Fourth Amendment rights against unreasonable search and seizure[[2]](#endnote-2). There are endless examples of police departments, technology giants, and advertisers using this technology over the last two decades that call into question the ethics of image classification. Along with privacy violations, very many of the classification models have proven to be biased, which we will also be exploring in this paper.

1. Background
2. Data

The dataset we worked with consists of 200MB of face images. Each row of the data is a face with pre-labelled gender, ethnicity, and age. The last column of the dataset is already-vectorized pixel data of each image. This means we avoid the processes of pooling, compression, and vectorization of images.

One limitation starting out is that this data only accounts for four ethnicity labels: White, Black, Asian, Indian, and Others. Therefore, we are not able to distinguish ethnicities such as Hispanic, Middle Eastern, Pacific Islander, etc. with just this data. Additionally, we chose to combine Asian and Indian images into the same “Asian” group.

* 1. EDA

Below are plots displaying the distribution of the different classes for each of the age, race, and gender features of the data.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Methods

Experimentation

For age and race subgroups, a leave-one-out method:

**Train**

**Validate**

**Leave out one**

**Leave out one**

**Test**

**All**

For gender, create skew:

**Train**

**80% Male**

**20% Female**

**Test**

**50/50**

**Validate**

**80% Male**

**20% Female**

**Train**

**80% Female**

**20% Male**

**Test**

**50/50**

**Validate**

**80% Female**

**20% Male**

We created synthetically imbalanced training data to evaluate the performance on several different subpopulations.

Results

Conclusions

Roles

References

1. https://www.nec.co.nz/market-leadership/publications-media/a-brief-history-of-facial-recognition/ [↑](#endnote-ref-1)
2. https://www.nytimes.com/wirecutter/blog/how-facial-recognition-works/ [↑](#endnote-ref-2)