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***

1. *Baseline Model 1: Logistic Regression*

We first implemented a logistic regression model to set up a baseline for the metrics. We chose the logistic regression model over the other models because it is not computationally expensive to train, and it is a proper tool for binary classification. We sampled 1000 images from the overall dataset for our training and testing purposes due to the size of the dataset. It was physically impossible to train on the entire dataset using only Google Colab Pro. We did some data cleaning before modelling by splitting datasets into train (80%), validation (10%), and test (10%). We flattened X as each image was originally represented by a 200 by 200 matrix. We then normalized the X matrix, which marked the completion of data preprocessing. As for the hyperparameter tuning of the logistic regression model, we used LASSO instead of the ridge regression because of the sparsity of the data. We then chose the corresponding liblinear solver for the model. So, the only hyperparameter that we are interested in tuning is C (the inverse of the regularization strength). We picked the optimal C according to the highest AUC score of the prediction on the validation set. The optimal AUC was found to be 0.84 and the average precision was 0.86. The metrics look decent, which means that the simple logistic regression has basic binary classification ability on gender and can serve well as the baseline model for us to later compare with the other models. During this elementary phase of the project, we also tried to explore dimension reduction methods to see if the performance will be negatively affected. We kept only the grayscale channel and removed the RGB channels of these images. But it turned out that dimension reduction did make the model perform worse and we chose not to utilize grayscale images for our study. Chart, line chart, scatter chart

Description automatically generated

1. *Baseline Model 2: Random Forest*

Since the size of the subset of the overall data is still enormous, we extensively relied on the power of Google Colab’s Tesla GPU for building the second baseline model. The random forest model is built via utilizing the GPU powered cuML module instead of the CPU powered Sklearn module to have the speed of training boosted by 45 times (see reference). The data processing steps are the same as what we did in the logistic regression notebook, except that the datasets were transformed to be GPU readable at the end of data processing. We ran a random forest model with no specified hyperparameters, and the validation accuracy turned out to be only 62%. This led us to think about that the random forest model is overfitted by the training data and thorough hyperparameter tuning and decision trees trimming should be performed. We first adopted the randomized search cross validation to try to narrow down the long list of hyperparameter candidates. One thing to notice is that tuning the cuML’s random forest model involves more hyperparameters. For instance, “n\_bins” refers to the maximum number of bins used by the split algorithm per feature. Increasing the number of bins on large and skewed input data can improve the accuracy. The randomized search cross validation algorithm chose the tree to be rather shallow and splits rather detailed. This first round of cross validation helped us to narrow down the range of hyperparameters that we should look into. We then used grid search cross validation to find the best combination of hyperparameters. The grid search tried to trim each decision tree in the forest by lowering the depth, but the accuracy still did not vary by much. So we performed another grid search cross validation that optimized on the AUC score instead of accuracy, just to boost the generalization performance. Even after this third round of hyperparameter tuning, the performance of the random forest model still did not meet our expectation. Both of the AUC and the AP scores were found to be below 0.8. Meanwhile, the accuracy of predicting gender is not much better than tossing a coin. Metrics have pointed us to ditch this model. We hypothesized that the random forest model inevitably learned too many details on each image and thus led to very poor generalization performances. We thus planned to treat this model as our second baseline model and continued to explore more options.

Chart, line chart

Description automatically generatedA picture containing chart

Description automatically generated

1. Final Model: Transfer Learning @Nansu

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

Random forest: [Accelerating Random Forests Up to 45x Using cuML | NVIDIA Technical Blog](https://developer.nvidia.com/blog/accelerating-random-forests-up-to-45x-using-cuml/)

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)