Gender Classification of Image Data

Godwin Anguzu | Haoliang Jiang | Emma Mavis | Nansu Wang

Machine Learning, Spring 2022

1. **Abstract**

We present an analysis of gender classification on image data that is imbalanced across race, gender, and age. The purpose is to emphasize how the imbalance greatly impacts model performance, despite the reality that image classification applications still operate on imbalanced data. We ran experiments across extremely skewed gender and racially imbalanced groups of image data. We first generated baseline models and narrowed down the choice of a neural network on the data set as a whole with no experimentation. Then we applied the best performing neural network model to various imbalanced data and compared the performance on the different subgroups. We found that model performance differs greatly across different subgroups, implying that equitable representation in data is of the utmost importance for model generalizability across subgroups.

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 in facial recognition when they began using computers to recognize the human face. 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**

Of the three specified races, more than half (54%) of the images were labelled White, 31% were labelled Asian, and only 4% were labelled Black. Additionally, females made up most of the images, 56%, and the remaining 44% were labelled male. We binned the age label into four groups: children, 0-13 years old; adolescents, 14-25 years old; adults, 26-45; and mature adults, 46 and above. Age group was composed of 33% children, 18% adolescents, 22% adults, and 27% mature adults.

From the above description and visualizing figures 1 – 3, we quickly notice an under representation of the black racial group, the male gender, and the adolescent age group. If this imbalance is not taken care of, it will affect the proper detection of these less represented groups. In order to demonstrate how drastic this performance capability is, we intentionally perform modelling with these imbalances and show you how attempting to classify these groups, even with the best models, leads to bias in classification of the less represented groups.

Figure 1: The bar plots show the distribution of race, gender and age   
group in image data

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**

Neural network solutions were tried for this problem. Since there are many well-defined neural network structures for computer vision tasks, we decided to use transfer learning workflow taking advantage of those pre-trained models. The transfer learning was learned from[[3]](#endnote-3).

The process contains two parts, preprocessing, training & fine-tuning. In preprocessing, all the values were rescaled to be in [-1, 1]. Then, prefetching and caching were adopted for faster speed. Each input photo was randomly augmented using random rotation and flips.



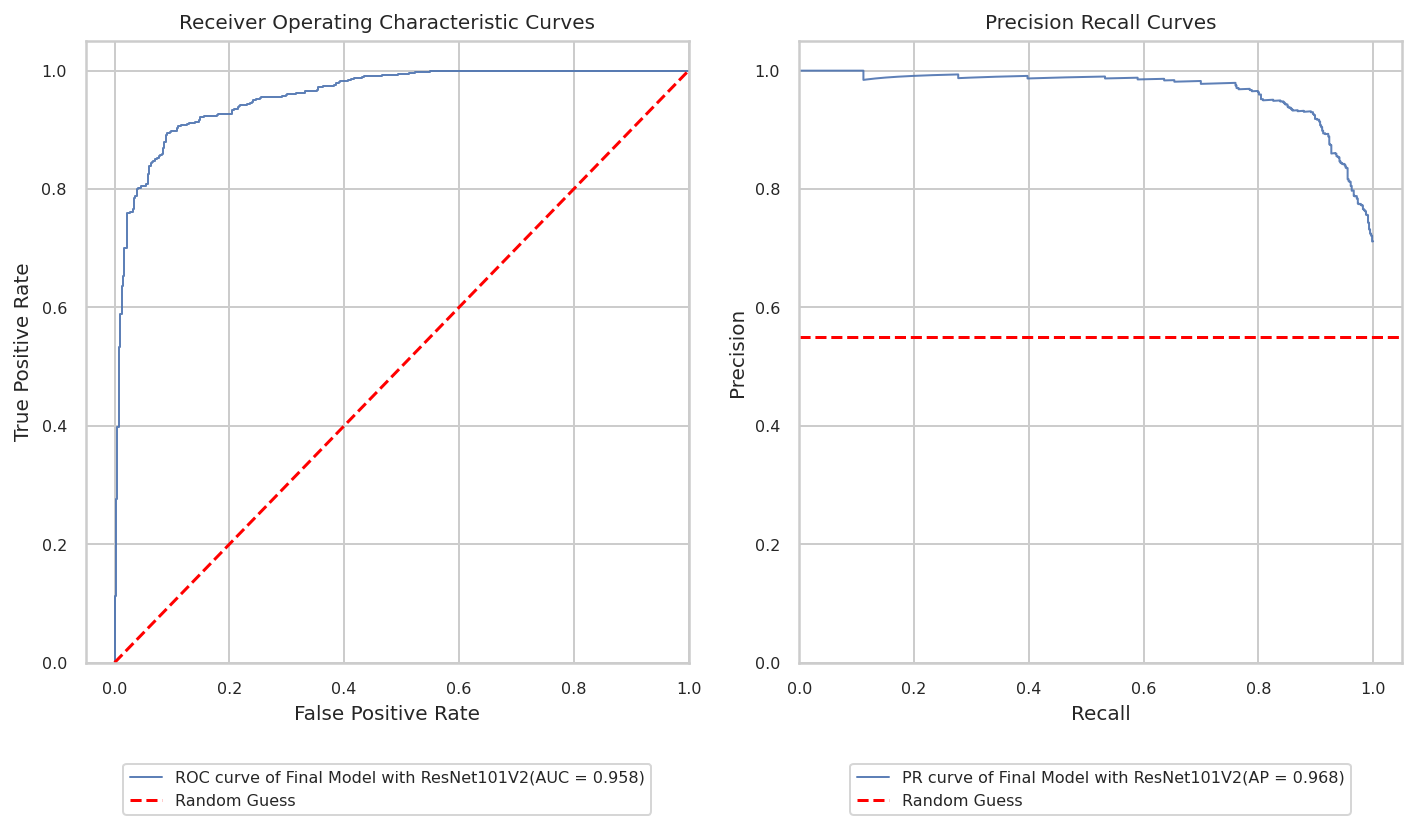
In training & fine-tuning, as shown in Fig. 1, the model was built around the pre-trained model to fit our input and output. Specifically, the top layer of the pre-trained model was excluded to fit our input size. After the pre-trained model, an average pooling layer, a dropout layer for regularization, and a fully connected layer to output were added.

When training, the pre-trained weights were frozen first and the model was trained until it converged. We prevented the large gradients from changing the pre-trained weights too much by first freezing them. Next, we fine-tuned the model for better performance. The pre-trained weights were defrosted and the model was trained using a small learning rate during a few epochs.



Eight pre-trained models were tried to be the base model. As shown in Fig. 2, ResNet101V2[[4]](#endnote-4) was the best both in AUC and AP. More parameters do not always mean better performance. Bias and variance tradeoff was considered here.

For hyper-parameters, dropout rates were tuned manually using a subset of random samples due to the computation cost. In the final model, the 0.1 dropout rate was selected. It usually took 20 epochs for the model to converge. When fine-tuning, the learning rate for fine-tuning was set to be 10-5, the optimizer is Adam[[5]](#endnote-5). In the final model, ten epochs of fine-tuning led to 0.0778 improvement in case of the binary accuracy in the validation set (from 0.7883 to 0.8661).



The final model achieved 0.958 AUC in the ROC curve and 0.968 AP in the PR curve. It outperforms other models that we explored in this project.

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)
3. fchollet, "Transfer learning & fine-tuning." Keras. Updated May 12, 2020 [Website]. Available: https://keras.io/guides/transfer\_learning/, Accessed on: April 12, 2022. [↑](#endnote-ref-3)
4. K. He, X. Zhang, S. Ren, J. Sun, "Identity mappings in deep residual networks". In: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham, vol. 9908, pp. 630-645, 2016. https://doi.org/10.1007/978-3-319-46493-0\_38 [↑](#endnote-ref-4)
5. "Keras Applications." Keras. [Website]. Available: https://keras.io/api/applications/, Accessed on: April 12, 2022. [↑](#endnote-ref-5)