# Algorithmic Bias in Facial Recognition Systems

Emma Dwight, Agasthya Pradhan Shenoy, and Mehul Smriti Raje

CS109b/Stat121b: Data Science, Spring 2018

## GOAL

Improve prediction accuracy of gender classification models using images of White, Black and Asian faces.

**Contributions:**

- Compare accuracy of models trained on skewed and balanced datasets
- Explore data augmentation techniques to counteract imbalance in training data
- Use transfer learning to determine whether pre-trained facial recognition models can improve their accuracy when re-trained on more diverse datasets.

## DATA SOURCES

**Adience Dataset:** 18,239 faces collected by the Open University of Israel.

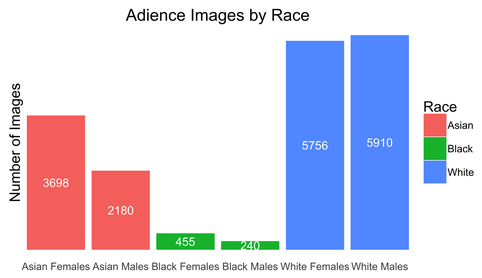


Figure 1: Distribution across races in Adience Dataset

**Selfie Dataset:** 46,836 selfie images collected by the University of Central Florida.

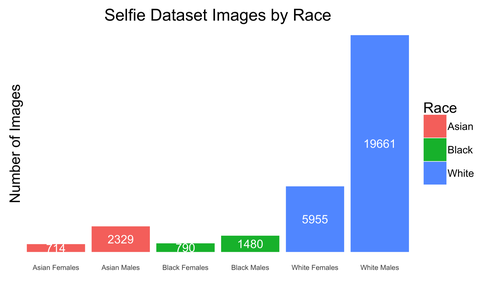


Figure 2: Distribution across races in Selfie Dataset

## MODEL

**CNN Architecture**

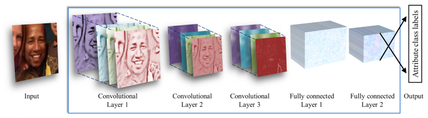


Figure 3: Illustration of the Adience Benchmark CNN architecture on which our model is based [1]

Building on the above model, our model adds two fully connected layers followed by relu and sigmoid activations to classify samples by gender.

## SAMPLING METHODS

- **Proportional Method:** Number of samples of each gender and race proportional to the skew in the Adience data.
- **Balanced Method:** Equal numbers of faces of each race and gender.

## DATA AUGMENTATION

The extreme imbalance in the training data limits accuracy, especially for the minority subgroups. We artificially produce more image data using augmentation techniques [2]. We experiment with gentle image augmentation using the horizontal flip, shear, and zoom options provided by keras, and with a more diverse range of transformations provided by the ImgAug package [3].

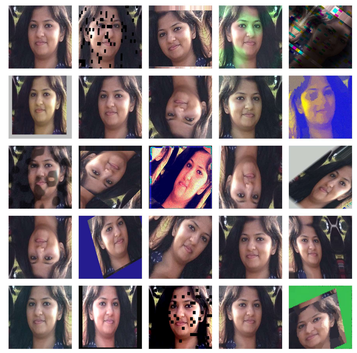

Figure 4: Example of Extreme Augmentation using ImgAug

## TRAINING

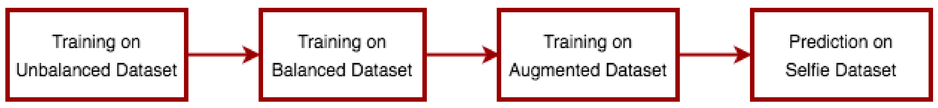


Figure 5: Algorithm to Improve Accuracy of Benchmark

## RESULTS

Training on balanced and augmented datasets can improve accuracy of the benchmark.

| Model Layer | Training Acc. | Validation Acc. |
|---|---|---|
| Unbalanced Data | 0.7980 | 0.6560 |
| Balanced Data | 0.9990 | 0.7380 |
| Augmented Data | 0.5015 | 0.5000 |

Table 1: Results at Each Hierarchy of Model

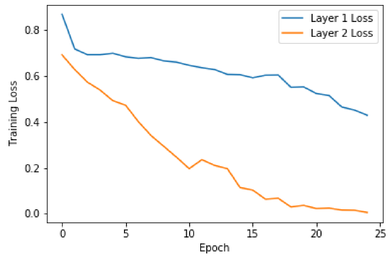


Figure 6: Training losses for First and Second Model Hierarchies

## CONCLUSIONS

Hierarchical model trained with balanced sampling from each race and gender boosts accuracy of gender classification. Adding another hierarchy to train on augmented data requires changes in existing architecture.

## FUTURE WORK

Future researchers ought to:

- Experiment with adversarial images to increase the robustness of their models
- Improve upon existing methods and software (eg. opencv) used to identify and crop faces in images, as we found these to succeed less often on faces of darker skin
- Consider over-sampling training images from minority groups in order to build models that work well for them (sample at rates not proportional to the population)
- Publish accuracy rates of their models for different racial groups in their test set

## REFERENCES

[1] G. Levi and T. Hassncer.
Age and gender classification using convolutional neural networks.
In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 34–42, June 2015.

[2] Luis Perez and Jason Wang.
The effectiveness of data augmentation in image classification using deep learning.
2017.
Available at `http://cs231n.stanford.edu/reports/2017/pdfs/300.pdf`.

[3] Alexander Jung.
Imgaug: a library for image augmentation in machine learning experiments.
Available at `https://github.com/aleju/imgaug`.