Algorithmic Bias in Facial Recognition: Exploring Racial and Gender Disparities through CNN Models

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Claire Yoon

Fall 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Rosanne Vrugtman, Department of Computer Science

Algorithmic Bias in Facial Recognition: Exploring Racial and Gender Disparities through CNN Models

CS4991 Capstone Report, 2023

Claire Yoon
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
ndq7xj@virginia.edu

ABSTRACT

Algorithmic bias in facial recognition technology has emerged as a significant concern since it consistently misidentifies certain races and genders, leading to undesirable outcomes. To investigate how facial recognition systems work and the potential factors that affect misidentification, I worked on a project titled *Celebrity Facial Recognition* with two fellow students. The project was designed to verify whether our machine learning models accurately identify and detect the correct person from a particular group of people with similarities in their appearances. By utilizing the Convolutional Neural Network (CNN) for image classification, web scraping and web crawling for data collection of six different groups (White male, White female, Black male, Black female, Asian male, and Asian female), and the PyTorch framework for transfer learning, we developed the project to train and evaluate the models. Once images were gathered from a search engine, the models were trained using input data and processed through multiple layers to detect the correct subject. The project resulted in the models detecting the correct person with about 80% validation accuracy across all the six groups, but lower validation accuracy among them occurred within the White male and Asian male groups. Future work on the project could include expansion of high-quality datasets and different configurations for equitable performance across diverse populations.

1. INTRODUCTION

Since facial recognition technology was pioneered in the 1960s by Bledsoe, Wolf, and Bisson, computer scientists whose efforts centered on training computers to recognize human faces [1], the technology has developed and is now used in various areas, including everyday uses from unlocking mobile devices to healthcare, education, and criminal justice. Because of its extensive versatility, benefits and convenience, the technology is implemented across many industries.

Although facial recognition technology provides numerous advantages like efficiency, customization, and cost reduction in a quick way to identify subjects, the technology has its flaws, including gender, race, and age biases due to statistical factors such as data sources including input data that are already biased or low quality. In addition, systemic biases can occur by propagating and generating information when machines are trained from the data sources [2]. Others can stem from human factors because when engineers design and train facial recognition systems, it might be easy for them to let unconscious biases in the systems without rigorous fairness assessment. The biased data sources and design decisions in development process lead to false identification, which brings negative impacts on minority groups.

2. RELATED WORKS

As algorithmic bias in facial recognition technology brings attention to society, there have been many efforts to research this issue. A key study by Buolamwini, a computer scientist and founder of the Algorithmic Justice League in 2018, examined the performance of facial recognition systems developed by Microsoft, Face++, and IBM. This study focused on how the AI-powered gender classification products perform differently based on the gender and skin type of individuals and evaluated the accuracy of the systems. As a result, the study revealed significant disparities in the algorithm's accuracy when classifying faces by gender, with error rates for darker-skinned females which are up to 34% higher than for lighter-skinned males [3]. This highlights the significance of diminishing bias and improving the identification of all faces regardless of race or gender.

Another research, addressed by Kaur, et al, discusses the rapid rise in the use of facial recognition technology, especially in security and criminal justice [4]. Although this emerging technology is promising, the authors point out certain obstacles. Currently, the algorithms have problems with variations in lighting conditions, pose, and occlusions. Moreover, the size and quality of the databases can significantly impact the software's accuracy, which can cause discrimination against certain groups of people. The authors emphasize the importance of transparency and accountability of facial recognition technology and call for broader societal discussions on the ethical implications of its deployment.

3. PROJECT DESIGN

This project was inspired from "Gender Shades" by Buolamwini. The purpose of this project was to develop machine learning models that accurately identify the correct person among a particular group of people with similarities such as race, gender, and age in their appearances. We also aimed to determine whether the models we develop are consistent with or different from prior research showing facial recognition systems misidentified Black and Asian faces 10 to 100 times more than White faces.

3.1 Data Collection

To develop and train machine learning models, data preparation should be the first step. We divided six different groups based on their race and gender: White male, White female, Black male, Black female, Asian male, and Asian female. Each group contains three individuals. To train the models for better performance, we chose celebrities as the representatives of each group to collect many high-resolution images. The celebrities were chosen based on similar ages and facial appearances. The selected celebrities are:

- White Male: Hugh Jackman, Jake Gyllenhaal, Ryan Reynolds
- White Female: Lucy Hale, Mila Kunis, Sarah Hyland
- Black Male: Danny Glover, Denzel Washington, Morgan Freeman
- Black Female: Alicia Keys, Jordin Sparks, Queen Latifah
- Asian Male: (BTS group members) Jimin, Jung Kook, V
- Asian Female: (Blackpink group members) Jennie, Jisoo, Rosé

For data collection of the celebrities, we used both web scraping and web crawling. Web scraping is the automated process of collecting structured information from the Internet. It extracts and duplicates data from any webpage it accesses. Web crawling is used by search engines to scan the Internet for pages according to the keywords the user inputs and to remember them through indexing for later use in search results. It navigates and reads pages for indexing. Deploying both web scraping and web crawling, 160 images were gathered for each celebrity. By a group, 360 images were used for training models and 120 images were used for validation/evaluation.

3.2 Algorithms and Methods

For creating facial recognition models, we chose Convolutional Neural Network (CNN) because CNNs are well-suited for image classification and tasks that involve the processing of pixel data. CNNs excel at finding patterns in images, enabling the recognition of objects, classes, and categories.

The processes of the project are as follows:

- Step 1: Gather different photos of each celebrity and compile them into a dataset
- Step 2: Train the models to generate a feature vector from the celebrity's image, which represents the unique characteristics of that face
- Step 3: (Once CNNs are trained) Allow models to compare that feature vector with images in its dataset to examine whether it matches the correct person

To reduce training time and improve performance within the timeline, we applied transfer learning, a machine learning technique used to conduct another related task. This technique allowed us to train only the last layer using a smaller dataset and ensure better generalization since neural pre-trained network models already learned a lot about features using extensive data. For this technique, we utilized PyTorch framework, specifically ResNet, a variant of the Residual Neural Network, as shown in Figure 1 [5].

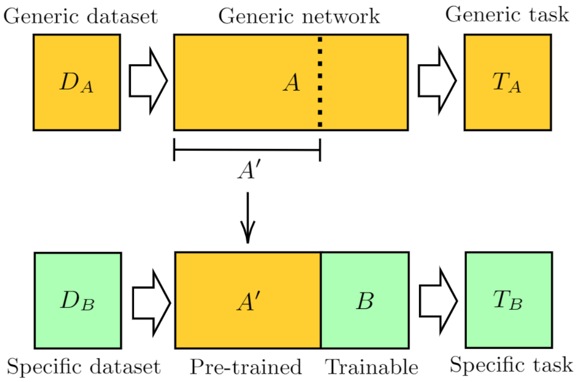

Figure 1. Transfer Learning

The following steps are the transfer learning process with PyTorch framework

- Step 1: Load the Data
  - Load data from folder with torchvision.dataset
  - Transform images to match network requirements such as resize, horizontal flip, normalize, and convert to tensor
- Step 2: Define Model
  - Use torchvision.models to load ResNet34 with pre-trained weight set to TRUE
  - Use CrossEntropyLoss for multi-class loss function and optimizer (SGD: Stochastic Gradient Descent) with learning rate of 0.001 and momentum of 0.9
- Step 3: Train and Test Model
  - Train and evaluate model with epoch of 10 (epoch refers the total number of iterations of all the training data in a cycle for training a machine learning model)
- Step 4: Results
  - Review results by comparing training loss with validation loss

4. RESULTS

This project enabled us to examine algorithmic bias in facial recognition technology by developing our machine learning models. Using PyTorch's resnet34 function with pre-trained weight set to True, we were able to get a relatively higher validation accuracy (refers to a metric that evaluates the performance of a model that was not used during the training phase) on each group of the model with a smaller dataset: 360 images for training models and 120 images for validating models. Overall, the six different models resulted in around 80% of validation accuracy with epoch number 10. Most groups' training loss and validation loss consistently decreased in every iteration except for the White Male and Asian Male groups. Our machine learning models showed different results than prior research revealed. Facial recognition systems are less accurate in identifying dark-skinned people, especially women. Comparing each group in our models, the lower validation accuracy occurred in the White Male group (70.0% in epoch = 10) and Asian Male group (72.5% in epoch = 10).

5. CONCLUSION

Examining inequality resulting from algorithmic bias in facial recognition technology is a significant issue in our society. The goal of this project was to develop machine learning models to evaluate our machine learning models and compare them with the results of prior research. We designed the project *Celebrity Facial Recognition* so people understand the importance of addressing this issue. Understanding the problem existing in facial recognition systems would allow our society to become more aware of the issue and ultimately devise comprehensive approaches to mitigate algorithmic bias in facial recognition technology. This project will help to emphasize that addressing algorithmic bias in facial recognition systems and the reflecting ruminative solution approach are important.

6. FUTURE WORK

This project was conducted with a limited timeline and computing resources. These included not having a GPU that could quickly test for gathering pictures and process training and evaluation of our models, as well as relatively smaller datasets. Future work will be needed to include enough computing resources and increase the number of subjects that can represent the groups as a whole. Although we were trying to select celebrities with similar appearances for each group, we acknowledge that they cannot represent the group as a whole. In the future, we can expand by adding more data with different age groups and appearances. Also, instead of using a pre-trained network through the framework, we can consider developing our own network to inspire other research and studies.

7. ACKNOWLEDGMENTS

This project was developed in collaboration with Beza Gashe and Stephanie Fissel. We would like to thank Professor Brian Wright, whose continuous support and guidance made this project possible.

REFERENCES

[1] A brief history of facial recognition - NEC New Zealand. Retrieved September 29, 2023 from https://www.nec.co.nz/market-leadership/publications-media/a-brief-history-of-facial-recognition/

[2] A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*.. Retrieved December 1, 2023 from https://dl.acm.org/doi/10.1145/3457607

[3] Gender Shades. Retrieved September 29, 2023 from http://gendershades.org/ http://journals.sagepub.com/doi/10.1177/0025802419893168

[4] Facial-recognition algorithms: A literature review. *Medicine Science and The Law*.. Retrieved from http://journals.sagepub.com/doi/10.1177/0025802419893168

[5] Transfer learning in hybrid classical-quantum neural networks. *arXiv: Quantum*

Physics.. Retrieved from https://quantum-journal.org/papers/q-2020-10-09-340/