Building Facial Recognition Using Factorial Encoding
Analysis of the Racial Bias in Google's Facial Recognition Software

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science University of Virginia
In Partial Fulfillment of the Requirements for the Degree Bachelor of Computer Science

By
Cooper Scher

December 9, 2022

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.

ADVISORS

Brianna Morrison, Department of Compute Science
Ben Laugelli, Department of Engineering and Society

Introduction

Reliable identification of persons and objects has been an evolving problem in a world where information flow and access are increasing exponentially. Identification technologies have been deployed in various contexts, such as user validation and data analysis (Zarkasyi et al., 2020). Thus, from unlocking a phone to helping authorities track down assailants, identification technology is instrumental in facilitating many societal interactions expected in daily life.

Facial recognition, as an identification technology, has been subject to considerable research as an algorithmic and technical problem. Other identification techniques such as passwords and fingerprinting are subject to fallibilities that facial recognition can overcome. It is much harder to beat a facial recognition software than a fingerprinting software, and faces don't need to be remembered (Zarkasyi et al., 2020). However, facial recognition algorithms have had issues with implicit biases along demographic lines, especially race (Gong et al., 2020). There has been an intensive focus from companies and academics to resolve these issues, but many facial recognition implementations still fail to perform with accuracy along all racial lines (Gong et al., 2020). To remedy the biases present in many current facial recognition algorithms, I will propose a facial recognition algorithm that uses an unsupervised adaptive synaptogenesis algorithm with factorial encoding to reduce bias.

While many facial recognition algorithm designs attempt to approach the subject without bias, there are deeper social and political elements that have facilitated biased results from current facial recognition algorithms despite a lack of algorithmic intent. One recent example is Google's reverse image search, which–using facial recognition to identify human images–was preferentially misclassifying darker-skinned faces as primates. I will use Google's failure as a case study to study some of these factors, including the sociopolitical biases present in training

datasets and the backgrounds of the researchers who design the algorithms (Lohr, 2022). Failing to understand these factors will stymy efforts to reduce bias as designers of facial recognition algorithms need to be cognizant of the non-technical context that can influence the resulting implementation.

To design a less biased facial recognition algorithm, the social and technical factors needed to be understood concurrently. Using a neural network based on adaptive synaptogenesis, I will implement an algorithm that is designed to handle incoming biased information. Further, I will use actor-network theory on the racial bias in Google's initial release of facial recognition to better understand how the amalgamation of actors involved in data creation, data processing, algorithm construction can lead to biased facial recognition results from an "unbiased" algorithm.

Technical Project Proposal

Facial recognition technologies that can uniquely identify users have been in development for nearly half a century. With the goal of streamlining processes such as identity confirmation and greatly increasing the capacity for image analysis research, facial recognition research took off in the fields of computer vision and artificial intelligence. Early approaches to the facial recognition problem utilized manual user-input of measurements among a set of facial features, such as eye size (Adjabi et al., 2020). With the advent of digital image processing and linear algebra-based techniques, facial recognition technology began to work using statistical methods such as principal component analysis (PCA) on digital representations of faces (Adjabi et al., 2020). Recently, machine learning and deep learning techniques have been applied to dramatically improve performance to human levels of accuracy, or 97% (Acien et al., 2019).

Nonetheless, even recent advances using deep learning have been subject to issues of bias due to input training datasets that teach biases to the algorithms (Acien et al., 2019).

The adaptive synaptogenesis neural network model is an unsupervised neural network algorithm based off the neural learning patterns of the human brain visual system. The algorithm is a feedforward network, which means it uses successive layers of neurons. In other words, the model works by a series of nodes that receive input from the previous layer, perform a calculation, and then sends output to the next layer. By training the model with many repetitions of inputs, specific outputs of the model become associated with specific inputs. Thus, a specific person's face or a group of people might become associated with the output of a specific set of neurons in the final layer, allowing the model to uniquely identify faces. The adaptive synaptogenesis algorithm is designed to randomly connect neurons between layers are until all the nodes are outputting at the same rate (Thomas et al., 2015). This property is seen in the brain where the goal is to efficiently represent information while minimizing the energy needs to fire neurons. Further, this allows the model to distribute information equally among neurons (Bartlett, 2007). This leads to a representation known as factorial encoding, where, in the ideal case, information is perfectly separated into subcomponents that can accurately reconstruct the input from the output (Bethge, 2006).

This technical project seeks to use the adaptive synaptogenesis neural network model to create a facial recognition algorithm that is more resistant to incoming data bias. Previous implementations have been able to deal with imbalanced input sets and still accurately distinguish the desired features. This is because the algorithm is decorrelating information between neurons, which means that larger input categories, for example certain races being more frequent in a dataset, will not get a larger proportion of neurons. This is especially helpful for

detecting minority classes, for example demographics who may appear less often in a facial dataset. To implement the algorithm, first a public dataset of facial images would be acquired. Next, the image data would have to be processed into numerical vectors, through a process called flattening, to allow the algorithm to identify features. Then, the adaptive synaptogenesis model algorithm would be modified to accommodate the processed data input as well as parameter-tuned for facial features. Modifications will include the substitution of code meant for singular numerical inputs to handle images. Finally, the output of the model would be tested and validated from an out-of-sample dataset to ensure intended performance.

The PubFig dataset contains over 50,000 images of 200 people and is not populationally representative of the population. On the other hand, the Princeton face images database contains 575 individual faces in a representative distribution of the population. These databases could be used to train and validate the model for robustness even with biased data input. The adaptive synaptogenesis algorithm created by (Thomas et al., 2015) will be used along with processing tools in matlab to create the modified algorithm. This project will be completed by myself over a semester of CS 3991.

STS Project Proposal

In 2011, Google, a company famous for its ubiquitous search engine, added functionality for reverse image search. Normally, users would enter text queries to see content of various types such as websites, images, or videos, but with the reverse image search, user could use images to search for weblinks and images. This feature had utility for expanding the search functionality of the website for users and was met with approval from users who found the feature entertaining and useful (Simonite, 2018). To implement this functionality, Google developed recognition algorithms that were able to profile images and determine search terms

associated with the image. Part of this design was facial recognition software of human images to identify specific persons in a reverse image search. However, in 2015, Google image-labeling technology was criticized by software engineer Jacky Alciné, who showed that Google's recognition software was classifying his black friends as gorillas (Waelen, 2022). In fact, many terms, associated with African Americans, such as "black man" were met with such problematic results with public blowback that it forced Google to remove some search functionality from the service (Simonite, 2018). After the scrutiny, it was no longer possible to search for gorillas in the reverse image search.

While it is easy to attribute the issues in the reverse image search to poor algorithm design or racial animus among the designers, this perspective doesn't address the systematic factors such as the bias in the generation and availability of input datasets. Even a perfectly unbiased algorithm will struggle when it is trained on a dataset that is biased. In particular, the google reverse image search algorithm worked exceedingly well in most image contexts. The algorithm based on image features such as color, deformation, and visual similarity was highly successful except when the features were shared between images that weren't tested by engineers, termed corner cases. While the confidence in search results was consistently above 90%, even with the problematic terms such as gorillas, the small error rate stacked onto billions of searches led to a high level of troublesome misclassifications (Simonite, 2018). Thus, even though the algorithm was very accurate and created without apparent racial intent, small corner cases issues led to a public fiasco.

The social and economic incentives of an actor like Google would indicate a desire for a non-biased outcome; however, the resulting product indicates that the systems and connections that lead to bias are not clear-cut. Given this apparent contradiction, I propose to study this

network through the lens of actor-network theory (ANT). ANT is a sociotechnical framework that attempts to view how social and technical forces can affect each other, specifically by construction of a network of technical and non-technical components, termed actors. One key feature is that each interaction between actors is not expected to be equal, leading to heterogeneous network, for example the relationship between algorithm developers and the data that is available for algorithm development. Another advantage of ANT is that actors can be abstract complicated processes, such as Google's algorithmic development process, as blackboxes, obviating the need to study each highly technical process.

I propose that bias is not just the result of an algorithm but also the input data context. The google images search engine pulls images from websites and online resources to show results for users, from which the recognition software is applied for reverse image search. This means that while the network builders, software engineers, were not designing an algorithm that would misclassify those of a certain race, the input data was taken from an environment that exhibited those biases (Waelen, 2022). From a lower sample size of available black faces on the internet, it is much easier to misclassify features to images that may have similar underlying mathematical features that more defined categories such as faces of other races. Thus, the actors, internet users, were important for the engineers to consider in the manufacturing of their reverse image search algorithm since it was affecting the performance of their algorithm in a way that was particularly damaging for the company.

To make this case, it would be beneficial to use actor-network theory to look at how the interactions between the software engineers, the company itself, the users, and the people who upload images to website have contributed to a system where facially unbiased algorithms can produce biased results. A profiling of the catalog of Google images would help quantify how

diverse and representative queries are of human faces. The internet users who uploaded the photos may have played a large and silent role in the bias of the Google algorithm, so, through the lens of ANT, analyzing the training data these users created will be illustrative.

Conclusion

The deliverable for this technical project proposed in this paper is a working algorithm that will be trained on a small set of faces with a biased and unbiased data input. The goal is to show a robust output that is free of considerable bias despite a training dataset that isn't representative of the population. By using the factorial encoding of the unsupervised adaptive synaptogenesis neural network algorithm, the effects of a biased data input can be reduced. The STS research paper will attempt to analyze the agents involved in the creation of biased data and how engineers respond to this challenge to see why one of the preeminent machine learning corporations was unsuccessful in avoiding bias in its search results. The ANT-based analysis is constructed with the goal of more concretely identifying the why and how input data results in biased algorithms amongst developers without apparent biased intentions. Even for an algorithm–such as the neural network suggest above–that is intended to avoid bias, identifying and avoiding biased data input will be instrumental in a resulting non-biased algorithm. Taken together, the technical report of a potential non-biased algorithm along with the actor-network theory analysis of the Google failure will help to provide a pathway for facial recognition technology that does not discriminate in its effectiveness.

Words: 2086

References

Acien, A., Morales, A., Vera-Rodriguez, R., Bartolome, I., & Fierrez, J. (2019). Measuring the gender and ethnicity bias in deep models for face recognition. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 584–593. https://doi.org/10.1007/978-3-030-13469-3_68

Adjabi, I., Ouahabi, A., Benzaoui, A., & Taleb-Ahmed, A. (2020). Past, present, and future of Face Recognition: A Review. *Electronics*, *9*(8), 1188. https://doi.org/10.3390/electronics9081188

Bartlett, M. S. (2007). Information maximization in face processing. *Neurocomputing*, *70*(13-15), 2204–2217. https://doi.org/10.1016/j.neucom.2006.02.025

Bethge, M. (2006). Factorial coding of natural images: How effective are linear models in removing higher-order dependencies? *Journal of the Optical Society of America A*, *23*(6), 1253. https://doi.org/10.1364/josaa.23.001253

Gong, S., Liu, X., & Jain, A. K. (2020). Jointly de-biasing face recognition and demographic attribute estimation. *Computer Vision – ECCV 2020*, 330–347. https://doi.org/10.1007/978-3-030-58526-6_20

Lohr, S. (2022). Facial Recognition Is Accurate, If You're a White Guy*. *Ethics of Data and Analytics*, 143–147. https://doi.org/10.1201/9781003278290-22

Simonite, T. (2018, January 11). *When it comes to gorillas, Google Photos remains blind*. Wired. Retrieved October 27, 2022, from https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Thomas, B. T., Blalock, D. W., & Levy, W. B. (2015). Adaptive synaptogenesis constructs neural codes that benefit discrimination. *PLOS Computational Biology*, *11*(7). https://doi.org/10.1371/journal.pcbi.1004299

Waelen, R. A. (2022). The struggle for recognition in the age of Facial Recognition Technology. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00146-8

Zarkasyi, M. I., Hidayatullah, M. R., & Zamzami, E. M. (2020). Literature review : Implementation of facial recognition in society. *Journal of Physics: Conference Series*, *1566*(1), 012069. https://doi.org/10.1088/1742-6596/1566/1/012069