TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

Fakultät Informatik

Discrimination in Algorithms (Face Recognition)

Intercultural Communications

Vorgelegt von: Ronny Pollak
Matrikelnummer: 3694422
Studiengang: Master Informatik
Dozent: Wolfgang Jockusch
Abgabedatum: 27.01.2023

Contents

List of Figures

List of Tables

1 Introduction

Discrimination in algorithms is a growing concern in the field of artificial intelligence (AI) and machine learning. Discrimination can occur in a number of ways, including bias in the training data or the algorithm itself. One specific area where discrimination has been identified is in facial recognition technology. This technology uses algorithms to analyze images of faces and match them to a database of known individuals. This essay will discuss the ways in which discrimination can occur in algorithms, as well as providing specific examples of discrimination in facial recognition technology.

2 Training and test data

To understand how algorithms can discriminate a group of people you first have to understand how these algorithms work. When building an AI algorithm it has to be trained on a big amount of data. This is called the training data. In AI, training data is a set of data used to train an algorithmic model. In 2.1 we can see a simplified figure of a neural net as blackbox that is being trained on the training data. The model uses this data to learn patterns and relationships in the data, which it can then use to make predictions or decisions on new, unseen data.

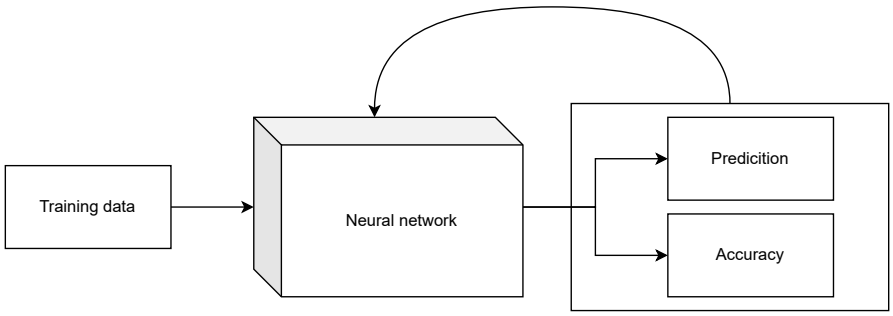

Figure 2.1: Neural network as blockbox with training data

On the other hand, is the test data. In 2.1 we can see a simplified figure of a neural net as blackbox that uses the test data to make predictions or decisions and compare them to the true values in the test data to evaluate its accuracy and reliability. In general, the training data is used to optimize the model and the test data is used to evaluate the model. [Kha+18, p. 32-33]

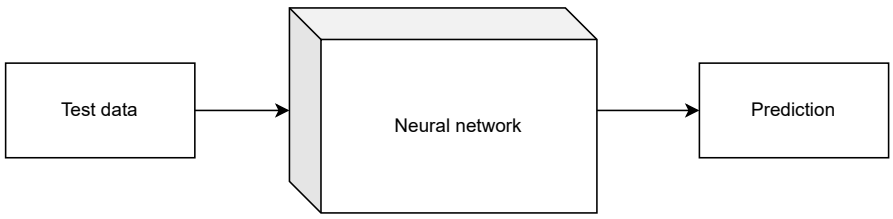

Figure 2.2: Neural network as blockbox with test data

3 Issues with data

AI bias is an anomaly in the output of machine learning algorithms, due to the prejudiced assumptions made during the algorithm development process or prejudices in the training data.

There are many problems concerning the data that can make algorithms unreliable and lead to bias. One of them is poorly selected data. The designers of the algorithmic system may decide that certain data is important for the outcome of the algorithm and other data isn't. This can lead to disadvantages for a certain group of people. [CM16, p. 7]

Another problem is the selection bias. Selection bias occurs when the set of data inputs to a model is not representative of a population, which can lead to conclusions that favor certain groups over others. [CM16, p. 8] For example, facial recognition AI algorithms are often trained on data that includes more examples of white people's faces than other races, leading to difficulty in recognizing faces of other races. In a study by Joy Buolamwini and Timnit Gebru the authors use the Fitzpatrick Skin Type classification system to characterize the gender and skin type distribution of two facial analysis benchmarks, IJB-A and Adience. They find that these datasets are overwhelmingly composed of lighter-skinned subjects, and introduce a new facial analysis dataset that is balanced by gender and skin type. They evaluate three commercial gender classification systems using their dataset and show that darker-skinned females are the most misclassified group, with error rates of up to 34.7%. The maximum error rate for lighter-skinned males is 0.8%. [BG18, p. 1]

In many databases that algorithms in law enforcement use for identification of suspects people of colour are overly represented and thus matched more frequently. This leads to a disproportionate number of both true and false accept. [BL19, p. 323-324] This leads to further discrimination when innocent people are stopped, searched or even arrested. Algorithms working with this kind of data reflect the hidden and historical biases in society that get transferred trough the training data. [CM16, p. 8]

4 How to fight it

To prevent discrimination in algorithms, there are a number of steps that can be taken. One step is to ensure that the training data used to train algorithms is representative of the population it will be used on. This can be achieved by using diverse and inclusive data sets, or by using techniques such as data augmentation to make the training data more representative. Data augmentation is a technique used to increase the amount of training data by adding modified copies of existing data or newly created synthetic data. The aim of this technique is to reduce overfitting and promote better generalization when training machine learning models. [Nan+, p. 2] In the following Figure 4.1 an example of a image data augmentation on an existing image of a person can be seen. This image gets modified in different ways to add more data.

Figure 4.1: Example data augmentation [Sin20]

Bibliography

[BG18] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Conference on Fairness, Accountability and Transparency. ISSN: 2640-3498. PMLR, Jan. 21, 2018, pp. 77–91. URL: https://proceedings.mlr.press/v81/buolamwini18a.html (visited on 01/20/2023).

[BL19] Fabio Bacchini and Ludovica Lorusso. "Race, again: how face recognition technology reinforces racial discrimination". In: *Journal of Information, Communication and Ethics in Society* 17.3 (Jan. 1, 2019). Publisher: Emerald Publishing Limited, pp. 321–335. ISSN: 1477-996X. DOI: 10.1108/JICES-05-2018-0050. URL: https://doi.org/10.1108/JICES-05-2018-0050 (visited on 01/14/2023).

[CM16] DJ Patil Cecilia Muñoz Megan Smith. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. • . Executive Office of the President, 2016.

[Kha+18] Salman Khan et al. *A Guide to Convolutional Neural Networks for Computer Vision*. Synthesis Lectures on Computer Vision. Cham: Springer International Publishing, 2018. ISBN: 978-3-031-00693-7 978-3-031-01821-3. DOI: 10.1007/978-3-031-01821-3. URL: https://link.springer.com/10.1007/978-3-031-01821-3 (visited on 01/20/2023).

[Nan+] Loris Nanni et al. "Feature transforms for image data augmentation". In: *arXiv preprint arXiv:2201.09700* ().

[Sin20] Manmeet Singh. *Face Data Augmentation Techniques*. Medium. May 25, 2020. URL: https://manmeet3.medium.com/face-data-augmentation-techniques-ace9e8ddb030 (visited on 01/20/2023).