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

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 Figure 2.1 we can see a simplified illustration of a neural network 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.

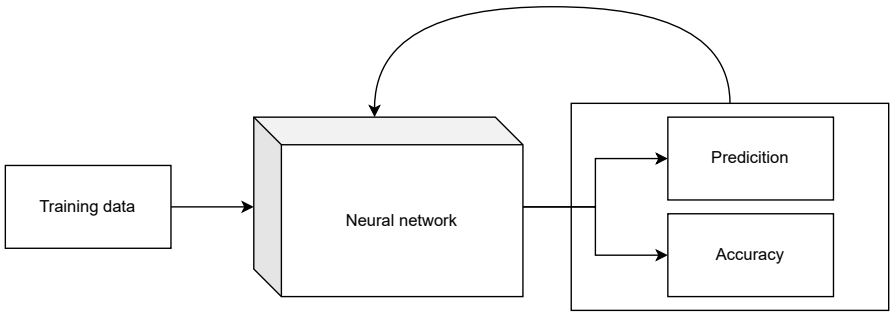

Figure 2.1: Neural network as blockbox with training data

On the other hand, is the test data. In Figure 2.1 we can see a simplified illustration of a neural network 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]

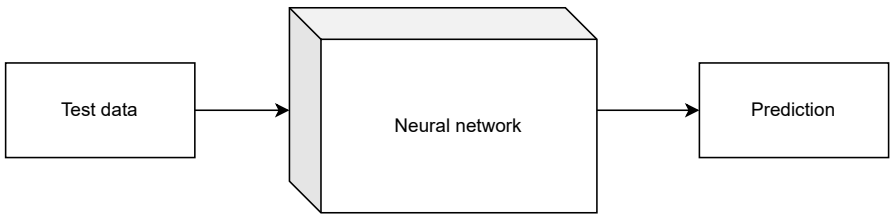

Figure 2.2: Neural network as blockbox with test data

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. As Muñoz, Smith, and Patil [MSP16, p. 7] describe it, poorly selected data is "where the designers of the algorithmic system decide that certain data are important to the decision but not others. (...) Such issues can be regarded as qualitative errors, where human choices in the selection of certain datasets as algorithmic inputs over others are ill-advised (...) resulting in potentially discriminatory effects."

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. [MSP16, p. 8] For example, the training data for facial recognition AI algorithms is often composed of a higher number of faces from white people than from other races. This leads to difficulty in recognizing the faces of people from other races. In a study by Buolamwini and Gebru [BG18] 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]

People of color are disproportionately represented in the databases used by law enforcement for suspect identification, leading to a higher frequency of matches and a disproportionate number of true and false acceptances. [BL19, p. 323-324] This can result in further discrimination when innocent individuals are stopped, searched, or arrested. These algorithms perpetuate the hidden, historical and systemic biases present in society that are transferred through their training data. [MSP16, p. 8]

4 Solution approaches

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+22, 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]

In the case of facial recognition technology, researchers and engineers can use techniques such as cross-dataset evaluation to test the accuracy of facial recognition algorithms on a diverse range of individuals. Cross-dataset evaluation is a method used to evaluate the performance of machine learning models on data that is different from the data that was used to train the model. This is important because a model that performs well on the training data may not perform well on new, unseen data. Cross-dataset evaluation helps to identify how well a model is likely to generalize to new data, and can be used to identify potential issues such as overfitting or bias in the training data. This can help to identify and address any issues of bias or discrimination in the algorithm. [ARLC22, p. 1-2] [Che+20, p. 3679-3680]

Pessach and Shmueli [PS22, p. 8-.10] state that generally there are three fairness-enhancing mechanisms in machine learning:

- **Pre-process** mechanisms involve changing the training data before it is fed into an machine learning algorithm, such as changing labels, reweighing instances, or modifying feature representations to make the classification fairer.
- **In-process** mechanisms involve modifying the machine learning algorithm during the training time, such as adding regularization terms, constraints, or adjusting decision tree split criteria to account for fairness.
- **Post-process** mechanisms are techniques used to adjust the predictions or threshold for classification made by a machine learning model after it has been trained, in order to reduce bias in the model's predictions. These techniques involve modifying the output of the model in some way, rather than changing the model itself.

Promoting diversity and inclusivity in the field of AI and machine learning is crucial in preventing discrimination in algorithms. This includes actively seeking out and encouraging individuals from underrepresented groups such as women, people of color, and people with disabilities to enter the field, and providing them with the necessary resources and support to succeed. This can be achieved through initiatives such as mentorship programs, targeted recruitment efforts, and providing access to education and training opportunities. Creating a culture of inclusivity within the field is important, by fostering an environment where all voices are heard, respected, and valued, and actively working to address and eliminate discrimination and bias within the field.

Regular audits and evaluations of algorithms are also critical in preventing discrimination. These audits and evaluations can help to identify any issues of bias or discrimination in the algorithms, and ensure that they are performing as intended. The evaluations can also be used to track the progress of the algorithm over time and identify areas for improvement. Additionally, regular audits and evaluations can help to build trust and transparency with stakeholders, including customers, regulators, and the general public.

5 Conclusion

In conclusion, discrimination in algorithms is a growing concern in the field of AI and machine learning. Discrimination can occur in a number of ways, including bias in the training data or the algorithm itself. Facial recognition technology is one specific area where discrimination has been identified, as studies have found that facial recognition algorithms are less accurate when analyzing images of people with darker skin tones. Preventing discrimination in algorithms is a complex problem that requires a multifaceted approach. It is important to use diverse and inclusive data sets, and techniques such as fairness enhancing mechanisms, cross-dataset evaluation, and transparent algorithms. It is also crucial for policymakers to be aware of the issues of discrimination in algorithms and work to prevent discrimination in algorithms. Additionally, it is important to promote diversity and inclusivity in the field of AI and machine learning, to promote accountability and transparency, to use human feedback, and to engage with communities and stakeholders. By taking these steps, we can work towards ensuring that algorithms are fair, equitable, and do not perpetuate discrimination.

Bibliography

[ARLC22] Said Al-Riyami, Alexei Lisitsa, and Frans Coenen. "Cross-Datasets Evaluation of Machine Learning Models for Intrusion Detection Systems". In: *Proceedings of Sixth International Congress on Information and Communication Technology*. Ed. by Xin-She Yang et al. Vol. 217. Series Title: Lecture Notes in Networks and Systems. Singapore: Springer Singapore, 2022, pp. 815–828. ISBN: 9789811621017 9789811621024. DOI: 10.1007/978-981-16-2102-4_73. URL: https://link.springer.com/10.1007/978-981-16-2102-4_73 (visited on 01/21/2023).

[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).

[Che+20] Yiran Chen et al. "CDEvalSumm: An Empirical Study of Cross-Dataset Evaluation for Neural Summarization Systems". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Findings 2020. Online: Association for Computational Linguistics, Nov. 2020, pp. 3679–3691. DOI: 10.18653/v1/2020.findings-emnlp.329. URL: https://aclanthology.org/2020.findings-emnlp.329 (visited on 01/21/2023).

[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).

[MSP16] Cecilia Muñoz, Megan Smith, and DJ Patil. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. • . Executive Office of the President, 2016.

[Nan+22] Loris Nanni et al. "Feature transforms for image data augmentation". In: *arXiv preprint arXiv:2201.09700* (2022).

[PS22] Dana Pessach and Erez Shmueli. "A Review on Fairness in Machine Learning". In: 55.3 (2022). ISSN: 0360-0300. URL: https://doi.org/10.1145/3494672.

[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).