A REPORT

On

BIAS in ML SYSTEMS

By

Aswin Benedict 2019A4PS0579P

Varkeychan Jacob 2017B5A70828P

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

(DECEMBER 2021)

Table of Contents

Introduction

Bias is a phenomenon that occurs when the algorithm produces results that are systematically prejudiced due to assumptions in the learning process. Simply put, it means that a biased ML output occurs due to partial training data. But this doesn't mean that; the data itself isn't inherently biased, but the real reason is that humans are interfering in the learning processes of ML models. The underlying biases are reflected in the form of inaccurate results in ML systems.

We will also be looking into bias in image recognition, labeling, and problems caused by it using a case study. For example, a photograph of a woman smiling in a bikini is labeled a slattern, slut, slovenly woman. A young man drinking beer is categorized as an alcoholic. A child wearing sunglasses is classified as a failure, loser, non-starter, unsuccessful person. You're looking at the "person" category in a dataset called ImageNet, one of the most widely used training sets for machine learning. We will also discuss the case when ImageNet was forced to remove 600,000 images from its database due to exposed racial bias in the program's AI system.

In this report, we discuss and focus on the different types of bias, the problems caused due to the presence of bias, how to identify bias, and techniques to mitigate bias.

Bias

2.1 Types of bias

Let's look into some of the primary and most common types of Bias in ML:

- **Measurement Bias:** is linked to problems with the accuracy of the training data and how it is measured or assessed. An experiment containing inaccurate measurement or collecting data the wrong way will create measurement bias and thus a biased output. For example, when testing a feature on a mobile app available both for Android and iPhone users, if you experiment only with android users, the results cannot be truly reflective, thus introducing measurement bias in the experiment.
  Thus, from a statistical point of view, Bias can be defined as a systemic deviation caused by the inaccurate estimation or sampling process.

- **Recall Bias:** The rate of how many unseen points a model labeled accurately over the total number of observations is defined as recall in ML. This error occurs in the labeling stage, and it appears when labels are irregularly given based on superficial observations. It is also called the false positive rate. For example, imagine a group of people sharing the hours of sleep they had the previous week; since they can't tell the precise amount, the estimation can take away from the actual values, resulting in recall bias.

- **Data Bias:** The simple definition of data bias is that the available data does not represent the population. Except for data generated by carefully designed and randomized experiments, most generally produced datasets are biased. For example, let's look into the Bias in reviews on Twitter or amazon, Wikipedia entries, etc. Only a tiny proportion of people contribute to this content, and their opinions and preferences are unlikely to reflect the population's opinions as a whole.

- **Algorithmic Bias:** refers to those biases added by the algorithm itself but were not present in the input data. The algorithm will also produce biased output if the input data is biased. It is difficult to define how an algorithm should proceed even after all the possible biases were found.
  "Tag recommendations" is an example of algorithmic bias. Imagine an algorithm where a user uploads a photo and adds various tags. A tag recommendation algorithm then suggests tags used in other images based on collaborative filtering. The user chooses the best-fit tag, enlarging that set of tags. This sounds simple, but a photo-hosting website should not include such functionality. This is because the algorithm needs data from people to improve, but as people use recommended tags, they add fewer tags of their own, picking from among available tags while not adding new ones. In short, the algorithm is killing itself.

- **Bias on User Interaction:** occurs from two sources, the user interfaces and the user's own self-selected biased interaction. The first case is the presentation bias, where whatever is presented to the user is selected, whereas everything else will be ignored. One good example is the recommendation system, like YouTube recommendation where about a hundred videos are recommended for the user based on his interaction, but this number is minuscule when compared to the millions that could be offered.

  Another bias due to the user interface is known as the **position bias**. The first focusing point can vary depending on the user's cultural background. For example, in western culture, we read from top to bottom, and we are biased to look first towards the top left corner, thus causing that corner to get more attention than the rest. **Ranking bias** is also a similar bias where bias occurs at the ranking of the websites by the web search engine, prompting the top-ranked website to get more attention.

2.2 Problems caused by bias, A case study

According to a recent survey, bias is one of the top two issues in AI/ML today, affecting nearly half of industry workers. Only 15% of AI/ML teams are working on this problem.

It is known that to build AI or an ML model, data is required; supervised ML systems designed for object and face recognition are trained on numerous amounts of data within many datasets that are made up of many discrete images. But if we look at the images used for training the ML model, we find a bedrock of shaky and skewed assumptions. To see these biases at work, let's look at ImageNet, one of the most iconic training sets of all time. The idea behind ImageNet was to map out the entire world of objects; they used an army of workers to sort and label an average of 50 images per minute into thousands of categories. ImageNet consisted of over 15 million labeled images classified into more than 20 thousand categories when the project was completed. It was the colossus of image recognition for ML for a long time. But soon, the flaw of ImageNet came to light, the absence of the category "person."

In ImageNet, for example, the category "human body" falls under the branch: Natural Object > Body > Human Body.

- Human body
    - "male body"
    - "Person"
    - "juvenile body"

- “adult body”
  - “adult female body”
  - “adult male body.”
- “female body.

Here we can see the implicit assumption that only male and female bodies are natural. There is also a category in ImageNet for the term “Hermaphrodite” situated in Person > Sensualist > Bisexual > alongside the classes “Pseudohermaphrodite” and “Switch Hitter.”. Some LGBTQ-themed books were also classified under “Abnormal Sexual Relations, Including Sexual Crimes.” From all these examples, we can see the extent of racial and sexist bias in ImageNet labeling. It was so bad that at some point if you had searched for American citizens, only fair-skinned individuals’ images were shown in the images section.

If we go further into the depths of ImageNet's Person categories, humans' classifications take a dark turn. There are categories for Bad Person, Call Girl, Drug Addict, Closet Queen, Convict, Crazy, Failure, Flop, Hypocrite, Jezebel, Kleptomaniac, Loser, Melancholic, Nonperson, Pervert, Prima Donna, Schizophrenic, Second-Rater, Spinster, Streetwalker, Stud, Tosser, Unskilled Person, Wanton, Waverer, and Wimp. There are many racist slurs and misogynistic terms. This went unnoticed since ImageNet was primarily used for object recognition. So, the Person category was rarely discussed, nor has it received enough public attention. However, this complex architecture of images of real people, tagged with often offensive labels, has been publicly available on the internet for over a decade. It provides an actual example of the complexities and dangers of human classification. One such example is a photograph showing a

dark-skinned child wearing a dirty cloth and clutching a stained doll that was simply labeled as a “toy.” But some labels are just nonsensical. A woman sleeps in an airplane seat, and her right arm protectively curls around her pregnant stomach. The image is labeled “snob.” A photoshopped picture shows a smiling Barack Obama wearing a Nazi uniform, his arm raised and holding a Nazi flag. It is labeled “Bolshevik.”

From the above examples, it is understandable how serious this bias is and its demographical, cultural, racial, gender, and historical implications on the internet and its users. Thus, ImageNet removed 600,000 images of people stored in its database due to racial, sexual, and cultural biasing.

Another such distinct example is that of the instance of Tay, Microsoft's AI chatbot under development. Microsoft created Tay to connect with 18–24-year-olds on Twitter, get informal speech training, and react in a near-human manner. Cybercriminals trained Tay to repeat racist, sexist, and anti-Semitic remarks within hours using its content-neutral algorithms.

2.3 Techniques to mitigate bias

There has been a recent hike in research and developments on fairness and bias in ML models. But achieving it is not that easy since it is a complex and multifaceted concept that depends on context, culture, and history. Remember that it is impossible to attain fairness across all definitions simultaneously. Therefore, even though we can try to mitigate bias and achieve fairness, it is impossible to achieve fairness in all metrics.

Some of the techniques to mitigate bias are as follows:

- **Choosing the correct learning model:** There are two sorts of learning models, each with its own set of advantages and disadvantages. In a supervised model, the stakeholders that produce the dataset have complete control over the training data. Ascertain that this group of stakeholders is evenly distributed and that they have attended unconscious bias training. Unsupervised models, on the other hand, rely on the neural network to detect bias tendencies. This implies that there should be some difference between the input data and the output result and bias prevention strategies so that the neural network can learn to distinguish between biased and non-biased data.

- **Using the suitable training dataset:** Machine intelligence is only as good as the data used to train it. The training data you give the neural network must be comprehensive, balanced, and free of humans' biased predispositions. It should also replicate real-world conditions like

demographic makeup. A good rule of thumb is to avoid reusing datasets — data from a place with a varied ethnic population, for example, cannot be used to a region with a larger one race population, and vice versa.

- **Performing the data processing mindfully:** There are three forms of data processing in machine intelligence: pre-processing, in-processing, and post-processing. When pre-processing datasets, bias might come in during formatting before they're fed into the neural network. In this phase, any data that potentially induce bias should be eliminated. Because the input is altered as it flows through the neural network, the weighting of the neural nodes must be right to avoid a biased output with in-processing. Finally, when interpreting data for human consumption in the post-processing step, make sure there is no bias.

Toolkits to Mitigate Bias

- **AI Fairness 360** (AIF360) is an extensible open-source toolkit for detecting, understanding, and mitigating algorithmic biases. It was designed with two main goals, ease of use and extensibility. Converting raw data into a fair model should be as easy as possible for the user.
- **Fairness Measures** is another toolkit that provides several fairness metrics like the difference of means, disparate impact, and odds ratio. They also offer some datasets that can only be accessed with the owners' explicit permission.
- **FairTest** approaches detecting biases in a dataset by checking for associations between predicted labels and protected attributes. This method is also helpful in identifying the regions of input space where the algorithms can have high errors.
- **Aequitas** is another auditing toolkit for data scientists and policymakers; it has a Python library and an associated website to upload data for bias analysis. It offers several fairness metrics, including demographic or statistical parity and disparate impact, along with a "fairness tree" to help users identify the appropriate metric to use for their particular situation.

Conclusion

From the reference papers provided, it is evident that making ML models effectively and ethically duplicating human decision-making processes is tough. The danger of bias is a significant problem for ML training and application. From all the examples mentioned above, we can see the severity and impact of the issues caused by bias. So mitigating or preventing bias in ML models becomes an integral part of modeling an ML system, even though fairness is a multi-faceted context-dependent construct that defies simple definition.

References

[1] Baeza-Yates, R. Bias on the Web. Commun. ACM 61, 6 (June 2018), 54–61.

[2] Ali, M., Sapiezynsk, P., Bogen, M., Korolova, A., Mislove, A., Rieke, A. Discrimination through optimization: how Facebook's ad delivery can lead to biased outcomes. In Proceedings of the ACM on Human-Computer Interaction 3 (2019); https://dl.acm.org/doi/10.1145/3359301

[3] Buolamwini, J., Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of Machine Learning Research 81 (2018), 1–15;

http://proceedings.mlr.press/v81/buolamwini18a/ buolamwini18a.pdf

[4] Crawford, K., Paglen, T. Excavating AI: The politics of images in machine learning training sets. The AI Now Institute, New York University, 2019;

https://www.excavating.ai

[5] Raji, I., Buolamwini, J. Actionable auditing: investigating the impact of publicly naming biased

performance results of commercial AI products. In Proceedings of the 2019 AAAI/ACM Conf. AI, Ethics, and Society, 429–435; https://dl.acm.org/ doi/10.1145/3306618.3314244

[6] Small, Z. 600,000 images removed from AI database after art project exposes racist bias.

Hyperallergic, 2019;

https://hyperallergic.com/518822/600000-imagesremoved-from-ai-database-after-art-projectexposesracist-bias/

[7] Barocas, S., Hardt, M., Narayanan, A. Fairness and machine learning: limitations and opportunities,
2019; https://fairmlbook.org
[8] Bellamy, R.K.E. et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. 2018, arXiv; https://arxiv.org/abs/1810.01943
[9]Zehlike, M., Castillo, C., Bonchi, F., Hajian, S., and Megahed, M. Fairness Measures: Datasets and software for detecting algorithmic discrimination, 2017. URL http://fairness-measures.org/
[10].Adebayo, J. A. ”FairML : Toolbox for diagnosing bias in predictive modeling”. Master’s thesis, Massachusetts Institute of Technology, 2016. https://github. com/adebayoj/fairml.
[11]Stevens, A., Anisfeld, A., Kuester, B., London, J., Saleiro, P., and Ghani, R. Aequitas: Bias and fairness audit, 2018. URL https://github.com/dssg/ aequitas. Center for Data Science and Public Policy, The University of Chicago.