Sources of Bias in Machine Learning Models and Methods to Mitigate Them

A Research Paper submitted to the Department of Engineering and Society

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

Srinivasa Josyula
Spring 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

Srinivasa Josyula
Richard Jacques, Department of Engineering and Society

Introduction

In the last ten years, the world has seen an explosion of machine learning applications. Machine learning is found everywhere from voice assistants and search engines to self-driving cars and everything in between. At the core of every machine learning algorithm is a plethora of data from which the model learns and then is able to make predictions on previously unseen data. Therefore, everything the model "knows" comes from the information in the dataset. Moreover, due to the increase in computing power over the years there has been a focus on creating larger models and feeding these models with ever larger datasets. With this increasing complexity, it is often hard to see how the model variables or dataset information are used in the final prediction. To explore these issues, I chose my technical project to be in the general area of information theory and how it relates to machine learning. I am interested in pursuing this topic because it goes against the prevailing philosophy in machine learning that more data is better. Employing more useful data can lead to faster training and lead to computational savings (Lazzaro, 2021).

In computer science, information has a precise mathematical definition, but for my use case it can be thought of as the contribution of a data point to the machine learning model's effectiveness (Shannon Entropy). Recently, a new metric was introduced to measure the information in a dataset when trained on a specific model family. This metric, coined V-information, is an extension of previous works based in information theory. V-information also extends to pointwise V-information which is the information gained from each data point. In my technical project, I will be taking a closer look at this new metric and evaluating its robustness in various scenarios. This work will help identify areas in a dataset that have a substantial impact on the model and could potentially introduce unwanted bias into the model. This will help researchers improve their datasets to ultimately provide equal treatment to all users.

As the field of machine learning has grown over the last decade, there has been a growing sense of concern over these models stemming from various issues. A big issue that has surfaced in the last few years has been that machine learning models have given results that discriminate against certain groups. For example, in some instances facial recognition software has not worked for people of color (The Regulatory Review, 2021). In other cases, voice assistants have worked better for people speaking in some accents of English. There are many ways in which this bias can creep into the models. In some instances, the developers of the model might be unaware of the users who will use this model which might lead them to create a flawed application. In other cases, the data that is used to train the model is biased and results in the model being biased. I am interested in this second issue because it relates to my technical project on information theory and how machine learning models extract information from data. It is quite easy to point out that biased data is the issue, but it is much harder to determine why certain datasets are biased. Especially in a world where datasets can span millions to billions of rows, the bias problem is especially hard to track down. This issue is important because technology is meant to unify individuals across class and racial divides, not drive them further. A recent paper shows that machine learning bias can even have an impact on healthcare, where a patient's history might cause the model to behave differently (Huang, 2022). Due to the far-reaching impact of machine learning bias, I want to explore how data has affected the way in which machine learning models behave, and how can we make sure that the data we input into the model is not biased? Additionally, I will also explore the work that has been done in identifying biased models and the solutions proposed in fixing them.

Literature Review

In this section, I will discuss several papers that are relevant to my research. First, we need to understand the definition of information in computer science. There is a precise mathematical definition of information where information is measured in bits. For simplicity, we can consider a coin flip which has 1 bit of information. This is because flipping a coin will either result in heads or tails. The first paper, "A Theory of Usable Information Under Computational Constraints" extends the idea of information to machine learning dataset to a new metric called V-information (Xu, Y., 2020). This metric is calculated by training the model with the actual dataset and with a null dataset and determining the information discrepancy between the two models. In most natural language processing cases, the null dataset is simply an empty string. By comparing these two models, it is possible to determine which datasets provide the most useful information compared to an empty dataset. A paper was shortly published after the original called, "Understanding Dataset Difficulty with V-Usable Information" (Ethayarajh, K., 2022). This paper extended the idea of V-information, to pointwise V-information which can be used to calculate the information provided by each datapoint rather than just the entire dataset. Moreover, it pioneered the idea that rather than testing various models on one dataset, we can assess the effectiveness of various datasets on one model. This work also helps determine which slice of the dataset could be most useful for training. The ideas presented in this section will be critical to understand how researchers can identify sections of data that are biased. This can have huge social implications for applications with algorithmic bias.

These works did not consider the implications of using v-information as a way to analyze the bias in datasets. In this paper, I will discuss how v-information can be used by software developers and healthcare professionals.

Methodology

To analyze bias in machine learning datasets, I narrowed down my research question into three major aspects. What type of bias exists in machine learning models? How do datasets contribute to this bias? How can we work to identify and decrease bias? To answer these questions, I will be using qualitative data from various articles. They will come from primary sources because I will explore different instances of bias in machine learning from previous research. The data will be descriptive because I will be exploring bias that already exists in models rather than do experiments to explore machine learning bias. By closely exploring different cases of machine learning bias in software, I will be able to gain a better understanding of my research question.

I will use the reading and synthesis method to understand this topic by reading various texts on the issue of machine learning bias and understanding what leads to biased datasets. I seek to understand if models or datasets are deliberately made to be biased for certain groups to gain an advantage. By reading various texts I hope to understand the common pitfalls in the network that lead to biased models and how to amend them.

I will explore this topic through the Actor-Network Theory (ANT) framework. As part of this theory, the relationship between everything in a network including humans, technology and inanimate objects is explored. In my case, the technology includes the various machine learning models and the datasets used to train them. The inanimate objects are the devices through which humans interact with these models, such as through an app or website. The humans in this network include both the end users served by the machine learning models and the developers that were part of designing these models. There needs to be a proper understanding of all the moving parts to make sure that the users in this network feel accurately represented (Cresswell, K., 2010).

Analyzing Various Cases of Bias

The first case of bias I will explore is in facial recognition technology. Most of us have used facial recognition software to unlock our phone or authenticate purchases. However, facial recognition is also used extensively by the government for surveillance, airport screening, and employment decisions. A popular study in 2018 called "Gender Shades" explored the racial bias in facial recognition software developed by IBM and Microsoft. The research done in this study found that the software had a 34% higher error rate in darker-skinned women than light-skinned men. This was also confirmed in a study done by the National Institute of Standards and Technology (NIST). It found that across 189 different algorithms, women of color had the least accurate results. These results have huge implications for law enforcement because people of color are disproportionately identified incorrectly.

The most obvious source of this bias is the lack of diverse datasets and overrepresentation of white and male participants. In the justice system however, the Black people are overrepresented in mugshots which created a feed-forward loop that leads to racist-policing. Moreover, most default camera settings do not capture darker skin tones properly resulting in lower quality images for people of color. This results in weaker training results for people of color. In a technical sense, identifying these flawed datasets and making changes to make them more equitable will create models with less bias. However, due to the considerable number of methodological biases (female, dark-skinned, young) that these algorithms must account for, it is entirely possible for some groups to be underrepresented. Therefore, there must also be a legislative push to monitor and decrease the reliance on these algorithms in public use settings. The repercussions of using a flawed model are much higher in the justice system than on a phone. There must be laws that hold companies accountable for creating equitable algorithms

and slowly decrease the reliance on these algorithms until concrete methods are found to effectively counter this bias. Although facial recognition technology holds massive potential for use in the criminal justice system, it is also necessary to be aware of its shortcomings. Until this technology becomes more sophisticated to deal with the nuances of society, we must learn to decrease our reliance on it (Najibi, Alex, 2020).

Another prevalent example of machine learning bias is in voice assistants. Quite often, these softwares are much better at recognizing some accents. One recent study even showed that all major voice assistants are around twice as likely to transcribe black speakers incorrectly compared to white speakers. This is because the software effectively censors phrases that it does not understand because it was never trained on other similar phrases. This issue is remarkably similar to the one faced by facial recognition software. The lack of representative datasets results in software that only works effectively in certain use cases while working with various levels of accuracy for the rest of the use cases. Once again, the hypothetical solution to this problem would be to create more inclusive datasets that capture a wider range of English accents. However, this can be quite an expensive and extensive task, considering there are close to 200 different dialects/accents of English. Moreover, for many companies such as Google and Microsoft it is not financially feasible to create such large datasets and re-train their models. They are more likely to create models that cater to a wide range of users. It is entirely possible that we must invent new methods where machine learning algorithms can learn how to identify various dialects from just a few training examples. However, voice assistant software is not used for critical applications compared to facial recognition software. Therefore, if creating an all-encompassing software is not feasible, companies should focus on creating more user-friendly software that informs the user if they are not understood. Once again, we must learn how to

balance the usefulness of voice assistants with their clear drawbacks. Companies must learn how to be transparent about the bias that exists in the models (Lopez-Lloreda, C., 2020).

Another area in which machine learning models fail to equally treat various demographics is in healthcare. There are numerous technologies ranging from artificial intelligence (AI) chatbots to cancer detection softwares that aim to leverage the power of machine learning to help patients. However, over time as these models are continually trained on new data, they become increasingly biased because of the data that is fed into them. For example, some systems underestimate the severity of illness in Black patients, so they are given lower priority in treatment. In other instances, female patients are more likely to be misdiagnosed for heart disease. The exact reason for these inequalities is hard to pinpoint because these algorithms read data from a variety of sources. When researchers attempt to increase the fairness in their algorithms, they seek to achieve a mathematical balance between all the groups in the dataset. In the case of cancer detection, decreasing the threshold for disadvantaged individuals would cause some "low-risk" cases to be classified as "high-risk". This would result in more people from the disadvantaged group coming to clinics to get more comprehensively evaluated. This would only partially solve the issue because it would still decrease the overall accuracy of the model because a lot more individuals would be falsely screened for cancer. In many of these cases, trying to create a fairer algorithm balances out the diverse groups at the cost of one group's accuracy declining. This shows us that merely going after mathematical fairness will not always benefit each group in the dataset. Technical solutions can patch specific issues, but not fix the underlying issues that exist in the system. A real solution will involve a long-term collaboration between healthcare providers and software developers to improve access to healthcare technology and create more diverse datasets. There is also a need to continually

monitor the performance of healthcare algorithms on various groups to ensure equality (Watcher, S., 2023).

After analyzing these various cases of machine learning algorithms, it is clear that bias is prevalent in many applications. In all these cases, some groups were disproportionately represented in the dataset which caused the model to fail poorly on some groups. These issues occur when machine learning models are used for complex societal problems. The issues are so complicated that datasets do not appropriately encompass the topic. Often, a technical solution is only enough to patch the issue but not sufficient for a long-term solution. That requires collaboration from multiple parties to identify the societal and software issues. It also requires constant monitoring to ensure that training the models does not introduce new biases. We also need to consider the feasibility of training large models with fair data because there are so many distinct groups to account for. In this case, companies and researchers should be transparent about the extent to which their algorithms work effectively.

Conclusion

There has been a steady rise in machine learning algorithms in recent years. These algorithms have been used in almost every facet of life to improve efficiency and increase access to resources. Some everyday examples of machine learning include facial recognition, digital assistants, and speech recognition. Machine learning has also been used in the medical and justice fields. All these algorithms learn from existing data to make decisions about a new instance. Because of the complexity of these tasks, the data is often biased causing the machine learning models to also exhibit signs of bias. This bias can sometimes be intentional and other times it highlights already existing issues within our community. In this paper, I explored the bias in facial recognition software, voice assistants, and medical analysis software. This analysis

proved that this bias is mainly caused by flawed datasets that disproportionately represented some groups over others. However, it was less clear why the datasets were flawed. In some cases, the developers neglected to train their models on some groups. This was true in the case of facial recognition where dark skinned individuals were neglected. In other cases, the problem space was too large to feasibly create an equitable dataset. This was encountered in the case of voice assistants where the data required to encapsulate all the different accents would be too large. Finally, the bias in medical diagnosis software was caused by underlying issues in the healthcare system that led to the algorithms becoming biased over time. Therefore, in some cases a technical change could fix the issue. This can include expanding an existing dataset or model. In other cases, the companies and researchers must be transparent about their software and inform the users of its shortcomings. This will put users at ease about the software and give them better information about which applications they can use. There is also a need to invent new methods that can efficiently train small datasets. This will take us a step further in tackling problems where extremely large datasets are currently required to create an accurate model. The most common case of machine learning bias exists due to underlying issues in the system. In the healthcare industry, there is an existing problem of disproportionate access, especially among people of color. These issues creep into the datasets and eventually make the models act in a biased manner. A simple technical solution will not work because healthcare datasets are continually updated so existing biases would once again become incorporated into the model. A long-term solution needs to include collaboration between the healthcare industry and software developers to create improved access to healthcare in conjunction with equitable datasets.

My technical project can help with this long-term process by allowing developers to pinpoint areas of the dataset that have a substantial impact on the model. This can be achieved by

analyzing the v-information in each dataset to determine areas of high impact. This can also be relayed to healthcare professionals who can analyze how the data was introduced into the model. Using this information, they can strive to improve those aspects of the data in their clinical setting. Ultimately, this will result in more equitable data being produced in healthcare settings which will eventually lead to fair models. This approach can also be used in other instances where machine learning models are used to solve complex societal challenges.

References

Cresswell, K. M., Worth, A., & Sheikh, A. (2010, November 1). *Actor-network theory and its role in understanding the implementation of information technology developments in healthcare - BMC Medical Informatics and decision making*. BioMed Central. Retrieved October 31, 2022, from https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-10-67

Ethayarajh, K., Choi, Y., & Swayamdipta, S. (2022). Understanding Dataset Difficulty with V-Usable Information. *ICML*. https://doi.org/10.48550/arXiv.2110.08420

Huang, J., Galal, G., Etemadi, M., & Vaidyanathan, M. (2022). Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review. *JMIR Medical Informatics*, *10*(5). https://doi.org/10.2196/36388

Lazzaro, S. (2021, June 21). *Machine Learning's rise, applications, and challenges*. VentureBeat. Retrieved October 31, 2022, from https://venturebeat.com/ai/machine-learnings-rise-applications-and-challenges/

Lopez-Lloreda, C. (2020, October 1). *How speech-recognition software discriminates against minority voices*. Scientific American. Retrieved April 7, 2023, from https://www.scientificamerican.com/article/how-speech-recognition-software-discriminates-against-minority-voices/

Najibi, Alex. (2020, October 26). *Racial discrimination in face recognition technology*. Science in the News. Retrieved April 6, 2023, from https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

The Regulatory Review. (2021, August 6). *Facing bias in facial recognition technology*. The Regulatory Review. Retrieved October 31, 2022, from https://www.theregreview.org/2021/03/20/saturday-seminar-facing-bias-in-facial-recognition-technology/#:~:text=According%20to%20the%20researchers%2C%20facial,particularly%20vulnerable%20to%20algorithmic%20bias.

Shannon entropy. Shannon Entropy - an overview | ScienceDirect Topics. (n.d.). Retrieved October 31, 2022, from https://www.sciencedirect.com/topics/engineering/shannon-entropy

Watcher, S., Mittelstadt, B., & Russell, C. (2023, February 8). *Health care bias is dangerous. but so are 'fairness' algorithms*. Wired. Retrieved April 7, 2023, from https://www.wired.com/story/bias-statistics-artificial-intelligence-healthcare/

Xu, Y., Zhao, S., Song, J., Stewart, R., & Ermon, S. (2020). A Theory of Usable Information Under Computational Constraints. *ICLR 2020*.https://doi.org/10.48550/arXiv.2002.10689