PERSONALIZING FEDERATED LEARNING USING META-LEARNING

SOCIAL EQUITY ANALYSIS OF MACHINE-LEARNING BASED HIRING TOOLS

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 Science in Computer Science

By

Matthew Whelan

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

Joshua Earle, Department of Engineering and Society

Aidong Zhang, Department of Computer Science

Introduction

From *Frankenstein* to *Ex Machina*, mainstream popular culture simultaneously romanticizes and agonizes over the prospect of machines representing human qualities. Machines have overtaken humans in various core tasks, including analyzing data and recognizing trends. This is precisely why machine learning has turned into a useful tool in many of the industries that power our society today. Machine learning means building a model based on 'training' data to make predictions or decisions without being explicitly programmed to do so (Brown, 2021). But, this means that high quality data is vital for model performance, and for many models this can include sensitive information, which presents many security concerns.

With the proliferation of data, namely sensitive data, the world has experienced a rise in cyber crime, and a subsequent rise in data security methods. One prevalent method is storing data in inaccessible locations to eliminate the possibility of unauthorized access (Ruehle, 2021). Federated learning–a growing framework within the machine learning space–can protect sensitive data through such storing mechanisms, without compromising the ability to train a machine learning model. Therefore, federated learning represents a viable solution to critical privacy concerns while still processing data efficiently.

The technical research will focus on the current state of the federated learning field and analyze an industry leading platform, including its practical applications in real world use cases. My research question is: how can we craft a framework that supports data sharing in a distributed environment using principles of federated learning? The coupled STS research will investigate the racial and gender bias that machine learning brings into society and actions that can be taken to reduce this prejudice. My STS research question is: what measures can we take to reduce the racial and gender bias that machine learning brings into society? The motivation of this research is to consider both the benefits and drawbacks of machine learning, and how it may allow for a more equitable future.

Technical Project

The technical project is part of an ongoing research project funded by a grant from the National Science Foundation under Professor Aidong Zhang. As stated previously, federated learning offers an enticing method of protecting data by leaving it stored on edge devices and running machine learning algorithms locally. In this way, a central server can communicate with these edge devices, but instead of sending data, devices can send updates to the model parameters back to the server. The server then evaluates the updates sent from all the devices and averages them to improve the shared model. This process continues in an iterative fashion until training is complete.

Federated learning is already being researched in-depth and implemented by various organizations and institutions. Notably, Google is testing federated learning on the Google Keyboard (Gboard) for Android devices. The Gboard gives suggestions powered by the Google search engine based on the user's input. For example, if the user types in the name of a popular restaurant, the Google listing of that restaurant is returned with information like the address and phone number. When the Gboard displays a suggested query, the device stores information about the current context and whether the user clicked on the suggestion (McMahan, 2017). The model then communicates with the server and uses the device suggestion history to improve predictions, but only if the device is charging, connected to WiFi and idle to prevent any impact on performance. However, there are limitations to this process. Edge devices generally have much lower performance, as they have connections with higher latency and lower throughput. Thus, federated learning presents a complicated framework to work around in practice.

Many existing FL libraries are not sufficient for developing an integrated federated learning model. FedML, a leading FL platform, proposes an efficient and easy-to-use platform for developing and evaluating federated learning models. It addresses current needs in the industry in 4 key areas. The first area is computing in a variety of environments: on-device training for Internet of Things (IoT) devices, distributed computing and single-machine simulation. The second area is support of diverse federated learning configurations, as FedML offers a client-oriented programming interface to enable different network topologies. Figure 1 shows how different topologies, the structure of learning networks, can be customized with FedML.

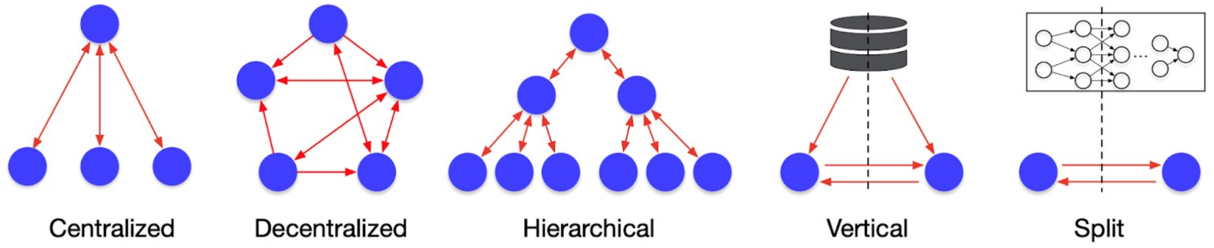

Figure 1: Various topology definitions in federated learning

FedML meets these diverse requirements by facilitating communication to any number of nodes during training. The third area is standardized implementations of status quo federated learning algorithms to emphasize code reusability and simplicity for the developer. The final area is standardized benchmarks with well-defined metrics and baseline results to ensure fair algorithm performance evaluation (He, 2022, pg. 3). These areas are often neglected by top federated learning libraries–alternatively, FedML offers these advantages in an open source format with constant development in tandem with the federated learning community. Therefore, although federated learning presents many challenges to developers, FedML offers an efficient and flexible means for implementing and evaluating federated learning algorithms.

STS Project:

Bias in Image Detection

Machine learning is increasingly used as an intelligent prediction tool around the world. No matter the application, models are fueled by data, so training on thorough, descriptive data is essential. The "learning" in machine learning means how a model absorbs characteristics of the dataset it is trained on. Therefore, if that data is biased, then the system very well could exhibit that same bias when it makes decisions. A study from researchers at MIT investigated services designed to analyze faces and identify characteristics like age and sex from companies like Microsoft and IBM. When tasked to identify sex, the companies had an error rate of no more than 1% for lighter skinned men and women, but for darker skinned women, this error soared to 21% and 35% for Microsoft and IBM, respectively (Buolamwini & Gebru, 2018, pg. 8). These biases arise precisely because of the flawed datasets used to train these models, which likely did not have equal representation of individuals across characteristics like sex and race.

This trend might not be limited to just these models, as a large portion of online available datasets are shown to have significant bias in their data. Buolamwini & Gebru (2018) also found that popular datasets, including National Institute of Standards and Technology (NIST) constructed dataset, had around 80% of training examples of lighter skinned people, with darker skinned females only representing 4.4% of the overall dataset (pg. 3). If datasets underrepresent certain social groups in training examples, it will only be harder to make accurate predictions for these social groups. Thus, models trained with this inadequate data cannot correctly learn how to identify marginalized groups. New research proposes a promising solution: "diversity in training data has a major influence on whether a model is able to overcome bias" (Zewe, 2022, pg. 1).

Diversity is an umbrella term for providing more descriptive data; in terms of image detection, this means showing faces from different viewpoints, so the network is better able to generalize to new faces. So, a solution for unfair data is to include all relevant social groups, as well as multidimensional data with enough perspectives for a model to truly understand what they are looking at.

Bias in Hiring Algorithms

Bias in machine learning models does not appear exclusively in image detection. Regrettably, models can reinforce stereotypes through active systems of power. This transpires in hiring algorithms, as Amazon's secret AI recruiting tool was found to discriminate against women. This algorithm penalized resumes that contained the word "women," as in "women's chess club," perpetuating the trend of women being underrepresented in tech hiring (Dastin, 2018, pg. 1). This is frightening not only due to its blatant discrimination, but because it is nearly impossible to see such instruments in action. AI recruiting tools are used by many companies to sift through thousands of candidates, but they are all non-user facing. Again, biased data, trained on inequalities in the tech industry, festers in poor, prejudiced models. These models, unlike facial recognition, have the power to decide the employment status of applicants, which could create a vicious cycle of unemployment and poverty. Therefore, to ensure machine learning systems are fair and accurate, creators must pay extremely close attention to the quality and diversity of the training data utilized. During design, creators must also examine the data's exclusion of potentially relevant social groups to avoid propagating existing biases.

Frameworks and Methodologies

To understand this idea more, this project will apply the Social Construction of Technology (SCOT) theory to investigate machine learning's impact on different social groups.

SCOT dictates that different social groups have different perceived problems with artifacts, but one specific stage of interest is the concept of interpretive flexibility. Pinch & Bijker (1987) define this as “flexibility in how people interpret artifacts but also in how artifacts are designed” (pg. 34). Clearly, social groups are impacted differently because of the particular design of recent models. In the characteristics of race and gender, I will investigate why these design flaws cause women and darker skinned people, in particular, to be disadvantaged in many models. In my project, the primary method of investigation will be reading and synthesizing previous literature particularly relevant to my topic. This will be the most effective way to conduct my research because there is a plethora of literature available online surrounding machine learning and its pitfalls. The timeline of this project is to first consider outside texts relevant to my project, then formulate arguments that can delve deeper into the nuanced biases of machine learning.

Key Texts

The first guiding text for my STS project is Joy Buolamwini and Timnit Gebru’s study on intersectional accuracy disparities in gender classification models. These researchers offer analysis of status quo models from big companies like Microsoft and IBM and how they are exhibiting racial and gender bias (Buolamwini & Gebru, 2018, pg. 3). They also look closer at the most used datasets and how many of them are not equitable for all social groups.

The second text is Adam Zewe’s article about research of data diversity and how that affects bias in machine learning models. This text will help me offer solutions to prevent biased models based on the type of data offered to the model (Zewe, 2022, pg. 1).

The third text is Jeffrey Dastin’s article on Amazon and other large corporations’ hiring algorithms using biased data. This will help me support my argument on biased models with an example outside of image detection, as hiring algorithms are more visible weapons of mass destruction as they directly prevent job opportunities to certain social groups (Dastin, 2018, pg. 2).

The final text is Sara Brown’s piece on machine learning and its functionality to society. She argues that data is the most important part of the process, with model selection as a close second; she explains general use cases for machine learning and also the major limitations in its applications to society. (Brown, 2021, pg. 1).

Conclusion

Machine learning is growing quickly in terms of usage and development in the discipline. This paper has focused on how the most crucial component of machine learning is data. Likewise, there are serious uncertainties going forward about both the privacy of data and the quality of data in a system. Federated learning is gaining traction as a solution to data privacy issues, and a platform that introduces new capabilities and fills the gaps in the industry is FedML. Moreover, the other core issue is the quality of data. Increasing the diversity of data, as well as the equality of groups represented, can have a profound impact on the fairness and accuracy of a machine learning model. These two guiding frameworks can assure that the core rights of privacy and social justice are maintained in our society. With the right oversight, these principles can ensure that machine learning will lead to a better tomorrow.

References

Brown, S. (2021, April 21). *Machine Learning, explained*. MIT Sloan. Retrieved October 10, 2022, from https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Buolamwini, J., & Gebru, T. (2018, February 4). *Gender shades: Intersectional accuracy disparities in commercial gender.* MIT Media Lab. Retrieved October 11, 2022, from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. Retrieved October 11, 2022, from https://www.reuters.com/article/us-amazon-comjobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

He, C., Li, S., So, J., Zeng, X., Zhang, M., Wang, H., Wang, X., Vepakomma, P., Singh, A., Qiu, H., Zhu, X., Wang, J., Shen, L., Zhao, P., Kang, Y., Liu, Y., Raskar, R., Yang, Q., Annavaram, M., & Avestimehr, S. (2020, November 8). *FedML: A research library and benchmark for Federated Machine Learning*. Retrieved October 10, 2022, from https://arxiv.org/abs/2007.13518

McMahan, B., & Ramage, D. (2017, April 6). *Federated learning: Collaborative machine learning without centralized training data*. Google AI Blog. Retrieved October 10, 2022, from https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Pinch, T. J. & Bijker, W. (1987). The Social construction of facts and artifacts. In The Social construction of technological systems: New directions in the sociology and history of technology. Cambridge, MA: MIT Press.

Ruehle, V., Sim, R., Yekhanin, S., Chandran, N., Chase, M., Jones, D., Laine, K., Köpf, B., Teevan, J., Kleewein, J., & Rajmohan, S. (2021, November 9). *Privacy preserving machine learning: Maintaining confidentiality and preserving trust*. Microsoft Research. Retrieved October 10, 2022, from https://www.microsoft.com/en-us/research/blog/privacy-preserving-machinelearning-maintaining-confidentiality-and-preserving-trust/

Zewe, A. (2022, February 21). *Can machine-learning models overcome biased datasets?* MIT News. Retrieved October 27, 2022, from https://news.mit.edu/2022/machine-learning-biased-data-0221