CEUR-WS.org/Vol-2812/RDAI-2021_paper_9.pdf

Addressing Unique Fairness Obstacles within Federated Learning

Annie Abay, Ebube Chuba, Yi Zhou, Nathalie Baracaldo, Heiko Ludwig
[1]IBM Research, Almaden
650 Harry Road
San Jose, California 95120

Abstract

Efforts to reduce social bias in machine learning has increased in the past several years. As data privacy concerns grow, finding techniques to train private, debiased machine learning models becomes increasingly important. Federated Learning (FL) has emerged as a popular privacy-preserving machine learning strategy. FL, however, by not providing complete access to training data, brings with it a unique set of difficulties in bias mitigation that have yet to be explored. In this paper, we delve into these difficulties, and how they can affect bias measured in federated learning models.

Introduction

Machine learning applications are used throughout the world to make decisions on matters with real-world consequences [1]. These decisions can affect whether a patient gets one medical treatment over another [11], which job applicant gets an interview [14], or which loan application gets approved [2, 7]. Over the past several years, attention has been brought to the potential social bias machine learning algorithms learn and subsequently induce onto the applications they are involved in [4]. Bias mitigation efforts, as a result, have swelled.

Many bias mitigation efforts focus on centralized machine learning [3, 8, 17], where all training data is stored and processed in a single place. However, as machine learning users and efforts grow exponentially, the issue of data privacy becomes more and more pertinent [15]; centralized machine learning often does not address these concerns. Directives to protect individual data abound, such as the Health Insurance Portability and Accountability Act (HIPAA), European General Data Protection Regulation (GDPR), New York's Stop Hacks and Improve Electronic Data Security (SHIELD) Act, and more. As privacy-preserving techniques have grown, machine learning approaches such as *federated learning* (FL) [13, 10] have emerged.

FL maintains data privacy by allowing multiple parties to collaboratively train a model without sharing their data. FL is already utilized in real-world settings, i.e. Google Gboard with cell phones, enterprise frameworks such as IBM Federated Learning [12].

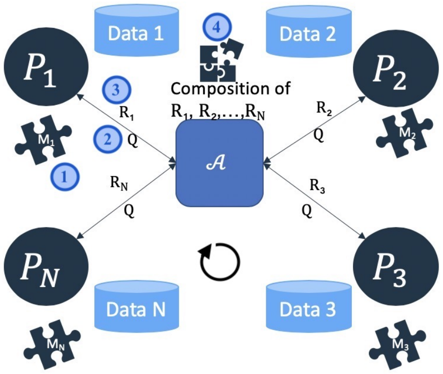

Figure 1: An overview of federated learning

As seen in Figure 1, during each round of training, parties (P_n) train a local model (M_1), and are queried (Q) by the *aggregator* (A) to send *model updates* (R_n); these updates can be model parameters or a party's local model. The aggregator utilizes these to update the global model, usually following a specific model fusion strategy. This strategy, called the *fusion algorithm*, takes the *model updates* ($R_1 ... R_n$) as input and can combine them in a myriad of ways, based on the design. It may simply average each party's weights, or perform a more involved aggregation, then return the global model. This process repeats for several rounds until a certain accuracy or maximum number of training rounds has been reached.

Designed for centralized machine learning, most bias mitigation approaches measure and reduce undesired bias with respect to a sensitive attribute, such as *race* or *sex*, in the training dataset. Unfortunately, existing techniques are not directly applicable to the federated learning setting. As federated learning bars full access to the training dataset, finding methods to address bias without directly examining sensitive information is an open challenge. FL is unique, and as the number bias mitigation methods in centralized machine learning grow, it is increasingly apparent that there are additional considerations to be made about how bias affects federated learning models.

In this paper, we investigate four sources of bias, three unique to training in the federated learning setting. These are *traditional bias sources*, *party selection and subsampling*,

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

data heterogeneity, and *fusion algorithms*. We hope this discussion will stimulate research in this nacient field.

Causes of Bias

Traditional Causes of Bias

Each party in a federated learning process trains their local model similarly to how it would be trained in a centralized machine learning. Factors, such as prejudice, underestimation, and negative legacy [9], that have been recognized as causing discrimination in centralized machine learning, apply to federated learning as well.

However, new sources of bias occur when training in a FL fashion. Through the model updates shared at each round, each party will introduce its own biases to the global model.

In the following, we introduce unique challenges to federated learning algorithms, setting and processing.

Data Heterogeneity

Unlike in centralized machine learning, federated learning faces a unique challenge related to data heterogeneity. Parties each have a local dataset that they use to train their local model. It is very possible that each party's subgroup of data differs greatly from the overall data composition from all parties involved. Additionally, in FL parties might be capable of dynamic participation, where they can drop out of a FL process and rejoin later on for various reasons, i.e. connectivity restraints. Overall and relative data composition thereby may be constantly changing, which affects how the global model learns bias.

For example, say a chain of hospitals wants to use federated learning to train an image classifier for detecting heart disease. Each hospital, in a different location, trains their local model with their patients' data. A hospital in a predominantly minority neighborhood of a larger, predominantly non-minority city is likely to have a very different set of patients in its local dataset, relative to the overall composition of the hospital chain's set of patients.

Fusion Algorithms

When an aggregator incorporates local model updates into the global federated model, it follows a strategy dictated by a *fusion algorithm*. Fusion algorithm designs vary, and influence the bias measured in the final model. For example, some are designed to equally incorporate model updates from parties, while others may perform a weighted average based on party size (i.e. parties with larger datasets influence the global model more than parties with smaller datasets) [13]. Depending on the application of the federated learning task, this could have negative effects on sensitive groups. In recent work, researchers have proposed examining parties' contributions to the global model to decide, or to what extent, their model updates will affect the global model. Many solely examine how model accuracy is affected, which does not inform about the effect on model bias. Some robust aggregation methods will leave out party replies that are dissimilar from the rest altogether [5, 16], which easily can exclude minority groups.

In the same hospital example as above, some hospitals training together have very different dataset sizes. This could be based on some hospitals being located in areas where the population is socioeconomically disadvantaged, thereby those hospitals have less patients that can afford an expensive medical procedure; these hospitals' datasets would be smaller. Involvement in a federated learning process that rewards larger party size with more global model influence would diminish these hospitals' contributions to the global model, and incorrectly give the impression that the model is comprehensively learning from the hospital chain's set of patients. This example can be easily reproduced for situations where other sensitive attributes like age, sex, or race can affect whether or not a user's data is systemically kept out of a federated learning task.

Party Selection and Subsampling

Parties involved in a federated learning process are queried by the aggregator at each round of training to send their *model updates*, which the aggregator will then incorporate into the global federated model. However, not all parties might be involved in every round of training [13, 6]. There are two main FL use cases, either where parties are as small as cell phones, and the number of parties in such FL systems is vast, or when parties are much larger entities and the number of parties is fewer. Especially in FL tasks with a large number of parties, the aim may be to satisfy a quota of parties' updates to begin the next round of training. Depending on the federated learning task, different attributes, some bias-correlated, can affect whether a party is included in a training round.

Consider a scenario where a company wants to train a model to improve the user experience in its cell phone app, and engages users in a FL process. In this example, each phone is a party, and the question of which model updates are included is dependent on network speed. Faster devices, which are likely to be newer, are likely to be represented at disproportionately higher rates than slower devices. Likewise, devices in geographic regions with slower networks may be represented at disproportionately lower rates. Inclusion here is correlated with socioeconomic status, and is a systemic source of bias.

Conclusion

Bias identification and mitigation in centralized machine learning has been discussed in recent years. Since then, federated learning has emerged and been rapidly put into practice, due to its potential to mitigate privacy and regulatory concerns related to machine learning. The shift from centralized to federated learning comes with challenges that stem from limited access to training data, alongside other barriers. Overcoming these obstacles is key to understanding new sources and causes of bias in FL settings. We have identified four causes of bias affecting federated learning, three of which are unique to the FL process. Successful solutions to mitigate bias should consider multiple components of the FL process, and ultimately ensure data privacy is maintained. We hope this paper inspires further exploration in this area.

References

[1] Angwin, J.; Larson, J.; Mattu, S.; and Mirchner, L. 2016. There's software used across the country to predict future criminals. And its biased against blacks. URL https://github.com/propublica/compas-analysis. Accessed: 2019-09-30.

[2] Bartlett, R.; Morse, A.; Stanton, R.; and Wallace, N. 2019. Consumer-lending discrimination in the FinTech era. Technical report, National Bureau of Economic Research.

[3] Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv preprint arXiv:1810.01943* .

[4] Bhandari, E. 2016. Big Data Can Be Used To Violate Civil Rights Laws, and the FTC Agrees. URL https://www.aclu.org/blog/free-future/big-datacan-be-used-violate-civil-rights-laws-and-ftc-agrees.

[5] Blanchard, P.; Guerraoui, R.; Stainer, J.; et al. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, 119–129.

[6] Chai, Z.; Ali, A.; Zawad, S.; Truex, S.; Anwar, A.; Baracaldo, N.; Zhou, Y.; Ludwig, H.; Yan, F.; and Cheng, Y. 2020. TiFL: A Tier-based Federated Learning System. *arXiv preprint arXiv:2001.09249* .

[7] Fuster, A.; Goldsmith-Pinkham, P.; Ramadorai, T.; and Walther, A. 2020. Predictably unequal? the effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets (October 1, 2020)* .

[8] Kamiran, F.; and Calders, T. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* .

[9] Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware Classifier with Prejudice Remover Regularizer. *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* .

[10] Konečnỳ, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* .

[11] Lee, S.; Mohr, N. M.; Street, W. N.; and Nadkarni, P. 2019. Machine learning in relation to emergency medicine clinical and operational scenarios: An overview. *Western Journal of Emergency Medicine* 20(2): 219.

[12] Ludwig, H.; Baracaldo, N.; Thomas, G.; Zhou, Y.; Anwar, A.; Rajamoni, S.; Ong, Y.; Radhakrishnan, J.; Verma, A.; Sinn, M.; Purcell, M.; Rawat, A.; Minh, T.; Holohan, N.; Chakraborty, S.; Whitherspoon, S.; Steuer, D.; Wynter, L.; Hassan, H.; Laguna, S.; Yurochkin, M.; Agarwal, M.; Chuba, E.; and Abay, A. 2020. IBM Federated Learning: an Enterprise Framework White Paper V0.1.

[13] McMahan, H. B.; Moore, E.; Ramage, D.; Hampson, S.; et al. 2016. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629* .

[14] Mujtaba, D. F.; and Mahapatra, N. R. 2019. Ethical considerations in ai-based recruitment. In *2019 IEEE International Symposium on Technology and Society (ISTAS)*, 1–7. IEEE.

[15] Reinsch, W. A.; and Lepczyk, A. 2018. A Data Localization Free-for-All?

[16] Yin, D.; Chen, Y.; Ramchandran, K.; and Bartlett, P. 2018. Byzantine-robust distributed learning: Towards optimal statistical rates. *arXiv preprint arXiv:1803.01498* .

[17] Zafar, M. B.; Valera, I.; Rodriguez, M. G.; and Gummadi, K. P. 2015. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259* .