CA689 Practicum - 2021/2022
Dublin City University, Faculty of Engineering and Computing
Stefano Marzo stefano.marzo2@mail.dcu.ie
Royston Pinto royston.pinto2@mail.dcu.ie

27 February 2022

State of the Art: An Enhanced Data Catalog to Support Trustworthy Federated Learning

1 Introduction and motivations

In the era of big data and the incredible development of artificial intelligence (AI) many new disruptive technologies are coming to light. Business, healthcare, science and even individuals are entrusting AI more and more. The need to handle huge volumes of data has led to the rapid development of dataset management platforms that offer features to comply with e.g. European GDPR regulations, and at the same time allow data management in terms of storage, retrieval, sharing, and more. Artificial intelligence, which is increasingly responsible for decision-making processes and predictions, feeds on data in so-called learning routines to become more accurate and reliable. In a classical declination, an AI model becomes data-independent after the learning process. This allows e.g. predictions to be made on new unseen data. This is possible because in many cases AI can be seen as a structure with an internal state that is updated in the learning stage, and is maintained during the production stage of the AI. This state does not keep track of every single data record used for learning, but retains instead a generalised model describing all the data that has been used for learning. Classically, AIs were produced and trained on local datasets, implying that data collection had to take place prior to the training stage. Organisations developing AI models should therefore have full access to data. With growing awareness of the importance of data, especially personal and sensitive data, European Union are enacting laws to increasingly regulate their storing and processing. Alongside this, new guidelines have been drawn up for artificial intelligence models that respect legal, ethical and moral aspects e.g. the Ethics guidelines for trustworthy AI authored by the High-Level Expert Group on Artificial Intelligence (AI HLEG) on behalf of the European Commission. The increasing attention to data protection enabled to explore new frontiers of machine learning related to artificial intelligence technologies. Federated Learning (FL) is a paradigm for training AI models that does not require the organisations developing the AI technology to have any access to the data needed for learning. This is achieved through the inversion of the learning life-cycle in which the

state of the AI is transmitted to remote devices e.g. smartphones, which update the model with local data and send the updated AI state back to the central server. This framework eliminates the need for data management, with clear advantages in terms of privacy and security. Nevertheless, FL introduces many other concerns regarding the quality of the predictive models produced. Devices receiving status updates from an AI model could deliberately modify it maliciously or enter false data, so-called data-poisoning practices. In addition to these problems, there are also a number of other concerns that can have an enormous impact on the outcome of a predictive model. Scientist and investigators are exploring the most relevant open challenges in the field of FL regards the fairness of the predictive models, scientific research has identified multiple factors that may affect federated AI models, including e.g. under-representation of population groups. The latter emerged problem may have multiple causes, e.g. the presence of minority groups in a certain target areas for FL. One can think of this problem as the introduction of indirect bias into a machine learning system. Although fairly reliable results are expected, it is possible that the state of AI is such that the influence of minorities is underestimated. This can create unexpected consequences in many decision-making processes e.g. in health care and business, hence it is necessary to explore new paradigms to mitigate the effect of indirect bias. The statistical reason that leads to distortions in learning systems in a federated environment is that different groups of individuals may have features that follow very different distributions. Data that reflect the above definition are called non-independent and identically distributed (non-iid) and scientific community is currently putting a considerable effort in developing solutions to deal with them. The purpose of this document is to give an overview of the context and the existing technologies that have been implemented to mitigate the effects of non-iid data in a FL scope and to investigate to what extent is it possible to develop a tool for data catalogs that allows to provide useful metrics regarding the unfairness of a certain dataset and hence enhance trustworthiness.

2 Related Works

2.1 FL fairness challenges

As (Kairouz et al., 2019) mention, the outcomes of machine learning algorithms can be unintended, and if applied e.g. to decision-making processes involving individuals, they can be even undesirable. A key aspect of FL fairness research is the correct representation of groups of individuals in a certain area. As federated learning takes place through the use of individual digital devices in the possession of the users, the results may be distorted by many reasons e.g. latest generation devices will be faster and may have

more influence, this can lead to inequality in terms of a user's socio-economic status. If the inequality is reflected on a particular group of people characterised by the same sensitive features (gender, ethnicity, etc.), providing e.g. a different number of incorrect predictions, they would not fulfil the requirements for individual (Dwork et al., 2011) and demographic fairness (Barocas et al., 2019). New techniques such as Distributionally-Robust Optimisation (DRO) have been proven effective in minimising inequality effects in a population of individuals by reducing the number of negative outcomes (Hashimoto et al., 2018; Taskesen et al., 2020). Also multicalibration (MC) and multi-center analysis (MCA) have been successfully employed to balance fairness in terms of outcomes for subgroups of individuals as reported in (Hébert-Johnson et al., 2018; Xie et al., 2021). All these techniques are effective in handling non-iid data.

2.2 Fairness quantification

Fairness can be expressed in various declinations, e.g. demographic parity (Dwork et al., 2011), equal opportunities (Hardt et al., 2016), inequality in treatments (or mistreatments) (Zafar et al., 2017) etc.. Identifying the possible causes of unfairness leads to the investigation of a method described in (Taskesen et al., 2020) which involves a quantitative measure of unfairness underlying a certain hypothesis.

2.3 Zero Knowledge Technologies

(Chen et al., 2021) shows how it is possible to take measurements that can be useful to characterise a certain populations' subgroups retaining no direct knowledge of them. (Chen et al., 2021) has been used to mitigate the effect of data poisoning, i.e. the malicious practise that involves corrupting federated model updates. (Chen et al., 2021) presents the so-called ZeKoC solution based on the Zero Knowledge Proof (ZKP) principles (Goldreich & Oren, 1994). ZKP technologies enable *fairness through unawareness*.

2.4 Presenting EdcTFL

Enhanced data catalog to support Trustworthy Federated Learning (EdcTFL) emerges from the need to provide a tool in the form of a plug-in for data catalogues that enables the unfairness of a given federated learning model to be measured with the help of measurements on clusters carried out in an unaware manner while respecting the privacy and security of user data. Studying a use case involving demographic disparity and the presence of minority groups within a target zone, federated learning algorithms will be employed in combination with the collection of metadata from the target population made

unrecognisable, then protected, through ZKP.

Table 1: Evaluation of the experiments, trec_eval output

Approach	FL	DC	non-iid	unfairness quantification	bias mitigation
DRO	✓	×	✓	×	✓
MC	✓	×	✓	✓(Pleiss et al., 2017)	✓
MCA	✓	×	✓	×	✓
UQ	×	×	✓	✓	✓
ZKP	✓	✓	–	–	✓(Unawareness)
ZeKoC	✓	×	✓	×	✓
EdcTFL	✓	✓	✓	✓	✓

References

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning* [http://www.fairmlbook.org]. fairmlbook.org.

Chen, Z., Tian, P., Liao, W., & Yu, W. (2021). Zero knowledge clustering based adversarial mitigation in heterogeneous federated learning. *IEEE Transactions on Network Science and Engineering*, *8*(2), 1070–1083. https://doi.org/10.1109/TNSE.2020.3002796

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2011). Fairness through awareness. *CoRR*, *abs/1104.3913*. http://arxiv.org/abs/1104.3913

Goldreich, O., & Oren, Y. (1994). Definitions and properties of zero-knowledge proof systems. *Journal of Cryptology*, *7*(1), 1–32.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Hashimoto, T. B., Srivastava, M., Namkoong, H., & Liang, P. (2018). Fairness without demographics in repeated loss minimization. *ICML*.

Hébert-Johnson, U., Kim, M., Reingold, O., & Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. *International Conference on Machine Learning*, 1939–1948.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K. A., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., ... Zhao, S. (2019). Advances and open problems in federated learning. https://arxiv.org/abs/1912.04977

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. *Advances in neural information processing systems*, *30*.

Taskesen, B., Nguyen, V. A., Kuhn, D., & Blanchet, J. (2020). A distributionally robust approach to fair classification.

Xie, M., Long, G., Shen, T., Zhou, T., Wang, X., Jiang, J., & Zhang, C. (2021). Multi-center federated learning.

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics*, 962–970.