# Efficient Federated Learning: Strategies to Enhance Model Efficiency, Address Data Heterogeneity, and Optimise Communication Schemes

## Seán Ó Fithcheallaigh

# Introduction to Federated Learning

Federated learning (FL) is a relatively new distributed machine learning paradigm that allows researchers to train models on large datasets while preserving the data's privacy and, of course, the source that generated the data, an essential requirement for many industries. Federated learning does this by providing a decentralised system that can be used to train neural networks on constrained devices (referred to as clients) using the data that the constrained device has gathered, where traditionally, the data would have been transmitted to a central server for training. When the client has trained a model on the stored data, the client can transmit those model updates back to the central server. When the server has gathered updates from enough clients, these updates can be brought together, and a global model can be trained, the model updates of which are then transmitted to the clients to use [1].

Researchers and industry professionals employ various methods to build FL systems, with one of the most popular being Federated Averaging [2]. The Federated Averaging algorithm works by a central server selecting a subset of connected devices and sending them to the current global model. Each device that receives the global model will then use it to train the data help locally on the device. A key element is the optimisation method used, such as gradient descent, to update the model parameters. The updated models from all the devices are sent back to the central server, which will aggregate them by taking the average of the weights it receives, and this, in turn, becomes the new global model.

Another approach used in FL is FL with differential privacy (DP), as discussed in [3]. In this work, the authors discuss how DP is a new framework where artificial noise is added to client updates before aggregating. The authors call this “noising before model aggregation” (NbAFL). The researchers in this work noted a tradeoff between the model convergence and the privacy levels offered. As the convergence gets better, the privacy levels decrease.

Researchers have also developed the idea of meta-learning within FL. With meta-learning, the idea is to learn and extract knowledge from previous tasks that can be transferred to a well-initialised model that can quickly adapt to new tasks. This approach prevents overfitting and allows for quick adaptation to any inconsistent data distributed as part of a wider FL system. Several federated-meta learning algorithms exist, such as FedMeta, GraphFL, and Fed4Rec [4].

The fact that federated learning is an emerging field of study means that many open challenges can still be addressed during a research project. For example, researchers still face challenges with the tools at hand. As well as the issues already mentioned, another issue is that some algorithms, for example, the FedAv, can be slower to converge when compared with centralised deep learning networks and less stable in heterogeneous situations [5]. Another challenge, as noted in [5], is that while FL offers privacy for the data, there can still be leaks, which can be detected by analysing the differences in the uploaded parameters (i.e., the weights). As well as data leaks, there is the potential for the data to be attacked (data poisoning), and the models to be attacked (model poisoning) [6].

Another challenging area is that of communication between the server and the clients. Clients might be situated in places that generate unreliable connections between the client and the server or where communications become expensive; improving communications protocols and overheads benefits federated learning [7]. Another area for further work can be found in the operation and performance of clients. Clients can be ineffective for many reasons (unstable networks, energy constraints, reduced/limited computational capabilities), and these inefficiencies can reduce the overall performance of the FL network [4].

Another big challenge for FL that can impact many models is the reduced performance when dealing with heterogeneous data (non-IID) from various devices or environments [7]. This is a particular problem because, in a real-world setting, heterogeneity of data and devices is more common.

Some federated learning algorithms can also be computationally expensive; this can be the case in federated meta-learning, which is compared to traditional meta-learning approaches, and then there is the challenge around correct labelling of the client datasets, which is important for supervised learning tasks [8].

# Problem Statement

Since the introduction of the Internet of Things (IoT), we live in a world where sensors have become commonplace, and the number of connected devices is projected to nearly double from today’s levels by 2030 [9]. Increasingly, sensors are being embedded in everyday objects (i.e., watches, cars, buildings, devices to monitor a person's health), and these devices are generating massive amounts of data on a wide range of systems and devices.

As privacy concerns around data come increasingly into focus, federated learning is seen as a promising approach to allow the data gathered by a constrained device, or edge device, to be trained on that device. The model updates can then be sent back to the central server, where the updates from many edge devices can be brought together to train a new global model, which will be transmitted back to the edge devices. This gives federated learning the ability to keep the data it uses private and on the device itself, removing the need for the data to be transmitted back to a data centre for processing, as would be the case in more traditional machine learning processes.

As has been said above, federated learning is a relatively new area of study, and several challenges must be addressed. The proposed research will address some of these challenges and add to the existing knowledge base in federated learning.

# Potential Application Areas

There is a wide range of areas where federated learning could provide benefits. One area of interest would be fraud detection within the banking system. The work in [10] looks at using federated learning to improve credit card fraud detection. The authors of [11] look at telecom fraud detection and discuss building a federated learning framework for telecom fraud detection using serverless microservices.

Thanks to the privacy benefits of federated learning, healthcare is another application area. [12] discusses the benefits of federated learning in areas related to COVID-19, mammograms, predicting the quality of sleep a person gets, and smart health systems.

Another application area is telecommunications, which is an essential aspect of federated learning. The work in [13] looks at implementing federated learning in large IoT networks and the challenges faced when implementing such a network, such as imbalanced and statistically heterogeneous data, as well as issues that arise when dealing with a range of constrained devices.

# Aims and Research Questions

The overarching aim of this research is to advance the understanding and application of federated learning wide ranging, distributed environments. The research will focus on improving and enhancing model efficiency, overall system security, and communication efficiency.

To meet the research aim, a methodical process will need to be followed where one stage leads to the next, and so on. An outline of this process is given below:

1. What is the current state of the art concerning federated learning:
   1. Identify the existing challenges, as well as the limitations, of federated learning.
   2. Detail this understanding in a comprehensive literature review.
2. From this understanding of the state of the art, identify several areas that require compressive research and analysis. Focusing on the areas of model efficiency, data heterogeneity and system communications, several research questions can be developed:
   1. With the ever-increasing use of IoT devices, can FL algorithms or tools be designed to operate specifically on resource-constrained devices to mini mise their computational and communication costs?
   2. Real-world data will typically always have some element of variability in it that can reduce the model's accuracy. Can we develop FL algorithms that are adaptable to non-IID data sets?
   3. Can adaptive communication strategies for FL be developed that dynamically adjust their transmission frequency and allowable bandwidth based on factors such as model complexity, the type of data, or the transmission network conditions?
   4. Can FL be resilient to unreliable network conditions introducing latency and packet loss while ensuring that model training is reliable, and updates remain efficient and robust?
   5. How can we design efficient aggregation algorithms for FL that minimise the data exchanged between clients and servers while not seeing a reduction in model accuracy?
3. Develop a framework that is novel and addresses the current gaps or deficiencies in the state of the art. This framework would act as a test bed which can be used to meet a number of research objectives. These objectives are:
   1. Develop a federated learning algorithm that significantly improves the models' accuracy when trained on non-IID data. The new algorithm will be compared to other state-of-the-art algorithms to rate the performance.
   2. Develop an optimisation algorithm for federated learning that uses quantisation and compression techniques to reduce the model size, thereby reducing communication overheads. The algorithm will be tested against benchmark datasets to rate the performance.
   3. Design a federated learning communication system that can adapt to heterogenous network conditions, reducing communication overheads. The new system will have to ensure the privacy and security of the data and model and show increased energy efficiency performance when compared to similar systems.
4. Demonstrate the usefulness of the framework in several scenarios, which are informed by real-world applications. While gathering empirical data, the framework should be evaluated for, among other things:
   1. Model performance.
   2. System Security.
   3. Data privacy.
   4. Communication efficiency

Developing points 2(a)-2(c) from above a bit further, looking at potential techniques, could be helpful, as these will form the basis of the research questions.

One potential technique which could be used to **improve the performance** of the FL model could be allowing multiple training rounds on the client device before transmitting these updates back to the server. This can have the benefit of producing more accurate updates, as well as reducing the frequency of updates. Another potential method to improve the model efficiency would be using learning rate schedules or employing adaptive learning rates, which adjust based on the performance of individual clients. Methods to improve model robustness could be investigated to prevent overfitting. Another item which could make the models more efficient and improve the communication scheme is optimising the network topology. With optimisation, the goal is to reduce the number of communication rounds and the amount of data being transmitted while still achieving the desired performance of the system. Other techniques can be employed to **improve model communication schemes**, such as model compression via quantisation or pruning. Quantisation can be achieved by reducing the precision of the model (say, from 32-bit floating point to 8-bit floating point or from floating point to integer values). Pruning is removing unwanted weights from the model to reduce its size. The ability to stop when convergence has occurred would increase the system's efficiency and reduce the number of rounds. Another approach would be to develop a system to determine which clients provided the most meaningful updates. Over time, those updates could take a greater weight, and the clients with less meaningful updates would be selected to send updates back to the server on a less frequent timescale. They could be removed from the system altogether, reducing communication overheads. **Data heterogeneity** continues to be a problem in FL. One approach to dealing with this could be using weighted averages, where different weights are assigned to clients during the aggregation phase based on several factors that would be selected for the situation. Another potential solution could be implementing a system to detect and remove outliers. These outliers could be from within the dataset from a particular client or the whole dataset from a particular client. In doing this, the data will be made more uniform. But care should always be taken when removing data.

# Project Plan

Please see Figure 1 below for the proposed project timeline.

A diagram of a process

Description automatically generated

Figure 1: Proposed Project Timeline

# References

|  |  |
| --- | --- |
| [1] | Hard, Andrew; et. al., "Federated Learning for Mobile Keyboard Prediction," 2018. [Online]. Available: https://arxiv.org/abs/1811.03604. [Accessed 21 10 2023]. |
| [2] | H. B. McMahan, E. Moore, D. Ramage, S. Hampson and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20 th International Conference on Artificial Intelligence and Statistics (AISTATS) 201*, 2017. |
| [3] | K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farhad, S. Jin, T. Q. S. Quek and H. V. Poor, "Federated Learning with Differential Privacy: Algorithms and Performance Analysis," Arxiv, 2019. |
| [4] | F. Liu, M. Li, X. Liu, T. Xue, J. Ren and C. Zhang, "A Review of Federated Meta-Learning and Its Application in Cyberspace Security," *Electronics,* vol. 12, no. 15, p. 3295, 2023. |
| [5] | Y. Zhang and D. Tang, "A differential privacy federated learning framework for accelerating convergence," in *2022 18th International Conference on Computational Intelligence and Security (CIS)*, Chengdu, China, 2022. |
| [6] | S. Tyagi, I. Singh Rajput and R. Pandey, "Federated learning: Applications, Security hazards and Defense measures," in *2023 International Conference on Device Intelligence, Computing and Communication Technologies, (DICCT)*, Dehradun, India, 2023. |
| [7] | G. Pradeep Reddy and Y. V. Pavan Kumar, "A Beginner's Guide to Federated Learning," in *2023 Intelligent Methods, Systems, and Applications (IMSA)*, Giza, Egypt, 2023. |
| [8] | L. Li, Y. Fan, M. Tse and K.-Y. Lin, "A review of applications in federated learning," *Computers & Industrial Engineering,* vol. 149, 2020. |
| [9] | L. Sujay Vailshery, "Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2023, with forecasts from 2022 to 2030," Statista, July 2023. [Online]. Available: https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/. [Accessed 23 11 2]. |
| [10] | K. Bian and H. Zheng, "FedAvg-DWA: A Novel Algorithm for Enhanced Fraud Detection in Federated Learning Environment," in *2023 4th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, Hangzhou, China, 2023. |
| [11] | S. Talluri, Q. Zhang and R. Chen, "A Cloud-Native Federated Learning Architecture for Telecom Fraud Detection," in *NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium*, Miami, FL, USA, 2023. |
| [12] | S. Moon and W. Hee Lee, "Privacy-Preserving Federated Learning in Healthcare," in *2023 International Conference on Electronics, Information, and Communication (ICEIC)*, Singapore, 2023. |
| [13] | W. Ni, J. Zheng and H. Tian, "Semi-Federated Learning for Collaborative Intelligence in Massive IoT Networks," *IEEE Internet of Things Journal,* vol. 10, no. 13, pp. 11942 - 11943, 2023. |