Federated Learning Method To Train Machine Learning Model With Data Privacy

Arun Poochelvan
Dept of Business Management
Ligs University

Abstract- *Companies are constantly looking for new technology for data accessibility in such a way that data is encrypted. One such is the federated learning. Federated learning is a closed loop system involving the statistical training of models keeping the data localized. It can address issues of data privacy. There are many approaches currently in use and are mainly horizontal federated learning and vertical federated learning. Here a feature set is formed is formed by combining all the datasets from various sources. But the labels are only for a single party. Data Privacy can be more efficient in the horizontal one than vertical federated learning as it can be more challenging. Reason for this is the limitation with just two parties. To close this dead end FedV framework is used for secured data privacy by using machine learning models such as linear models and logistic regression [1][7].*

Keywords- Data Privacy, Federated learning, Regression, Decentralized, Machine learning, Central server.

I. INTRODUCTION

Federated learning enables training statistical models based on decentralized devices. A new paradigm of machine learning is to train the data and to decentralize it. In comparison to the traditional centralized techniques which only require data sets in a single server. Federated learning reduces data security and privacy by data localization [6]. It protects user privacy by decoupling and machine learning model aggregation at centralized server. The main purpose is to learn the global model without losing out on data. Even if the data is anonymously stored in a server, it is prone to risk of data leakage [7].

FL allows training data without collecting raw input from the users. It is used in large scale to train the model. For example, query search on Google can be combined with FL, quality improvements can be made with enhancing user privacy. This is an example of federal learning where there was access to training data. It could also be used in multi-tasking and parallel optimization. Language models have been trained using FL [3].

There are two approaches to federated learning – horizontal and vertical. The difference is in the data that is available to each party. In the horizontal FL each of the party has access to the full-fledged set and labels. This enables training of local models in its own datasets. All the datasets are aggregated to create a global model.

In vertical FL each party does not have a complete set of features and hence cannot train by data localization. All the data that is received is curetted to form a complete feature without exposing data and the model is trained in such a way that it is leak proof. FL is best suited for task labels that don't have any human intervention and the training data is extremely large to be feasibly collected. Fully decentralized FL is still under progress. It focuses on the learning process [1].

Block chain can make a fully decentralized FL. This could be done by providing the data directly to the block chain. By doing so, the users can download an updated version. This reduces the global models to the updates written to the block of chain which could be continuously updated [2].

Advantages:

1. The central model contains the average model of all devices
2. The central model is connected to devices to optimize weight
3. It updates on its own using models from local devices
4. Only minimal information is transmitted
5. Reduces the risk of de anonymity
6. Model is ephemeral
7. Trusted model aggregator
8. Higher standards of privacy
9. Latency is minimal

II. PROBLEM FORMULATION

The federated learning problem is subject to constraints so that it can generate data stored and processed locally. To minimize the objective function detection of object. One being the traditional way of detecting is by using

deep learning networks. The other one is the one which sees the object as a regression problem. It uses a single network to return.

min w F(w)
where F(w) := m ∑ k=1 pkFk (w)
m is the total number of devices
pk ≥ 0 and ∑k pk = 1
Fk is local objective function for kth device.
The local objective -Fk (w) = 1 nk ∑ nkjk=1 fjk (w; xjk, yjk)

FedV uses cryptosystem (FE) that allows computing without giving out the inputs over a function. FE is a key encryption family that enables the computation of a function that takes up input in form of cipher text. This can be trusted by the third-party members who set up the cryptosystem in the first place and hence it is safe. Functional encryption for inner product (FEIP) allows the computation between encrypted private data and public data. This is taken in the form of x and y vectors. The private encrypted data can be accessed by the third-party audits. There are two possible schemes: single input functional encryption and multi-functional encryption [1][6].

A cryptography method called additive secret sharing could also be used. This involves calculations of parameters and provides an encrypted calculation. Information is collected from various devices in the global cloud, and it goes from encryption to decryption when the devices are connected. It protects privacy and vulnerability for loss of data [1][6].

III. REVIEW OF LITERATURE

[1] Proposes an approach, FedV. It has an aggregator, set of parties, and third-party authority crypto infrastructure for encryption. The function of the aggregator is to coordinate with the training process amongst parties. Each of the parties has an active party and multiple passive parties having a main dataset and functional subset. FedV safeguards privacy and enables one on one communication which reduces the training period and the amounts of data to be transferred.

[5] Uses a case study where NVIDIA introduced FL on its driving platform. As we know that the geography a place is not constant, they have to train their models individually for all circumstances. The local results are sent to FL server.

[6] Discusses about that optimization methods that allow local uploading and low client participation. Federating Averaging method is used which is based on averaging stochastic gradient descent (SGD). It can handle nonconvex problems and heterogeneous or mixed data. It also proposes another method by using Bayesian network. It can also be used for non-convex models for large, federated learning.

[7]These talks about how Google query can authenticate using FL. This system consists of a trained baseline model and a triggering model. Its objective is to improve query click through rate by taking the suggestions from baseline model and removing suggestions from triggering model.

IV. PRIVACY

Sharing information or leaking of data during the process of training is a constant threat to privacy. Privacy motivated the need of keeping raw data safe and secured [5].

Privacy preserving is studied in machine learning and federated learning. Differential privacy is the most widely used approaches. Differential privacy where a change in one input gives only a small difference in output. This gives an idea whether the device can be used for learning. In the learning methods a common way is using Gaussian noise. But adding more noise reduces accuracy [6].

In Federated privacy learning, setting has its own challenges. Global privacy requires constant up gradation in models at each round. These are private for third party authority. Federated learning has advantages over centre training. Anonymous data set can put client privacy at risk. The date transferred from FL has minimum updates for accuracy. Machine learning models main goal of to protect privacy of users. This is evident in our daily life like aggregation, security, OTPs (one time password) [6].

V. METHODOLOGY

The method implemented involves certain steps.

Consider an example where a bank has to collaborate with a credit card company for a fraud detection machine learning mobile. It would proceed without using a decentralized model. Obviously both parties need profit and also reliability. This calls for FL using machine learning [5]. For machine learning

Differential privacy where a change in one input gives only a small difference in output. This gives an idea whether the device can be used for learning. In the learning methods a common way is using Gaussian noise. But adding more noise reduces accuracy.

Homo morphic encryption where computing is done on encrypted data.

Secure multiparty computation which enables multiple parties to compute on function without the loss of input information from any party involved.

VI. IMPLEMENTATION

Environmental conditions- The device used for training needs to be charges and requires charging, uninterrupted Wi-Fi. The devices can be charged overnight. Many users do not have reliable network. Device specifications- device should be minimum 2GB without a necessary RAM, android operating and SDK level 21+this causes skew in the training.Successful training clients-configuration of the devices s to respond in training.Evaluation and training client overlap-the client selected for training and evaluation rounds are not mutually exclusive but there might be an overlap in a small subset of training population.

The server constraints are as follows Goal client count-it is the target number of clients in training, Minimum client count-it is the minimum clients for training, Training period- frequency of rounds of training, Report window-report from the clients with model update, Minimum reporting fraction- the fraction clients who have reported back with an update [1][7].

VII. APPLICATIONS

Federal learning holds great potential in many sectors, its principle can be used in machine learning-healthcare tasks and computer vision models. This can be used in hospitals where it can predict the outcome of the patient like the complications, detection of variations or diseases. It is based on the medical sensors and the electronic health records [2].

FL can be used in marketing, business, pricing products. It will reduce the overall development costs. It can improve the accuracy of algorithms and they can be personalized. It helps boost the income and hence the economy [2].

FL can also be used to detect fraud analytics and money laundering in financial world. Banks and credit card companies could use this facility to find fraudulent financial transactions. In manufacturing and mining FL between suppliers and purchasers improves maintenance. It improves the overall supply to chain performance [2].

Federation learning in smart phone devices at homes has benefactor results of predicting and reducing power consumption, device usage. It can also be extended to water and waste management. This also paves way for personalized devices [2].

VIII. RESULTS AND DISCUSSION

The basic model gives a score which is compared to a threshold value; by selecting various thresholds operating points are collected. The environmental conditions require unmetered networks and Wi-Fi. In many places this becomes a major issue as they do not have uninterrupted network or a stable power. The device is restricted to 2 GB ram thought this factor did not really matter.

The skew in the training is around 80% of the devises were responding to the training in each of the round. This skew tends towards a big end product as they are more stable. Lower ends are highly unstable. Evaluation and training client overlap-the client selected for training and evaluation rounds are not mutually exclusive but there might be an overlap in a small subset of training population [7].

Communication seems to be necessary in FL. Optimizing the function can reduce the lack of precision. Divide and conquer and one-shot communication scheme can be used but the former is massive and statistical. One shot has been proposed but not yet theorized.

To create realistic system for FL it is necessary to ensure all ways to protect privacy without any data leakage and reduce the transfer of huge data. It should be greater in accuracy and provide a wide range of communication [6].

The devices are not completely dedicated to the task and are not active in any iteration step. Federated labels may or may not be labelled [7].

IX. CHALLENGES

The legal implications of data privacy will slow down the implementation of FL which benefits the healthcare sectors. also, the healthcare sectors will not make such a huge investment in the technology. As a result, these calls for cost effectiveness in FL. They will not take the risk of doing so and would rather prefer manual work. The industrial sector hastaken advantage of the federated learning and improved its efficiency by machine learning techniques. The success of it lies from both ends and requires some resources as well as motivation [2][6].

Even if the company benefits from the Fl venture there is no guarantee that that resulting model will address the

company's purpose. Everyday federated learning has been tremendously used and it continues to show its efficiency and security. Some of them have internet of things, cloud computing as well [6].

X. CONCLUSIONS

Federated learning has gained a lot of popularity in recent years on machine learning for data privacy. Privacy is a major concern and FL promises good results keeping it safe and secure [5].

As the importance of security and machine learning accelerates new technologies in federated learning hold a good promise. It is able to address many issues such as training details, statistical models based on decentralized devices [1].

Federal learning using machine learning paradigm. Where statistical models are trained to safeguard privacy. Its applications and challenges have been discussed. The applications give way to explore and improve user experience in privacy advantaged manner [6][7].

REFERENCES

[1] Xu, R., Baracaldo, N., Zhou, Y., Anwar, A., Joshi, J., & Ludwig, H. (2021). FedV: Privacy-Preserving Federated Learning over Vertically Partitioned Data. arXiv preprint arXiv:2103.03918.
[2] https://medium.com/bcggamma/federated-learning-the-next-big-step-ahead-for-data-sharing-2ae32d375309
[3] https://medium.com/@mu.ammad.ud.din/federated-learning-the-rise-of-privacy-by-design-machine-learning-and-ai-7cf4e196bbc9
[4] https://medium.com/omdena/data-privacy-and-federated-machine-learning-d04da8599896
[5] https://medium.com/@ODSC/how-you-can-use-federated-learning-for-security-privacy-ee0c99cf54b3
[6] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.
[7] Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., ... &Beaufays, F. (2018). Applied federated learning: Improving google keyboard query suggestions. arXiv preprint arXiv:1812.02903.