Privacy-Preserving in FinTech using Deep Learning with Federated Learning in Cryptocurrency

Shailendra Sharma¹, Bonthu Kotaiah², Dr Samarth Singh³, K. V. Daya Sagar⁴, Dr S Durga⁵, Nidhi Sree K⁶
¹Doctoral Researcher at University of Delhi, Ghaziabad, U.P., India- 201010, vkatpalia@gmail.com
²ASSISTANT PROFESSOR, DEPARTMENT OF CS & IT, MAULANA AZAD NATIONAL URDU (A CENTRAL) UNIVERSITY, GACHIBOWLI, HYDERABAD, TELANGANA, kotaiah.bonthu@manuu.edu.in
³Associate Professor, Shri vishwakarma skill univ, govt of Haryana, samarthsinghsamarth@yahoo.com
⁴Associate Professor, Electronics and Computer Engineering, Koneru Lakshmaiah Education Foundation, vaddeswaram, Guntur, Andhra pradesh, India, sagar.tadepell@gmail.com
⁵Assistant Professor, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, sdurga1234@gmail.com
⁶Bhavans Vivekananda College for Science Humanities and Commerce, Hyderabad, nidhisreebvvc1@gmail.com

Corresponding Email: vkatpalia@gmail.com

Abstract

Recent developments in deep learning techniques have produced significant improvements in long-standing AI jobs like drug discovery, gene analysis, and speech and image recognition. Although deep learning has numerous benefits, the identical training dataset that has made it so reliable also raises serious privacy concerns to address these privacy concerns, McMahan et al. created Federated Deep Learning, a together distributed deep learning paradigm for the mobile devices (FDL). Deep learning and distributed computation are essentially combined in FDL, where several parties get involved into the process of distributed training and parameter server records track of a deep learning model that needs to be built. The central parameter server first distributes a pre-trained model globally on a common set of data to each participant. Then, in each cycle, each party utilizes a local dataset to train and improve the current global model. The gradients are gathered from each party by the central parameter server, which then utilizes them to build a new global oriented model for upcoming iteration. At final stage the different parties and the central parameter of server repeat the aforementioned procedure till the global modelling attains a particular accuracy or ideal convergence.

Keywords: Federated learning, Blockchain technology, deep learning, privacy.

INTRODUCTION

It has been described that machine learning and data science play a crucial role in science. Large amounts of data incorporated in research projects are related with the advancement of science through data analysis. Consequently, the need for privacy is a fundamental principle for the majority of persons. Globally, data collection has become an increasingly important idea. This is evident from the development of data-collecting software and applications. It has also been suggested that social media platforms acquire data from their users to improve their services. Significantly increased data usage has raised the quantity of data that requires confidentiality. Therefore, consumers and politicians have concentrated on concepts connected to privacy. The General Data Protection Regulation (GDPR) is among the global measures adopted for data protection. The move was motivated by the necessity of fostering the required forms of development and guaranteeing that there are notable accomplishments regarding protected data. In 2017, Google introduced the concept of federated learning.

Federated Learning

Google developed the concept of federated learning in 2017, with the intention of supporting data scientists with their work. Federated Learning is a technique for training models using a huge corpus of decentralized data. Federated Learning operates on a scalable production framework for autonomous analysis and machine learning across a variety of devices and domains. The major purpose of federated learning was to protect confidentiality and enhance the efficiency of the data scientists' job.
Thus, scientists might share models for statistical data analysis on decentralized devices and servers containing local data sets. This approach provided scientists with access to more effective models that facilitate their work.

Due to the decentralized nature of the built systems, there was no official oversight of the statistical models that the scientist provided. The concept enabled the majority of scientists to be astounded by the possibilities of the federated learning system as a whole. In addition to this, data scientists were not compelled to send data online or to the cloud, which was one of the primary benefits. The only requirement was that they receive the statistical models from the federated learning platform and use them to evaluate their data.

A comparison of the novel decentralized system with federated knowledge and the conventional centralized machine learning approaches demonstrates a considerable improvement in maintaining data security for various users. Through these tools, research teams and data scientists can now protect the confidentiality of their data. Therefore, the proliferation of technological systems has increased the demand for their maintenance. These data types are promoted on the basis that they eliminate any concerns regarding data privacy. The confidence instilled in researchers determines the confidence with which they conduct their whole research.

Financial Machine Learning and Blockchain

The financial services industry has been one of the main drivers of the technological advancement inside the blockchain ecosystem. Greenwich Associates estimates that the industry spends $1.7 billion annually on blockchain as organizations go from proof of concept to full implementation. Few, however, comprehend the consequences of incorporating machine learning as a supporting tool. To better evaluate probable outcomes, let's start by assessing the advantages and disadvantages of this nascent integration.

Advantages of Integration

Numerous advantages result from the use of blockchain technology and artificial intelligence. In brief, by using machine learning's analytical capabilities, the built-in security features of blockchain can be reinforced. The ability to process massive amounts of data securely and effectively in the financial services industry may be extremely valuable to institutions and end users.

Protection of Payment Networks

The capacity of blockchain technology to act as a borderless payment network is one of its key advantages. As a decentralized solution, numerous blockchain protocols have been developed to support frictionless transactions with cheap transaction fees. In reaction to the exorbitant fees and sluggish processing times typical of centralized banking institutions, these solutions were developed. Apart from this important application of the blockchain technology, widespread adoption is hampered by persistent concerns of security. Theft and fraud are very common with blockchain transactions because all that is needed is a pair of public and private keys. However, irregularities in account activity can be swiftly identified with the help of machine learning's improved potential, leading to human intervention. Both financial service providers and their clients are protected by this extra security step. Security issues can also be fixed with the help of alternative AI technologies like biometrics and behavioural analysis.
Figure 1: Blockchain Technology

Federated Learning Privacy and Security Issues

FL has been described as presenting a number of obstacles for its users. Continuous usage of federated learning, FL, presents a number of privacy and security issues. The users are affected by these obstacles, which prevent them from executing the essential applications and services of the federated learning tools with precision. The overall strategy focuses on privacy concerns that may affect users. In a distributed machine learning environment, the presence of unprocessed data poses the greatest privacy risk.

Other difficulties have been related with the sharing of model data as a training method.

The method has conditioned the system and equipment to become more valued and accessible. The primary emphasis remains on the potential leak, which may be caused by the employed machine. The leak could harm the available individuals and have a considerably greater impact on all the accessible persons. The risks associated with these methods include the persistence of data loss and the loss of sensitive text content. These challenges highlight the need to enhance the federated learning system and safeguard people's safety. Several techniques have previously been devised to ensure that privacy has been protected and the highlighted risks and difficulties have been addressed.

Differential privacy was conceptualized as a random approach for ensuring output distribution. The strategy focuses on promoting the elements contained and restricting access to system samples. It permits restricted access to the type of sample utilized in the machine learning procedure. Consequently, the learning methods are safeguarded, ensuring that differentiated privacy has been attained. The strategy can focus on privacy principles, showing that the implemented system is somewhat effective.

SMG, or secure multiparty computation, has also been utilized as a method. This strategy focuses on a collaborative computational concept that ensures the included functions are not dependent on the required sort of development. It fosters a certain collaboratively agreed-upon purpose, preventing leakage. The method does the necessary learning and analysis while preserving confidentiality and privacy by preventing leaks. Since then, the procedure has been reported as being quite effective.
Preserving Pii Data in Fintech

FinTech is among the technologies that value cyber security the most. FinTech is comprised primarily of technologies and software geared toward financial applications. The increased use of these technologies is mostly driven by the increasing demand for their development. Incorporating innovations into the field of finance has enhanced the field over the past few decades. This strategy has been linked to greater effectiveness in the financial sector. More individuals have incorporated technologies into banking sectors to increase security.

FinTech necessitates a substantial level of security measures. Security is a major worry for the majority of financial geniuses. The requisite level of security ensures that individuals can safely deposit funds in many institutions and send funds as necessary. These technologies are essential for business owners since they foster the appropriate type of community growth. Consequently, the technology linked with financial applications is an indispensable component of the global economy. Its uses have been defined as having a specific significance for firms’ growth. To ensure that the entire sector of finance has been analyzed and is secure for all types of transactions, promoting-security-related algorithms have been developed. It has also led to the reduction of fraud across the majority of financial departments.

![Diagram](source: OECD staff illustration)

**Figure 2: Financing transaction using blockchain**

According to Estevez, FinTech arose in the 21st century and was initially implemented in the backend systems of large financial institutions. The technologies proven to be highly effective, resulting in a widespread appreciation for them in consumer-oriented services. FinTech is currently affiliated with the education, retail banking, fundraising, non-profit, and investing sectors. Each of these businesses and industries now achieves a distinct form of expansion by ensuring efficient funding. The financial sectors of these firms have been infused with trust as a result of the general applications of this form of development.

AI integration in financial products based on blockchain

In recent years, there has been a rise in the usage of the distributed ledger technology (DLTs), as the blockchain, across a variety of businesses, but primarily in the banking industry. Applications developed on blockchain are rapidly growing due to automation and disintermediation, as well as the stated plus points of speed, efficiency, and a level transparency in such cutting-edge technologies that may offer (OECD, 2020). The spread adoption of DLTs in the field of finance sector might be fueled by initiatives to improve disintermediation efficiency, includes in the market of security (issue and the post-trade/clearing , settlement of
transaction), doing payments (digital currencies of central bank and fiat-backed stable coins), and asset tokenization in general. The responsibilities and business models of financial operators may also alter as a result of this (e.g. custodians).

The industry is promoting the confluence of AI and DLTs in this blockchain-based finance as a means to improve the performance of such systems, as greater automation may enhance the promised efficiencies of blockchain-based systems. To support claims of convergence between the two technologies, however, it doesn't appear that blockchain-based applications have implemented AI to a sufficiently high level just yet. What is actually happening in practice, as opposed to a convergence, is the integration of DLT solutions into specific AI processes as well as the integration applications of AI into particular blockchain systems, for the usage cases like risk management (e.g. for data management).

The latter technique uses DLTs to provide data to a machine learning model, utilising the immutable and disintermediated characteristics of blockchains. Additionally, it enables the communication of private data on a no-knowledge basis without breaking any rules pertaining to confidentiality or privacy. The usage of DLTs in AI processes could potentially enable users of these systems to monetise their own data that is utilised through models of ML and other AI-driven systems (e.g. IoT). The development of as AI uses cases that is spread by the technology's ability to enhance automation and disintermediation gain of efficiency in DLT-based systems and networks.

Theoretically, by transferring the burden curation of data from the third-party nodes to autonomous, automatic AI-powered based systems, which are harder to manipulate, the employment of AI in DLT-based systems could improve the robustness of information capture and exchange. The chain's data inputs may now be of higher quality as a result. Usage of AI can, in particular, improve the efficiency of off-chain nodes from outside parties, such as the dubbed "Oracles" nodes that add external data to the network. There is a danger that the network will get inaccurate or insufficient data feeds from performing below the bar or even the susceptible third-party off-chain nodes when Oracles are deployed in DLT networks. AI inference directly on-chain, which would eliminate the need for information providers from outside the chain, like Oracles, the deployment of AI may possibly lead to even further disintermediation. It could act as a safeguard against cyberattacks and the manipulation of the those third-party data presented to the networks by checking the veracity and integrity of the data offered by the Oracles.

Objectives

Protection of Privacy

In FPPDL, we presumptively presume that parties don't trust one another or any outside parties. Therefore, while training a collaborative modelling without the assurance protection of privacy, parties might not be ready to disclose their information. In FPPDL, every party uses Differentially Private GAN (DPGAN) to published differentially private and local in place of exchanging the original data and parameters of model while the first benchmarking phase, samples are used for mutual evaluation. They then use the previously mentioned three-layer onion-style encryption technique to encrypt the shared gradients in order to safeguard privacy during group deep model training.

Fairness

Our major objective in this situation is to distributed various iterations of the final level of FL model among participants according on their involvements. Fairness via Awareness is the concept of fairness that is most relevant to our purpose. According to this theory, people who are comparable in terms of a similarity metric established for a certain job ought to experience similar results. Parties with higher contributions need to receive greater rewards than parties with lower contributions. Furthermore, we make it obvious that low-contribution parties are not evil; instead, they follows a protocol genuinely and attempt to access the data of other parties despite contributing little to nothing or negatively.

Methodology

An architecture built on blockchain
We use Blockchain 2.0 to create a private Blockchain that is only accessible to the involved parties in order to create the privacy-preserving deep learning algorithm into the decentralised architecture for FPPDL. FPPDL gives the peer-to-peer design of Blockchain as opposed to the present server-based architecture, enabling each participant to maintain its modularity while collaborating with others. Additionally, each party retains total control over its own data rather than handing it over to the central server. Additionally, Blockchain has the innate capability to automatically coordinate each party's joining and leaving, further promoting the federation's independence and modularity. Blockchains improve robustness due to the absence of a single point of failure.

In init block and the block of operation are the two sorts of blocks that we build here for the FPPDL blockchain.

![Figure 3: An illustration of a genesis block construction](image)

**Figure 3: An illustration of a genesis block construction**

![Figure 4: An illustration of an operation block structure](image)

**Figure 4: An illustration of an operation block structure**

Fppdl Implementation

The section explains the two-stage implementation process used by the FPPDL to uphold privacy and fairness. The techniques for doing so include initialising local reliability values, levels of sharing, and the points by initial benchmarking in the privacy-preserving collaborative deep learning phase.

Benchmarking

Before collaborative model training begins, the suggested first benchmarking technique tries to evaluate each participant's local training data by reciprocal assessment without even looking at the raw unfiltered data. The technique works as the follows: each participant trains a DPGAN with local training dataset to provide synthetic examples. Given the limited privacy budget used in DPGAN, these generated samples will only show a few implicit density estimates. They will not reveal the genuine sensitive examples or the true distribution of data. Without releasing labels, each participant publishes false samples that were each separately created and shared.
Figure 5: Two stage implementations in FPPDL

Level and Points of Sharing Initialization

The indicator of the level of sharing is the maximum count of samples or gradients that one party may share with other parties. The formula allows an automatic determination of party i’s permissible sharing level. \( \lambda_i = u_i / |D_i| \) The formula allows an automatic determination of a reasonable sharing amount for party I. \( \lambda_i = u_i / |D_i| \) Following is how points are initialised:

\[
P_i = \lambda_i * |w_i| * (n - 1)
\]

Collaborative Deep Learning with Privacy Preserving

The algorithm provides a concise summary of the procedures for privacy-preserving of collaborated deep learning in ever round of communication, includes how to upgrade the points as per upload/download, instructions to safeguard the protection of individual model updates utilizing onion-style encryption, trailed by boundary and neighbourhood validity update, and tenable party set support by the Blockchain framework.

The quantity of points party I obtains during each communication cycle has a significant impact on the budget of download of party I or \( d_i \) in particular. In other words, party I will not have enough points to cover the gradients offered by other parties if \( d_i \) is greater than \( p_i \). Additionally, \( d_i \) can be produced dynamically based on the existing \( p_i \) points in each communication round. Although \( d_i = p_i \) is initialised for simplicity in each communication cycle, the number of gradients that can be downloaded depends on the requester’s local credibility list and the sharing levels of the requested parties.
Encryption of a three-layer system

All parties have access to the encrypted gradients, though, as each party is required to keep a different encrypted gradient that will be transferred to a different party on Blockchain for commitment. This problem can be solved by combining homomorphic authentication with public-key encryption. However, due to the enormous dimension of the released gradient vector, the gradient vector's encryption is costly in terms of processing and the communication.

Here we proposed a three-layered onion-shape encryption scheme. The first layer uses homomorphic encryption using symmetric key keystream $k_j$ to secure local model gradients.

Update on Local Credibility

In each session of collaborative learning, every partner chooses and shares a subset of DPGAN samples at random based on their own sharing level. Each party determines the local level credibility of the variations parties relied on the labels that were provided, and at the time of the current round, evaluates this local credibility using its updated local model.

Results of Experiment

Fairness of our FPPDL over the MNIST dataset and the distributed framework, using various model topologies

Table 1: Results

<table>
<thead>
<tr>
<th>Setting 2</th>
<th></th>
<th>Setting 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Distributed</td>
<td>W1 DL</td>
<td>Distributed</td>
</tr>
<tr>
<td></td>
<td>CNN</td>
<td>MLP</td>
<td>CNN</td>
</tr>
<tr>
<td>$P_4$</td>
<td>-0.69</td>
<td>0.31</td>
<td>0.88</td>
</tr>
<tr>
<td>$P_1$</td>
<td>0.21</td>
<td>-0.16</td>
<td>0.77</td>
</tr>
<tr>
<td>$P_3$</td>
<td>-0.03</td>
<td>0.03</td>
<td>0.78</td>
</tr>
<tr>
<td>$P_5$</td>
<td>-0.17</td>
<td>-0.05</td>
<td>0.76</td>
</tr>
</tbody>
</table>

We only compare the collaborative fairness of our FPPDL with the distributed framework utilising DSSGD as parties in the centralised framework are unable to access the trained global model and parties don’t interact in the standalone architecture. The fairness of the distributed system and our FPPDL were determined using the MNIST and SVHN datasets, which had various frameworks, party numbers, and the settings. The high positive fairness values, the majority of which are above 0.5, demonstrate that FPPDL achieves a respectable level of fairness and support the premise of fairness: the party with more training data and less privacy creates a higher accuracy. The distributed framework, in contrast, exhibits inadequate fairness, with values that are consistently lower than those of the FPPDL and, in some cases, even negative values, highlighting the lack of justice in the distributed framework. This is because the distributed architecture enables all participants to derive models that perform as well, regardless of how much one participant contributes.

Discussion

Collaboration and data augmentation: We use data augmentation to increase local dataset sizes to aid credibility initialization, allowing DPGAN to create trustworthy samples while maintaining a reasonable privacy budget. Data augmentation, on the other hand, is meant to enhance the amount of training data utilising knowledge already existing in the local training data, boosting the generalizability of the local model while failing to generalise to previously undiscovered data. Given this, it is obvious why parties still need to work together for increased value even after data augmentation. It cannot depict diffusion over
the globe. By using DPGAN, not just the security of the authentic data also the privacy of the enhanced data which are comparable to the original data is preserved.

Security and Fairness: Onion-style encryption with three layers provides stronger privacy protection without sacrificing convenience. In the initial benchmarking, parties generate DPGAN samples relied on their local training data, that is later compared by other parties’ which standalone models; in the process of collaborative learning, each party randomly selects and shares a subset of DPGAN samples as per individual sharing level at every round of the communication, and then upgrades the local authenticity of other parties

Future Scope

This paper suggests FPPDL, a decentralised deep learning architecture that protects privacy and considers fairness. Our updated framework demonstrates the following characteristics: It (1) investigates the research problem of the collaborative fairness in the deep learning for the first time by introducing the concept of local authenticity and points of transaction, which are initialised by initial benchmarking and upgraded during privacy-preserving collaborative deep learning; (2) resolves inherently the pertinent problems in server-oriented frameworks; and (3) inspects Blockchain for decentralisation. The experimental results highlight the utility of our proposed framework by demonstrating that our FPPDL routinely outperforms the standalone framework and achieves accuracy comparable to the centralised and distributed selective SGD frameworks without differential privacy.

REFERENCES

Agreement Scheme. In 2021 IEEE Globecom Workshops (GC Wkshps) (pp. 1-6). IEEE.
42. Rawal, B. S., Manogaran, G., Singh, R., Poongodi, M., & Hamdi, M. (2021, June). Network augmentation by dynamically splitting the switching function in SDN. In 2021 IEEE International Conference on Communications Workshops (ICC Workshops) (pp. 1-6). IEEE.


