Design of a Fair Distributed Computing Platform based on Distributed Ledger Technology and Performance Measurements

Bo-Yan Liao[1] , Jia-Wei Chang[2] and Hao-Shang Ma[3]

[1] Department of Computer Science and Information EngineeringNational Taichung University of Science and Technology, Taichung City, Taiwan
[2] Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, Taichung City, Taiwan
[3] Department of Computer Science and Information Engineering, National Taichung University of Science and Technology, Taichung City, Taiwan
z775357@gmail.com , jwchang@nutc.edu.tw , hsma@nutc.edu.tw

Abstract. We propose a fair distributed computing platform based on Distributed Ledger Technology (DLT) and performance measurements. The platform integrates DLT and federated learning, enabling users to train machine learning models on their local devices without compromising their privacy by sharing their data with a central server. Instead, only the trained model weights are uploaded to a central server for aggregation. To address privacy concerns associated with federated learning, we integrate various privacy-preserving methods, such as differential privacy, model pruning, and homomorphic encryption, into the platform framework. These techniques help protect user privacy while improving model accuracy. To address the non-IID data problem in federated learning, we use performance measurements to balance the training workload among users, and blacklist malicious users while incentivizing participation. DLT ensures the security and integrity of the platform by validating and recording all data transactions on the ledger. Overall, the proposed platform has the potential to revolutionize machine learning model training by making it more efficient, secure, fair, and transparent.

Keywords: Federated learning, InterPlanetary File System (IPFS), Blockchain.

1 Introduction

As the popularity of smart connected devices such as smartphones, smart homes, and wearables continues to grow, people's lives are increasingly dependent on these intelligent devices. However, the data generated by these devices is mostly stored in various devices, forming multiple Isolated Data Island that cannot be effectively utilized. Moreover, since data involves personal privacy, users do not want to transmit their data to a central server for training, thereby exposing personal privacy information. To solve these problems, the Google AI team proposed the federated learning framework in 2016, which can effectively use distributed data for model training while protecting personal privacy. In federated learning, users can train models on their

local devices without transmitting sensitive data directly to the central server. Through federated learning, not only can the problem of Isolated Data Island be solved, but also user privacy can be protected, which has become an attractive research topic for enterprises and researchers in various fields.

Despite significant progress in addressing privacy and data ownership issues faced by centralized machine learning, the application of federated learning still faces many challenges. One of the key challenges is how to effectively prevent free-riders or malicious users, and improve the performance and accuracy of federated learning under Non-IID data, by optimizing the order of aggregation weights and using clustering methods. Additionally, as deep learning continues to expand into various fields, the value of model weights cannot be ignored. This paper further explores the fair trade of model weights by establishing a public trading platform based on blockchain and InterPlanetary File System (IPFS) technology. This platform enables each user to safely share their trained model weights with other researchers, while also receiving corresponding virtual currency rewards. Through this approach, not only can the efficiency of model training be effectively improved, but users can also better control their data and model weights, thus protecting personal privacy. Furthermore, the fair trading platform can promote collaboration between different fields to jointly solve various real-world problems.

2 Related work

2.1 Performance Measurements for AI Applications

In the frameworks of Federated Learning and Swarm Learning, a large number of users participate in the collaboration of the model freely. Participants locally process their own privacy data using technologies such as homomorphic encryption and differential privacy before iterating the model. After completion, the model is uploaded to the server for aggregation, and the participants receive rewards [1]. In this mechanism, participants may modify their local data to obtain more rewards, so that the trained weights can better generalize the model, while attackers may use fake data to maliciously attack the model. Therefore, a fair valuation method is needed to evaluate the quality of the weights provided by the participants. In the figure below, different evaluation methods are used for various fields and tasks in machine learning. Therefore, We combines tools such as classification report, F1 score, and valuation to give a rating of the contribution value to the participants, as shown in Table 1.

Domain	Task	Commonly used evaluation metrics
Computer Vision	Object Detection Models	AP、mAP
	Multi-Object Tracking Models	MOTA、MOTP、MT、ML、IDs、FM、IDF1
	Image Data Augmentation	UIQM， PCQI
Natural	Machine Translation	GLUE、ROUGE、METEOR、

Language Processing	Models	CIDEr
	Document Summary Evaluation	METEOR、GLUE、Edmundson、ROUGE
	Code Generation Transformers	BigCloneBench、Defect Detection
Audio	Music source separation	SI-SNRi、SDRi，SDR
	Speaker Diarization	Accuracy、F1 score
	Speech Recognition	SER、S.Corr、WER/CER

2.2 Public and Incorruptible Transaction Platform using Distributed Ledger Technology (DLT)

To verify the information of participants and record the complete information of model updates for achieving full fairness, the simplest way is to use DLT-related technologies, like Blockchain has the characteristics of immutability and irreversibility, so it can be used to verify whether the relevant information of participants is correct and record the entire process on the chain.

Blockchain

Blockchain is a concept proposed by Satoshi Nakamoto in the Bitcoin white paper in 2008. It heavily utilizes cryptography and consensus mechanisms. Blockchain is composed of blocks linked together, and each block contains an encrypted hash of the previous block, transaction records, and a timestamp. The encrypted hash is calculated by a designated party through a consensus mechanism, which requires a certain amount of resources. Popular consensus algorithms include Proof-of-Work (PoW) and Proof-of-Stake (PoS). This architecture is tamper-resistant. If a transaction record is modified, the content of subsequent blocks should also change accordingly. Therefore, this technology is widely used in various fields to protect data from tampering.

Distributed Ledger

A distributed ledger is a technology that records all transaction contents on the chain through blockchain technology and can verify the authenticity of transactions through miners. Because blockchain has the characteristics of fairness and immutability, it is very suitable for recording transactions.

Consensus Algorithm

In order to verify whether a transaction has been tampered with, miners verify each block on the chain, and the transaction is considered complete and recorded on the blockchain only when most miners reach consensus. Common consensus algorithms include Proof of Work (PoW), Proof of Stake (PoS), Tangle, etc.

Smart Contract

The concept of smart contracts was first proposed by Nick Szabo in 1994, but it wasn't until 2015 that they began to be widely used on the Ethereum platform. Smart contracts are programs that are stored on the blockchain and cannot be tampered with. When certain conditions are met, the program will automatically exe-

cute, just like a contract being enforced in a court of law. Smart contracts are very fair and free from tampering concerns, and their execution can be enforced.

(1) IPFS

IPFS is a peer-to-peer (P2P) decentralized file system that can split a file into several file fragments and store them on various nodes. The file fragments are scattered across multiple nodes, and the integrity of the file is ensured by hash values. The location of each node and its hash value are recorded in a Distributed Hash Table (DHT). Since the status of each node cannot be verified during use, each file is typically backed up on multiple nodes to prevent single-point-of-failure issues.

(2) Decentralized identity (Identity verification)

Decentralized identity means that the identity is not verified through a third party. On the blockchain, you can only prove that you are you through a public key, which is usually not linked to the real world, and cannot prove the authenticity of the user. In recent years, with the trend of blockchain, many decentralized applications (DAPPs) have emerged, and many public and private chains have been developed. However, there is no intersection between these chains, so when users use DAPPs, they have to register a new identity on the corresponding chain. In recent years, some banks have gradually introduced FIDO (Fast IDentity Online), a standard released by the FIDO Alliance, which combines public key and biometric technologies to establish identity verification mechanisms.

3 Method

We found that the order of aggregation is critical for the accuracy of the global weights in Non-IID data. Therefore, inspired by clustering methods, a new aggregation method is proposed. In this method, the similarity between users is first calculated, and appropriate clustering and aggregation orders are assigned based on the model weights trained for each user. To implement this aggregation method,We conducted ablation experiments to investigate whether various clustering methods applied to the aggregation of federated learning can increase the accuracy of the globally aggregated model under Non-IID data.

Density-based clustering methods do not require the number of clusters to be specified in advance. We adopted the DBSCAN algorithm to classify each data point as either noise or a member of a cluster. This algorithm accurately classifies clusters of various shapes.

Inspired by the research of Miao, Yinbin, et al. [2], We further compares the weights uploaded by users with the weights calculated by the global model using cosine similarity, and aggregates similar weights from bottom to top. At the same time, the aggregation method is designed to minimize the standard deviation of the similarity between gradients during aggregation and maximize the number of gradient weight models in aggregation. Finally, the accuracy of the weights for the task is evaluated to determine whether to include these weights in the next aggregation. Through multiple

bottom-to-top aggregations, weights that are closer in distance will be aggregated, gradually forming larger clusters of positive weights, while malicious weights will be blocked and discarded.

A record should be kept for each upload of weights and aggregation, and the history can be traced at any time. With blockchain technology, each upload of weights or aggregation of the global model by a user will be recorded on the blockchain. The information that should be recorded on the blockchain, as shown in Figure 1, includes the training mode, time, uploading user, ID of the previous global model, the user who trained the model or which weights were used for aggregation, time spent, and the location where the weights are finally stored in IPFS. With this information, we can always find out on the blockchain which people have operated on the model. If any malicious attacks are detected, the attackers can be blocked or added to a black-list.

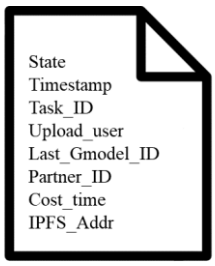

Fig. 1. : Transaction Information Structure Recorded on Blockchain

In order to protect the value of the entire global model, We first adds a watermark to the local model weights on individuals' devices, which records the private key of each contributor. When there are doubts about users' usage rights, verification can be performed through the extraction of weights or backdoor attacks. Under the FedIPR framework proposed by Li et al. [3], unless the model is destroyed, the results of the watermark verification cannot be altered, thereby effectively preventing unauthorized commercialization of the trained model by others.

It is difficult for small companies or individuals to have a commercially viable model as it requires high costs for storing data and training the model, which could amount to millions of dollars per year. Additionally, collecting user data for training purposes also incurs significant costs and time. Therefore, we aims to establish a fair trading platform as shown in Figure 2 where users can freely join to trade or train model weights. Evaluating the value of a model is an important issue. The primary costs of training a model are server costs, the value of user data, and the cost of training on user devices. Therefore, the equivalent relationship between models can be calculated through the download volume of each model, and the real value that a user possesses can be calculated through the value of data that they provide.

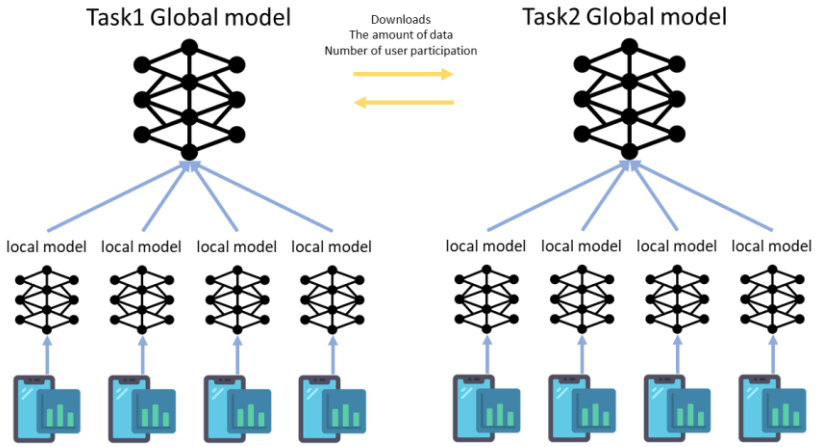

Fig. 2. Fair transaction design for decentralized deep learning task trading platform.

4 Results

We discuss the challenges faced by the federated learning framework and propose a fair-trading platform based on blockchain and InterPlanetary File System (IPFS) technology. Our platform allows users to safely share their trained model weights with other researchers and receive virtual currency rewards. We also discuss the need for a fair valuation method to evaluate the quality of the weights provided by participants and present various evaluation metrics for different fields and tasks in machine learning. Lastly, we explore the use of blockchain technology to verify the information of participants and record the complete information of model updates to achieve full fairness.

Acknowledgement

This work was partially supported by the National Science and Technology Council, Taiwan, R.O.C. [grand number NSTC 111-2221-E-025 -008].

References

1. Warnat-Herresthal, Stefanie, et al. "Swarm learning for decentralized and confidential clinical machine learning." Nature 594.7862 (2021): 265-270.
2. Miao, Y., Liu, Z., Li, H., Choo, K. K. R., & Deng, R. H. (2022). Privacy-preserving byzantine-robust federated learning via blockchain systems. IEEE Transactions on Information Forensics and Security, 17, 2848-2861.
3. Li, Bowen, et al. "Fedipr: Ownership verification for federated deep neural network models." IEEE Transactions on Pattern Analysis and Machine Intelligence (2022).