Building Communication Efficient Asynchronous Peer-to-Peer Federated LLMs with Blockchain

Sree Bhargavi Balija[1*], Amitash Nanda[1*] , Debashis Sahoo[2]
[1]Dept. of Electrical and Computer Engineering, University of California San Diego
[2]Dept. of Computer Science and Engineering, University of California San Diego
{sbalija, ananda, dsahoo}@ucsd.edu

Abstract

Large language models (LLM) have gathered attention with the advent of ChatGPT. However, developing personalized LLM models faces challenges in real-world applications due to data scarcity and privacy concerns. Federated learning addresses these issues, providing collaborative training while preserving the client's data. Although it has made significant progress, federated learning still faces ongoing challenges, such as communication efficiency, heterogeneous data, and privacy-preserving methods. This paper presents a novel, fully decentralized federated learning framework for LLMs to address these challenges. We utilize different blockchain-federated LLM (BC-FL) algorithms, effectively balancing the trade-off between latency and accuracy in a decentralized-federated learning environment. Additionally, we address the challenge of communication overhead in peer-to-peer networks by optimizing the path for weight transfer and mitigating node anomalies. We conducted experiments to evaluate memory usage and latency in server and serverless environments. Our results demonstrate a decrease in latency by $5X$ and a 13% increase in accuracy for serverless cases. Comparisons between synchronous and asynchronous scenarios revealed a 76% reduction in information passing time for the latter. The PageRank method is most efficient in eliminating anomalous nodes for better performance of the global federated LLM model. The code is available on GitHub (https://github.com/Sreebhargavibalijaa/Federated_finetuning_LLM-s_p2p_environment).

Introduction

Large language models (LLM) have demonstrated prominent capabilities in solving complex problems (Radford et al., 2018). Consequently, this has propelled significant research interests and applications in healthcare domain as addressed in (He et al., 2023). LLM models heavily depend on large volumes of data, while applications involving sensitive environments such as healthcare incur scarce and confidential data and require maintaining privacy (Thirunavukarasu et al., 2023). Moreover, data sharing and collaboration are limited due to privacy concerns and commercial competition. To address this issue, recent efforts have focused on a continual learning approach to leverage clinical data across hospitals for sepsis prediction (Amrollahi et al., 2022). Addressing the challenges of building personalized LLM models is a significant issue. One of the promising approaches to handling LLM is integrating federated learning (Bharati et al., 2022) and blockchain technologies (Zhu et al., 2023).

Federated Learning (FL) enables the collaborative training of machine learning models across multiple decentralized devices and has been pivotal in harnessing diverse data sources while preserving privacy. FL implementation involves two broad methods: centralized and decentralized. In the first method, a central server acts as a trusted coordinator among different clients, and model weights are shared between each client and the central server. The latter uses a peer-to-peer (P2P) framework in a fully decentralized way. The natural language processing capabilities reached a new height with the development of a Generative Pre-trained transformer (GPT) series (Kalyan, 2023). However, with the increase in model size and complexity, challenges arise in securely aggregating weights from dispersed clients. Blockchain technologies known for their robust security and decentralized ledger systems can handle such challenges in a federated LLM setup. By leveraging blockchain to transfer weights between local clients, we establish an efficient communication of weights and model updates in a decentralized environment.

In this research, we conducted several experiments to explore the trade-off between accuracy vs latency for server and serverless cases. During these experiments, we employed different anomaly detection methods like PageRank, DBSCAN, and modified Z-scores to detect anomalies in the client nodes and remove them to increase the latency of the network. We further personalized client's data into IID and non-IID and prune the anomaly nodes.

Method

In this section, we will provide the motivation behind this work with problem formulation and an overview of the proposed framework.

Motivation and Contribution

The iterative nature of the FL does not eliminate the network congestion and privacy threats problem completely (Kavitha et al., 2022) that occurs during the transfer of weights.
Indeed, for complex models, largescale applications, and high-frequency updates, the communication overhead is not negligible and needs to be addressed as stated in (Chhetri et al., 2023). FL prevents visibility into the local dataset and operation process of individual nodes due to privacy concerns. As a result, such systems are susceptible to abnormal actions from the nodes. Identifying such anomalous nodes is crucial, as such abnormalities can diminish the system's accuracy and efficiency. Several research studies (Thudumu et al., 2020) have been proposed to decrease the number of bits transferred for each worker update, but removing anomaly nodes, which contribute more to the network congestion problem, is novel. Our proposed research addresses this in a decentralized, federated learning setup and identifies anomaly nodes using methods like DBSCAN, PageRank, and Modified z scores.

Our framework uses a Hyperledger sawtooth distributed ledger for synchronous handling of weights between the initial client node and other remaining nodes. Blockchain's distributed nature enhances the resilience of the federated LLM network against attacks and operation failures. Page rank anomaly detection involves identifying the anomalies based on the connectivity patterns and identifies anomalies based on the ranks assigned to each node (Chung, 2014). DBSCAN is a density-based clustering algorithm that marks data points not belonging to any cluster as outliers (Khan et al., 2014). Modified Z-scores are variations of standard Z-scores and are sensitive to outliers that can significantly affect the mean and standard deviation (Tiwari & Maurya, 2023).

Research Gaps

In the realm of peer-to-peer networks, there is a significant gap in the research on how to balance the trade-off between accuracy and latency while maintaining efficient communication. Developing advanced optimization techniques to simultaneously enhance both latency and accuracy is a critical yet underexplored area. There are challenges in scalability, as existing research needs to adequately address how the networks can maintain effective trade-offs with expanding size and complexity. Also, the influence of different network topologies on this trade-off is yet another unexplored area that significantly impacts performance. Additionally, the integration of energy-efficient strategies in maintaining the balance has not been sufficiently explored. Further, most previous works are based on synchronous transmission (Lee & Ko, 2023), while asynchronous transmission for federated LLM is limited.

BC-FL Algorithm

Considering a network represented by a graph $G = (V, E)$, where V is the set of nodes (client devices) and E is the set of edges (connections between clients). Each edge $e_{ij} \in E$ between nodes i and j has an associated bandwidth B_{ij}. The problem involves finding an efficient way to distribute a global model across this network from the first client to the remaining clients and performing latency vs accuracy trade-off. We have applied the Bellman-Ford algorithm (Barkund et al.) to find the shortest paths in a network from a node, represented as $T = (V_T, E_T)$, where $V_T \subseteq V$ and $E_T \subseteq E$. However, when anomalies are detected in the network leading to the removal of nodes, the network's structure, or the graph T, becomes altered. This modification requires re-applying the shortest path algorithm to accurately reflect the changes in the network's topology. The Bellman-Ford algorithm would need to be rerun on this modified network to determine new shortest paths from a given node, considering the removal of affected nodes and any resultant changes in path lengths or connectivity.

Problem Formulation

We have considered Independent and identically distributed (IID) and non-IID cases. Our framework deals with a network optimization problem that aims to minimize the total network latency. The total network latency is defined as the sum of a fixed delay for global model calculation (D_G)

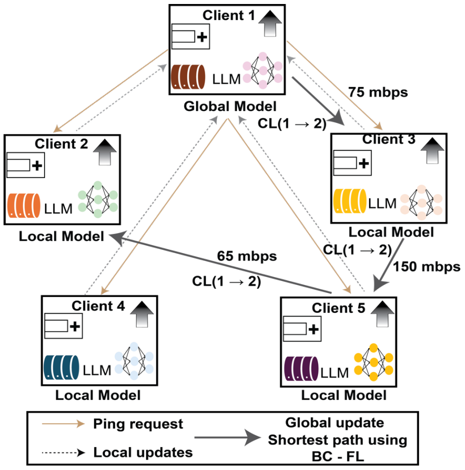

Figure 1: Overview of the proposed framework.

and the maximum latency from a given node to all other nodes in a selected subset of a directed graph. The latency between two nodes is defined as the ratio of model size to the bandwidth of the channel connecting them. We consider the network as a directed graph with weights as latency. The objective is to minimize the total network latency asynchronously. Overview of the framework is shown in Figure 1. Latency between nodes i and j , $L_{i,j}$ is calculates as:

$$L_{i,j} = \frac{M}{B_{i,j}} \tag{1}$$

where M represents size of the model updates and B is the channel bandwidth.

For a node i, maximum latency to nodes in subset S and total network latency for subset S:

$$L(i,S)_{max} = max_{i\in S}\, L_{i,j} \tag{2}$$

$$L(S)_{total} = D_G + max_{i\in S} L(i,S)_{max} \tag{3}$$

where S represents subset of nodes after removing anomaly nodes obtained from different methods to minimize latency, D_G is the delay of global model calculation.

Summary of the approach and its implications in network optimization:

$$L(S)_{total} = (D_G + max_{i\in V_T\setminus H}\,(L(H\rightarrow i)) \tag{4}$$

where $L(H\rightarrow i)$ is the latencies along the path from H to different nodes i in the graph T'.

The problem of asynchronous way formulated as:

$$Minimize\ (D_G + max_{i\in V_T\setminus H}\ (L(H\rightarrow i))) \tag{5}$$

The problem of synchronous way formulated as:

$$Minimize\ (D_G + (\sum_{e_{i,j}\ \in Path\ (H\rightarrow i)} L_{i,j})) \tag{6}$$

$$s.t.\ T' = MST(G\ Anomaly\ Nodes) \tag{7}$$

$$L_{i,j} = \frac{S}{B_{i,j}}, \forall e_{i,j} \in\ E_T \tag{8}$$

where MST is minimum spanning tree. The results provide a comparative analysis using our framework.

Experiment and Results

Datasets

Medical datasets are rare and difficult to find due to HIPPA privacy regulations, so we have used open-source data scraped from mtsamples. The dataset consists of six columns: a short description of transcription, medical specialty classification of transcription, transcription title, complete transcription, and keywords. We have considered only two columns, a description of the full transcript and medical specialty, for our experimentation purpose. There are 40 medical specialties, such as bariatrics, immunology, surgery, etc.

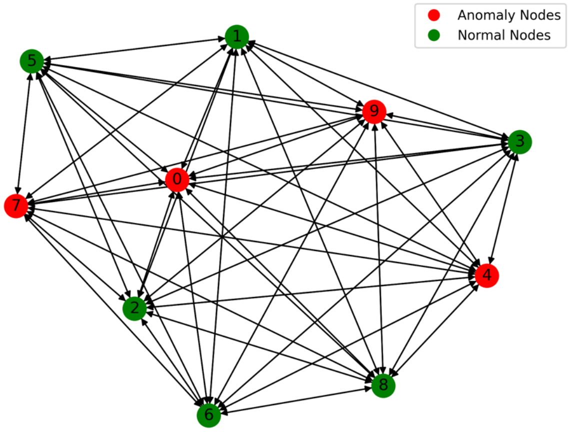

Figure 2: Anomaly nodes simulation using PageRank.

The dataset is divided into 12,000 training data samples and 3,000 testing data samples.

Implementation Details

We shuffled the dataset randomly and divided it into different sets distributed among multiple clients and then utilized the flower framework with a pre-trained model for LLM, biobert-v1.1 to determine the latencies and accuracy. Our proposed framework for the serverless case provides information on accuracy, latency, and memory usage. In the case of non-IID datasets, we divided the dataset into multiple shards based on the number of clients. In both cases, we used a learning rate of 1e-5 and a batch size of 16 with the AdamW optimizer for the training process. We trained this process for 20 epochs, resulting in 20 global accuracies. We simulated anomaly nodes using various anomaly detection algorithms. Following the removal of these anomaly nodes, we recalculated the shortest path for transferring weights between clients as shown in Figure 2.

Results

As shown in Figure 3, we have visualized the graph for accuracy vs number of epochs and provided a comparative study on accuracy variations for server and serverless case (IID and non-IID). We can clearly observe that accuracies are higher in the serverless than the server case. In Figures 4-5, we have illustrated bar plots for accuracies and latencies for each number of worker clients (5, 10, 20) for IID cases. We can clearly see that latencies are less, and accuracies are high for decentralized case as compared to centralized. In Figures 6-7, we have plotted the information passing time from initial client to remaining all clients in synchronous and asynchronous way. Based on the method, we removed the anomalous nodes and reran the bellman-ford algorithm to find the shortest time path for passing the information.

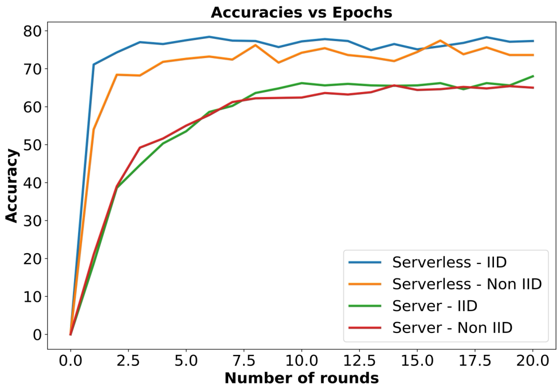

Figure 3: Accuracy vs Epochs comparison.

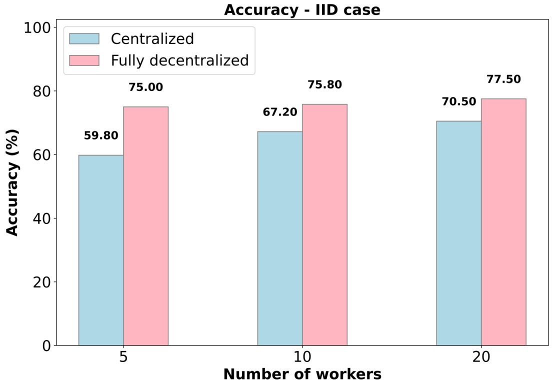

Figure 4: Comparison of accuracies with workers.

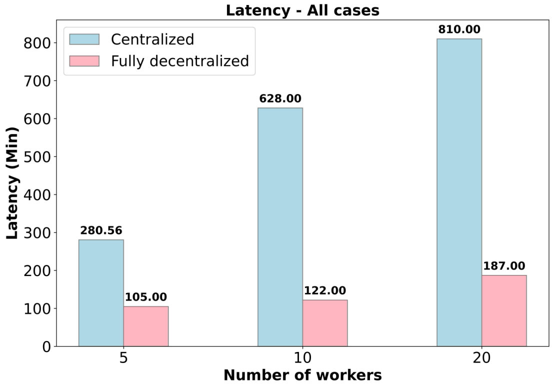

Figure 5: Comparison of latencies with workers.

Method	Anomaly nodes	Asynchronous time (sec)	Accuracy	
			IID	Non IID
DBSCAN	-	9.3	75	73
PageRank	0,4,7,9	7.44	74	67.8
Modified Z scores	8,9	9.3	74	72.6

Table 1: Latency vs Accuracy trade-off.

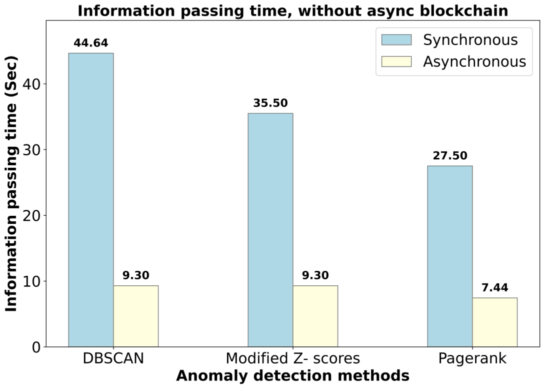

Figure 6: Comparison of information passing time.

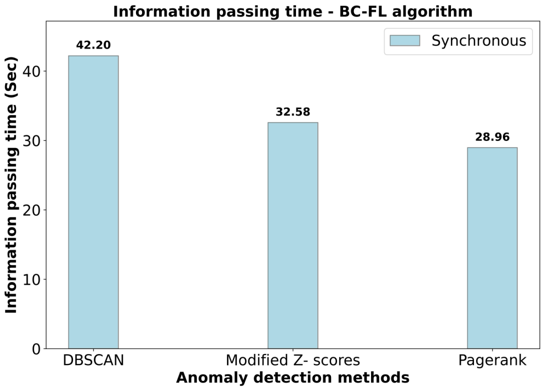

Figure 7: Comparison of information passing time BC-FL.

Later we have noted that passing time is less for the PageRank method (removed anomaly nodes: 8, 9) compared to other methods with little loss in accuracy. As shown in Table 1, we report the latency vs accuracy trade-off.

Conclusion

We have observed that federated LLM models constructed using non-IID data have low accuracy and higher latency than models built with IID data, potentially due to insufficient fine-tuning data and data heterogeneity. A fully decentralized system is more efficient than a centralized one in lines of accuracy (13% ↑) and latency (5X ↓). The asynchronous information passing time is longer for the DBSCAN method compared to other methods and from the latency vs accuracy trade-off analysis, we concluded that PageRank is the best method to identify anomaly nodes for good accuracy and latency. Future experiments will focus on improving accuracy in non-IID cases, with a specific strategy.

References

Amrollahi, F., Shashikumar, S. P., Holder, A. L., & Nemati, S. 2022. Leveraging clinical data across healthcare institutions for continual learning of predictive risk models. *Scientific reports*, *12*(1), 8380.

Barkund, S. H., Sharma, A., & Bhapkar, H. Survey of Shortest Path Algorithms.

Bharati, S., Mondal, M., Podder, P., & Prasath, V. 2022. Federated learning: Applications, challenges and future directions. *International Journal of Hybrid Intelligent Systems*, *18*(1-2), 19-35.

Chhetri, B., Gopali, S., Olapojoye, R., Dehbashi, S., & Namin, A. S. 2023. A Survey on Blockchain-Based Federated Learning and Data Privacy. 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC),

Chung, F. 2014. A Brief Survey of PageRank Algorithms. *IEEE Trans. Netw. Sci. Eng.*, *1*(1), 38-42.

He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., & Cambria, E. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Kalyan, K. S. 2023. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 100048.

Kavitha, T., Pandeeswari, N., Shobana, R., Vinothini, V., Sakthisudhan, K., Jeyam, A., & Malar, A. J. G. 2022. Data congestion control framework in Wireless Sensor Network in IoT enabled intelligent transportation system. *Measurement: Sensors*, *24*, 100563.

Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. 2014. DBSCAN: Past, present and future. The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014),

Lee, J., & Ko, H. 2023. Energy and Distribution-Aware Cooperative Clustering Algorithm in Internet of Things (IoT)-Based Federated Learning. *IEEE Transactions on Vehicular Technology*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. 2018. Improving language understanding by generative pre-training.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. 2023. Large language models in medicine. *Nature medicine*, *29*(8), 1930-1940.

Thudumu, S., Branch, P., Jin, J., & Singh, J. 2020. A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, *7*, 1-30.

Tiwari, A., & Maurya, R. K. 2023. Anomaly Detection in Time Series Data: Exploring Algorithms and Methods. *Advancement of Computer Technology and its Applications*, *7*(1), 37-46.

Zhu, J., Cao, J., Saxena, D., Jiang, S., & Ferradi, H. 2023. Blockchain-empowered federated learning: Challenges, solutions, and future directions. *ACM Computing Surveys*, *55*(11), 1-31.