Optimizing Communication and Performance in Federated Learning for Large Language Models

InSeo Song
Computer Engineering
Gachon University
Seongnam, Republic of Korea
z8086486@gachon.ac.kr

KangYoon Lee*
Computer Engineering
Gachon University
Seongnam, Republic of Korea
keylee@gachon.ac.kr

***Abstract*— LLMs(Large Language Models) has shown excellent performance in natural language processing and generation tasks. However, most LLMs learn by aggregating large amounts of data to a central server in a centralized environment, which causes problems such as data privacy infringement risk, high computing resource requirements, and increased operating costs. To solve these limitations, federated learning, which enables local learning without transmitting sensitive data to a central server, is attracting attention. In this study, we propose a new federated learning-based LLM Fine-Tuning strategy using the FlowerTune framework. The proposed methodology maximizes communication efficiency while ensuring data privacy in a distributed environment such as the medical domain, and verifies its performance based on pre-trained LLM. This study conducted experiments in 10-round and 100-round settings using three medical datasets: PubMedQA, MedMCQA, and MedQA, and its main contributions are as follows. It demonstrated high performance by achieving an average accuracy of 63.26% even in limited communication rounds (10 rounds). We applied a LoRA-based parameter-efficient fine-tuning technique to maximize learning efficiency and solve the resource constraint problem between clients. The empirical findings indicate that the advocated approach effectively addresses the constraints associated with centralized learning methodologies, thereby attaining superior performance while simultaneously minimizing communication overhead. This investigation posits the feasibility of integrating federated learning and LLMs within contexts where data confidentiality is paramount, such as in the medical field, and establishes a groundwork for subsequent inquiries into effective federated learning strategies.**

Keywords—Federated Learning, Large Language Models, Fine-Tuning

I. INTRODUCTION

LLMs(Large Language Models) have recently demonstrated exceptional efficacy in tasks pertaining to natural language processing and generation. Nevertheless, the majority of current LLMs employ extensive datasets within a centralized learning framework for both pre-training and fine-tuning. This centralized learning paradigm is subject to several limitations:

1. Data privacy: In scenarios involving sensitive information, such as medical or financial data, there exists a risk of privacy infringement during the transfer of data to a central server, thereby resulting in the exclusion of significant datasets from the learning process[1].

2. Computational and resource constraints: Centralized learning necessitates a high-performance central server alongside substantial storage capabilities for the aggregation and processing of large-scale datasets. This requirement poses challenges, particularly in edge devices or resource-constrained environments. The movement and processing of large-scale data incur considerable resource consumption, lead to degradation in learning speed, and escalate operational costs[2].

To address these issues, federated learning has garnered considerable attention. Federated learning circumvents the transfer of data to a central server by conducting local learning on individual clients and subsequently sharing only model updates (parameters) with the server. This methodology safeguards data privacy while facilitating effective model learning within a distributed framework[3][4].

The present study implements the FlowerTune-based LLM Fine-Tuning strategy as an innovative approach aimed at enhancing the performance of LLMs in a federated learning context. FlowerTune affords an efficient learning trajectory in a federated learning environment, and we have validated its efficacy and applicability through the fine-tuning of LLMs within the medical domain. Specifically, we contribute the following:

1. Validation of the performance of federated learning-based LLMs with minimal communication costs: This study experimentally investigates the correlation between the number of communication rounds and model performance. By contrasting the configurations of 10 and 100 rounds, we establish that this strategy can attain remarkable performance in federated learning-based fine-tuning with reduced communication expenses.

2. Proposal of an optimization methodology for federated learning-based LLM Fine-Tuning: We have implemented an optimization strategy that takes into account the data from each client, the number of communication rounds, and the Parameter-Efficient Fine-Tuning (PEFT) technique, and we have empirically assessed the means to enhance efficiency within a federated learning environment through this approach.

3. Examination of considerations pertinent to the application of federated learning environments: We have synthesized the primary technical challenges that emerge when utilizing federated learning libraries such as FlowerTune in conjunction with pre-trained LLMs from Hugging Face, as well as the considerations necessary for their resolution.

The structure of this paper is delineated as follows. In the Related Works section, we compile pertinent research and scrutinize the methodologies adopted in prior studies concerning LLM learning techniques founded on federated learning. In the Methodology section, we articulate the design of this study and provide a comprehensive description of the LLM Fine-Tuning methodology predicated on FlowerTune. In the Experiments and Result Analysis section, we analyze the experimental framework and primary findings, as well as deliberate on the implications of federated learning-based fine-tuning on LLM performance. Ultimately, we encapsulate the conclusions drawn from the study and propose directions for future research.

II. Related Works

Federated learning is characterized by inherent constraints regarding performance deterioration and communication expenditures, which arise from data heterogeneity and disparities in resources among clients. In order to address these constraints, numerous scholarly investigations have endeavored to identify effective methodologies for the amalgamation of federated learning and LLMs.

Kim et al. (2023) [5] proposed a methodology to markedly enhance transfer efficacy by assimilating pre-trained LLM into federated learning and training and transferring solely the adapter layer while constraining the transformation layer and embedding layer through the adapter mechanism. This investigation realized transfer efficacy and performance preservation across diverse datasets in the NLP and CV domains, and significantly broadened the applicability of LLM in federated learning contexts.

Jiang et al. (2024) [6] proposed a Customized Wireless Federated Learning methodology to enhance the learning of LLM in wireless network contexts. This investigation introduced PFIT (Local Model Personalization predicated on Reinforcement Learning) and PFTT (Efficient Personalization through the Integration of LoRA and Global Adapters) to tackle data heterogeneity and significant communication overhead. Consequently, it posited the potential for personalizing the fine-tuning strategy of LLM in Federated Learning environments, thereby augmenting performance and communication efficiency.

III. Methodology

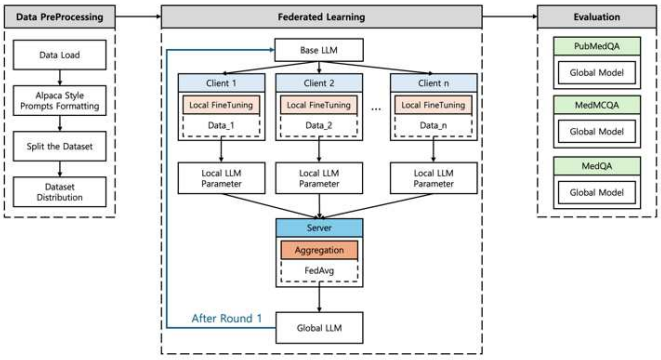

Fig. 1. Federated Learning Process.

A. Base Model and Data Processing

In this investigation, we employed Bio-Medical-Llama-3-8B, a pre-trained extensive language model, as the foundational model and executed fine-tuning through federated learning specifically tailored to the medical domain. The dataset utilized was medical_meadow_medical_flashcards, and the subsequent data preprocessing procedures were implemented to facilitate its processing.

Data ingestion and preprocessing: The dataset was segmented in accordance with the number of clients to facilitate distributed learning across the clients. The dataset was reformatted into an Alpaca-style prompt configuration to adapt the model for tasks structured as medical inquiries and responses.

Dataset segmentation and allocation: The entirety of the dataset was partitioned either uniformly or non-uniformly among clients to accurately reflect the characteristics inherent to the distributed environment.

B. Federated Learning Process

The federated learning methodology was executed as depicted in Figure 1.

1. Initial model dissemination: The central server disseminates the pre-trained Base LLM to 20 client entities.

2. Local Fine-Tuning: Each client engages in localized model Fine-Tuning employing the designated dataset. Fine-Tuning enhances parameter efficiency through the implementation of LoRA.

3. Model parameter update and aggregation: Each client transmits the locally refined model parameters to the central server. The server consolidates the model parameters received from the clients utilizing the FedAvg algorithm, thereby producing a global model.

4. Distribution of the global model and iterative process: The resultant global model is redistributed to the clients, initiating the subsequent round of learning. This iterative process is reiterated for a predetermined number of rounds, and this investigation implemented two distinct methodologies: 10 rounds and 100 rounds.

C. Evaluation

The investigation assessed the efficacy of LLM utilizing the PubMedQA[7], MedMCQA[8], and MedQA[9] datasets. At this juncture, 4-bit quantization[10] was implemented for computational and memory optimization. As a metric for evaluation, the model's precision was quantified on each dataset, and the findings were disseminated based on the overarching model.

IV. Experiments And Result Analysis

Algorithm 1 Federated Learning with FlowerTune

Input:
M_0: Pre-trained model (`ContactDoctor/Bio-Medical-Llama-3-8B`)
D_i: Local dataset of client i
N: Number of clients
R: Total number of training rounds
Hyperparameters: B (Batch size), η (Learning rate), LoRA Parameters (r, α), q (Quantization level)

Output:
Global model M_G

Initialization:
1. The server distributes the initial model M_0 to all clients $C_1, C_2, ..., C_N$.
2. Each client prepares local data D_i and initializes hyperparameters B, η, r, α, q.

For $t = 1$ **to** R **do**:

1. **Local Training:**
   Each client i performs local fine-tuning:
   $$M_i = FineTune(M_0, D_i, B, \eta, LoRAParameters)$$
   The updated local parameters ΔM_i are sent to the server.
2. **Aggregation on Server:**
   The server aggregates the parameters from all clients using the FedAvg algorithm:
   $$M_G = FedAvg(\{\Delta M_i\}_{i=1}^{N})$$
3. **Global Model Distribution:**
   The updated global model M_G is distributed to all clients for the next round.

Evaluation:
Every $t = 10, 20, ..., R$ rounds, evaluate M_G on PubMedQA, MedMCQA, and MedQA datasets to monitor performance (e.g., accuracy).

Return:
The final global model M_G after R training rounds.

The empirical investigations conducted in this research are comprehensively delineated in Algorithm 1 and adhere to a systematic approach for the refinement of a substantial pre-trained language model employing FlowerTune within a distributed client-server framework. This methodological approach guarantees optimal learning efficacy while simultaneously safeguarding data confidentiality. Algorithm 1 can be succinctly encapsulated as follows:

Upon initialization, the primary server disseminates the pre-trained language model M_0 to all collaborating clients $(C_1, C_2, \ldots, C_n)$ to commence the learning process. Each client organizes a local dataset D_i and establishes hyperparameters including batch size B, learning rate n, and LoRA parameters r, a.

In the context of localized training, each individual client engages in the refinement of the pre-trained model M_0 by employing the dataset D_i. Through the implementation of the Low-Rank Adaptation (LoRA) mechanism, exclusively the lightweight adapter layer undergoes training, whereas the transformer layers and embedding layers remain in a static state.

Aggregation on the central server entails the transmission of updated local model parameters from client devices to the central repository. The server employs the Federated Averaging (FedAvg) algorithm to compute the average of the client parameters, thereby producing a comprehensive global model denoted as M_G.

The Global Model Distribution entails the dissemination of the revised global model M_G to all participating clients in preparation for the subsequent training iteration.

Assessment, the efficacy of the global model can be assessed as a consequence of a particular iteration during the educational process, and in this investigation, the efficacy of the global model was quantified as a consequence of the terminal iteration utilizing the PubMedQA, MedMCQA, and MedQA datasets.

TABLE I. 10 ROUND PERFORMANCE RESULTS

	PubMedQA	**MedMCQA**	**MedQA**	**Average**
Accuracy	70.60%	57.68%	61.50%	63.26%

TABLE II. 100 ROUND PERFORMANCE RESULTS

	PubMedQA	**MedMCQA**	**MedQA**	**Average**
Accuracy	71.60%	52.47%	52.08%	58.80%

In the aggregate of Round 10, the datasets PubMedQA, MedMCQA, and MedQA exhibited accuracy rates of 70.60%, 57.68%, and 61.50% respectively, culminating in a mean accuracy of 63.26%. Conversely, in the cumulative analysis of Round 100, a marginal enhancement in accuracy was noted for PubMedQA (71.60%), whereas a decline in performance was evident in the other evaluated metrics.

The rationale underlying the superior performance observed over 10 rounds, as opposed to that over 100 rounds, can be attributed to the hyperparameters established within this research, which were specifically tailored to minimize communication costs while accommodating a limited number of rounds. Notably, hyperparameters including the learning rate, batch size, and LoRA parameters during the local learning phase were meticulously optimized, taking into account the distinctive

data characteristics and communication limitations inherent to each client. This strategic optimization facilitated rapid convergence and effective integration of the global model during the initial stages of model updating, thereby attaining high performance even with a reduced number of rounds.

Through this, we have established that elevated performance can be attained with a reduced number of iterations through an optimal fine-tuning methodology, and communication efficacy can be assured in a federated learning context. Furthermore, we have fine-tuned the pre-trained LLM in a federated learning framework to demonstrate superior performance in the medical sector, indicating the feasibility of training models without centralized transmission of sensitive medical information by preserving data locally and advancing with the learning process.

V. Conclusion

This research introduces a FlowerTune oriented Fine-Tuning strategy as an innovative methodology aimed at enhancing the efficacy of pre-trained LLMs within a federated learning framework. This approach illustrates the feasibility of efficiently training models while safeguarding sensitive information from being relayed to a central server in the healthcare sector, thereby showcasing both efficiency and performance improvements in a federated learning context.

We propose a methodology that can sustain elevated efficacy even in a constrained communication milieu, and systematically examine the correlation between communication expenses and efficacy by contrasting 10 round and 100 round configurations. Furthermore, we assess an optimization approach that enhances fine-tuning efficiency in a federated learning context by employing a lightweight parameter-efficient fine-tuning technique, LoRA, and addresses challenges such as data disparity among clients. The experimental findings indicate that the proposed methodology demonstrates superior performance across three medical domain datasets.

Although this investigation illustrates the capacity of pre-trained LLM in federated learning-centric fine-tuning, additional inquiry is requisite to enhance the model efficacy and efficiency in FL contexts:

1. Addressing data heterogeneity and imbalance necessitates the exploration of methodologies like client weight adjustment and dynamic learning rate adjustment.

2. Confirmation of relevance to diverse sectors: Although this investigation centered on the healthcare sector, additional inquiry is requisite to assess the efficacy and performance of federated learning-based fine-tuning in alternative high-privacy sectors such as jurisprudence, finance, and pedagogy.

This research posits the feasibility of attaining superior performance with minimal communication expenses while preserving data confidentiality through the integration of federated learning and extensive language models, it is anticipated that this development will facilitate the broader application of large-scale language models within federated learning contexts.

Acknowledgment

This research was supported by the National Research Foundation of Korea (NRF-2022R1F1A1069069).

References

[1] P. H. R. Emerick, S. C. Sampaio, B. L. Dalmazo, A. Riker, A. V. Neto, and R. Immich, “Enhancing Privacy in Healthcare: A Multilevel Approach to (Pseudo)Anonymization,” 2024 International Wireless Communications and Mobile Computing (IWCMC), May 27, 2024.

[2] M. Al Zaabi and S. Alhashmi, “Big data security and privacy in healthcare: A systematic review and future research directions,” Information Development, vol. Q2, April 2024.

[3] J. Moon, S. Yang, and K. Lee, “FedOps: A Platform of Federated Learning Operations With Heterogeneity Management,” IEEE, vol. 12, 2023.

[4] Q. Tan, S. Wu, and Y. Tao, “Privacy-Enhanced Federated Learning for Non-IID Data,” Mathematics, vol. Q2, September 2023.

[5] G. Kim, J. Yoo, and S. Kang, “Efficient Federated Learning with Pre-Trained Large Language Model Using Several Adapter Mechanisms,” Mathematics, vol. 11, no. 21, p. 4479, October 2023.

[6] F. Jiang, L. Dong, S. Tu, Y. Peng, K. Wang, K. Yang, C. Pan, and D. Niyato, “Personalized Wireless Federated Learning for Large Language Models,” arXiv preprint arXiv:2404.13238, April 2024.

[7] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu, “PubMedQA: A Dataset for Biomedical Research Question Answering,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2567–2577, November 2019.

[8] A. Pal, L. K. Umapathi, and M. Sankarasubbu, “MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical Domain Question Answering,” arXiv preprint arXiv:2203.14371, March 2022.

[9] D. Jin, E. Pan, N. Ouattfolle, W. Weng, H. Fang, and P. Szolovits, “What Disease Does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams,” arXiv preprint arXiv:2009.13081, September 2020.

[10] A. Trusov, E. Limonova, D. P. Nikolaev, and V. Arlazarov, “4.6-Bit Quantization for Fast and Accurate Neural Network Inference on CPUs,” Mathematics, vol. Q2, February 2024.