Knowledge Transfer via Compact Model in Federated Learning (Student Abstract)

Jiaming Pei[1], Wei Li[1], Lukun Wang[2]

[1]School of Computer Science, The University of Sydney, Australia
[2]Shandong University of Science and Technology, China
jpei0906@uni.sydney.edu.au, weiwilson.li@sydney.edu.au, wanglukun@sdust.edu.cn

Abstract

Communication overhead remains a significant challenge in federated learning due to frequent global model updates. Essentially, the update of the global model can be viewed as knowledge transfer. We aim to transfer more knowledge through a compact model while reducing communication overhead. In our study, we introduce a federated learning framework where clients pre-train large models locally and the server initializes a compact model to communicate. This compact model should be light in size but still have enough knowledge to refine the global model effectively. We facilitate the knowledge transfer from local to global models based on pre-training outcomes. Our experiments show that our approach significantly reduce communication overhead without sacrificing accuracy.

Introduction

Communication efficiency is crucial for federated learning systems (Pei et al. 2022). The use of complex deep learning models intensifies communication challenges, stemming from the model's granularity or incremental data from clients. Large size models increase the volume of parameters to be exchanged, exacerbating communication overhead, and consuming more time and energy for synchronization. This is particularly problematic in bandwidth-constrained or real-time environments and can lead to inefficient power consumption in edge devices (Jiang et al. 2022). Optimizing model sizes without compromising learning capacity is thus essential for scalable federated learning.

Strategies like neural network pruning (Abdi et al. 2023), knowledge distillation (Guo et al. 2023), and parameter compression (Said, Pourreza, and Le 2022) aim to reduce communication overhead. Pruning eliminates redundant model parameters but often requires fine-tuning due to data heterogeneity, increasing computational and communication costs. Knowledge distillation allows smaller models to mimic larger ones but can compromise client data privacy in a federated setting. Parameter compression, while reducing model size, can degrade accuracy if applied aggressively.

To address the limitations of these methods, we propose a new framework, illustrated in Figure 1, to optimize the knowledge transfer in federated learning without increasing communication overhead. Our approach uses a compact global model for server-client communication, updated via prediction loss and probability from local models. The compact global model is an efficient and performance-retentive version of the original model, designed to minimize communication overhead. Our method is distinct from neural network pruning as it preserves the global model architecture, preventing the loss of crucial nodes or weights. Unlike conventional knowledge distillation techniques that centralize teacher model parameters, our method empowers local teacher models to directly update student models on individual clients. This decentralized approach significantly reduces communication overhead, leading to a more efficient and scalable knowledge distillation process. Furthermore, our framework stands apart from parameter compression techniques as it does not compromise key information by reducing the size of weight representations. In essence, our framework achieves efficient knowledge transfer through the alignment of predictive probability distributions, enhancing communication efficiency and safeguarding data privacy.

Methodology

In a federated learning system, each of the N participating clients possesses a local model θ_n and a corresponding dataset D_n. Before communication commences, clients train their local models using their own data and perform pre-training to refine the model parameters, as shown in Step 1 of Figure 1. To reduce communication overhead, the server initializes a compact model Θ with random parameters, ensuring conciseness compared to the local models θ_n while maintaining the same architecture. The compact model Θ acts as a communication intermediary between the server and the clients. Once the initialization of Θ is completed, the server distributes it to all participating clients, initiating communication upon the completion of θ_n training.

Step 2 outlines the single-round knowledge transfer process, iterated until the global model Θ converges. During each round, compact models Θ_n on clients absorb knowledge from local models θ_n and are aggregated into an updated global compact model Θ on the server for the next communication round. We update Θ_n by computing two losses: (1) the prediction loss of Θ_n on client n according to label information, and (2) the KL divergence be-

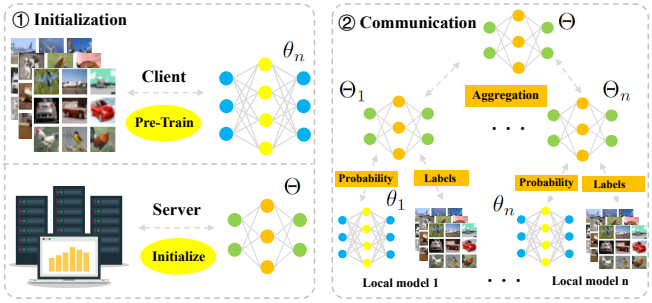

Figure 1: Overview of our framework. The process begins with Step 1, followed by iterative repetitions of Step 2 until convergence is achieved.

tween the softmax outputs of Θ_n and θ_n. This KL loss computes the logarithmic probability divergence, capturing differences between models as:

$$\Theta_n^{\varepsilon} = \Theta_n^{\varepsilon-1} - \lambda_1 F1(\Theta_n^{\varepsilon-1}) - \lambda_2 F2(\Theta_n^{\varepsilon-1}, \theta_n) \quad (1)$$

where $F1$ calculates the loss value of prediction from label information and $F2$ calculates the difference in probability distribution of the prediction results between θ_n and Θ_n in ϵ epoch. After updating, each client uploads the updated model Θ_n to server. The server then aggregates the compact models Θ_n to Θ with the optimization objective after receiving all Θ_n:

$$\min_{\theta \in \mathbb{R}^d} F(\Theta) = \sum_{n=1}^{N} \frac{D_n}{D} F_n(\Theta_n) \quad (2)$$

This minimizes the weighted overall loss across clients based on their relative dataset sizes D_n/D. Our approach efficiently extracts knowledge from local models into a compact global model, reducing communication costs while preserving personalized client data characteristics.

Experiment

We evaluated our framework on FMNIST and CIFAR10 using CNN and VGG9 models, with a compact same-architecture server model. After local training, the server model communicates and updates parameters. We compared against FedAvg, PrunFL (Jiang et al. 2022), and FEDGEN (Jiang et al. 2023) on accuracy, communication overhead, and training time. The a parameter of Dirichlet distribution was used to realize different levels of non-IID.

Our method outperforms baselines in terms of both accuracy and efficiency across various datasets, as shown in Table 1. Notably, our approach achieves high efficiency, characterized by minimal communication overheads and often reduced training times compared to benchmark methods. This enhanced efficiency stems from our compact server model, which significantly reduces communication costs. Furthermore, our framework strategically extrapolates salient features without incurring additional communication

FMNIST -α=0.2			
	Accuracy (%)	Overhead (Mb)	Time (s)
Ours	**89.21**	**9.01E+01**	**7.24E+02**
FedAVG	83.45	1.00E+02	1.65E+03
PrunFL	87.45	9.55E+01	1.35E+03
FEDGEN	88.54	1.12E+02	9.06E+02
CIFAR10 - α=0.2			
	Accuracy (%)	Overhead (Mb)	Time (s)
Ours	**72.16**	**4.01E+04**	**7.51E+03**
FedAVG	61.25	7.70E+04	2.17E+04
PrunFL	68.78	6.54E+04	8.91E+03
FEDGEN	71.75	6.45E+04	7.91E+04

Table 1: Performance comparison on two data sets.

by aligning predictive probabilities between global and local models. Overall, our paradigm demonstrates promising potential for federated learning, particularly in scenarios where communication efficiency and model accuracy are conflicting priorities.

Conclusion

Our approach leverages the advantages of both large client-side models and compact server-side models, harmonizing their performance through prediction probability distribution alignment. This innovative strategy presents a solution to communication overheads and data privacy concerns. However, its efficacy is contingent on robust client-side data. Future work will explore refining the alignment process and ensuring adaptability to real IoT environment.

References

Abdi, A.; Rashidi, S.; Fekri, F.; and Krishna, T. 2023. Efficient Distributed Inference of Deep Neural Networks via Restructuring and Pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 6640–6648.

Guo, Z.; Zhang, C.; Fan, Y.; Tian, Y.; Zhang, C.; and Chawla, N. V. 2023. Boosting graph neural networks via adaptive knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7793–7801.

Jiang, Y.; Wang, S.; Valls, V.; Ko, B. J.; Lee, W.-H.; Leung, K. K.; and Tassiulas, L. 2022. Model pruning enables efficient federated learning on edge devices. *IEEE Transactions on Neural Networks and Learning Systems*.

Jiang, Z.; Xu, Y.; Xu, H.; Wang, Z.; Liu, J.; Chen, Q.; and Qiao, C. 2023. Computation and Communication Efficient Federated Learning With Adaptive Model Pruning. *IEEE Transactions on Mobile Computing*.

Pei, J.; Yu, Z.; Li, J.; Jan, M. A.; and Lakshmanna, K. 2022. TKAGFL: a federated communication framework under data heterogeneity. *IEEE Transactions on Network Science and Engineering*.

Said, A.; Pourreza, R.; and Le, H. 2022. Optimized learned entropy coding parameters for practical neural-based image and video compression. In *2022 IEEE International Conference on Image Processing (ICIP)*, 661–665. IEEE.