TransFL: Selective Transformer Layer Sharing for Efficient Federated Learning

Soojin Kim, Minjung Kim, JaeYeon Park
Department of Mobile Systems Engineering, Dankook University
Yong-in, Republic of Korea
soojin0929@dankook.ac.kr, minjung.kim@dankook.ac.kr, jaeyeon.park@dankook.ac.kr

***Abstract*—Recent advances in vision tasks have shifted from CNN-based spatio-temporal models to Transformer-based architectures such as ViT. However, efficiently applying these models in federated learning (FL) remains underexplored. In this work, we propose TransFL, a framework that selects and shares only important Transformer layers based on gradient magnitude. This method reduces communication overhead while preserving performance.**

***Index Terms*—Federated Learning, Transformer, ViT**

I. Introduction

Transformer-based architectures have recently emerged as a dominant paradigm in computer vision, surpassing traditional convolutional neural networks in a variety of tasks. Vision Transformer (ViT) and its variants have demonstrated strong performance in image classification and representation learning by modeling long-range dependencies and capturing global context [1]. As these models become increasingly popular, there is a growing need to adapt them to real-world settings, including decentralized and privacy-sensitive environments. Federated learning provides a promising framework for such deployment by enabling collaborative training without sharing raw data [2]. However, most existing federated learning algorithms are tailored for CNN-based architectures and assume full model sharing, which is not scalable when applied to large Transformer models [3]. To make Transformer-based learning more efficient in federated settings, a selective and communication-efficient strategy is needed.

II. Motivation

The development of Transformer models for computer vision marks a significant shift away from conventional spatio-temporal CNNs [4], [5]. This evolution brings new opportunities but also raises practical challenges, especially in federated learning environments where clients must coordinate without revealing their private data [6]–[8]. Transformer models such as ViT are substantially larger and more complex than typical CNNs, making full model sharing during federated training costly and often infeasible. Moreover, existing federated learning methods do not account for the fact that some Transformer layers contribute more to model performance than others. In practice, certain layers play a dominant role in learning transferable representations, while others are less influential. Despite this, most frameworks continue to share all layers uniformly, which leads to unnecessary communication and computational waste. To address this inefficiency, we propose a new approach that selectively shares only the most impactful Transformer layers. Leveraging gradient magnitude as a signal to identify which layers are most active during learning, our framework aims to reduce overhead while maintaining competitive accuracy.

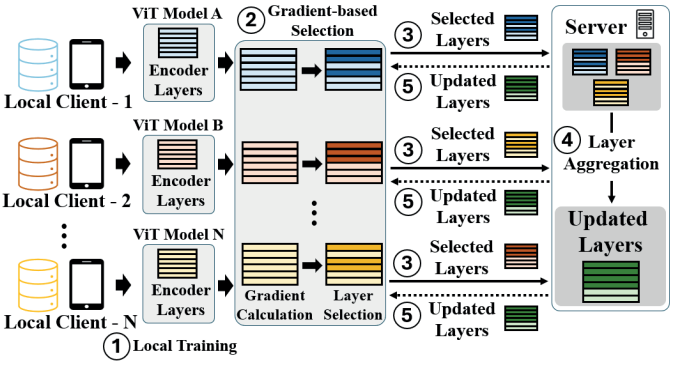

Fig. 1: Overall architecture of TransFL

III. Framework

We introduce TransFL, a federated learning framework designed specifically for Transformer-based models. The key idea is to reduce communication costs by sharing only the most informative layers across clients. At the beginning of each communication round, every client performs local training on its private dataset using a ViT model. During this process, the client computes the gradient norm for each Transformer layer ℓ to estimate its contribution to learning. Specifically, for a given layer ℓ with parameters θ_ℓ, we calculate the average gradient norm as follows.

$$G_\ell = \frac{1}{|\theta_\ell|} \sum_{i=1}^{|\theta_\ell|} \left\| \frac{\partial \mathcal{L}}{\partial \theta_{\ell,i}} \right\| \tag{1}$$

Where $\mathcal{L}$ is the local loss function and $|\theta_\ell|$ is the number of parameters in layer ℓ. This value G_ℓ reflects how much the layer contributes to loss reduction. Once training is complete, the client ranks all layers by their gradient magnitudes G_ℓ and selects the top-k layers with the highest values. These layers are considered the most informative and are chosen for synchronization. The selected layers are sent to the central server, which performs aggregation using a standard FedAvg [3] approach. After aggregation, only the updated parameters of

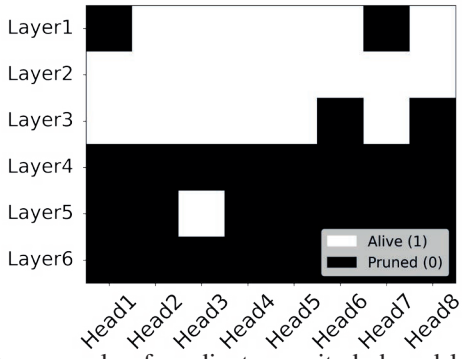

Fig. 2: An example of gradient magnitude-based layer selection

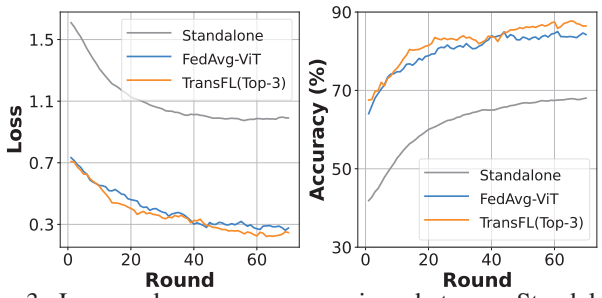

Fig. 3: Loss and accuracy comparison between Standalone, FedAvg, and TransFL on CIFAR-10

the selected layers are broadcast back to the clients. Layers not selected remain local and are not modified during this round. This selective communication strategy allows clients to focus bandwidth on the most critical parts of the model, enabling scalable and efficient training. Notably, TransFL does not require modifications to the Transformer architecture and is fully compatible with existing training pipelines.

IV. Results

We proposed TransFL, a federated learning framework for Transformer-based models that selectively shares important layers based on gradient magnitude. This approach reduces communication overhead while preserving competitive performance. To evaluate the effectiveness of TransFL, we conducted experiments on the CIFAR-10 dataset without data augmentation in a federated learning environment. Training was performed over 70 communication rounds, with 5 randomly selected clients participating in each round. Each client used a local batch size of 64. As the backbone model, we adopted a Vision Transformer (ViT) with 6 Transformer layers.

In the first experiment, we visualized the layer selection behavior of clients during federated training. Specifically, we generated a heatmap with attention heads on the x-axis and Transformer layers on the y-axis to analyze which parts of the model were consistently selected. As shown in Figure 2, layers 1, 2, and 3 were selected by a client, indicating that these layers contributed more significantly to local training. This confirms that TransFL can effectively identify and focus on highly activated layers during the learning process.

In the second experiment, as shown in Figure 3, we compared FedAvg and top-3 selective layer sharing of TransFL with a standalone baseline, where each client trains its ViT model independently without using FedAvg. The results show that both FedAvg and TransFL achieved significantly higher local accuracy (82.67% and 86.71%) than the standalone model (70.73%). Notably, even when the standalone model was trained with data augmentation techniques such as random cropping, jittering, and flipping, its accuracy improved to 81.83%, but still remained lower than FedAvg. These results highlight the effectiveness of federated knowledge sharing, even when only a small subset of Transformer layers is communicated. Furthermore, thanks to the gradient-based layer selection, TransFL reduced communication cost from 18.2 MB per round (i.e., full-model sharing) to just 9.16 MB when sharing only the top-3 layers. These suggest that TransFL not only improves model performance but also offers practical efficiency for real-world federated learning scenarios.

V. Conclusion

We proposed TransFL, a federated learning framework that efficiently trains Transformer-based models by selectively sharing only the most informative layers based on gradient magnitude. Unlike full-model aggregation, TransFL cuts communication cost by nearly half while preserving accuracy. Experiments on CIFAR-10 with Vision Transformer show that TransFL outperforms standalone training and achieves lower bandwidth usage, making it a practical solution for real-world federated learning with large Transformer models.

Acknowledgment

This research was supported by the MSIT, Korea, under the National Program for Excellence in SW, supervised by the IITP in 2024 (2024-0-00035).

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[4] J. Park, K. Lee, N. Park, S. C. You, and J. Ko, "Self-attention lstm-fcn model for arrhythmia classification and uncertainty assessment," *Artificial Intelligence in Medicine*, vol. 142, p. 102570, 2023.

[5] J. Park, H. Cho, R. K. Balan, and J. Ko, "Heartquake: Accurate low-cost non-invasive ecg monitoring using bed-mounted geophones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–28, 2020.

[6] J. Park and J. Ko, "Fedhm: Practical federated learning for heterogeneous model deployments," *ICT Express*, vol. 10, no. 2, pp. 387–392, 2024.

[7] J. Park, K. Lee, S. Lee, M. Zhang, and J. Ko, "Attfl: A personalized federated learning framework for time-series mobile and embedded sensor data processing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 3, pp. 1–31, 2023.

[8] Y. Shin, K. Lee, S. Lee, Y. R. Choi, H.-S. Kim, and J. Ko, "Effective heterogeneous federated learning via efficient hypernetwork-based weight generation," in *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, 2024, pp. 112–125.