IEEE Hong Kong 6G Wireless Summit

11-12 September 2024
Hong Kong, China

IEEE ComSoc IEEE Communications Society | IEEE

Department of of Electronic and Computer Engineering, HKUST

Federated Low-Rank Adaptation for Large Language Model Fine-Tuning Over Wireless Networks

Zixin Wang, from Prof. Khaled B. Letaief's Research Group

- Low-rank adaptation (LoRA) is an emerging fine-tuning method for personalized large language models (LLMs) due to its capability of achieving comparable learning performance to full fine-tuning by training a much smaller number of parameters. Federated fine-tuning (FedFT) combines LoRA with federated learning (FL) to enable collaborative fine-tuning of a global model with edge devices, leveraging distributed data while ensuring privacy. However, limited radio resources and computation capabilities of edge devices pose critical challenges on deploying FedFT over wireless networks.

- We propose a split FedFT framework to separately deploy the computationally-intensive encoder of a pre-trained model at the edge server while reserving the embedding and the task modules at the edge devices, where the information exchanges between these modules are carried out over wireless networks. By exploiting the low-rank property of LoRA, the proposed FedFT framework reduces communication overhead by aggregating the gradient of the task module with respect to the output of a low-rank matrix.

- To enhance learning performance under stringent resource constraints, we formulate a joint device scheduling and bandwidth allocation problem while considering average transmission delay. By applying the Lyapunov technique, we decompose the formulated long-term mixed-integer programming (MIP) problem into sequential sub-problems, followed by developing an online algorithm for effective device scheduling and bandwidth allocation.

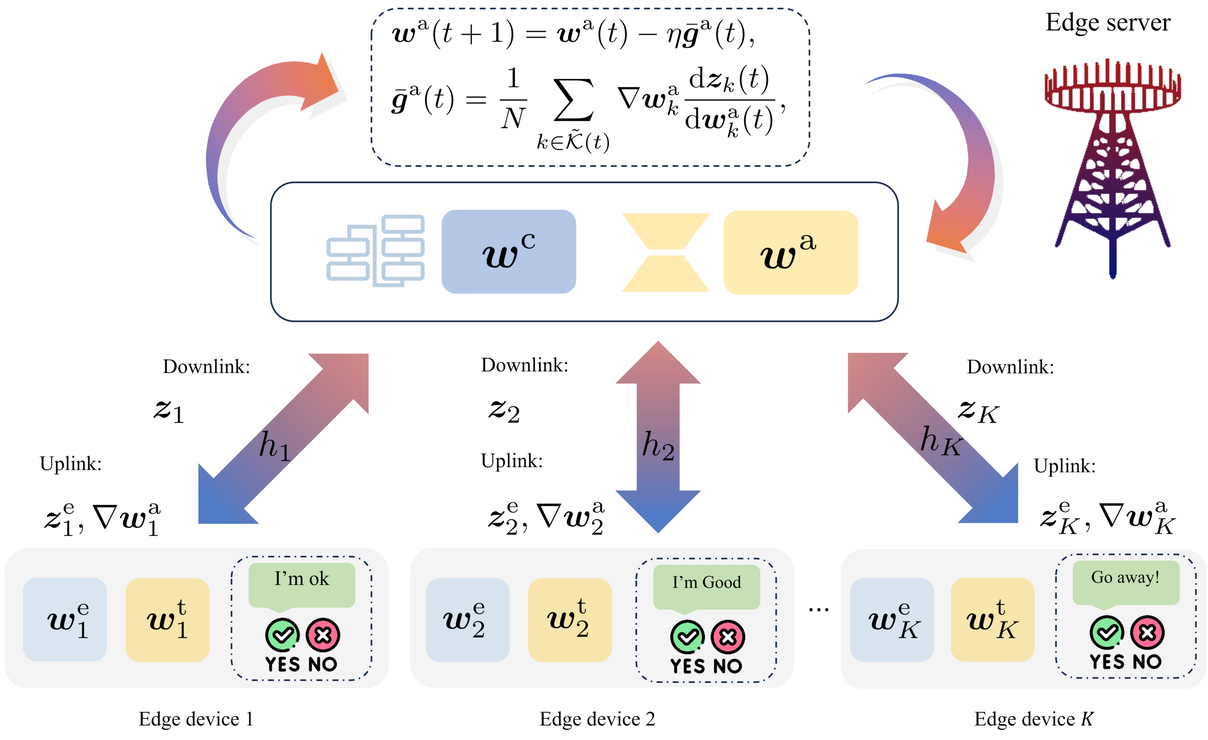


Illustration of the communication process for the proposed FedFT framework.

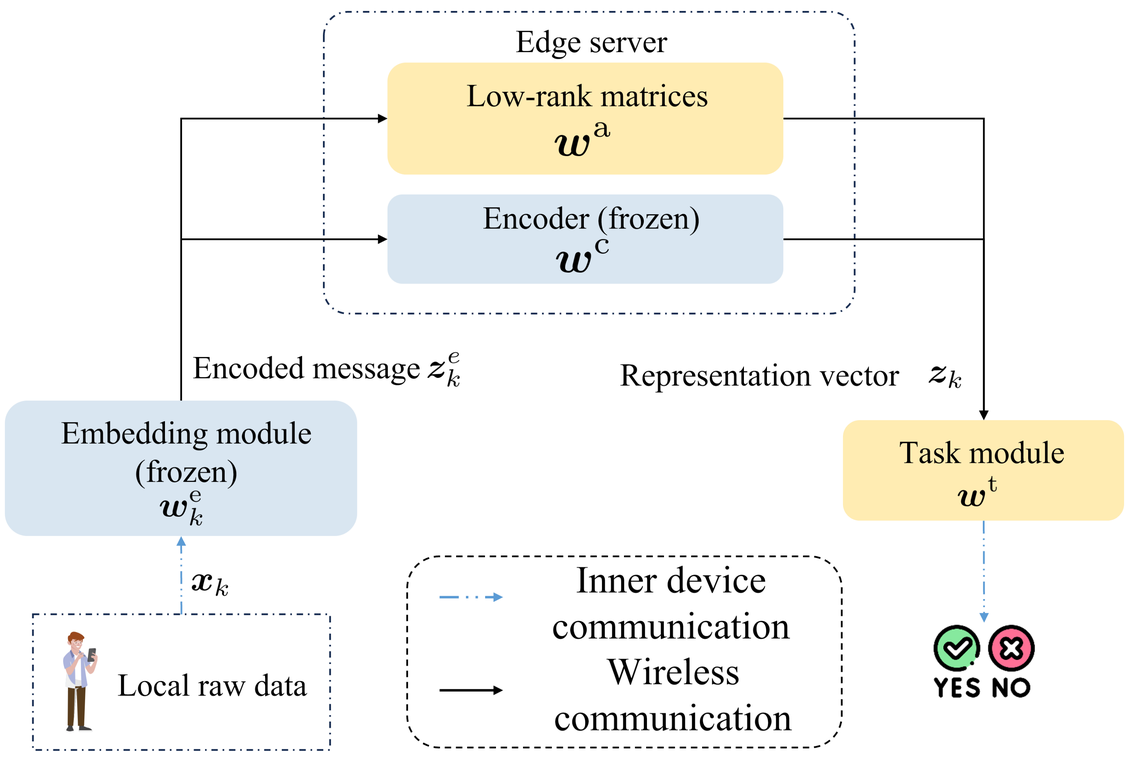
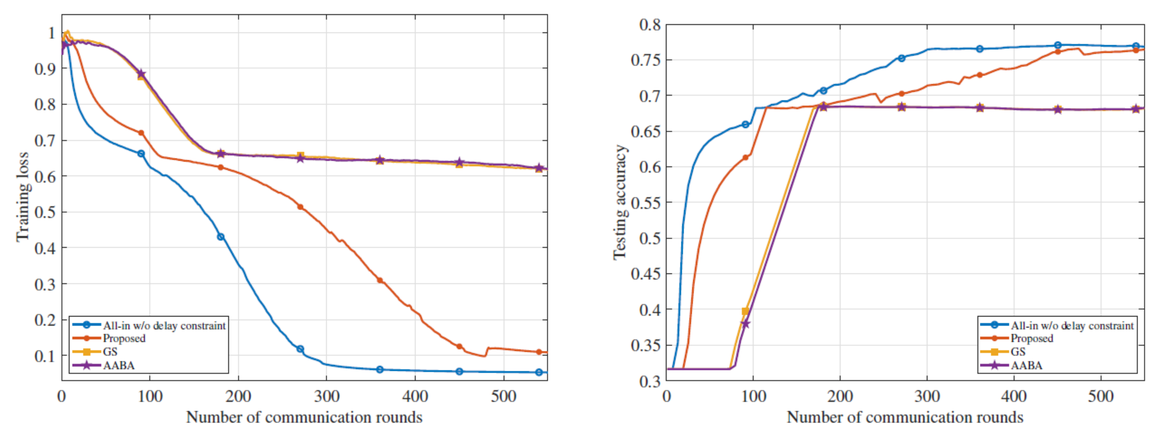
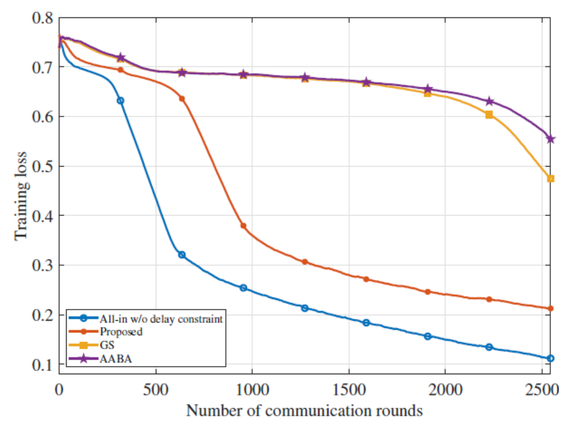


Illustration of the forward inference for the proposed FedFT framework.

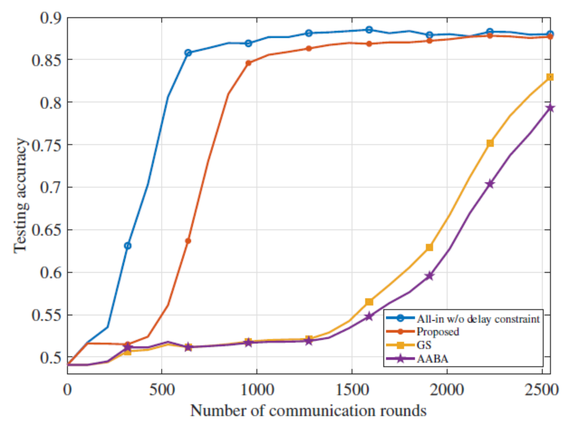


(c) Training loss (MRPC)

(d) Test accuracy (MRPC)



(a) Training loss (SST-2)



(b) Test accuracy (SST-2)

- We adopt **BERT** as the pre-trained model with its more than **100M** parameters and apply LoRA to the value, key, queue matrices and the dense layers in the pooler of the encoder. Specifically, we set the rank of the low-rank matrices in LoRA as 8, and the associated weight parameter as 16. As a result, the number of trainable parameters is **454K** and is **0.42%** of the total number of parameters. We adopt the SST-2 and the MRPC datasets to evaluate the learning performance of the proposed FedFT framework for natural language understanding.

- Beyond the best performance for the All-in scheme, the proposed online algorithm outperforms the rest of the benchmarks with a significant gap from the perspective of training loss and testing accuracy.

Related Works

1. **Z. Wang**, Y. Zhou, Y. Shi, and K. B. Letaief. "Federated Low-Rank Adaptation for Large Language Model Fine-Tuning Over Wireless Networks", accepted for publication in *IEEE Global Commun. Conf. (GLOBECOM),* Cape Town, South Africa, Dec., 2024.
2. **Z. Wang**, Y. Zhou, Y. Shi, and K. B. Letaief. "Federated Fine-Tuning for Pre-Trained Foundation Models Over Wireless Networks", submitted to *IEEE Trans. Wireless Commun.*, 2024.

Acknowledgment

This work was supported in part by the Hong Kong Research Grants Council under the Areas of Excellence Scheme Grant AoE/E-601/22-R