Heterogeneous LoRA for Federated Fine-tuning of On-Device Foundation Models

Yae Jee Cho[1*], **Luyang Liu**[2], **Zheng Xu**[2], **Aldi Fahrezi**[2], **Gauri Joshi**[1]
[1]Carnegie Mellon University, [2]Google Research
yaejeec@andrew.cmu.edu, {luyangliu,xuzheng,aldifahrezi}@google.com,
gaurij@andrew.cmu.edu

Abstract

Foundation models (FMs) adapt well to specific domains or tasks with fine-tuning, and federated learning (FL) enables the potential for privacy-preserving fine-tuning of the FMs with on-device local data. For federated fine-tuning of FMs, we consider the FMs with small to medium parameter sizes of single digit billion at maximum, referred to as *on-device FMs (ODFMs)* that can be deployed on devices for inference but can only be fine-tuned with parameter efficient methods. In our work, we tackle the data and system heterogeneity problem of federated fine-tuning of ODFMs by proposing a novel method using heterogeneous low-rank approximations (LoRAs), namely HETLORA. First, we show that the naive approach of using homogeneous LoRA ranks across devices face a trade-off between overfitting and slow convergence, and thus propose HETLORA, which allows *heterogeneous ranks* across client devices and efficiently aggregates and distributes these heterogeneous LoRA modules. By applying rank self-pruning locally and sparsity-weighted aggregation at the server, HETLORA combines the advantages of high and low-rank LoRAs, which achieves improved convergence speed and final performance compared to homogeneous LoRA. Furthermore, HETLORA offers enhanced computation efficiency compared to full fine-tuning, making it suitable for federated fine-tuning across heterogeneous devices.

1 Introduction

The emerging foundation models (FMs) (Bommasani et al., 2022; Zhou et al., 2023; Radford et al., 2021; Devlin et al., 2019; OpenAI, 2023; Google, 2022; Touvron et al., 2023; Brown et al., 2020; Google, 2022; Driess et al., 2023; Google, 2023) have shown remarkable zero/few shot learning capabilities, performing well on a variety of tasks including text/image generation with prompts, language translation, solving math problems, and conversing in natural language. Standard FMs, however, demand costly resources for directly fine-tuning their entire parameter space. To tackle this issue, many recent works have proposed different parameter-efficient fine-tuning (PEFT) methods of FMs such as prompt tuning (Lester et al., 2021), utilizing adapters (Houlsby et al., 2019), or low-rank adaptation (LoRA) of the original model (Hu et al., 2021) which freezes the original pre-trained parameters of the FM and train additional, smaller number of parameters instead.

These PEFT methods, however, assume that i) FMs are deployed to and trained with the data of a *single* machine/client for adaptation to the downstream task and that ii) the client has enough resources to even fit a standard FM of hundred billion size for, at least, inference. In practice, there are frequently cases where we are interested in fine-tuning FMs for on-device private data that is distributed across multiple devices (clients). For instance, sensitive and private data such as medical information or law-related documents may be hard to collect centrally in a private manner and fine-tuning of the FMs may need to be done at the edge (Manoel et al., 2023; Shoham and Rappoport, 2023; Zhang et al., 2023c).

In our work, we focus on such federated fine-tuning scenarios, where we train a set of parameters collaboratively across clients to obtain a global set of parameters that can be plugged in to the FM for the targeted downstream task. Note that federated fine-tuning is orthogonal to personalization of FMs in federated learning (FL) (Guo et al., 2023), which

*Work done while at Google Research. Corresponding authors: yaejeec@andrew.cmu.edu, luyangliu@google.com