

Pluto and Charon: A Time and Memory Efficient Collaborative Edge AI Framework for Personal LLMs Fine-Tuning

Bei Ouyang^{★1}, Shengyuan Ye^{★1}, Liekang Zeng², Tianyi Qian¹, Jingyi Li¹, Xu Chen^{†1}

¹School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

²IoT Thrust and Research Center for Digital World with Intelligent Things, HKUST (GZ), Guangzhou, China
{ouyb9,yeshy8,qianty,lijy573}@mail2.sysu.edu.cn,liekangzeng@hkust-gz.edu.cn,chenxu35@mail.sysu.edu.cn

ABSTRACT

Large language models (LLMs) have unlocked a plethora of powerful applications at the network edge, such as intelligent personal assistants. Data privacy and security concerns have prompted a shift towards edge-based fine-tuning of personal LLMs, away from cloud reliance. However, this raises issues of computational intensity and resource scarcity, hindering training efficiency and feasibility. While current studies investigate parameter-efficient fine-tuning (PEFT) techniques to mitigate resource constraints, our analysis indicates that these techniques are not sufficiently resource-efficient for edge devices. Other studies focus on exploiting the potential of edge devices through resource management optimization, yet are ultimately bottlenecked by the resource wall of individual devices.

To tackle these challenges, we propose **Pluto and Charon (PAC)**, a time and memory efficient collaborative edge AI framework for personal LLMs fine-tuning. PAC breaks the resource wall of personal LLMs fine-tuning with a sophisticated algorithm-system co-design. (1) Algorithmically, PAC implements a personal LLMs fine-tuning technique that is efficient in terms of parameters, time, and memory. It utilizes Parallel Adapters to circumvent the need for a full backward pass through the LLM backbone. Additionally, an activation cache mechanism further streamlining the process by negating the necessity for repeated forward passes across multiple epochs. (2) Systematically, PAC leverages edge devices in close proximity, pooling them as a collective resource for in-situ personal LLMs fine-tuning, utilizing a hybrid data and pipeline parallelism to orchestrate distributed training. The use of the activation cache eliminates the need for forward pass through the LLM backbone, enabling exclusive fine-tuning of the Parallel Adapters using data parallelism. Extensive evaluation based on prototype implementation demonstrates that PAC remarkably outperforms state-of-the-art approaches, achieving up to 8.64 \times end-to-end speedup and up to 88.16% reduction in memory footprint.

KEYWORDS

Edge intelligence, large language model, parameter-efficient fine-tuning, pipeline parallelism, data parallelism, parallel processing

★: Equal contributions. †: Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ICPP '24, August 12–15, 2024, Gotland, Sweden
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1793-2/24/08
<https://doi.org/10.1145/3673038.3673043>

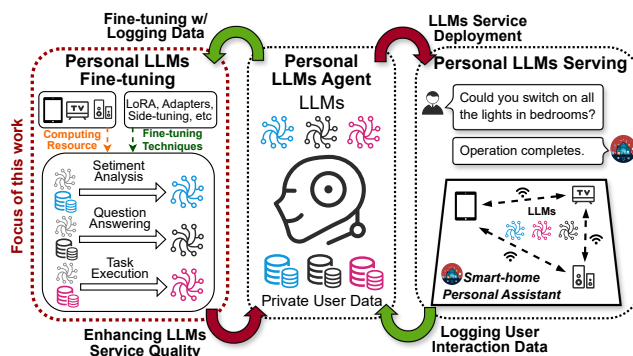


Figure 1: An illustration of hosting personal LLM-based intelligent agents within a smart home.

ACM Reference Format:

Bei Ouyang^{★1}, Shengyuan Ye^{★1}, Liekang Zeng², Tianyi Qian¹, Jingyi Li¹, Xu Chen^{†1}. 2024. Pluto and Charon: A Time and Memory Efficient Collaborative Edge AI Framework for Personal LLMs Fine-Tuning. In *The 53rd International Conference on Parallel Processing (ICPP '24)*, August 12–15, 2024, Gotland, Sweden. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3673038.3673043>

1 INTRODUCTION

Large language models (LLMs) [13, 20, 22] have ushered in a revolution in machine intelligence, owing to their exceptional capabilities in a wide range of machine learning tasks. While born on datacenter warehouse, LLMs have quickly sunk to edge devices and facilitated a range of intelligent applications at the network edge, such as *intelligent personal assistants* (IPAs) which are software agents that can augment individuals' abilities, complete complicated tasks, and even satisfy emotional needs. A recent survey [14] targeting LLM-based IPAs has revealed that over 80% of industry experts believe that, owing to the sensitive and privacy-critical nature of user data, personal LLMs should be fully (or primarily) hosted at the edge in order to enable privacy-preserving model personalization and serving. Figure 1 illustrates the scenario of hosting a personal LLM-based intelligent agent within a smart home. A personal LLM agent provides users with high-performance, privacy-preserving intelligent services. Meanwhile, the agent also tracks user interactions, learns from experiences, and extracts knowledge to fine-tune the personal LLMs and further enhance the service quality.

While the serving of LLMs on edge devices has been made feasible through careful engineering [6, 26, 28], fine-tuning these models remains significantly challenging due to the resource-intensive nature of LLM training. Towards alleviating the resource challenges, some research works [4, 17] have explored parameter-efficient fine-tuning (PEFT) techniques, such as Adapters [9] and LoRA [10],

which modify less than 2% of the model's parameters, thereby reducing resource requirements. Although these techniques are highly parameter-efficient, our analysis observes that they are not resource-efficient enough for edge environments. The inefficiency stems from Adapters and LoRA embedding trainable structures in the LLM backbone, necessitating complete backward passes through the LLMs during backpropagation. As we will empirically show in §2, fine-tuning a popular LLM of T5-Base (0.25B) by Google with PEFT techniques can only reduce computational overhead by up to 30% compared to full model fine-tuning. In practice, fine-tuning the T5-Base on a typical edge device (e.g., NVIDIA Jetson Nano [1]) still demands a minimum of 72.6 minutes per training epoch employing LoRA. Moreover, on-device fine-tuning is severely hindered by the memory wall of a single device. Predominant techniques require substantial memory to accommodate both model parameters and intermediate results. The observed memory expense for fine-tuning LLMs like T5-Large (0.74B), which exceeds 7.1 GB with LoRA and 6.8 GB with Adapters, is often unaffordable as typical mobile devices only possess 4-12GB DRAMs in total to run both system software and applications.

Other leading researchers have explored designing sophisticated resource management mechanisms (e.g., CPU-DSP co-execution [25], memory budget adapting [5, 23]) to leverage native resources, but are still bottlenecked by the intrinsic resource shortage of single device. To break the resource wall of a single device, we alternatively observe that prevalent edge environments like smart homes usually comprise a group of trusted idle devices beyond a single terminal (e.g., phones and smart-home devices). These accompanying devices are typically in physical proximity and can be associated as a resource augmentation for in-situ personal LLMs fine-tuning.

As motivated, in this paper, we introduce **Pluto and Charon (PAC)**, a time and memory efficient collaborative edge AI framework for personal LLMs fine-tuning. PAC's contribution goes beyond merely leveraging distributed edge devices, instead it breaks the resource wall of in-situ personal LLMs fine-tuning with a sophisticated algorithm-system co-design:

- **(Algorithm)** We evaluate two predominant PEFT techniques, Adapters and LoRA, and reveal that although parameter efficient, these techniques do not achieve sufficient resource efficiency. In light of the side-tuning [34] techniques, we employ not only parameter but also time and memory-efficient personal LLMs fine-tuning techniques with Parallel Adapters, which provides a dedicated gradient "highway" for the trainable parameters. Additionally, our Parallel Adapters stand out from other PEFT techniques by preserving the invariant intermediate activations from the LLM backbone for any given input sequence. By reusing these cached activations across multiple epochs, PAC increases resource efficiency and reduces fine-tuning latency by eliminating repetitive forward propagation through the LLM backbone.

- **(System)** We leverage edge devices in physical proximity and associate them as an edge resource pool for in-situ personal LLMs fine-tuning. Our fine-tuning process can be divided into two phases: (1) For the first epoch, the LLMs backbone, augmented with Parallel Adapters, is fine-tuned across multiple edge devices. To enhance scalability and training throughput, a hybrid parallelism approach that combines the merits of both data and pipeline parallelism is employed by PAC as a principle to manage collaborative training

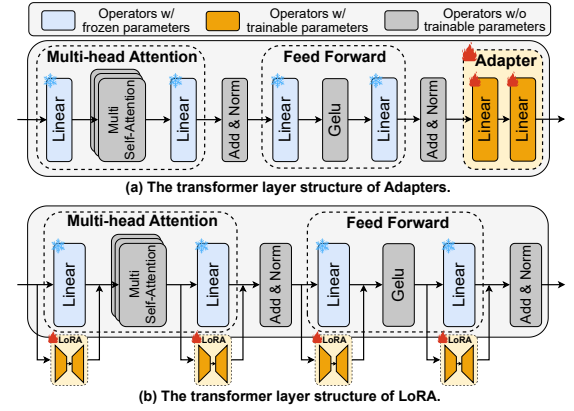


Figure 2: Illustration of the model structures with two PEFT.

across multiple edge devices. (2) In subsequent fine-tuning epochs, the activation cache obviates the need for forward propagation through the LLM backbone, allowing for the exclusive fine-tuning of our Parallel Adapters using data parallelism.

We implement PAC in realistic testbeds with a cluster of edge devices. Extensive evaluations across three LLMs demonstrate that PAC not only accelerates fine-tuning up to 8.64× faster than existing state-of-the-art methods but also significantly lowers the peak memory footprint by up to 88.13%, without sacrificing model performance. The main contributions are summarized as follows.

- We carry out extensive measurement studies on predominant PEFT techniques on resource-constrained edge devices and demonstrate that they are not sufficiently resource-efficient.
- We design a not only parameter but also resource efficient LLM fine-tuning technique for resource-limited edge environments.
- We propose a time and memory efficient collaborative edge AI framework PAC for the in-situ fine-tuning of personal LLMs, which combines sophisticated algorithm-system co-design.
- We implement PAC and evaluate it in realistic edge testbeds. Experimental results show up to 8.64× fine-tuning acceleration and 88.16% memory reduction without sacrificing performance compared to state-of-the-art methods.

2 MOTIVATION AND PRELIMINARIES

2.1 Transformer-Based LLMs and Fine-Tuning

Transformer-Based LLMs. Transformer-based LLMs have gained prominence in various language-related applications due to their impressive performance. These models consist of multiple Transformer layers, each comprising two main components: the Multi-head Attention and the Feed Forward block. The Multi-head Attention block utilizes linear layers to generate query (Q), key (K), and value (V) matrices for each attention head, allowing for independent self-attention computations. The outputs of these attention heads are then concatenated and processed through a final linear layer. The Feed Forward block involves two linear operations that increase the hidden size from h to $4h$ and then reduce it back to h .

Personal LLMs Fine-Tuning. The training of LLMs typically consists of two stages: pre-training and fine-tuning. Before being deployed for specific tasks, language models are often pre-trained

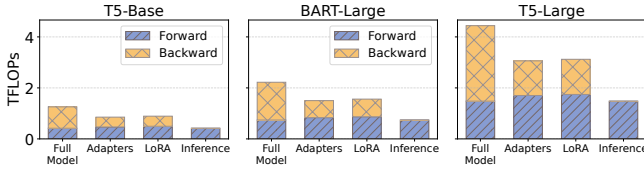


Figure 3: The comparison of floating point of operations (FLOPs). Mini-batch size: 16; sequence length: 128.

Techniques	Trainable Parameters	Memory Footprint (GB)			
		Weights	Activations	Gradients	Total
Full	737M (100%)	2.75	5.33	2.75	10.83
Adapters	12M (1.70 %)	2.80	4.04	0.05	6.89
LoRA	9M (1.26%)	2.78	4.31	0.04	7.13
Inference	/	2.75	/	/	2.75

Table 1: The breakdown of memory footprint. "Activations" contain the intermediate results and optimizer states. Model: T5-Large; mini-batch size: 16; sequence length: 128.

on extensive text datasets containing vast linguistic data. The pre-training process enables the model to acquire a general understanding of linguistic structure and patterns that are widely applicable. The fine-tuning adapts the pre-trained model to various, concrete downstream language tasks such as intelligent personal assistants. During actual deployment, the data required for fine-tuning is often generated at the user end, which can carry significant concerns regarding data security and privacy. In recent years, in-situ learning on edge devices [5, 15, 19, 23] has emerged as a promising approach for customizing LLMs while preserving user data fully in-situ.

Full model fine-tuning updates all parameters of an LLM for a specific downstream task. However, it is impractical for adapting an LLM to multiple distinct downstream tasks, as each target task would require maintaining a separate LLM with whole parameters. Some leading researchers have proposed parameter-efficient fine-tuning (PEFT) techniques [9, 10, 12, 16] which adapt a small subset of the LLM parameters or a set of newly added parameters for each new task. Adapters [9] and LoRA [10] are two of the most widely used PEFT techniques. Figure 2 illustrates how the transformer layer structure incorporates these two techniques. Specifically, adapters are compact bottleneck modules inserted at the end of each transformer layer. Similarly, LoRA injects trainable low-rank matrices into a frozen pre-trained model. These decompose the weight matrix parameter updates into two learnable low-rank matrices. Extensive experiments have demonstrated that these PEFT techniques can achieve performance comparable to full fine-tuning.

Although these PEFT techniques can greatly reduce the number of trainable parameters (around 98%), our analysis has revealed that they do not significantly decrease the computational and memory requirements during training. Figure 3 illustrates the floating point of operations (FLOPs) of different fine-tuning techniques and inference. Adapters and LoRA exhibit a limited reduction in computation (around 30%). Table 1 summarizes the memory footprint breakdown for T5-Large. Although Adapters and LoRA minimize the gradient memory footprint by restricting the number of trainable parameters, the memory consumed by activations still constitutes substantial overhead, with a maximum reduction of only 36%. The reason is that both Adapters and LoRA introduce trainable structures within the LLM backbone, such as at the end of each transformer block

or as bypasses to linear layers. Computing gradients for trainable parameters via backpropagation involves traversing the LLM backbone, compromising the efficiency of PEFT techniques due to the additional computational overhead and memory required to maintain considerable intermediate activations in LLM backbone.

2.2 Personal LLMs Fine-Tuning with Resource-Constrained Edge Devices

On-device fine-tuning enables leveraging idle resources at the edge while fully preserving user data privacy [5, 15, 19, 23]. This paradigm is widely adopted in privacy-sensitive edge computing applications. However, the resource-intensive nature of LLMs fine-tuning presents two significant challenges for resource-limited edge devices: **(1) The computational capabilities of edge devices are constrained.** Edge devices often face stark computational constraints compared to the powerful accelerators available in cloud datacenters. The Jetson Nano [1], a specialized platform for edge AI, peaks at a mere 0.47 TFLOPS, a tiny fraction of the 312 TFLOPS achievable with NVIDIA's A100 GPU typically found in data centers. Fine-tuning a T5-Base model with Adapters on a single Jetson Nano requires an epoch time of 72.6 minutes, which is 175.5× longer than that in a NVIDIA A100 GPU, showing the fundamental contradiction between intensive LLM fine-tuning workload and constrained on-board resources. **(2) On-device fine-tuning is hindered by the memory wall.** As shown in Table 1, fine-tuning the T5-Large model incurs a peak memory footprint that is often unaffordable for edge devices. For instance, although PEFT techniques such as Adapters and LoRA adjust only approximately 2% of the parameters, they still require substantial memory 6.89 GB and 7.13 GB respectively. Compared to full model fine-tuning, which requires over 10 GB, these techniques reduce memory usage by only 36%, often insufficient for typical mobile devices with 4-12 GB DRAM to run system software and applications.

To break the resource wall of a single edge device, in our work, we alternatively observe that prevalent edge scenarios usually comprise a group of trusted idle devices beyond a single terminal. These accompanying devices are typically located in close physical proximity, such as being connected to the same local area network (LAN), and can be utilized as a resource augmentation for in-situ LLMs fine-tuning acceleration. While several pioneering research works [24, 28] have delved into collaborative edge computing to overcome resource limitations faced by edge devices, the majority of these works primarily focus on LLMs inference. Other studies [4, 27] employing federated learning for fine-tuning LLMs with collaborative edge devices primarily address the dissolution of data silos, rather than resource augmentation within LANs.

3 SYSTEM OVERVIEW

PAC is a time, memory and parameter efficient collaborative framework for personal LLMs fine-tuning across multiple edge devices. PAC first equips the target LLM with our Parallel Adapters module (Step ①). PAC profiler fine-tunes the LLM using a calibration dataset on edge devices to record the runtime profile required for parallelism planning (Step ②). PAC planner then takes the profiling results as input and generates planning configurations, including LLM partitioning points and device grouping strategies (Step ③).

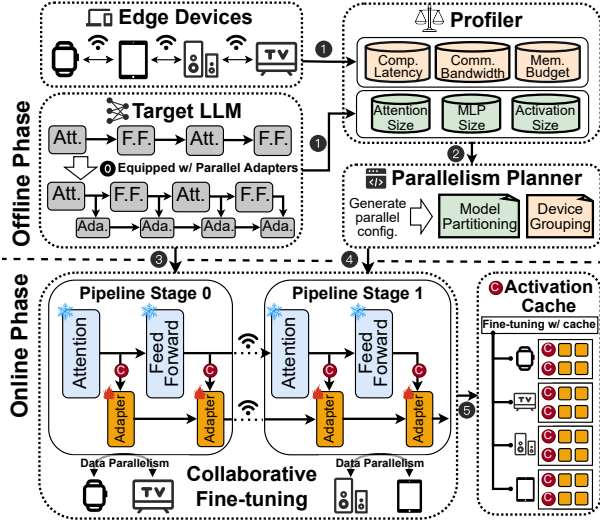


Figure 4: PAC workflow.

We configure the Parallel Adapters as trainable while freezing the LLM backbone parameters (Step ③). The parallel configurations generated by the PAC planner are then applied to the edge devices, enabling time, memory, and parameter-efficient hybrid data and pipeline parallelism fine-tuning of the target LLM (Step ④). Since the LLM backbone parameters remain fixed, the intermediate activations generated by the backbone model are invariant for a given input sequence. The PAC maintains a cache of these invariant activations. Through leveraging the cached activations, the efficiency of the fine-tuning process can be accelerated (Step ⑤).

4 TIME, MEMORY AND PARAMETER EFFICIENT FINE-TUNING ALGORITHM

4.1 Fine-Tuning LLMs with Parallel Adapters

Observation and Key Insight. As discussed in §2, while techniques such as LoRA [10] and Adapters [9] reduce the number of parameters that need to be updated during fine-tuning, they do not significantly reduce the computational and memory requirements during the training on edge devices. This is because the parameters being updated are still inside the LLM backbone. To calculate the gradients for backpropagation, the full backward passes through the entire pre-trained model are still necessary, as illustrated in Figure 5(a) and (b). In the research field of AI, side-tuning [34] is a specialized fine-tuning technique. It adds a trainable side network that runs in parallel to the backbone model, with the side network’s representation summed with the backbone’s output in the final layer. Crucially, side-tuning only updates the side network, without backpropagating through the backbone model.

Parallel Adapters Architecture. In light of side-tuning, we employ a time and memory efficient personal LLMs fine-tuning technique with Parallel Adapters. The overall structure is illustrated in Figure 5(c). Specifically, we decouple conventional Adapters [9] from the LLM backbone, avoiding their integration at the end of each transformer layer. Instead, we provide a dedicated parallel highway for our trainable adapters network, which takes intermediate activations from the backbone transformer as input and generates the final predictions. In this way, backpropagation through

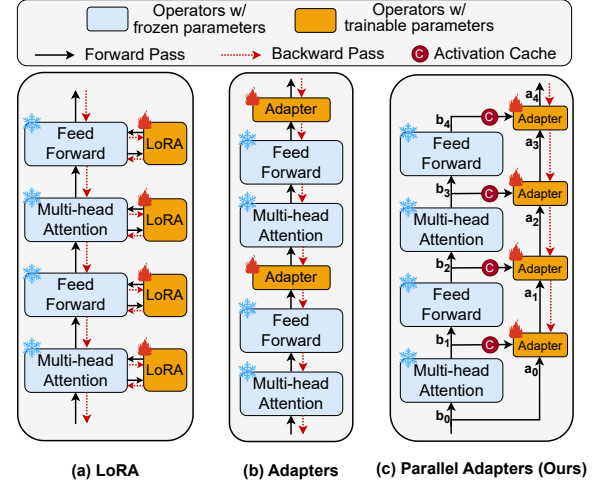


Figure 5: Comparison between LLMs fine-tuning with LoRA, Adapters, and our Parallel Adapters.

the LLM backbone is free, reducing memory demands for massive activations and computational burdens, thereby enhancing time and memory efficiency over techniques like Adapters and LoRA. Our adapters module demonstrates comprehensive compatibility with established LLM fine-tuning adapters architectures, including the use of linear layers for upward and downward projections as well as trimmed lightweight versions of the backbone transformer [7, 9, 21]. To ensure the lightweight and resource-efficient nature of our parallel network, the hidden dimension of our Parallel Adapters will be r , where $r \ll d$. Specifically, considering an LLM backbone, composed of L layers, and thus L intermediate outputs b_1, b_2, \dots, b_L , each contains n tokens with a hidden dimensionality of d , $b_i \in \mathbb{R}^{n \times d}$. We denote the embedding input sequence as $b_0 \in \mathbb{R}^{n \times d}$. Assuming adapters are inserted after every layer of backbone, Parallel Adapters consist of L adapters, yielding L intermediate outputs a_1, a_2, \dots, a_L , $a_i \in \mathbb{R}^{n \times r}$. We denote $a_0 = \mathbf{W}_{down} b_0$, where $\mathbf{W}_{down} \in \mathbb{R}^d \rightarrow \mathbb{R}^r$. We learn function f_i for i -th adapter of our Parallel Adapters, which operate on these intermediate outputs.

$$a_i = f_i(b_i, a_{i-1}). \quad (1)$$

Our evaluation in §6 reveals that parallel adapters can achieve comparable model performance to mainstream fine-tuning techniques while being more resource-efficient and better suited for resource-constrained edge environments.

4.2 PAC Activation Cache for Parallel Adapters

Observation and Opportunities. Leveraging Parallel Adapters substantially diminishes the computational and memory demands by circumventing backward propagation through the LLM backbone. However, for edge environments with limited resources, forward propagation calculations on the backbone of LLMs also require substantial computational resources. Figure 3 demonstrates that the computational overhead for forward propagation constitutes 54% and 56% of the total overhead when fine-tuning the T5-Large with Adapters and LoRA, respectively.

To minimize the computational demand, we identify two distinct opportunities for utilizing Parallel Adapters in in-situ fine-tuning of LLMs: (1) During the pre-training phase of LLMs, due to

the vast volumes of data involved, researchers typically train for only one epoch, meaning each sequence input is processed by the model a single time. However, in typical in-situ LLM fine-tuning scenarios, users often utilize small datasets collected from their specific context, repeatedly training the models with these inputs until achieving model convergence. (2) When employing parallel adapters to fine-tune LLMs, the parameters of the LLM backbone remain fixed. Unlike other PEFT techniques, the LLM backbone operates independently of the intermediate outputs generated by Parallel Adapters. Consequently, for a given input sequence, the activations generated by the LLM backbone are always invariant.

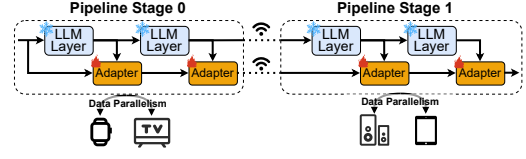
Fine-Tuning Parallel Adapters with PAC Activation Cache. Our key idea leverages the frozen parameters of the backbone model, enabling the caching of activations produced during the forward propagation of the same input sequence, thereby facilitating their reuse across multiple epochs [4]. As discussed in §4.1, the parallel adapters are a lightweight, separate network that takes the intermediate activations from the backbone transformer as input and generates predictions. During the first epoch, when processing a new input sequence, we cache all the input activations required by the Parallel Adapters that are obtained from the LLM backbone, as illustrated in Figure 5(c), highlighted by the red circle. In subsequent fine-tuning epochs using the same input sequence, we can skip the forward propagation through the LLM backbone entirely, since the required activations have already been cached. The combination of Parallel Adapters and activation caching allows efficient fine-tuning of the LLMs without the need for both forward and backward propagation through the backbone network, thereby (1) significantly accelerating the fine-tuning process and (2) reducing the memory footprint by allowing the release of the memory space occupied by the LLM parameters.

5 COLLABORATIVE EDGE AI SYSTEM FOR EFFICIENT PERSONAL LLMs FINE-TUNING

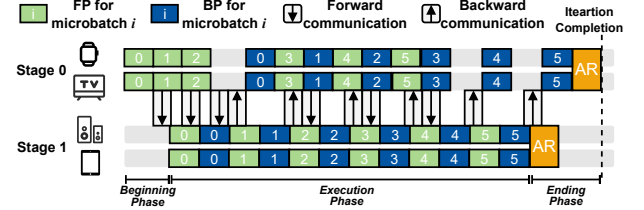
In PAC, we leverage edge devices in physical proximity and associate them as a resource pool to boost in-situ fine-tuning. Specifically, the fine-tuning procedure comprises two phases: (1) In the initial epoch, the backbone of LLMs, enhanced with Parallel Adapters, undergoes fine-tuning across multiple edge devices through a blend of data and pipeline parallelism (§5.1); (2) In subsequent epochs, the activation cache eliminates the necessity for forward propagation within the backbone, thereby enabling the exclusive fine-tuning of our Parallel Adapters utilizing data parallelism (§5.2).

5.1 Resource-Efficient Collaborative Orchestration for LLMs Fine-Tuning

Observation of Data and Pipeline Parallelism at the Edge. When collaborating on LLM fine-tuning among edge devices, the principle question is which type of parallelism should be used. The most common way to train models in parallel is *data parallelism* (DP) [8]. However, DP necessitates that each device maintains a replica of the entire model, a requirement difficult to meet for LLMs with extensive parameter sizes, often surpassing the capacity of a single device. *Pipeline parallelism* (PP) [30] is further proposed to address this problem. In PP, the model is partitioned into multiple consecutive stages and each stage is mapped to a separate device.



(a) The LLM transformer layers is partitioned into two stages, where both Stage 0 and 1 are replicated on a device group with two devices for intra-stage data parallelism.



(b) Fine-tuning pipeline of 6 micro-batches. The numbers in the cells represent micro-batch ids. AllReduce (AR) is performed in both Stage 0 and 1 for model synchronization.

Figure 6: An instance of hybrid parallelism in PAC.

Consequently, PP enables the training of increasingly large models by deploying more devices. Nonetheless, PP encounters scalability constraints as the addition of edge devices results in more stages. This not only results in a significant presence of pipeline bubbles but also amplifies the impact of inter-stage communication latency, thereby hindering efficiency. The above observation motivates us to employ a *hybrid parallelism* (HP) architecture that incorporates the best of both DP and PP, so as to achieve superior performance and scalability in resource-constrained edge environments.

Hybrid Parallelism Architecture in PAC. As illustrated in Figure 6(a), PAC first divides an LLM into multiple *stages* where each contains a *stage model* composed of a set of consecutive transformer layer. Edge devices are allocated into several device groups, each comprising one or more devices. PAC maps each stage to a group, with the stage model replicated across all devices within that group. Throughout the fine-tuning process, a mini-batch is divided into several *micro-batches* for concurrent processing to enhance parallelism. If a device cluster hosts multiple devices, micro-batches are further subdivided. Each device is responsible for executing the forward (FP) and backward passes (BP) for its assigned stage model and aggregates gradients across all micro-batches for every mini-batch. Upon completing a mini-batch, gradient synchronization within each device group is achieved through *AllReduce*. Since the majority of parameters in LLMs are frozen, AllReduce synchronizes only the lightweight parallel adapters, ensuring a swift process. We adopt the *one-forward-one-backward* (1F1B) micro-batch scheduling [18] which schedules the BP early to release the activation memory produced by FP for reuse. Figure 6(b) depicts a well-structured hybrid parallelism, encompassing FP, BP, and inter-stage communication.

Profiling. To enable parallelism planning, PAC profiler first fine-tunes the target LLM using calibration datasets to record the runtime profile required for planning. We define $t_f^{d,l}(\beta)$ and $t_b^{d,l}(\beta)$ as the FP and BP execution times for layer l on device d with batch size of β , respectively. u_d denotes the memory budget of device d . The size of output activations, input gradients, and weight parameters in bytes will also be collected to calculate memory footprint.

Planning Algorithm for Hybrid Parallelism. The global throughput of a pipeline is determined by the execution time of the

slowest stage. Consequently, our algorithm endeavors to partition the model into balanced stages. We consider an LLM consisting of L layers and denote \mathcal{D} as an ordered set of all devices involved in planning, while $\mathcal{D}_n = \{d_0, \dots, d_{n-1}\}$ as the subset of first n devices in \mathcal{D} . $W(x \rightarrow y, \mathcal{D}_n, s)$ denote the time taken by the slowest stage in the optimally balanced sub-pipeline between layer x to y with \mathcal{D}_n , when divided into s stages. To solve this partitioning problem, we break the pipeline into sub-pipelines and leverage the idea of dynamic programming. The formula of the dynamic programming algorithm can be written as:

$$W(0 \rightarrow y, \mathcal{D}_n, s) = \min_{0 \leq q < y} \min_{1 \leq m < n} \max\{W(0 \rightarrow q, \mathcal{D}_{n-m}, s-1), T(q+1 \rightarrow y, \{d_{n-m}, \dots, d_{n-1}\})\}, \quad (2)$$

where the first term inside the max is the time of the optimally balanced sub-pipeline between layers 0 to q with $n-m$ devices. The second term represents the time required by the single stage comprising layers $q+1$ to y across m devices. The notation $T(x \rightarrow y, \mathcal{D}_n)$ denotes the time required for a single stage to execute FP and BP in a data-parallel manner across the device group \mathcal{D}_n :

$$T(x \rightarrow y, \mathcal{D}_n) = \begin{cases} +\infty, & \text{if } \exists d \in \mathcal{D}_n, m_d > u_d, \\ \max_{d \in \mathcal{D}_n} \sum_{l=x}^y \left[t_f^{d,l} \left(\frac{B}{n} \right) + t_b^{d,l} \left(\frac{B}{n} \right) \right], & \text{else,} \end{cases} \quad (3)$$

where M is the number of micro-batch and B is the micro-batch size. The peak memory footprint of device d , denoted as m_d is the sum of the memory usage of the LLM parameters, parameter gradients, and activations. Without out-of-memory (OOM) exceptions, total data-parallel execution time is determined by the slowest device. If OOM occurs, the time will be set to positive infinity. During the dynamic programming, we will record pipeline planning configurations, including LLM segmentation points and device groupings.

Upon the completion of dynamic programming process, we obtain a set of balanced partition configurations for various number of pipeline stages: $\{W_s | \text{config. of } W(0 \rightarrow L, \mathcal{D}, s), s \in \{1, 2, \dots, |\mathcal{D}|\}\}$. The next step is to determine the optimal number of stages. Using recorded configurations, we can profile FP and BP execution time of stage i in W_s as $e_f^s(i)$ and $e_b^s(i)$. Similarly, forward and backward communication time between stages i and $i+1$ are represented as $c_f^s(i)$ and $c_b^s(i)$. $AR^s(i)$ represents the AllReduce time of stage i in W_s . As shown in Figure 6(b), we can divide per mini-batch training of W_s into three phases: *beginning phase*, *execution phase*, and *ending phase* with corresponding times denoted as L_b^s, L_e^s, L_n^s :

$$L_b^s = \sum_{i=1}^{s-1} [e_f^s(i) + c_f^s(i)], \quad L_e^s = M \cdot (e_f^s(s) + e_b^s(s)), \quad (4)$$

$$L_n^s = \max_{i \in \{1, \dots, s\}} (AR^s(i) + \sum_{j=i}^{s-1} (e_b^s(j) + c_b^s(j))), \quad (5)$$

$$\min_s (L_b^s + L_e^s + L_n^s). \quad (6)$$

Our algorithm aims to minimize this total latency by optimally determining the number of stages s . We remark that our parallelism planning is an offline procedure that runs once before deployment. The time complexity for our dynamic programming algorithm exhibits an upper bound of $O(L^2|\mathcal{D}|^2)$. In our experiment, the whole planning time is within three seconds on an edge device.

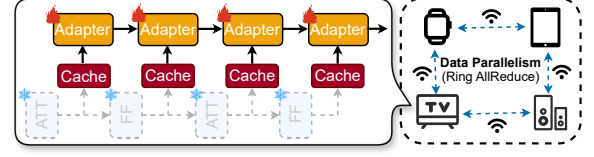


Figure 7: An instance of fine-tuning with activation cache.

5.2 Cache-Enabled Collaborative Edge Fine-Tuning of Parallel Adapters

Data-Parallel Fine-Tuning for Parallel Adapters The computationally lightweight nature of the Parallel Adapters precludes the use of pipeline parallelism to fine-tuning with activation cache, as it would result in unoverlapping inter-stage communication latency. Therefore, we employ data parallelism to exclusively fine-tune our Parallel Adapters. Specifically, after the first training epoch, the activation cache for all samples is already collected. We then perform collective communication to redistribute the Parallel Adapters parameters and locally cached activations across all devices, ensuring each device receives the complete set of adapter parameters and corresponding activations. The devices then utilize this shared information to fine-tune the parallel adapters in a data-parallel manner. In our experiments, fine-tuning the BART-Large model on the MRPC dataset for three epochs, the redistribution of parameters and activations only contributed to approximately 8% of the total training time. Notably, the overhead of this process can be further amortized over additional training epochs. An instance of personal LLMs fine-tuning with activation cache is depicted in Figure 7.

Storage Cost Analysis. Employing activation caching can reduce the computational requirements of forward propagation; however, it incurs additional storage overhead for activations. Specifically, the storage overhead is $s \times h \times l$ per sequence, where s denotes the sequence length, h represents the transformer's internal feature dimension, and l corresponds to the number of transformer layers. For T5-Base model, the activation caching requires less than 1 GB to store the activations for 500 training samples with sequence length of 30. Such cost is no more than 1% of the storage of a modern mobile device, e.g., hundreds of GB. During fine-tuning, the activation cache is reloaded from disk per micro-batch, a process that takes no more than tens of milliseconds on embedded flash storage. The cache will be cleared once the fine-tuning process finishes.

6 EVALUATION

6.1 Implementation and Setups

Implementation of PAC. We have fully implemented the prototype framework of PAC and baselines with $\sim 2,000$ LoC in Python atop Pytorch [2]. PAC's idea is also portable and can work well with other lightweight ML frameworks such as MNN [11] and TF-Lite [3]. Our Parallel Adapters is a lightweight version of the backbone model. The size of Parallel Adapters is determined by the reduction factor k . All weights and hidden state dimensions of the Parallel Adapters are $\frac{1}{k}$ times the corresponding weights and hidden states of the backbone model. In our experiments, the reduction factor k is set to 8. The weights of the Parallel Adapters are initialized based on structural pruning, using the weights of the backbone model. We insert Parallel Adapters at the end of each transformer layer.

Table 2: Training durations (in hours) for different methods: 3 epochs for MRPC and STS-B, and 1 epoch for SST-2 and QNLI.

Fine-tuning Techniques	Baseline Methods	T5-Base				BART-Large				T5-Large			
		MRPC	STS-B	SST-2	QNLI	MRPC	STS-B	SST-2	QNLI	MRPC	STS-B	SST-2	QNLI
Full Model	Standalone	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	Eco-FL	0.45	0.71	2.74	4.32	2.41	3.78	14.56	22.98	OOM	OOM	OOM	OOM
	EDDL	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
Adapters	Standalone	1.21	1.9	7.29	11.51	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	Eco-FL	0.39	0.61	2.35	3.71	0.54	0.85	3.27	5.16	2.75	4.31	16.59	26.19
	EDDL	0.34	0.53	2.06	3.25	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
LoRA	Standalone	1.21	1.89	7.28	11.49	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	Eco-FL	0.41	0.64	2.45	3.87	0.55	0.87	3.33	5.26	2.73	4.28	16.48	26.02
	EDDL	0.31	0.48	1.86	2.94	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
Parallel Adapters	PAC (Ours)	0.14	0.22	1.34	2.12	0.29	0.45	2.69	4.25	0.69	1.09	8.88	14.02

Table 3: Comparison of final performance between different fine-tuning techniques across four datasets. We report the average of F1 score and accuracy for MRPC. We use Pearson-Spearman Correlation as the metric for STS-B. For SST-2 and QNLI, we report accuracy. The mean value is the average performance of Full Model, Adapters and LoRA.

Fine-tuning Techniques	T5-Base				BART-Large				T5-Large			
	MRPC	STS-B	SST-2	QNLI	MRPC	STS-B	SST-2	QNLI	MRPC	STS-B	SST-2	QNLI
Full Model	89.71	90.94	94.03	93.08	88.16	91.10	95.64	94.40	92.78	91.08	95.30	93.30
Adapters	88.73	90.51	93.58	93.04	86.63	90.24	94.93	93.27	91.86	90.58	96.10	94.07
LoRA	86.27	90.73	93.69	93.30	87.46	90.36	95.23	94.48	90.27	92.08	95.53	94.18
Mean Value	88.24	90.73	93.77	93.14	87.42	90.57	95.27	94.05	91.64	91.25	95.64	93.85
Parallel Adapters (Ours)	88.24	90.43	93.46	93.25	87.71	90.54	95.25	93.68	91.7	91.57	95.76	93.7
Difference from Mean	+0.00	-0.30	-0.31	+0.11	+0.29	-0.03	-0.02	-0.37	+0.06	+0.32	+0.12	-0.15

Models and Datasets. We evaluate PAC with three typical transformer based LLM with parameters ranging from 0.25B to 0.74B, as detailed in Table 4, which are widely considered for IPA and edge deployments [14, 32]. All experiments were performed under conditions using Float32 precision to ensure fine-tuning performance. We employ two variants of the T5 model [20], specifically T5-Base and T5-Large with differing parameter sizes. We also compare PAC with baseline methods with BART-Large [13] as the backbone for our parallel adapters. We evaluate our fine-tuned LLMs with four tasks from GLUE benchmark. The four tasks evaluate models on multiple diverse tasks over sentiment analysis (SST2), similarity and paraphrase (MRPC, STS-B) and natural language inference (QNLI).

Edge Environment Setup. We evaluate PAC across a realistic edge platform consisting of multiple NVIDIA Jetson Nano [1], widely recognized as prevalent off-the-shelf edge devices. Each device is equipped with a 128-core NVIDIA Maxwell GPU and 4GB unified memory. We simulate common network conditions in edge environments (e.g., smart homes) by setting the intra-cluster network bandwidth to 1000Mbps.

Baseline Methods. We compare PAC with both single-device method and the state-of-the-art collaborative edge training methods: (1) **Standalone** means fine-tuning LLMs on a single edge device. We compare with it to analyze the scalability performance of PAC. (2) **Eco-FL** [30] is a collaborative edge system that facilitates pipeline parallelism training across an edge device cluster within the same local area network, segmenting LLMs into sequential stages for processing in a pipeline fashion. (3) **EDDL** [8] employs conventional data parallel training across edge devices, distributing batch data among cluster devices for simultaneous processing.

Table 4: LLM model specifications used for experiments. "en-de" indicates encoder-decoder LLM structure.

Model	Structure	Layers	Heads	Hidden Size	Param. Count
T5-Base [20]	en-de	12	12	768	0.25B
BART-Large [13]	en-de	12	16	1024	0.41B
T5-Large [20]	en-de	24	16	1024	0.74B

Considering that the aforementioned baseline systems were not specifically designed for the fine-tuning of LLMs, we ensure a fair comparison by equipping these edge systems with various LLM fine-tuning techniques. These include full model fine-tuning and popular PEFT techniques. (1) In **Full model fine-tuning**, all the LLM parameters are updated for a downstream task. (2) **LoRA** [10] is a widely-used PEFT technique that decomposes the parameter update for a weight matrix into two trainable low-rank matrices. (3) **Adapters** [9] is another widely-used PEFT technique that injects small trainable modules at the end of each transformer layer.

6.2 End-to-end Performance

Table 2 and Table 3 summarize the end-to-end performance comparisons between PAC, the single-device method, and state-of-the-art collaborative edge training methods. To ensure fair comparisons, these baseline methods are enhanced with prevalent PEFT techniques, including Adapters and LoRA. Fine-tuning the smaller datasets, MRPC and STS-B, is conducted over three epochs, with the latter two epochs benefiting from the PAC activation cache. Conversely, for larger datasets such as STS-2 and QNLI, a single epoch of fine-tuning is sufficient to achieve satisfactory performance.

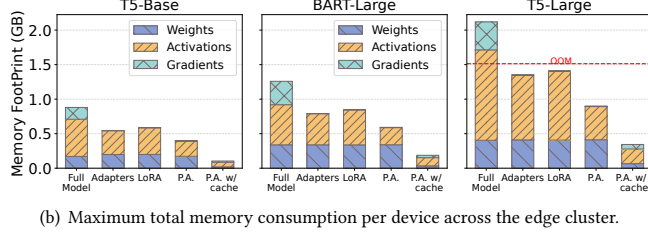
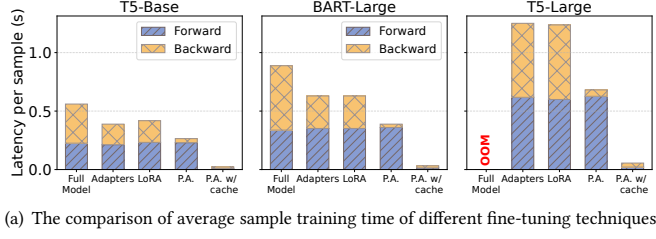


Figure 8: Comparison of different fine-tuning techniques. P.A. indicates our Parallel Adapters technique. Mini-batch size 16; sequence length: 128.

PAC significantly speeds up the training process while preserving convergence performance. PAC achieves an acceleration ranging from $1.21\times$ to $5.44\times$ on SST-2 and QNLI without utilizing activation cache. In comparison to Standalone and EDDL, these two baselines often encounter Out-of-Memory (OOM) issues, even when integrating PEFT techniques such as LoRA and Adapters. This issue stems from the training requirement for each edge device to host the entire target model. Particularly for T5-Large, a single Jetson Nano is inadequate to accommodate LLM parameters, not to mention the intermediate activations. Compared to Eco-FL, PAC achieves an acceleration ranging from $1.21\times$ to $5.41\times$ on SST-2 and QNLI, without utilizing activation cache. Parallel Adapters not only alleviate the memory footprint of LLM parameters but also intermediate activations. Eco-FL's pipeline parallel strategies allow each edge device to host only a portion of the model parameters. However, these devices still bear a substantial memory footprint from intermediate activations, even when employing PEFT technologies such as LoRA and Adapters. Therefore, the Eco-FL approach necessitates the use of smaller micro-batch sizes or a reduction in the number of micro-batches simultaneously input into the pipeline. This results in decreased concurrency in pipeline parallelism and lowers the training throughput. Moreover, our hybrid parallelism merges the benefits of both data and pipeline parallelism, providing an expanded search space for parallel architectures to accommodate complex edge environments. Our method enables the identification of the most efficient parallel configuration with maximum throughput within the constraints of available resources. With the integration of our activation cache mechanism, PAC achieves speedups of up to $8.64\times$ on the MRPC and STS-B datasets. As discussed in §4.1, our Parallel Adapters constitute a lightweight, independent network. We can skip both the forward and backward passes through the LLM backbone, since the required activations have already been calculated and stored. Consequently, training overhead can be markedly reduced in the second and third fine-tuning epochs.

Table 3 displays the performance of various full model and PEFT fine-tuning methods on four datasets after training. Fine-tuning

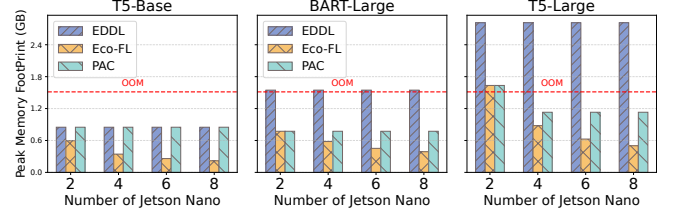
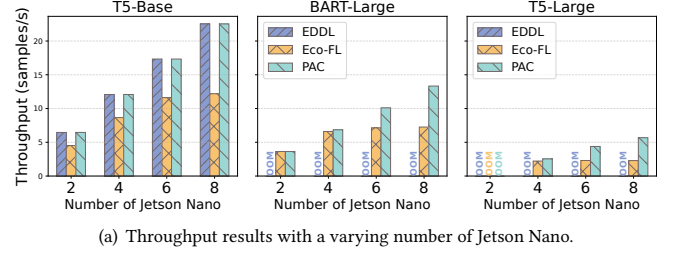


Figure 9: Comparison of different fine-tuning techniques.

Model	Number of Jeton Nano			
	2	4	6	8
T5-Base	[2]	[4]	[6]	[8]
BART-Large	[1-1]	[2-2]	[3-3]	[4-4]
T5-Large	[1-1]	[2-1-1]	[2-2-2]	[3-3-2]

(a) PAC's device grouping results.

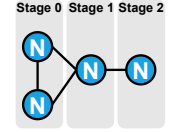


Figure 10: Device grouping results of PAC's hybrid parallelism for experiments in Figure 9. "N" indicates Jetson Nano.

involves 3 epochs for the smaller MRPC and STS-B datasets, and 1 epoch for the larger SST-2 and QNLI datasets. We can observe that PAC achieves comparable or superior performance to full model fine-tuning and PEFT techniques across various models and datasets. The largest discrepancy in mean performance metrics between PAC and these methods is only -0.37 , a negligible difference. Notably, PAC frequently outperforms these methods and achieves the highest performance on the SST-2 dataset with the T5-Large model.

6.3 Significance of Parallel Adapters at the Edge

We conducted experiments to assess the time and memory efficiency of Parallel Adapters at the edge. In this section, we perform data parallelism for Parallel Adapters with activation cache across 8 devices and hybrid parallelism for other fine-tuning techniques without 1F1B micro-batch scheduling. "Activations" contain the intermediate results and optimizer states. Figure 8 illustrates that Parallel Adapters outperform other fine-tuning techniques regarding both time and memory efficiency.

Parallel Adapters markedly reduce per-sample training time. Figure 8(a) presents the average sample training time across different fine-tuning techniques. Without activation cache, Parallel Adapters can reduce the average sample training time by 31.94% to 56.24% compared to baseline methods, primarily owing to a substantial decrease in backward propagation overhead. Both Adapters and LoRA incorporate trainable structures into the backbone model, thus necessitating backpropagation across the entire backbone model for gradient computation of these parameters. Consequently, regarding backward time, Adapters and LoRA can only achieve approximately a 49% reduction compared to full

fine-tuning. In contrast, backpropagation through the backbone model is unnecessary with Parallel Adapters, leading to a more substantial reduction in backward time, nearly 92% compared to full fine-tuning. Moreover, Parallel Adapters With activation cache mechanism can further decrease the average sample training time up to 96.39%. These results demonstrate the substantial reduction in training time achieved by Parallel Adapters.

Parallel Adapters yield a substantial reduction in memory usage. Figure 8(b) depicts the breakdown of the memory footprint for different fine-tuning techniques. We report the peak memory consumption per device across edge clusters. Without activation cache, Parallel Adapters can reduce memory usage by 25.27% to 60.49%. Adapters and LoRA demonstrate parameter efficiency but do not exhibit significant memory efficiency. While these techniques notably decrease the memory footprint of gradients by reducing the number of trainable parameters, the memory usage associated with activations remains considerable. However, intermediate activations often become the primary memory bottleneck during training, especially with larger batch sizes. For PEFT techniques such as Adapters and LoRA, activation memory can be reduced by up to 28.15% compared to full fine-tuning across the three models. In contrast, Parallel Adapters achieve a more significant reduction, reaching up to 58.87%. With activation cache, Parallel Adapters can decrease the peak memory footprint from 74.57% to 88.16% compared to baselines. This is because it's sufficient to store only the lightweight Parallel Adapters, eliminating the need to host the entire LLM backbone in memory.

6.4 Analysis of Collaborative Edge Fine-Tuning

We perform an ablation study to understand the contribution of hybrid parallelism and activation cache in our system design.

Comparison PAC with EDDL and Eco-FL. To explore the scalability advantages of PAC's hybrid parallelism over Eco-FL's pipeline parallelism and EDDL's data parallelism, we compared the throughput of these methods when training collaboratively across 2 to 8 edge devices. The batch size was consistent with the number of devices, and the sequence length of each sample was fixed at 128. We implement Eco-FL and EDDL using the Parallel Adapters technique to ensure a fair comparison. Note that none of the three methods utilizes activation cache.

Figure 9(b) illustrates the maximum per device memory footprint of model weights across edge cluster. For EDDL, each device must host a complete LLM, preventing the reduction of the parameters' memory footprint through scaling up the number of devices. Therefore, as shown in Figure 9(a), the EDDL method exhibits OOM errors with both the BART-Large and T5-Large models. Conversely, PAC and Eco-FL utilize pipeline parallelism, partitioning the model into multiple stages with each handled by different devices. This approach allows for scaling the number of devices to reduce the peak memory footprint. PAC's hybrid parallelism offers a broader search space for parallel strategies compared to Eco-FL's pipeline parallelism. Our planning algorithm for PAC is capable of identifying more efficient hybrid parallel configurations within memory constraints, enhancing resource utilization. Although PAC may incur higher memory overhead in some instances, it achieves greater system throughput. Specifically, when compared to Eco-FL, PAC exhibits an increase in throughput from 39.50% to 84.79%.

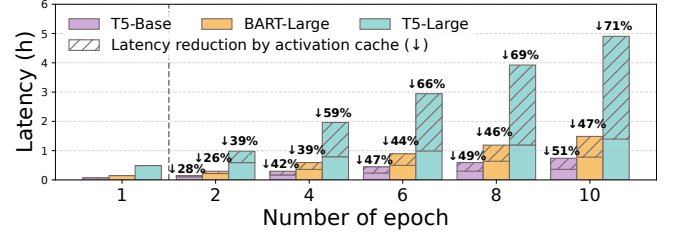


Figure 11: Fine-tuning time with PAC. Time without activation cache is represented by bars. The corresponding reduction in time achieved utilizing activation cache is represented by shaded areas. Dataset: MRPC.

To more clearly illustrate the parallel strategies adopted by PAC, we present the device grouping configurations for PAC across various LLMs and numbers of devices in Figure 10. On the left, a table displays all the grouping results across three models. On the right, an instance is shown where a model is divided into three stages, with two devices handling the first stage to perform data parallelism. Specifically, when fine-tuning BART-Large with eight devices, EDDL encounters OOM issues because a single Jetson Nano cannot accommodate a complete BART-Large. Eco-FL addresses this problem by dividing the model into eight stages and employing straight pipeline parallelism for training. On the contrary, our PAC approach divides BART-Large into two stage models, with each stage replicated across four devices. This configuration significantly reduces the number of stages in the pipeline, thereby minimizing inter-stage data dependencies and communication latency, which in turn enhances the pipeline's concurrent efficiency. These results demonstrate that our hybrid parallel approach offers a larger search space for parallel configurations, providing enhanced scalability and robustness across varying numbers of devices and workloads.

Comparison of PAC with and without activation cache. We further investigated how our activation cache mechanism benefits the required fine-tuning latency. By leveraging activation cache, the fine-tuning latency per epoch can decrease up to 79.51%. Figure 11 shows the fine-tuning latency reduction as the number of epochs increases with the use of the activation cache mechanism. We can observe that by leveraging the activation cache, fine-tuning latency is reduced by 26% to 71%. Moreover, this reduction in latency increases with the number of epochs. For example, with the T5-Large model, training for two epochs reduces latency by 39%, whereas training for ten epochs increases the reduction to 71%. This reduction can be attributed to the fact that the Parallel Adapters constitute a lightweight, independent network, resulting in a significant decrease in training cost compared to the LLM backbone. We can bypass both the forward and backward passes through the LLM backbone since the necessary activations are already cached.

7 RELATED WORK

Parameter-Efficient Fine-Tuning for LLM. Prompt tuning [12] proposes to prepend the model input embeddings with a trainable tensor. Adapters tuning [9] adds domain-specific layers after attention and FFN layers in transformer. LoRA [10] decomposes the parameter update for a weight matrix into two trainable low-rank matrices. To further reduce the memory overhead, pioneering studies explore fine-tuning techniques that obviate the need for

backpropagation through the backbone model. Y-tuning [16] learns additional task-specific label representations, which are integrated with the output of the backbone model to circumvent backpropagation. LST [21] involves the use of pruned lightweight transformer structures from the backbone as a side network. E³VA [31] extends the concept of the side network into the realm of computer vision.

On-device DNN Fine-Tuning. POET [19] achieves the finetuning of a BERT model on embedded devices, optimizing for both training speed and energy consumption. Lin et al. [15] enable training directly on devices with a minimal memory requirement of only 256KB. Sage and Melon [5, 23] implement hybrid memory management and conservation strategies, including operator fusion and the use of a dedicated memory pool, to mitigate memory limitations. Additionally, Mandheling [25] incorporates mixed-precision training along with DSP offloading to enhance the speed of learning.

Collaborative Edge Computing for DNN Fine-Tuning. Federated Learning (FL) has been a promising paradigm in distributed machine learning that enables in-situ model fine-tuning. FwdLLM [27] designs a backpropagation-free fine-tuning FL protocol to enhance efficiency. AdaFL [4] proposes an FL framework for finetuning LLMs that features adaptable depth and width in its adapters modules. Breaking through the conventional paradigm of FL, Ye et al. [30, 33] devise a pipeline parallel architecture that facilitates the collaborative fine-tuning of DNNs across multiple edge devices. EDDL [8] adopts data parallelism training across embedded devices in a local area network. Asteroid [29] also employs HPP across multiple edge devices for DNN training, but it does not specifically address the parameter-efficient fine-tuning of LLMs.

8 CONCLUSION

This paper proposes PAC, a time and memory efficient collaborative edge AI framework for personal LLMs fine-tuning. PAC breaks the resource wall of personal LLMs fine-tuning with a sophisticated algorithm-system co-design, achieving an acceleration of 8.64× and 88.16% memory reduction compared to state-of-the-art methods.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful feedback. This work was supported in part by the National Science Foundation of China (No. U20A20159); Guangdong Basic and Applied Basic Research Foundation (No. 2023B1515120058, No. 2021B151520008); Guangzhou Basic and Applied Basic Research Program (No. 2024A04J6367).

REFERENCES

- [1] 2019. Jetson-Nano. <https://developer.nvidia.com/embedded/jetson-nano-developer-kit>.
- [2] 2019. PyTorch. <https://github.com/pytorch/pytorch>.
- [3] 2021. On-device training with tensorflow lite. https://www.tensorflow.org/lite/examples/on_device_training/overview.
- [4] Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. 2023. Efficient federated learning for modern nlp. In *MobiCom*. 1–16.
- [5] In Gim and JeongGil Ko. 2022. Memory-efficient dnn training on mobile devices. In *MobiSys*. 464–476.
- [6] Liwei Guo, Wonkyo Choe, and Felix Xiaozhu Lin. 2023. Sti: Turbocharge nlp inference at the edge via elastic pipelining. In *ASPLOS, Volume 2*. 791–803.
- [7] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608* (2024).
- [8] Pengzhan Hao and Yifan Zhang. 2021. Eddl: A distributed deep learning system for resource-limited edge computing environment. In *2021 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 1–13.
- [9] Neil Houlsby, Andrei Giurugu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *ICML*. PMLR, 2790–2799.
- [10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
- [11] Xiaotang Jiang, Huan Wang, Yiliu Chen, Ziqi Wu, Lichuan Wang, Bin Zou, Yafeng Yang, Zongyang Cui, Yu Cai, Tianhang Yu, et al. 2020. Mnn: A universal and efficient inference engine. *Proceedings of MLSys 2* (2020), 1–13.
- [12] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).
- [13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).
- [14] Yuan Chun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459* (2024).
- [15] Ji Lin, Ligeng Zhu, Wei-Ming Chen, Wei-Chen Wang, Chuang Gan, and Song Han. 2022. On-device training under 256kb memory. *NeurIPS* 35 (2022).
- [16] Yitao Liu, Chenxin An, and Xipeng Qiu. 2024. Y-tuning: An efficient tuning paradigm for large-scale pre-trained models via label representation learning. *Frontiers of Computer Science* 18, 4 (2024), 184320.
- [17] Xupeng Miao, Gabriele Oliaro, Xin hao Cheng, Mengdi Wu, Colin Unger, and Zhihao Jia. 2024. FlexLLM: A System for Co-Serving Large Language Model Inference and Parameter-Efficient Finetuning. *arXiv:2402.18789* (2024).
- [18] Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R Devanur, Gregory R Ganger, Phillip B Gibbons, and Matei Zaharia. 2019. PipeDream: generalized pipeline parallelism for DNN training. In *SOSP*. 1–15.
- [19] Shishir G Patil, Paras Jain, Prabal Dutta, Ion Stoica, and Joseph Gonzalez. 2022. POET: Training neural networks on tiny devices with integrated rematerialization and paging. In *International Conference on Machine Learning*. PMLR, 17573–17583.
- [20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR* 21, 140 (2020).
- [21] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *NeurIPS* 35 (2022).
- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS* 30 (2017).
- [23] Qipeng Wang, Mengwei Xu, Chao Jin, Xinran Dong, Jinliang Yuan, Xin Jin, Gang Huang, Yunxin Liu, and Xuanzhe Liu. 2022. Melon: Breaking the memory wall for resource-efficient on-device machine learning. In *MobiSys*. 450–463.
- [24] Yuanxin Wei, Shengyuan Ye, Jiazhi Jiang, Xu Chen, Dan Huang, Jiangsu Du, and Yutong Lu. 2024. Communication-Efficient Model Parallelism for Distributed In-situ Transformer Inference. In *DATE*. IEEE, 1–6.
- [25] Daliang Xu, Mengwei Xu, Qipeng Wang, Shangguang Wang, Yun Ma, Kang Huang, Gang Huang, Xin Jin, and Xuanzhe Liu. 2022. Mandheling: Mixed-precision on-device dnn training with dsp offloading. In *MobiCom*. 214–227.
- [26] Daliang Xu, Wangsong Yin, Xin Jin, Ying Zhang, Shiyun Wei, Mengwei Xu, and Xuanzhe Liu. 2023. Llmcd: Fast and scalable on-device large language model inference. *arXiv preprint arXiv:2309.04255* (2023).
- [27] M Xu, D Cai, Y Wu, X Li, and S Wang. 2024. Fwdllm: Efficient fedllm using forward gradient. (2024).
- [28] Shengyuan Ye, Jiangsu Du, Liekang Zeng, Wenzhong Ou, Xiaowen Chu, Yutong Lu, and Xu Chen. 2024. Galaxy: A Resource-Efficient Collaborative Edge AI System for In-situ Transformer Inference. *arXiv preprint arXiv:2405.17245* (2024).
- [29] Shengyuan Ye, Liekang Zeng, Xiaowen Chu, Guoliang Xing, and Xu Chen. 2024. Asteroid: Resource-Efficient Hybrid Pipeline Parallelism for Collaborative DNN Training on Heterogeneous Edge Devices. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 312–326.
- [30] Shengyuan Ye, Liekang Zeng, Qiong Wu, Ke Luo, Qingze Fang, and Xu Chen. 2022. Eco-FL: Adaptive federated learning with efficient edge collaborative pipeline training. In *Proceedings of the 51st International Conference on Parallel Processing*. 1–11.
- [31] Dongshuo Yin, Xueting Han, Bin Li, Hao Feng, and Jing Bai. 2023. Parameter-efficient is not sufficient: Exploring parameter, memory, and time efficient adapter tuning for dense predictions. *arXiv preprint arXiv:2306.09729* (2023).
- [32] Jinliang Yuan, Chen Yang, Dongqi Cai, Shihe Wang, Xin Yuan, Zeling Zhang, Xiang Li, Dingge Zhang, Hanzi Mei, Xianqing Jia, et al. 2023. Rethinking mobile AI ecosystem in the LLM era. *arXiv preprint arXiv:2308.14363* (2023).
- [33] Liekang Zeng, Shengyuan Ye, Xu Chen, and Yang Yang. 2024. Implementation of Big AI Models for Wireless Networks with Collaborative Edge Computing. *IEEE Wireless Communications* 31, 3 (2024), 50–58.
- [34] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. 2020. Side-tuning: a baseline for network adaptation via additive side networks. In *ECCV 2020, Proceedings, Part III* 16. Springer, 698–714.