Probing the Efficacy of Federated Parameter-Efficient Fine-Tuning of Vision Transformers for Medical Image Classification

Naif Alkhunaizi[†], Faris Almalik[†], Rouqaiah Al-Refai, Muzammal Naseer, and Karthik Nandakumar

Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE {naif.alkhunaizi, faris.almalik, rouqaiah.al-refai, muzammal.naseer, karthik.nandakumar}@mbzuai.ac.ae

Abstract. With the advent of large pre-trained transformer models, fine-tuning these models for various downstream tasks is a critical problem. Paucity of training data, the existence of data silos, and stringent privacy constraints exacerbate this fine-tuning problem in the medical imaging domain, creating a strong need for algorithms that enable collaborative fine-tuning of pre-trained models. Moreover, the large size of these models necessitates the use of parameter-efficient fine-tuning (PEFT) to reduce the communication burden in federated learning. In this work, we systematically investigate various federated PEFT strategies for adapting a Vision Transformer (ViT) model (pre-trained on a large natural image dataset) for medical image classification. Apart from evaluating known PEFT techniques, we introduce new federated variants of PEFT algorithms such as visual prompt tuning (VPT), low-rank decomposition of visual prompts, stochastic block attention fine-tuning, and hybrid PEFT methods like low-rank adaptation (LoRA)+VPT. Moreover, we perform a thorough empirical analysis to identify the optimal PEFT method for the federated setting and understand the impact of data distribution on federated PEFT, especially for out-of-domain (OOD) and non-IID data. The key insight of this study is that while most federated PEFT methods work well for in-domain transfer, there is a substantial accuracy vs. efficiency trade-off when dealing with OOD and non-IID scenarios, which is commonly the case in medical imaging. Specifically, every order of magnitude reduction in fine-tuned/exchanged parameters can lead to a 4% drop in accuracy. Thus, the initial model choice is crucial for federated PEFT. It is preferable to use medical foundation models learned from in-domain medical image data (if available) rather than general vision models. Code will be provided upon acceptance.

Keywords: Vision Transformers \cdot Parameter-Efficient Fine-tuning \cdot Out-of-Domain Transfer \cdot Federated Learning

1 Introduction

Transformer models pre-trained on large-scale data can serve as a foundation for a wide range of downstream tasks [2]. While many general vision foundation

models are available [8,23], developing generic medical foundation models is a challenge due to the diversity of imaging modalities and limited availability of well-annotated data [33]. Consider the scenario where a healthcare organization wants to learn transformer models for a range of medical image classification tasks such as chest x-ray disease classification [9,22], melanoma classification [5,24], and tumor categorization [6,18]. There are two main challenges in this problem setting. Firstly, the organization may not have sufficient training data for each task to learn task-specific models from scratch. This can be addressed by fine-tuning a model that is pre-trained on a large-scale, independent dataset (transfer learning) for the task(s) at hand [32]. Secondly, storing a separate model for each task is inefficient due to the large size of transformer models. Parameterefficient fine-tuning (PEFT) methods such as subset fine-tuning [27], adapter [4], low-rank adaptation (LoRA) [15], and prompt tuning [16] can mitigate this problem by fine-tuning only a small number of parameters for each task and storing the base model along with minimal task-specific parameters. Most PEFT methods exploit the inherently modular transformer architecture (characterized by a sequence of identical self-attention blocks processing a set of tokens).

In some medical imaging applications, even fine-tuning of pre-trained models may not be feasible when a hospital has data only from a few patients. However, a consortium of hospitals may be willing to collaborate to realize PEFT. This introduces the additional challenge of privacy because healthcare data is often regulated by strict privacy guidelines (e.g., GDPR, HIPAA), and it is not possible to pool data from multiple healthcare institutions centrally to enable machine learning. Federated learning (FL) [20] can enable multiple entities to train a model collaboratively without sharing raw data. However, regular exchange of parameters in FL can become a communication burden, especially if the models are large. Hence, the combination of FL and PEFT is an ideal solution that can effectively solve multiple issues (paucity of data, storage of multiple large models, communication cost, and data privacy) simultaneously [34].

In this work, we consider a Vision Transformer (ViT) [8] model pre-trained on natural images as an illustrative example and explore federated PEFT in a cross-silo setting (with a small number of clients), aiming to answer the following questions: (i) Which PEFT method works well in conjunction with FL and provides the best accuracy vs. efficiency trade-off? (ii) Can federated PEFT transfer well for out-of-domain (OOD) and non-IID (independent, identically distributed) data encountered in medical image analysis? To the best of our knowledge, this is the first study that attempts to systematically study various PEFT strategies for ViTs within the FL framework. Our main contributions are:

- 1. New federated variants of PEFT methods: We are the first to investigate visual prompt tuning (VPT) and low-rank decomposition of visual prompts (DVPT) in a federated setting. We also introduce a new federated subset fine-tuning approach called stochastic block attention (SBA). Finally, we also consider hybrid methods such as combining LoRA with VPT.
- 2. Analysis of federated PEFT methods: We demonstrate that there is indeed a substantial trade-off between parameter efficiency and model ac-

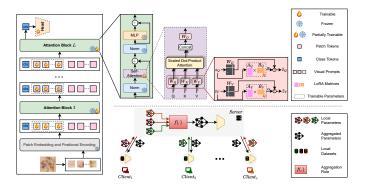


Fig. 1: Adaptation of Vision Transformer (ViT) model using federated PEFT methods. Only the parameters marked as *trainable* are exchanged between the clients and the server, while the frozen parameters are not communicated.

curacy in federated PEFT, especially for *out-of-domain tasks with non-IID* client distributions. Hence, one must proceed with caution when adapting a general vision model for medical image classification using federated PEFT.

2 Background and Related Work

Vision Transformer (ViT): A pre-trained ViT [8] can be considered as a feature extractor \mathcal{V}_{ψ} that maps a given input image x into a d-dimensional feature vector $\mathbf{f} \in \mathbb{R}^d$, where Ψ denotes the complete set of ViT parameters. For image classification, a classification head \mathcal{H}_n is typically trained to learn the mapping between **f** and the class label $y \in \{1, 2, \dots, K\}$, where K is the number of classes and η represents the parameters of the head \mathcal{H} . A ViT divides the **x** into S nonoverlapping patches and a linear patch embedding layer \mathcal{E}_{Λ} (with parameters A) is used to project each patch into \mathbb{R}^d , resulting in $\mathcal{T}_0 = \{\mathbf{t}_1, \cdots, \mathbf{t}_S\}$ patch tokens. Additionally, a learnable class token $(\tilde{\mathbf{t}}_0 \in \mathbb{R}^d)$ is prepended to the sequence of patch tokens to assimilate the information as the tokens pass through L transformer blocks (denoted by $\mathcal{G}_{\psi_{\ell}}$). The operations of each transformer block in a ViT can be represented as $\{\tilde{\mathbf{t}}_{\ell}, \mathcal{T}_{\ell}\} = \mathcal{G}_{\psi_{\ell}}(\{\tilde{\mathbf{t}}_{\ell-1}, \mathcal{T}_{\ell-1}\}), \ell \in [1, L]$. The class token output by the L^{th} (last) block (i.e., $\tilde{\mathbf{t}}_L$) can be considered as the final feature representation f. Each transformer block, in turn, consists of three types of parameters (Fig. 1) - ϕ_{ℓ} denotes the layer normalization parameters of the ℓ^{th} block, θ_{ℓ} denotes the weight matrices of the multi-head self-attention (MHSA) layer of the $\ell^{\rm th}$ block, and ω_{ℓ} represents the parameters of the multi-layer perceptron (MLP) of the ℓ^{th} block. For convenience, let $\Phi = \{\phi_\ell\}_{\ell=1}^L$, $\Theta = \{\theta_\ell\}_{\ell=1}^L$, and $\Omega = \{\omega_\ell\}_{\ell=1}^L$ denote the collection of normalization, MHSA, and MLP parameters of all the L blocks, respectively. Similarly, $\psi_l = \{\phi_\ell, \theta_\ell, \omega_\ell\}$ denote the set of all parameters of the ℓ^{th} block. Thus, ViT parameters can be summarized as $\Psi = \{\Lambda, \Phi, \Theta, \Omega\} = \{\Lambda, \psi_1, \cdots, \psi_L\}$ and ViT operations can be summarized as $\tilde{\mathbf{t}}_L = \mathcal{V}_{\Psi}(\mathbf{x}) = \mathcal{G}_{\psi_L}(\cdots(\mathcal{G}_{\psi_1}(\{\tilde{\mathbf{t}}_0, \mathcal{E}_A(\mathbf{x})\})))$. Since training a ViT from scratch

requires a large dataset due to the lack of inductive bias [19], fine-tuning has been the de-facto approach to adapt pre-trained ViTs for downstream tasks [3]. Parameter-Efficient Fine-Tuning (PEFT): PEFT methods achieve efficient adaptation of large pre-trained models [14,15,12] by learning only a limited number of parameters. Linear probing learns only the head parameters η and it represents the lower bound for all PEFT methods. In contrast, full fine-tuning involves updating all the ViT parameters (Ψ) in addition to the head (η) . Subset fine-tuning methods fine-tune only a chosen subset of the pre-trained model parameters such as the last few layers of the network (e.g., [13]) or the MHSA layer within each ViT block (e.g., [27]). Visual Prompt Tuning (VPT) [16] introduces a set of R learnable visual prompts before each ViT block, represented by $\mathcal{P}_v = \{\mathcal{P}_{v_\ell}\}_{\ell=1}^L \in \mathbb{R}^{L \times R \times d}$. The operations of each transformer block in a visually prompted ViT can be represented as $\{\tilde{\mathbf{t}}_\ell, _, \mathcal{T}_\ell\} = \mathcal{G}_{\psi_\ell}(\{\tilde{\mathbf{t}}_{\ell-1}, \mathcal{P}_{v_\ell}, \mathcal{T}_{\ell-1}\})$, $\ell \in [1, L]$. During fine-tuning, only the prompts \mathcal{P}_v are updated and the ViT parameters are unchanged. Low-Rank Adaptation (LoRA) [15] injects trainable lowrank matrices in parallel to the attention layer [14] while keeping the pre-trained model weights frozen. The MHSA parameters of a block ℓ can be considered as a collection of four weight matrices denoted as $\theta_{\ell} = \{\mathbf{W}_{O,\ell}, \mathbf{W}_{Q,\ell}, \mathbf{W}_{K,\ell}, \mathbf{W}_{V,\ell}\}.$ In LoRA, the updates to $\mathbf{W}_{Q,\ell}$ and $\mathbf{W}_{V,\ell}$ are decomposed into a pair of low rank matrices $\mathbf{A} \in \mathbb{R}^{r \times d}$ and $\mathbf{B} \in \mathbb{R}^{d \times r}$, where r represents the rank of the two matrices. Let $z_{*,\ell}$ and $\tilde{z}_{*,\ell}$ be the input and output, respectively, of an attention layer in the ℓ^{th} block. Then, LoRA operations can be summarized as:

$$\tilde{z}_{Q,\ell} = \mathbf{W}_{Q,\ell} z_{Q,\ell} + \alpha \mathbf{B}_{Q,\ell} \mathbf{A}_{Q,\ell} z_{Q,\ell},
\tilde{z}_{V,\ell} = \mathbf{W}_{V,\ell} z_{V,\ell} + \alpha \mathbf{B}_{V,\ell} \mathbf{A}_{V,\ell} z_{V,\ell}.$$
(1)

Here, $\mathcal{A} = \{\mathbf{A}_{Q,\ell}, \mathbf{A}_{V,\ell}\}_{\ell=1}^L$ and $\mathcal{B} = \{\mathbf{B}_{Q,\ell}, \mathbf{B}_{V,\ell}\}_{\ell=1}^L$ are the only learnable parameters, and α is a fixed scalar. Recently, PEFT methods have also been studied in the FL context. While textual prompt learning via FL was proposed in [11,31], a FL extension to LoRA was proposed in [1].

3 Federated Parameter-Efficient Fine-Tuning Methods

Problem Statement: We assume that a ViT feature extractor \mathcal{V}_{Ψ_0} that is already pre-trained on a large independent dataset is available at the server. The goal of the server is to collaborate with the C clients to fine-tune the pre-trained ViT feature extractor \mathcal{V}_{Ψ_0} and learn the task-specific classification head \mathcal{H}_{η} in a federated fashion while maximizing task-specific performance and minimizing the number of parameters that are tuned and exchanged. The server initializes η as η_0 and broadcasts both Ψ_0 and η_0 to all the clients before the collaboration begins. At the end of T collaboration rounds, the objective is to obtain Ψ_T and η_T , which are fine-tuned for the specified task. By minimizing the number of parameters that are tuned and exchanged, we seek to reduce both the communication costs between the clients and the server as well as the memory footprint required to store the task-specific parameters.

Vanilla Federated Learning (FedAvg): [20] Given an appropriate per-client loss function $\mathcal{L}^{(c)}(\Psi, \eta)$, the global loss function is defined as:

$$\mathcal{L}(\Psi, \eta) = \sum_{c=1}^{C} \frac{N^{(c)}}{N} \mathcal{L}^{(c)}(\Psi, \eta), \tag{2}$$

where $N = \sum_{c=1}^{C} N^{(c)}$ and $N^{(c)}$ is the number of training samples available at client $c \in [1, C]$. Starting from (Ψ_0, η_0) , $\mathcal{L}(\Psi, \eta)$ is iteratively minimized over T collaboration rounds. At the start of round t, client parameters are initialized as: $\Psi_{t-1}^{(c)} = \Psi_{t-1}$ and $\eta_{t-1}^{(c)} = \eta_{t-1}$, $\forall t \in [1, T]$. In round t, the clients obtain:

$$\Psi_t^{(c)}, \eta_t^{(c)} = \underset{\Psi, \eta}{\operatorname{arg\,min}} \ \mathcal{L}^{(c)}(\Psi, \eta). \tag{3}$$

At the end of round t, the server aggregates the client parameters as:

$$\Psi_t = \sum_{c=1}^{C} \frac{N^{(c)}}{N} \Psi_t^{(c)}, \ \eta_t = \sum_{c=1}^{C} \frac{N^{(c)}}{N} \eta_t^{(c)}. \tag{4}$$

The above formulation can be considered as the federated version of full fine-tuning. For federated linear probing, only η is updated and $\Psi_T = \Psi_0$.

3.1 Proposed Variants of Federated PEFT Methods

Federated Subset Fine-tuning: Inspired by [27], we unfreeze *all* MHSA parameters across all the L blocks and fine-tune Θ in a federated fashion. Henceforth, we refer to this method as **all blocks attention (ABA)** with $\{\Theta, \eta\}$ being the only trainable parameters. The optimization formulation for ABA is $\min_{\Theta, \eta} \mathcal{L}(\Psi = \{\Lambda, \Phi, \Theta, \Omega\}, \eta)$.

In the ABA method, clients must fine-tune and communicate the parameters of L MHSA layers, which is roughly a third of the parameters involved in full fine-tuning. To further improve parameter efficiency, we propose **stochastic block attention (SBA)**, which requires updating parameters of only a *single* MHSA layer in each collaboration round. Specifically, the server randomly samples a block ℓ^* in each round, where $\ell^* \in [1, L]$, and unfreezes its corresponding MHSA weights θ_{ℓ^*} . Then, all clients learn $\{\theta_{\ell^*}, \eta\}$ collaboratively as $\min_{\theta_{\ell^*}, \eta} \mathcal{L}(\Psi, \eta)$.

The SBA method involves learning only a fraction (1/L) of the ABA parameters in each round, resulting in better communication efficiency. However, SBA requires the same storage as ABA because all the MHSA layers get updated over different rounds. In both ABA and SBA, FedAvg is used for aggregation.

Federated VPT: When VPT is used, the augmented ViT parameters can be denoted as $\mathcal{V}_{[\Psi,\mathcal{P}_v]}$ and the objective is $\min_{\mathcal{P}_v,\eta} \mathcal{L}([\Psi,\mathcal{P}_v],\eta)$, where the visual prompts are again aggregated through FedAvg. To further reduce the number of exchanged parameters, clients can decompose the locally learned prompts [30] into low-rank matrices using singular value decomposition (SVD) [17]. We refer to this technique as **Decomposed Visual Prompts** (DVPT), where the prompts from all transformer blocks are concatenated to obtain a $(LR \times d)$

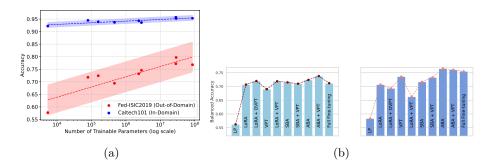


Fig. 2: (a) Accuracy vs. efficiency trade-off of various federated PEFT methods (Full Fine-tuning to LoRa shown in Table 1). The trade-off is more pronounced for OOD transfer (Fed-ISIC2019) compared to in-domain transfer (CalTech101). (b) Accuracy of federated PEFT methods on Fed-ISIC2019 with only 5 clients (excluding client 4), when (**Left**) base model is fine-tuned first with in-domain data (client 4 data) and (**Right**) base model is pre-trained using natural images. Clearly, the in-domain base model shows less performance variability.

matrix, which is decomposed as $\mathcal{A}_{vp} \in \mathbb{R}^{LR \times r_v}$ and $\mathcal{B}_{vp} \in \mathbb{R}^{r_v \times d}$. Here, r_v denotes the rank of visual prompt decomposition matrices \mathcal{A}_{vp} and \mathcal{B}_{vp} . **Federated LoRA**: When LoRA is used, the augmented ViT parameters can be denoted as $\mathcal{V}_{[\Psi,\mathcal{A},\mathcal{B}]}$ and the objective function is $\min_{\mathcal{A},\mathcal{B},\eta} \mathcal{L}([\Psi,\mathcal{A},\mathcal{B}],\eta)$. In federated LoRA, the server: (i) reconstructs back the weight update matrices sent by the clients as $\Delta \mathbf{W}_{Q,\ell}^{(c)} = \mathbf{B}_{Q,\ell}^{(c)} \mathbf{A}_{Q,\ell}^{(c)}$ and $\Delta \mathbf{W}_{V,\ell}^{(c)} = \mathbf{B}_{V,\ell}^{(c)} \mathbf{A}_{V,\ell}^{(c)}$, (ii) performs FedAvg based on these reconstructed matrices $\Delta \mathbf{W}_{Q,\ell}^{(c)}$ and $\Delta \mathbf{W}_{V,\ell}^{(c)}$, and (iii) applies SVD to the aggregated matrices to obtain the new global weight update matrices $\mathbf{B}_{Q,\ell}$, $\mathbf{A}_{Q,\ell}$, $\mathbf{B}_{V,\ell}$, and $\mathbf{A}_{V,\ell}$, which are sent back to the clients, where $\ell \in [1, L]$. Thus, federated LoRA provides communication efficiency for both the clients and the server, as well as involves less trainable parameters. The number of trainable parameters in LoRA is directly related to the rank r. Finally, another

key area of investigation in this work is understanding the impact of integrating multiple PEFT methods in federated settings to adapt pre-trained ViTs.

4 Results and Discussion

Datasets: We conducted experiments on Fed-ISIC2019 [26], HAM10000 [28], Caltech101 [10], and Flowers102 [21] datasets. While the first two datasets are from the medical imaging domain (OOD), the latter two correspond to in-domain scenarios. Fed-ISIC2019 also has non-IID data distribution.

Implementation Setup: We use the ImageNet [7] pre-trained ViT-B/16 model from timm library [29], with L=12 blocks, d=768, and S=196 patches. We use normal distribution with $\mu=0$ and $\sigma=0.1$ for LoRA initialization, with r=4 and $\alpha=2$. For VPT, we set R=50. For DVPT, we experimented

Table 1: Benchmarking of different approaches for federated fine-tuning of ViT. The number of exchangeable parameters (measured in *millions*) associated with each method is highlighted. Each experiment was repeated three times using different seeds, with the table reporting the mean and standard deviation.

		Exchange	able Parameters (\downarrow)	Balanced Accuracy (†)						
Method	Parameters	Number	Percentage (%)	HAM10000 (IID)	Fed-ISIC2019 (non-IID)	Caltech101 (IID)	Flowers102 (IID)			
Centralized	Ψ , η	-	-	0.805 ± 0.011	0.786 ± 0.015	0.964 ± 0.003	0.966 ± 0.006			
$\overline{\text{CentralizedVPT}}$	P_v , η	-	-	0.781 ± 0.008	0.746 ± 0.021	0.944 ± 0.002	0.966 ± 0.006			
Full Fine-tuning	Ψ , η	86.0	100	0.791 ± 0.025	0.768 ± 0.046	0.956 ± 0.004	0.970 ± 0.004			
ABA + VPT	Θ , η , P_v	28.815	33.5	$\textbf{0.812}\pm\textbf{0.019}$	$\textbf{0.797}\pm\textbf{0.001}$	0.946 ± 0.006	0.940 ± 0.002			
ABA	Θ , η	28.354	32.97	0.812 ± 0.024	0.772 ± 0.013	$\textbf{0.958}\pm\textbf{0.001}$	$\textbf{0.969}\pm\textbf{0.001}$			
$\overline{SBA + VPT}$	θ_l^*, η, P_v	2.829	3.29	0.792 ± 0.016	0.746 ± 0.014	0.942 ± 0.009	0.938 ± 0.009			
SBA	θ_l^* , η	2.368	2.75	0.782 ± 0.009	0.732 ± 0.009	0.948 ± 0.005	0.962 ± 0.002			
Lora + VPT	$\mathcal{A}, \mathcal{B}, \\ \mathcal{P}_v, \eta$	0.613	0.71	0.784 ± 0.018	0.729 ± 0.014	0.937 ± 0.006	0.947 ± 0.005			
VPT	P_v , η	0.466	0.54	0.789 ± 0.002	0.694 ± 0.013	0.939 ± 0.002	0.939 ± 0.005			
Lora + DVPT	$A, B, A_{vp}, B_{vp}, \eta$	0.231	0.27	0.772 ± 0.022	0.724 ± 0.011	0.940 ± 0.004	0.946 ± 0.008			
LoRA	A , B , η	0.152	0.18	0.770 ± 0.011	0.718 ± 0.004	0.949 ± 0.003	0.963 ± 0.006			
PromptFL	\mathcal{P}_t	0.026	0.03	0.384 ± 0.005	0.389 ± 0.003	0.929 ± 0.002	0.814 ± 0.006			
Linear Probing	η	0.005	0.006	0.714 ± 0.012	0.577 ± 0.015	0.924 ± 0.003	0.929 ± 0.011			
Local	Ψ, η	0.0	0.0	0.674 ± 0.019	0.291 ± 0.028	0.607 ± 0.068	0.458 ± 0.009			

with different rank values and set the rank r_v to 8. We run FL for T=200 collaboration rounds, employing an SGD optimizer with a learning rate of 10^{-2} , and a batch size of 32 using cross-entropy loss. We set the number of clients to C=6 and allow parameter exchange in every round. All experiments were implemented using PyTorch 2.1.0 and Nvidia A100 GPU. For more details on the datasets and experimental set-up, please refer to the supplementary material.

Results: Results are summarized in Table 1, where the methods are divided into three groups. The first two rows correspond to centralized training, where data from all clients gets pooled at one location (violating privacy constraints). This setting serves as a useful reference to understand the impact of FL. The middle group of methods exchange parameters related to the ViT in a federated setting. Among these methods, federated full fine-tuning of ViT (in yellow) is used as the baseline to assess the various PEFT methods. The third group (last three rows) corresponds to the case where no ViT parameters are exchanged.

Which federated PEFT method provides the best accuracy vs. efficiency trade-off? While the performance of ABA is comparable to the baseline across all datasets, SBA achieves good performance only on IID datasets and has a noticeable degradation in Fed-ISIC2019. Due to the non-IID nature of Fed-ISIC2019, the stochastic block selected at each round might lead to divergence in the training process in certain rounds. A similar trend was observed with VPT, LoRA, and linear probing. While most federated PEFT methods work well for the IID scenario, easily achieving up to three orders of magnitude decrease in the exchangeable parameters at a marginal cost to accuracy, they exhibit sub-optimal performance when there is statistical heterogeneity across clients.

Can federated PEFT transfer well for OOD tasks? Our main finding is that there is a trade-off between parameter efficiency and model accuracy in federated PEFT. While this trade-off is marginal for in-domain tasks (approximately 0.5% decrease in accuracy for every order of magnitude reduction in the number of parameters fine-tuned/exchanged), this trade-off becomes substantial for out-of-domain tasks with non-IID client distributions (approximately 4% decrease in accuracy for every order of magnitude reduction as shown in Figure 2a). Therefore, ABA is the best approach for OOD transfer, though it has less parameter efficiency. While existing wisdom is that PEFT can be achieved without compromising on model accuracy [25], we have demonstrated that the above claim is true only for in-domain tasks. For further validation, we first fine-tune the pre-trained ViT on client 4 of Fed-ISIC2019 and attempt to again fine-tune this new "pre-trained" model using the remaining 5 clients in a federated manner. Note that after the first fine-tuning, the classification head is discarded, but the feature extraction model is already familiar with the medical imaging domain. So, the second federated PEFT stage can be considered as in-domain transfer. As depicted in Fig. 2b (Left), there is little difference among the federated PEFT methods in this scenario, proving that they perform equally well for in-domain transfer. However, when the original "pre-trained" model is plugged back and collaboratively fine-tuned with the same 5 clients (excluding client 4), we observe significant variability in the accuracy (Fig. 2b (Right)).

Can a combination of PEFT methods further improve performance? We experimented with different combinations of PEFT methods by using ABA, SBA, and LoRA in conjunction with VPT. Note that since both attention fine-tuning (ABA and SBA) and LoRA attempt to update the attention weights, it does not make sense to combine them. The results show that combining VPT with attention fine-tuning is beneficial for OOD transfer, while it hurts in-domain transfer. This finding is confirmed by observing a similar trend when comparing the LoRA+VPT method with LoRA. Furthermore, combining LoRA with DVPT improves parameter efficiency by 3× while yielding almost similar results. Comparison with PromptFL: Federated learning of text prompts led to drastic performance degradation, particularly for OOD tasks (HAM10000 and Fed-ISIC2019), highlighting the relative superiority of federated VPT over PromptFL.

5 Conclusion

This work probed the efficacy of various federated PEFT methods to adapt pretrained vision transformers for medical image classification, focusing on achieving optimal performance while minimizing communication costs. Through extensive experimentation, we show that PEFT methods exhibit limited efficacy when applied to heterogeneous and out-of-domain datasets across participating clients. Hence, we recommend that it is preferable to start fine-tuning with in-domain medical foundation models (if available), rather than models pre-trained on natural images. Our findings also highlight the robustness of visual prompts over text prompts, especially when the task does not involve natural images.

References

- Babakniya, S., et al.: SLoRA: Federated Parameter Efficient Fine-Tuning of Language Models. In: NeurIPS Workshop (2023)
- 2. Bommasani, R., et al.: On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 (2022)
- Carion, N., et al.: End-to-end Object Detection with Transformers. In: ECCV. pp. 213–229 (2020)
- 4. Chen, S., et al.: AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition. In: NeurIPS. pp. 16664–16678 (2022)
- Cirrincione, G., et al.: Transformer-Based Approach to Melanoma Detection. Sensors 23(12) (2023)
- Dai, Y., Gao, Y., Liu, F.: TransMed: Transformers Advance Multi-Modal Medical Image Classification . Diagnostics 11(8) (2021)
- Deng, J., et al.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
- 8. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
- 9. Duong, L.T., et al.: Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning. Expert Systems with Applications 184 (2021)
- Fei-Fei, L., Fergus, R., Perona, P.: One-Shot Learning of Object Categories. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(4), 594–611 (2006)
- 11. Guo, T., et al.: PromptFL: Let Federated Participants Cooperatively Learn Prompts Instead of Models Federated Learning in Age of Foundation Model. IEEE Transactions on Mobile Computing (2023)
- 12. He, J., et al.: Towards a Unified View on Visual Parameter-Efficient Transfer Learning. In: ICLR (2022)
- 13. He, K., et al.: Masked Autoencoders Are Scalable Vision Learners. In: CVPR. pp. 16000–16009 (2022)
- 14. He, X., et al.: Parameter-Efficient Model Adaptation for Vision Transformers. In: AAAI. pp. 817–825 (2023)
- 15. Hu, E.J., et al.: LoRA: Low-Rank Adaptation of Large Language Models. In: ICLR (2022)
- 16. Jia, M., et al.: Visual prompt tuning. In: ECCV. pp. 709–727 (2022)
- 17. Klema, V., Laub, A.: The singular value decomposition: Its computation and some applications. IEEE Transactions on Automatic Control **25**(2), 164–176 (1980)
- 18. Lu, M., et al.: Smile: Sparse-attention based multiple instance contrastive learning for glioma sub-type classification using pathological images. In: MICCAI Workshop on Computational Pathology. pp. 159–169 (2021)
- 19. Lu, Z., et al.: Bridging the Gap Between Vision Transformers and Convolutional Neural Networks on Small Datasets. In: NeurIPS. pp. 14663–14677 (2022)
- 20. McMahan, B., et al.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: AISTATS. pp. 1273–1282 (2017)
- 21. Nilsback, M.E., Zisserman, A.: Automated Flower Classification over a Large Number of Classes. In: ICVGIP. pp. 722–729 (2008)
- 22. Okolo, G.I., Katsigiannis, S., Ramzan, N.: IEViT: An enhanced vision transformer architecture for chest X-ray image classification. Computer Methods and Programs in Biomedicine **226** (2022)

- Radford, A., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML. pp. 8748–8763 (2021)
- Sarker, M.M.K., et al.: TransSLC: Skin lesion classification in dermatoscopic images using transformers. In: Medical Image Understanding and Analysis. pp. 651–660 (2022)
- 25. Sun, G., et al.: Conquering the Communication Constraints to Enable Large Pre-Trained Models in Federated Learning. arXiv:2210.01708 (2022)
- 26. Terrail, J.O., et al.: FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings. In: NeurIPS (2022)
- 27. Touvron, H., et al.: Three things everyone should know about Vision Transformers. In: ECCV. pp. 497–515 (2022)
- 28. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data 5(11) (2018)
- 29. Wightman, R.: Pytorch image models (2019)
- 30. Xiao, Y., et al.: Decomposed Prompt Tuning via Low-Rank Reparameterization. In: Findings of EMNLP (2023)
- 31. Yang, F.E., Wang, C.Y., Wang, Y.C.F.: Efficient Model Personalization in Federated Learning via Client-Specific Prompt Generation. In: ICCV. pp. 19159–19168 (2023)
- Zamir, A.R., et al.: Taskonomy: Disentangling Task Transfer Learning. In: CVPR. pp. 3712–3722 (2018)
- 33. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical Image Analysis **91** (2024)
- 34. Zhuang, W., Chen, C., Lyu, L.: When Foundation Model Meets Federated Learning: Motivations, Challenges, and Future Directions. arXiv:2306.15546 (2023)

Supplementary Material

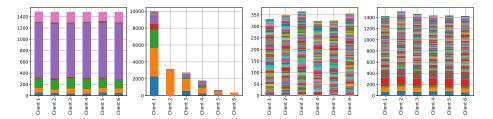


Fig. 3: From left to right, distribution of **HAM10000** (IID), **Fed-ISIC2019** (Non-IID), Flowers102 (IID), and Caltech101 (IID) datasets. Each stacked bar represents the number of training samples, and each color represents a class. Fed-ISIC2019 [26] contains 23, 247 samples across eight melanoma classes. HAM10000 [28] comprises 10,015 dermoscopic images categorized into 7 lesion types. We employ 80% (20%) train (test) split for both these datasets. Caltech101 [10] has 101 categories of natural images with a 50% (50%) train (test) split. Flowers102 [21] includes 102 categories with a 25% (75%) train (test) split.

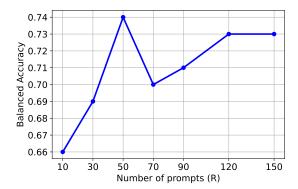


Fig. 4: Balanced accuracy with different number of prompts for the VPT method on Fed-ISIC2019 dataset. We found that optimal performance was achieved with R=50 prompts.

Table 2: Balanced accuracy of LoRA with different initialization methods, scale, and rank for the HAM10000 and Fed-ISIC2019 datasets. Following [15], we set all matrices in $\mathcal B$ to $\mathbf 0$. We observe that initializing $\mathcal A$ based on a Normal(0,0.1) distribution with r=4 and $\alpha=2$ represents the most effective trade-off between performance and the number of trained parameters associated with LoRA.

	Xavier		vier	Kaiming		ImageNet		Normal $\mu = 0, \sigma = 0.5$		Normal $\mu = 0, \sigma = 0.1$	
	Rank	4	8	4	8	4	8	4	8	4	8
HAM10000	Scale = 2 $Scale = 0.5$										$0.801 \\ 0.831$
FedISIC2019	$\begin{array}{c} {\rm Scale} = 2 \\ {\rm Scale} = 0.5 \end{array}$	$0.754 \\ 0.758$	$0.777 \\ 0.764$	$0.765 \\ 0.723$	$0.785 \\ 0.730$	$0.757 \\ 0.741$	$0.773 \\ 0.751$	$0.495 \\ 0.748$	$0.572 \\ 0.651$	0.757 0.746	$0.786 \\ 0.742$