Sparsity- and Hybridity-Inspired Visual Parameter-Efficient Fine-Tuning for Medical Diagnosis

Mingyuan Liu[1], Lu Xu[1], Shengnan Liu[1], and Jicong Zhang[1,2,(✉)]

[1]School of Biological Science and Medical Engineering,
Beihang University, Beijing, China
[2]Hefei Innovation Research Institute, Beihang University, Hefei, Anhui, China
jicongzhang@buaa.edu.cn

Abstract. The success of Large Vision Models (LVMs) is accompanied by vast data volumes, which are prohibitively expensive in medical diagnosis. To address this, recent efforts exploit Parameter-Efficient Fine-Tuning (PEFT), which trains a small number of weights while freezing the rest for knowledge transfer. However, they typically assign trainable weights to the same positions in LVMs in a heuristic manner, regardless of task differences, making them suboptimal for professional applications like medical diagnosis. To address this, we statistically reveal the nature of sparsity and hybridity during diagnostic-targeted fine-tuning, i.e., a small portion of key weights significantly impacts performance, and these key weights are hybrid, including both task-specific and task-agnostic parts. Based on this, we propose a novel Sparsity- and Hybridity-inspired Parameter Efficient Fine-Tuning (SH-PEFT). It selects and trains a small portion of weights based on their importance, which is innovatively estimated by hybridizing both task-specific and task-agnostic strategies. Validated on six medical datasets of different modalities, we demonstrate that SH-PEFT achieves state-of-the-art performance in transferring LVMs to medical diagnosis in terms of accuracy. By tuning around 0.01% number of weights, it outperforms full model fine-tuning. Moreover, SH-PEFT also performs comparably to other models deliberately optimized for specific medical tasks. Extensive experiments demonstrate the effectiveness of each design and reveal the great potential of pre-trained LVM transfer for medical diagnosis.

Keywords: Parameter-efficient fine-tuning · Medical diagnosis · Vision transformer · Sparsity and hybridity.

1 Introduction

With the support of the vision transformer and massive data, large visual models (LVMs) have achieved great success [27, 7]. However, when it comes to professional tasks such as medical diagnosis, the performance of LVMs is still insufficient. Training medical-specific LVMs is prohibitively expensive, due to the

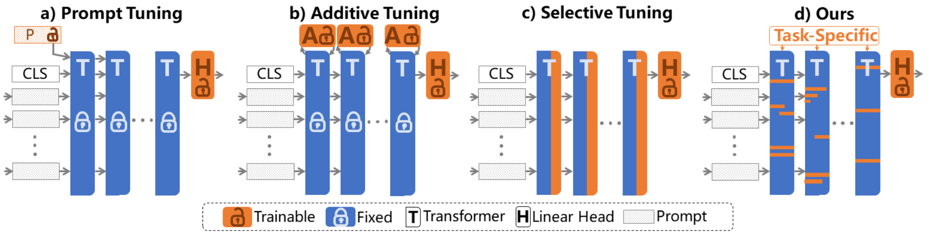

Fig. 1: During PEFT, instead of heuristically assigning trainable weights to fixed positions across various medical diagnostic tasks, we employ a data-driven approach to select key weights for each task, enabling more effective fine-tuning.

difficulties of acquiring a large volume of medical data [28, 23, 8]. To address this, Parameter-Efficient Fine-Tuning (PEFT) is proposed, which tunes only a small fraction of weights while freezing the rest [22, 12, 6]. It promotes effective knowledge transfer, reduces optimization difficulty, saves storage burden, and avoids over-fitting.

Recent PEFT efforts heuristically assign trainable weights to the same positions in LVMs. As shown in Fig. 1, they could be roughly divided into three categories: **(1) Prompt Tuning** introduces trainable tokens while maintaining the pre-trained weights frozen [16]. Techniques, such as the position of trainable tokens [13, 19] and operations performed on them [26], are explored. Furthermore, the generalization capabilities are validated across different tasks, such as image generation [29], image segmentation [21], and video understanding [14]. **(2) Additive Tuning** inserts new trainable modules (a.k.a, adapters) either between or alongside existing transformer blocks [11]. A representative example is AdaptFormer [4], which appends trainable encode-decoder modules to each transformer block. Moreover, unlike most designs that incur additional computational overhead during inference, LoRA [12] utilizes low-rank decomposition to integrate adapters into the original LVM, avoiding extra computational cost. **(3) Selective Tuning** trains a subset of weights within a LVM without changing the overall network structure, such as training all biases [33, 3], attention layers [30], or normalization layers [2], while keeping the remaining weights frozen. **However**, the aforementioned methods add trainable parameters in a fixed manner to the same locations, ignoring the diverse downstream tasks and image modalities, therefore leading to suboptimal performance.

Motivated by this, we first conduct a statistical exploration of weight changes between pre-trained and fully fine-tuned LVMs on medical diagnosis. Our analysis reveals two significant findings: **(1) Sparsity** indicates that a minority of key weights play a majority role in downstream adaptation. **(2) Hybridity** means the positions of these key weights partially overlap across tasks, including both task-specific and task-agnostic components. Based on the findings, we introduce a novel strategy called Sparsity- and Hybridity-inspired visual PEFT (SH-PEFT) for adapting pre-trained LVMs to medical diagnosis. SH-PEFT selects a

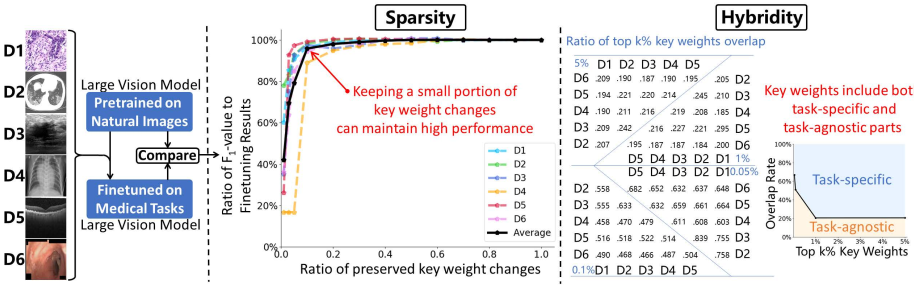

Fig. 2: We experimentally conclude the sparsity and hybridity nature of key weight distributions from six medical datasets, by comparing differences between pre-trained and medical fine-tuned CLIP model. **Sparsity** indicates that a few key weights largely impact performance, thereby motivating us to select important weights for tuning. **Hybridity** indicates that key weights contain both task-specific and task-agnostic parts, thereby prompting us to explore a hybrid strategy to locate key weights for more effective PEFT.

subset of weights that potentially introduce significant impacts on performance for tuning, based on their estimated importance. The importance of each weight is innovatively approximated based on its hybrid contributions: its task-specific role in minimizing loss for a specific downstream medical task, as well as its task-agnostic significance within a LVM. The estimation criterion is simple and effective, requiring only about 120 seconds for a dataset with 10k images, using a ViT-B/16 transformer, which is far less than the subsequent training time.

Our contributions are three folds: **(1)** Based on our statistical analysis, we reveal the sparsity and hybridity characteristics that exist in the process of transferring a large vision model to medical diagnosis. **(2)** We propose a novel Sparsity- and Hybridity-inspired Parameter Efficient Fine-Tuning (SH-PEFT), which estimates the importance of each weight by hybridizing both task-specific and task-agnostic strategies, and subsequently selects a small number of the most important weights for effective tuning. **(3)** Extensive experiments on six medical datasets with different modalities reveal the effectiveness of our SH-PEFT: it achieves state-of-the-art PEFT performance under a comparable number of trainable weights; it performs comparably with models deliberately designed for specific diagnostic tasks; its hybrid weight importance estimation strategy effectively enhances performance.

2 Method

2.1 Statistical Evidence of Sparsity and Hybridity

For constructing effective PEFT for medical diagnosis, we first explore the feasibility of tuning a few key weights and the proper location to introduce these trainable weights. As shown in Fig. 2, we experimentally reveal the existence of sparsity and hybridity nature of the key weight distributions, by analyzing weight differences between pre-trained and medical fine-tuned models.

Specifically, experiments are conducted on six medical datasets with different modalities, to maximize the applicability of our conclusions to various diagnostic tasks. Examples of six datasets are shown in Fig. 2. We choose CLIP [27] as the LVM due to its widespread application and outstanding performance among publicly available checkpoints. More details are elaborated in Sec. 3.1.

Sparsity indicates that a small ratio of key weights significantly impacts the performance of knowledge transfer, suggesting the potential for selective fine-tuning in medical PEFT. Given pre-trained weights $\mathcal{W}^{ori}$ of a model $\varPhi_{\mathcal{W}^{ori}}(\cdot)$ and fine-tuned weights $\mathcal{W}^{ft}$ on a medical task, $\varDelta\mathcal{W}^{ft} = |\mathcal{W}^{ft} - \mathcal{W}^{ori}|$ measures weight changes. We first identify top $k\%$ elements with the largest variations and directly replace weights in $\mathcal{W}^{ori}$ at positions of top $k\%$ by the corresponding weights from $\mathcal{W}^{ft}$, denoted as $\mathcal{W}^{ori \leftarrow ft@k\%}$. Then the performance of $\varPhi_{\mathcal{W}^{ori \leftarrow ft@k\%}}(\cdot)$ is directly validated, and results are shown in Fig. 2. Results show that keeping around 10% of the changes could maintain around 95% of the full fine-tune performance, revealing the feasibility of selective tuning for medical PEFT.

Hybridity indicates that key weights contain both task-specific and task-agnostic parts, suggesting the necessity of proposing a hybrid strategy for locating key weights. Given fine-tuned weights on different medical diagnostic tasks $\mathcal{W}^{ft}_{T_m}$ and $\mathcal{W}^{ft}_{T_n}$, we measure the positional overlap of the key weights between them at different $k\%$ (5%, 1%, 0.1%, and 0.05%). Results in Fig. 2 show that, in different tasks, most of the key weights do not overlap, indicating they are task-specific. Meanwhile, positions of a small portion of key weights are shared across different tasks, suggesting they are task-agnostic. This inspires us to hybrid both task-specific and task-agnostic strategies to explore the positions of key weights.

2.2 Hybrid Weight Importance Estimation for PEFT

Inspired by the aforementioned findings, we propose SH-PEFT to adaptively determine trainable weights in a model, by jointly considering their importance in both the specific task and the model structure, as shown in Fig. 3

Given a dataset $\mathcal{D}_t$, the learning objective is to minimize the empirical risk $E(\mathcal{D}_t, \mathcal{W})$ by updating weights $\mathcal{W}$ in a model. For a model with m layers, $\mathcal{W} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_m\}$ and the n-th weight in layer m is $\mathbf{w}_{m,n}$. Its importance $\mathbb{I}_{m,n}$ could be estimated by:

$$\begin{aligned} \mathbb{I}_{m,n} &= \mathbb{I}^{td}_{m,n} + \lambda \mathbb{I}^{ta}_{m,n} \\ &= |\varDelta E(\mathcal{D}_t, \mathcal{W}|\mathbf{w}_{m,n} \rightarrow \hat{\mathbf{w}}_{m,n})| + \lambda |\varDelta E(\mathcal{D}, \mathcal{W}|\mathbf{w}_{m,n} \rightarrow 0)| \end{aligned} \tag{1}$$

The first term $\mathbb{I}^{td}_{m,n}$ is task-dependent. It measures the change of empirical risk caused by the weight update from $\mathbf{w}_{m,n}$ to $\hat{\mathbf{w}}_{m,n}$ after training on $\mathcal{D}_t$. The second term $\mathbb{I}^{ta}_{m,n}$ is task-agnostic. It measures the empirical risk change by removing a weight $\mathbf{w}_{m,n}$ on arbitrary task $\mathcal{D}$, reflecting the significance of a weight $\mathbf{w}_{m,n}$

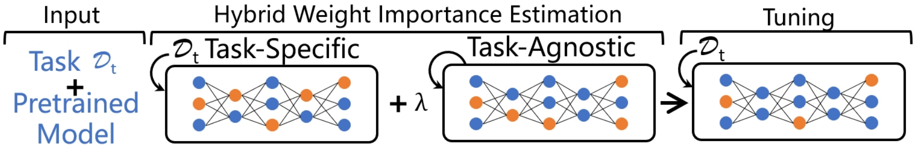

Fig. 3: Inspired by the sparsity and hybridity, we propose a novel SH-PEFT approach to fine-tune a few key weights for adapting pre-trained vision transformers to medical diagnosis. The key weights can be effectively and quickly identified by jointly considering their importance from both task-specific and task-agnostic perspectives.

in the model. λ balances the values between the two terms. However, the two terms are difficult to estimate. Because they require training or evaluating a model several times for each weight, which is computationally prohibitive.

To address this issue, we estimate the empirical risk changes in two ways. The first way is intuitive. We use the accumulation of gradients over several iterations to judge the trend of weight changes, i.e., $\mathbb{I}^{td-L1}_{m,n} = \Sigma^{B}_{b=1} \partial E_b / \partial \mathbf{w}_{m,n}$, where B is the number of iteration and $\partial E_b / \partial \mathbf{w}_{m,n}$ calculates the gradient on each mini-batch of weight $\mathbf{w}_{m,n}$. $L1$ means it is the first estimation method. The task-agnostic part is estimated by $\mathbb{I}^{ta-L1}_{m,n} = |\mathbf{w}_{m,n}|$, which is the absolute value of the weight. It is based on the assumption that a large weight can cause significant changes to the input features so that it plays a more significant role in maintaining the functionality of the model than a small weight. This assumption is previously applied in the model pruning tasks [24, 34] and is transferred to PEFT scenario by our SH-PEFT.

The second way is inspired by [10], where the task-dependent importance of $\mathbf{w}_{m,n}$ could be estimated by its first-order Taylor expansion of loss $\mathcal{L}$ in its vicinity range. It could be written as $\mathbb{I}^{td-L2}_{m,n} = \partial \mathcal{L} / \partial \mathbf{w}_{m,n} * (\hat{\mathbf{w}}_{m,n} - \mathbf{w}_{m,n})$. Since the estimation should be finished within a few forward passes, the weight differences could be abbreviated as its gradient, so that $\mathbb{I}^{td-L2}_{m,n} \approx (\partial \mathcal{L} / \partial \mathbf{w}_{m,n})^2$. To avoid inconsistent scaling between the two terms due to squaring, we also apply squaring to the task-agnostic part $\mathbb{I}^{ta-L2}_{m,n} = (\mathbf{w}_{m,n})^2$. Moreover, in both ways, before applying λ, the second term is weighted by $\Sigma \mathbb{I}^{td}_{m,n} / \Sigma \mathbb{I}^{ta}_{m,n}$ to balance their scaling differences.

After estimating the importance of all parameters, we set a threshold τ based on the number of trainable weights to be allocated. If the importance of a weight is larger than τ, the weight can be updated; if it is smaller, it remains unchanged. Therefore, the update strategy for each weight at step t+1 is:

$$\mathbf{w}^{t+1}_{m,n} = \mathbf{w}^{t}_{m,n} - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{w}^{t}_{m,n}} \mathbf{M}_{m,n} \text{ ,} \mathit{where}\ \mathbf{M}_{m,n} = \begin{cases} 1 & \mathit{if}\ \mathbb{I}_{m,n} > \tau \\ 0 & \mathit{else} \end{cases} \tag{2}$$

, where $\mathbf{M}_{m,n}$ is a binary mask of weight $\mathbf{w}_{m,n}$ and η is the learning rate.

It is worth noting that, although we gain inspiration from previous works [10, 24, 34], we have unique contributions. Specifically, our SH-PEFT differs from the most similar work, SPT [10], in three aspects: **1) The scope of weight selection**: SPT selects weights only from linear layers for data-specific tuning. However, other layers like layer-norm also play an important role in PEFT [20, 2]. Our SH-PEFT extends the selection scope to all operations, which enhances the model adaptability to downstream tasks. **2) The strategy of weight importance estimation**: SPT uses the square of the derivative to estimate the weight importance in a task-specific manner. In contrast, we draw inspiration from our discovered hybridity nature and jointly consider both task-specific and -agnostic factors, leading to more reliable weight estimation strategy for better PEFT. **3) The mode of usage**: SPT works in conjunction with other PEFT methods like adapters [11]. Differently, our SH-PEFT could work independently, which flexibly circumvents the shortage of other methods, such as the computational overhead introduced by an adapter.

3 Experiment and Result

3.1 Training Details

Six datasets with different modalities are used for our statistical analysis and quantitative evaluation in order to ensure the applicability of the conclusion and method to diverse medical diagnostic problems. They include: **D1**: Chaoyang [35] is a *pathological* dataset of the human colon. It has 4,021 training and 2,139 testing images, covering 4 categories: normal, serrated, adenocarcinoma, and adenoma. **D2**: Covid19-CT [32] is lung *CT* dataset for diagnosing Covid-19. It has 425/118/203 images for training/validation/testing, respectively. **D3**: BUSI [1] is an *ultrasound* image dataset for early diagnosis of normal, benign, or malignant breast cancer, with 559/79/160 images for training/validation/testing. **D4**: CXT3 [15] is a chest *X-ray* dataset from children including normal, bacterial, and viral cases, with 4,708/524/1,248 images for training/validation/testing. **D5**: OCT [15] is a retina *OCT* dataset, with 97,477/10,832/1,000 images for training/validation/testing. including choroidal neovascularization, diabetic macular edema, multiple drusen, and normal. **D6**: LIUMC [25] is a *colonoscopy* dataset for ulcerative colitis with 4 securities. It has 9,590 images for training and 1,686 images for testing.

Experiments are conducted on CLIP [27] pre-trained visual transformer of ViT-B/16 structure. This is because its innovative training approach, which connects images and text, has become a paradigm for other large-model training, as well as due to its excellent performance among publicly available checkpoints. During training, the final projection layer in vision transformer is replaced by a L_2 normalization and a linear layer. The training uses the SGD optimizer with batch size 64. The initial learning rate is 0.001 and is adjusted by CosineAnnealing. Each dataset is trained for 40k iterations, and F_1-value is the main measurement of the final performance. The key weights are selected within one training epoch. Unless otherwise specified, the SH-PEFT model uses $\mathbb{I}^{L2}$ strategy with $\lambda = 1$, and selects 1% of the parameters as trainable parameters in the following experiments

Table 1: Comparison with state-of-the-art PEFT methods, measured by F_1 (%). 'S', 'A', and 'P', denote selective, additive, and prompt tuning respectively.

Method (Pub'Year)	Type	$\mathcal{D}1$	$\mathcal{D}2$	$\mathcal{D}3$	$\mathcal{D}4$	$\mathcal{D}5$	$\mathcal{D}6$	Avg
Full Finetune	S	80.2	73.5	82.5	76.3	94.2	69.1	79.3
Linear Prob	S	68.3	76.4	76.0	79.6	81.9	60.4	73.8
Adapter-par (NeurIPS'22) [4]	A	75.5	78.6	84.1	77.7	93.2	67.4	79.4
SSF (NeurIPS'22) [20]	A	78.8	79.5	89.8	78.5	92.2	71.9	81.8
LoRa (ICLR'22) [12]	A	**81.9**	82.4	87.8	78.0	**96.3**	67.9	82.4
VPT-Deep (ECCV'22) [13]	P	70.4	78.2	75.1	78.0	82.3	64.2	74.7
VPT-Shallow (ECCV'22) [13]	P	74.2	78.8	81.6	80.3	90.0	69.3	79.0
FT-LN (Arxiv'23) [2]	S	74.2	75.5	79.6	76.4	87.2	68.1	76.8
BitFit (ACL'22) [33]	S	79.1	80.0	86.1	73.6	93.3	72.2	80.7
FT-Att (ECCV'22) [30]	S	81.1	81.6	88.8	79.4	95.8	70.1	82.8
SPT-LoRa (ICCV'23) [10]	S+A	81.8	80.2	90.2	77.3	96.0	69.5	82.9
SH-PEFT (Ours)	S	80.6	**83.0**	**90.5**	**81.0**	95.6	**72.7**	**83.9**

Table 2: Comparison with latest efforts on Chaoyang dataset.

Method	F_1%	ACC%
NSHE(TMI'22) [35]	76.5	83.4
PVB+L(ECCV'22) [17]	-	84.3
GSB(NN'24) [9]	-	82.5
SH-PEFT (Ours)	**80.6**	**84.8**

Table 3: Comparison with latest efforts on COVID19-CT dataset.

Method	F_1%	ACC%
SKNet(CVPR'19) [18]	76	77
ECAN (ECCV'20) [31]	74	75
ResGANet(MIA'22) [5]	81	80
SH-PEFT (Ours)	**83.0**	**83.3**

3.2 Comparison with State-of-the-art Methods

Superiority over PEFT methods: Table 1 shows our SH-PEFT outperforms PEFT methods of different types on six medical datasets measured by F_1-value (Due to space limitations, ACC and AUC results are shown in Supp.). It demonstrates that our method can effectively transfer general visual knowledge from LVMs to medical diagnosis, indicating the effectiveness of our flexible selective learning method and hybrid feature weight importance estimation strategy. Specifically, for fair comparisons, all models are trained following their official implementation and use the same hyper-parameters mentioned in Sec. 3.1.

Superiority over Domain-specific Methods: Table 2 and Table 3 show that our SH-PEFT achieves comparable results to recent deep learning methods that are optimized for these specific tasks (Due to space limitations, more results of other datasets are shown in Supp.). The outstanding outcomes indicate that our SH-PEFT is effective and designing medical-targeted PEFT could be a promising approach for better medical diagnostic applications.

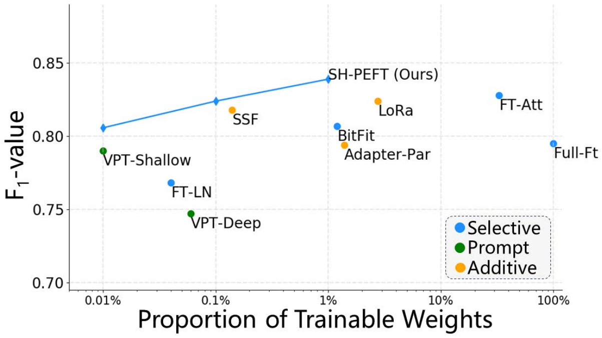

Fig. 4: Under the same ratio of trainable weights, SH-PEFT outperforms state-of-the-art PEFT methods in terms of the average F_1-value across six datasets.

Table 4: Ablation on hybrid estimation strategies, measured on the average of six datasets.

$\mathbb{I}^{ta}$	$\mathbb{I}^{td}$	F_1%	ACC%
✓		81.9	84.5
	L1	81.3	84.3
✓	L1	82.4	85.0
	L2	82.0	84.7
✓	L2	**83.9**	**86.3**

Table 5: Ablation on the balancing weight λ, measured on the average of six datasets.

λ	F_1%	ACC%
1.5	82.4	85.1
1.2	83.1	85.6
1.0	**83.9**	**86.3**
0.8	82.6	85.1
0.5	82.5	85.1

Superiority Under Comparable Number of Trainable Weights: Fig. 4 demonstrates that, compared to other methods, our SH-PEFT achieves better performance when the number of trainable parameters is similar. It is worth noting that tuning around 0.01% number of weights by SH-PEFT outperforms full model fine-tuning, indicating that our allocation of trainable parameters is effective.

3.3 Ablation Studies

Hybrid weight importance estimation: Table 4 demonstrates that task-specific and task-agnostic strategies could work complementarily to improve performance, indicating the effectiveness of our hybrid strategy.
Effectivevness of λ: Table 5 demonstrates the performance is relatively robust to the selection of λ, and the best performance is achieved when λ equals to 1.

4 Conclusion

We statistically reveal the characteristics of sparsity and hybridity when transferring general LVMs to medical diagnosis. Inspired by this, we propose SH-PEFT to allocate a small portion of trainable weights for tuning, based on their hybrid importance measured in both task-specific and task-agnostic manner. We validate the effectiveness of PEFT on six medical diagnosis datasets with different modalities. Results show that, with the same number of trainable weights, our SH-PEFT outperforms existing PEFT methods in terms of accuracy. Furthermore, the model fine-tuned by SH-PEFT outperforms deep learning models specifically optimized for diagnostic tasks, indicating the effectiveness of our strategy. Ablation studies further demonstrate the effectiveness of each design.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant No. 62394313, 62371024), the Beijing Natural Science Foundation (Grant No. Z200024), and the Industrialization Research Project of Hefei Innovation Research Institute, Beihang University (Grant No. BHKX-20-01).

Disclosure of Interests. The authors declare no competing interests.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Basu, S., Massiceti, D., Hu, S.X., Feizi, S.: Strong baselines for parameter efficient few-shot fine-tuning. ArXiv:2304.01917 (2023)
3. Cai, H., Gan, C., Zhu, L., Han, S.: Tinytl: Reduce memory, not parameters for efficient on-device learning. NeurIPS **33**, 11285–11297 (2020)
4. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: Adaptformer: Adapting vision transformers for scalable visual recognition. NeurIPS **35**, 16664–16678 (2022)
5. Cheng, J., Tian, S., Yu, L., Gao, C., Kang, X., Ma, X., Wu, W., Liu, S., Lu, H.: Resganet: Residual group attention network for medical image classification and segmentation. MedIA **76**, 102313 (2022)
6. Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.M., Chen, W., et al.: Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence **5**(3), 220–235 (2023)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
8. Dutt, R., Ericsson, L., Sanchez, P., Tsaftaris, S.A., Hospedales, T.: Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. ArXiv:2305.08252 (2023)
9. Gao, T., Xu, C.Z., Zhang, L., Kong, H.: Gsb: Group superposition binarization for vision transformer with limited training samples. Neural Networks **172**, 106133 (2024)
10. He, H., Cai, J., Zhang, J., Tao, D., Zhuang, B.: Sensitivity-aware visual parameter-efficient fine-tuning. In: ICCV. pp. 11825–11835 (2023)
11. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: ICML. pp. 2790–2799. PMLR (2019)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)

13. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: ECCV. pp. 709–727. Springer (2022)
14. Ju, C., Han, T., Zheng, K., Zhang, Y., Xie, W.: Prompting visual-language models for efficient video understanding. In: ECCV. pp. 105–124. Springer (2022)
15. Kermany, D., Zhang, K., Goldbaum, M.: Large dataset of labeled optical coherence tomography (oct) and chest x-ray images. Mendeley Data **3**(10.17632) (2018)
16. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. ArXiv:2104.08691 (2021)
17. Li, K., Yu, R., Wang, Z., Yuan, L., Song, G., Chen, J.: Locality guidance for improving vision transformers on tiny datasets. In: ECCV. pp. 110–127. Springer (2022)
18. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: CVPR. pp. 510–519 (2019)
19. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: ACL-IJCNLP. pp. 4582–4597 (2021)
20. Lian, D., Zhou, D., Feng, J., Wang, X.: Scaling & shifting your features: A new baseline for efficient model tuning. NeurIPS **35**, 109–123 (2022)
21. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: CVPR. pp. 7061–7070 (2023)
22. Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., Raffel, C.A.: Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. NeurIPS **35**, 1950–1965 (2022)
23. Matsoukas, C., Haslum, J.F., Söderberg, M., Smith, K.: Is it time to replace cnns with transformers for medical images? ArXiv:2108.09038 (2021)
24. Molchanov, P., Mallya, A., Tyree, S., Frosio, I., Kautz, J.: Importance estimation for neural network pruning. In: CVPR. pp. 11264–11272 (2019)
25. Polat, G., Kani, H.T., Ergenc, I., Ozen Alahdab, Y., Temizel, A., Atug, O.: Improving the computer-aided estimation of ulcerative colitis severity according to mayo endoscopic score by using regression-based deep learning. Inflammatory Bowel Diseases **29**(9), 1431–1439 (2023)
26. Qin, Y., Wang, X., Su, Y., Lin, Y., Ding, N., Yi, J., Chen, W., Liu, Z., Li, J., Hou, L., et al.: Exploring universal intrinsic task subspace via prompt tuning. ArXiv:2110.07867 (2021)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)
28. Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. MedIA p. 102802 (2023)
29. Sohn, K., Chang, H., Lezama, J., Polania, L., Zhang, H., Hao, Y., Essa, I., Jiang, L.: Visual prompt tuning for generative transfer learning. In: CVPR. pp. 19840–19851 (2023)
30. Touvron, H., Cord, M., El-Nouby, A., Verbeek, J., Jégou, H.: Three things everyone should know about vision transformers. In: ECCV. pp. 497–515. Springer (2022)
31. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: CVPR. pp. 11534–11542 (2020)
32. Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., Xie, P.: Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865 (2020)
33. Zaken, E.B., Goldberg, Y., Ravfogel, S.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In: ACL (Volume 2: Short Papers). pp. 1–9 (2022)

34. Zhang, Q., Zuo, S., Liang, C., Bukharin, A., He, P., Chen, W., Zhao, T.: Platon: Pruning large transformer models with upper confidence bound of weight importance. In: ICML. pp. 26809–26823. PMLR (2022)
35. Zhu, C., Chen, W., Peng, T., Wang, Y., Jin, M.: Hard sample aware noise robust learning for histopathology image classification. IEEE TMI **41**(4), 881–894 (2021)