

Mini but Mighty: Finetuning ViTs with Mini Adapters

Imad Eddine Marouf Enzo Tartaglione Stéphane Lathuilière
 LTCI, Télécom-Paris, Institut Polytechnique de Paris, France

imad.marouf@ip-paris.fr

Abstract

Vision Transformers (ViTs) have become one of the dominant architectures in computer vision, and pre-trained ViT models are commonly adapted to new tasks via finetuning. Recent works proposed several parameter-efficient transfer learning methods, such as adapters, to avoid the prohibitive training and storage cost of finetuning.

In this work, we observe that adapters perform poorly when the dimension of adapters is small, and we propose MiMi, a training framework that addresses this issue. We start with large adapters which can reach high performance, and iteratively reduce their size. To enable automatic estimation of the hidden dimension of every adapter, we also introduce a new scoring function, specifically designed for adapters, that compares the neuron importance across layers. Our method outperforms existing methods in finding the best trade-off between accuracy and trained parameters across the three dataset benchmarks DomainNet, VTAB, and Multi-task, for a total of 29 datasets.¹

1. Introduction

Transformers have gained increasing attention owing to their outstanding performance [6, 24, 40, 41]: Vision Transformers (ViTs) trained on large-scale datasets have demonstrated a remarkable ability to learn new tasks [6]. The most commonly adopted strategy to learn new tasks consists of fully or partially fine-tuning a pre-trained network; however, when dealing with multiple tasks, this approach necessitates training multiple separate models, which results in large storage costs.

Recently, Parameter-Efficient Training (PET) approaches have been developed to help large pre-trained models adapt to new tasks, with minimal added parameters [11, 15, 17]. Among these, adapters [14] and its variants [11, 19, 26] are frequently employed for Natural Language Processing (NLP) tasks. Adapters are small modules inserted into transformer blocks, which enable

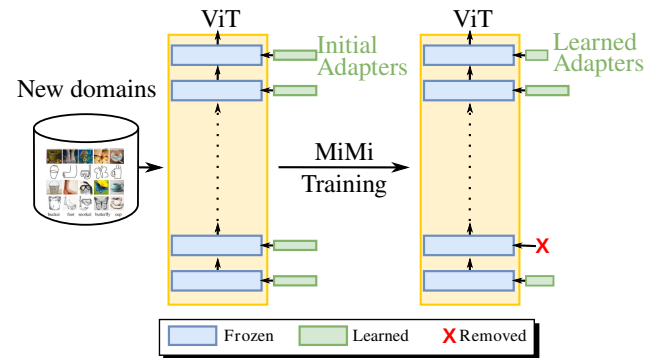


Figure 1. Layer-wise small blocks are injected into ViTs to efficiently adapt to new domains. MiMi estimates the best rank for each adapter weight, it reduces the number of parameters and removes completely injected adapters for some layers if necessary.

efficient adjustment of the data representation to the downstream task: they offer similar performance to full fine-tuning (*i.e.* updating all parameters) while requiring a very low number of trainable parameters [14, 34].

When it comes to vision tasks, PET approaches are mostly explored for convolutional neural networks [2, 27, 35, 36]. In contrast, several PET approaches have been proposed in NLP tasks: here, adapters are Multi-Layer-Perceptrons (MLPs) equipped with residual connections [14]. These multi-layer adapters can fit new tasks with enough representation power and the size of their hidden layers provides a simple trade-off between performance and parameter efficiency [14]. Nevertheless, they suffer from two weaknesses. First, the performance drops when the size of multi-layer adapters is too small [4] (as confirmed by our experiments -see Se. 4.1, and supplementary material-). Second, the optimal hyper-parametrization of adapters is complex: the hidden layer dimensions must be specified for every adapter in every layer, and its optimal size depends on the downstream task. Thus, these adapters cannot be employed where the available storage is limited.

In this work, we propose a training scheme named MiMi (Fig. 1) which addresses these two limitations. Our

¹Code is available: <https://github.com/IemProg/MiMi>

approach facilitates efficient parameter allocation by predominantly assigning additional parameters to layers that genuinely necessitate adaptation to the new task (Fig. 5). More specifically, we start by training adapters with high-dimensional hidden spaces and gradually decrease their dimensionality by identifying neurons that can be omitted in each adapter. Additionally, we introduce a novel scoring criterion to determine the layers where more adaptation is needed, which enables the comparison of a “neuron importance” among adapters in various layers.

Our work makes the following key contributions:

- We propose a novel iterative training scheme for learning small adapters for ViTs.
- We present a new scoring function that can effectively compare the significance of neurons across adapters. This approach enables us to estimate the optimal hidden dimension of adapters for ViTs automatically, which leads to a more efficient parameter allocation.
- Finally, we compare the proposed approach with multiple PET methods designed for both NLP and vision tasks using a total of 29 datasets. From these experiments, we draw several conclusions: (i) we demonstrate that our approach obtains the best performance in terms of accuracy among methods with similar numbers of parameters; (ii) our ablation study validates the positive impact of our adaptive strategy to automatically estimate the hidden dimension of adapters.

2. Related Work

Vision Transformers. Originally designed for NLP tasks, Transformers [42] have recently been adapted for vision tasks, such as image classification. Vision Transformers (ViTs) divide the image into patches, process them as token embeddings, and employ transformer encoders with self-attention to learn image representations [6]. ViTs have shown impressive performance, outperforming ConvNets in some cases [8]. However, their large parameter count leads to significant storage costs, limiting complete finetuning for each new task. This motivates our study. While Swin [24] is a widely adopted ViT due to its excellent performance across vision tasks, our approach of using tiny adapters can be applied to any ViT architecture (see Sec. 3).

Network Pruning. When referred to deep neural networks, pruning consists of reducing the number of parameters of a pre-trained model [5, 9]. It can be roughly categorized into two groups: (i) unstructured pruning, which removes the least significant weights (according to certain criteria like weight magnitude [10] or gradient magnitude [29]) without a specific structure to be followed; (ii) structured pruning, which focuses in removing model sub-structures, like channels [13, 39] or attention heads [28]. Pruning techniques usually reduce the number of parameters in a network trained for a specific task, while MiMi decreases the

number of parameters added through adapters that fit the model to a new task without altering the original model’s parameters. SparseAdapters (SA) [12], show that applying unstructured pruning to adapters [14] achieves comparable performance. In comparison to SA, our method incorporates a look-ahead strategy that considers the effects of up-sampling layers, while SA does not. Furthermore, MiMi employs structured pruning, whereas SA utilizes unstructured pruning techniques to reduce the size of adapters and remove them if necessary. GraSP [43] utilizes Hessian-gradient products for each layer, discarding weights with elevated scores in a single move, emphasizing those that improve gradient flow. Conversely, SNIP [21] determines layer gradients using sampled mini-batch data, assigning scores and eliminating weights with the highest scores in one step. However, both these approaches are not apt for pruning Adapters. Our experiments show that they do not perform well when applied to adapters.

Efficient Transformers Finetuning. ViTs’ lack of typical CNN inductive biases makes their finetuning on new tasks susceptible to overfitting [8, 23]. Additionally, the need to update all the parameters and store a separate model copy per task hinders scalability and real-world applicability. To tackle this, three types of approaches have emerged: (i) updating only newly added parameters [4, 14, 15, 17, 34]; (ii) sparse parameter updates [1, 15]; and (iii) low-rank factorization for weight updates [19]. While prompt techniques like VPT [17] achieve excellent performance, they lack flexibility for downstream tasks that differ significantly from pre-training [4].

Our work falls into the first category, building on adapters [14] for NLP tasks. However, we introduce a specific training algorithm enabling high performance, with small-size adapters for downstream tasks. Unlike previous adapter approaches [4, 14] with fixed-size adapters across layers [4], MiMi dynamically assesses adapter sizes and even removes them if necessary. By minimizing trainable parameters, MiMi enhances performance and reduces storage footprint in multi-task scenarios. Our preliminary results demonstrate that different layers require different adapter sizes, as highlighted in the supplementary material. In contrast, [16, 48] utilize Neural Architecture Search (NAS) to identify optimal PET configurations, facilitating an expansive configuration search space populated by various representative PET methods. However, this approach is notably computation-intensive and uses different PET modules. Our work is primarily concentrated on determining the appropriate adapter size for each layer and eliminating certain adapters when deemed essential.

3. Proposed Method

In this section, we start with the description of adapters [14] and discuss their practical benefits. Then, we

introduce MiMi, our method to estimate the hidden dimension for each adapter that can effectively maintain high performance with fewer parameters for memory efficiency.

3.1. Preliminaries

Our objective is to adapt a pre-trained ViT network for a new task by incorporating small modules called “adapters” into the existing layers. This adaptation process involves training the linear classifier (referred to as the “head”) and the adapter parameters while keeping the weights of the original model frozen. In our training procedure, we focus on describing the adapter parameters, and the linear classifier parameters are also learned simultaneously.

ViT architectures, such as the original ViT [6] or Swin [24], consist of layers with two main sub-layers: a multi-head self-attention (MSA) layer and a fully-connected layer (MLP). Layer normalization is applied before each of these sub-layers, and residual connections are employed to skip MSA and MLP. In our approach, we introduce two adapters after each sub-layer. The adapter is directly applied to the output of the corresponding sub-layer, as depicted in Fig. 2a. The internal structure of the adapters is illustrated in Fig. 2b.

Considering the i -th adapter added to our pre-trained ViT, let $\mathbf{h}_i \in \mathbb{R}^{M_i}$ denote its input, of size M_i . Following [14], adapters employ a first fully-connected layer that down-projects \mathbf{h}_i into $\mathbf{z}_i \in \mathbb{R}^{N_i}$ with some non-linear activation $\phi(\cdot)$. This layer is parametrized by a linear projection matrix $\mathbf{W}_i^{\text{down}} \in \mathbb{R}^{M_i \times N_i}$. Then, a second fully connected layer with parameters $\mathbf{W}_i^{\text{up}} \in \mathbb{R}^{N_i \times M_i}$ up-samples \mathbf{z}_i , producing as output $\mathbf{r}_i \in \mathbb{R}^{M_i}$. Finally, a residual skip-connection is employed inside the adapter module such that, if \mathbf{r}_i is close to zero, the whole adapter module degenerates to an identity function. To summarize, given the input vector \mathbf{h}_i , the output vector \mathbf{h}'_i is calculated as:

$$\mathbf{h}'_i = \mathbf{W}_i^{\text{up}} \cdot \phi(\mathbf{W}_i^{\text{down}} \cdot \mathbf{h}_i) + \mathbf{h}_i. \quad (1)$$

The total number of parameters in the adapter is equal to $2 \cdot N_i \cdot M_i$: since M_i is fixed, we generally choose N_i such that $N_i \ll M_i$ to obtain a low number of parameters. We define the compression rate σ_i of an adapter as $\sigma_i = \frac{M_i}{N_i}$.

Previous works [14, 34, 38] have employed adapters with a uniform hidden dimension N_i for all the adapters. However, this approach may not be optimal as early and late layers to the input of the model may focus on different types of patterns [4, 47] (see supplementary material). If we enable dynamic adjustment of the adapter’s hidden dimension N_i (or equivalently, σ_i) and determine their injection points, we enhance adaptation to downstream tasks effectively.

3.2. Overview of MiMi

Let \mathbf{W}^{ViT} be the initial parameters of the ViT model which are frozen through the whole adaptation process.

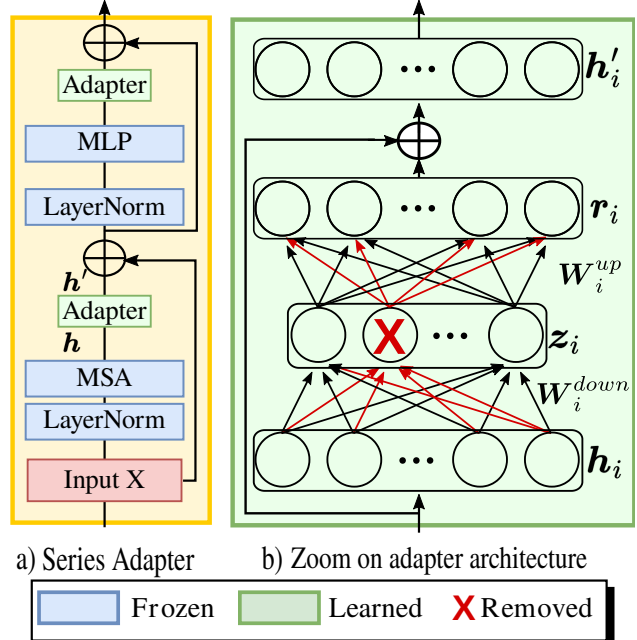


Figure 2. The adapter structure injected into ViT model (a), and our approach to adjust the adapter’s size (b). MSA and MLP are multi-head self-attention and feed-forward blocks, respectively.

Algorithm 1 MiMi

```

1: procedure MiMi( $\mathbf{W}^{\text{ViT}}, \mathbf{W}^{\text{ada}}, \rho, \sigma^{\text{target}}$ )
2:   Learn  $\mathbf{W}^{\text{ada}}$  ▷  $\mathbf{W}^{\text{ViT}}$  is frozen
3:   while  $\sigma < \sigma^{\text{target}}$  do
4:     Sort  $\mathbf{W}^{\text{ada}}$  according to  $\mathcal{I}^{ij}$  (Sec. 3.3)
5:      $\mathbf{W}^{\text{ada}} \leftarrow \text{top}(1-\rho)$  in  $\mathbf{W}^{\text{ada}}$  ▷ Selection
6:     Fine-tune  $\mathbf{W}^{\text{ada}}$  ▷  $\mathbf{W}^{\text{ViT}}$  is frozen
7:   end while
8:   return  $\mathbf{W}^{\text{ada}}$ 
9: end procedure

```

With MiMi our goal is to learn \mathbf{W}^{ada} , the set containing the adapter parameters $\mathbf{W}_i^{\text{ada}} = \mathbf{W}_i^{\text{up}} \cup \mathbf{W}_i^{\text{down}}$ of every i -th ViT sub-layer. In previous works [14, 36] $\mathbf{W}_i^{\text{ada}}$ is straightforwardly learned with stochastic gradient descent-based optimization; however, in our experiments (see Sec. 4) we show that this approach does not perform well in the case of tiny adapters (small N_i values). We start from the observation that, with the existing optimizers, sequentially training and pruning a large network is a successful strategy to find small networks with good performance, while directly training small networks usually suffers from optimization issues [7]. Therefore, we propose to start from large adapters and adopt an iterative pruning strategy that iteratively reduces their dimensions as detailed in Alg. 1. We initialize every adapter with a hidden dimension proportional to its input dimension. We start from compression

rates $\sigma_i = \sigma_0$ for every layer, where σ_0 is the initial compression rate. In our first training stage (line 2), we learn the adapter parameters \mathbf{W}^{ada} via cross-entropy loss minimization using stochastic gradient descent. Then, we estimate a score that measures the importance of each adapter’s neurons (more details will be provided in Sec. 3.3). This score is used to select the **neurons** that have the smallest impact on the adapter outputs; more precisely, we remove the bottom fraction ρ of neurons from \mathbf{W}^{ada} (line 5). The remaining ones will constitute the new adapter configuration, and the hidden space sizes N_i are updated accordingly. If the achieved average compression rate σ is still lower than the target σ^{target} , another compression iteration follows; otherwise, the achieved configuration will be returned and the method stops. Note that the total number of training cycles C is given by:

$$C = \left\lceil \frac{\log(\sigma^0) - \log(\sigma^{\text{target}})}{\log(\rho)} - 1 \right\rceil, \quad (2)$$

where $\lceil \cdot \rceil$ denotes the ceiling function. Therefore, our training scheme stops after a deterministic number of iterations that can be computed in advance. While we employ a stopping criterion based on a specific target compression rate, a target performance on a validation set could also be used.

3.3. Importance Score in MiMi

In this section, we present the importance score function that we use in our training algorithm. Our design of the scoring function is motivated by the observation that, if an entire row in $\mathbf{W}_i^{\text{down}}$ and an entire column in \mathbf{W}_i^{up} are equal to zero, then our adapter is strictly equivalent to one with a smaller dimension M_i . Therefore, we propose a novel scoring function to employ the sum of the L^1 norm of the corresponding row in $\mathbf{W}_i^{\text{down}}$ and the corresponding column in \mathbf{W}_i^{up} . More precisely, our importance score is formulated as follows:

$$\mathcal{I}^{ij} = \frac{1}{N_i + M_i} \left(\sum_{k=1}^{M_i} |W_i^{\text{down}}[j, k]| + \sum_{k=1}^{N_i} |W_i^{\text{up}}[k, j]| \right), \quad (3)$$

where $[\cdot, \cdot]$ denotes the matrix indexing operator. This importance score can be interpreted as a “look-ahead” strategy, where we observe, besides the output of a specific j -th neuron in the hidden space, also the impact of such an output in the next layer. Note that this formulation is based only on the magnitude of parameters belonging to the same neuron of down-sampling, and its corresponding up-sampling one, and not on the magnitude of activations. This makes the importance score more computationally efficient since activation-based scoring depends on the input images, and consequently, statistics should be gathered at the batch or the dataset level, inducing non-negligible computation overhead. Furthermore, this choice is empirically supported by

many works in the literature, like [3, 10, 30, 37]. A noteworthy element is that \mathcal{I}^{ij} is normalized by the total number of parameters associated with a specific dimension of the adapter: this enables fair comparison across adapters, despite different input and hidden layer sizes. More details behind the motivation of our choice for equation 3 are provided in the supplementary material.

4. Experiments

We provide here the details about the datasets and our experimental setup.

Datasets. We evaluate our methods using the protocol adopted by [23], which consists of ten datasets for image classification tasks divided into two benchmarks. The first benchmark is known as *DomainNet* [33]. It contains six different visual domains, which makes the finetuning experiments non-trivial. Since *DomainNet* does not have a labeled testing set, we use the validation dataset for testing, as in [33]. The second benchmark contains CIFAR-10/CIFAR-100 [20], Oxford Flowers102 [32], and SVHN [31], which are widely used as low-regime training datasets. Contrarily to *DomainNet*, these datasets are not single-task oriented but contain a larger variety of domains/tasks. We refer to them as belonging to the *Multi-task* benchmark. Additionally, we provide an evaluation of the *VTAB* benchmark [45], consisting of 19 diverse visual tasks (see supplementary).

Implementation Details. We follow the training protocol adopted by [23]. We conduct our experiments with the official pre-trained Swin-T [24] (~27M parameters) trained on ImageNet-1K. In all our experiments, we use the AdamW [25] optimizer with a cosine decay learning-rate scheduler for 80 epochs, preceded by 20 epochs of linear warm-up. In all the experiments, the images are resized to the same fixed resolution (224 × 224). With MiMi, ρ is set to 50%, namely we half the number of neurons in the adapters, at every 100 epochs.

4.1. Main results

We compare our proposed method MiMi with multiple PETs methods. We remark that all the baselines are obtained with a $C = 5$ cycles training, while MiMi will always have a lower or equal number of training cycles in Tab. 1. We include the following methods:

- *Full finetuning*: finetune all parameters of the model.
- *Att/MLP finetune*: we only tune the Attention/MLP layers and the classification head.
- *Linear-probe*: all parameters are frozen except for the task-specific classification layer.
- *Adapters* [14]: we add adapters with $\sigma = 32$ to have adapters with hidden dimensionality proportional to the input dimension M_i . We also include variants where the size of every adapter is fixed over all the

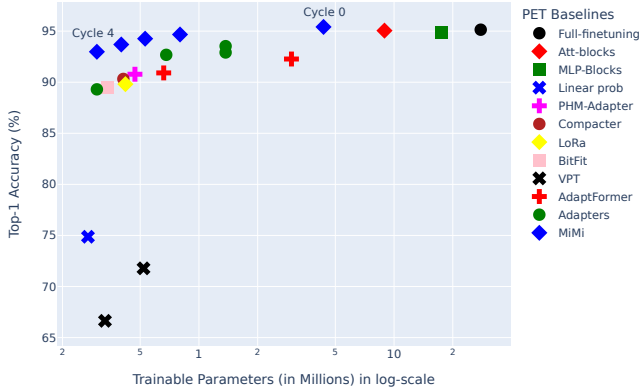


Figure 3. Evaluation of PET baselines mean top-1 accuracy on **Multi-task benchmark**. We observe MiMi (◆) maintains good performance when reducing the number of parameters, compared to other PET baselines.

layers: $N_i = 47$, and $N_i = 23$. These baselines are considered to emphasize the effect of parameter allocation throughout the layers on the final performance.

- *BitFit* [1]: only the biases are finetuned.
- *PHM-Adapter* [46]: the weights of the adapters are learned using parameterized hyper-complex multiplication layers (PHM) layers.
- *Compacter* [19]: adapter weights are learned using shared PHM layers.
- *AdaptFormer* [4]: introduces Adapters, but only after MLP block with a scaling parameter s applied to the output of the injected modules.
- *VPT* [17]: finetuning learnable parameters (i.e. prompts) injected into the embedding space for each layer in ViT.
- *SSF* [22]: aims to adjust the feature activation scaling and shifting its output.
- *Fact-TK* [18]: a tensorization-decomposition method to store the weight updates into a single 3D tensor.

Discussion. Fig. 3 visualizes the average accuracy versus the number of trainable parameters achieved for the Multi-task benchmark, while Table 1 reports the number of trained parameters and the average accuracy across datasets in the MultiTask benchmark. The detailed Tables for DomainNet and VTAB benchmarks are in the supplementary material. For all the benchmarks, the number of trained parameters is reported in millions, and the average top-1 accuracy on the datasets is reported in the rightmost column.

We observe that *full finetuning* achieves commendable performance, albeit demanding an extensive parameter-tuning for each dataset. In comparison, finetuning solely the *attention/MLP* layer proves remarkably effective among the vanilla finetuning baselines. However, this approach still necessitates a substantial number of task-specific pa-

Method	Params (M) ↓	Trained (%) ↓	C100	C10	F.	S.	Mean
Full finetuning	27.8	100	88.13	98.50	97.35	96.59	95.14
Att-blocks	8.93	32.14	88.03	98.41	97.79	95.99	95.05
MLP-blocks	17.54	63.12	88.44	98.47	96.50	96.14	94.89
MiMi (0 cycle) [†]	4.35	15.81	88.27	98.53	97.59	97.28	95.41
AdaptFormer-64	0.66	2.38	83.79	96.93	90.50	92.45	90.91
AdaptFormer-256	2.98	8.55	84.74	97.23	92.13	94.97	92.27
Adapters $N_i = 47$	1.37	4.90	85.04	97.52	92.72	96.35	92.91
Adapters $N_i = 23$	0.68	2.47	85.18	97.57	92.16	95.81	92.68
Adapters $\sigma = 32$	1.37	4.90	85.59	97.49	94.80	96.27	93.53
Linear prob	0.27	0.95	75.58	91.84	76.80	55.26	74.87
PHM-Adapter	0.47	1.72	84.17	96.48	89.18	93.32	90.78
Compacter	0.41	1.44	83.95	96.26	88.43	92.67	90.32
BitFit	0.34	1.22	83.56	96.14	87.85	90.29	89.46
VPT-deep (10 tokens)	0.33	1.20	67.69	90.99	22.77	85.11	66.64
VPT-deep (100 tokens)	0.52	1.88	72.53	93.03	34.88	86.70	71.78
Adapters $N_i = 1$	0.30	1.07	82.60	96.03	89.77	88.80	89.30
SSF	0.28	0.96	83.02	96.46	95.59	95.11	92.54
Fact-TK ₃₂	0.33	1.18	82.91	96.59	87.46	90.84	89.45
MiMi (1 cycle)	0.80	2.89	87.12	97.98	96.59	96.98	94.67
MiMi (2 cycles)	0.53	1.92	86.33	97.49	96.73	96.48	94.26
MiMi (3 cycles)	0.40	1.43	85.22	97.11	96.81	95.60	93.69
MiMi (4 cycles)	0.30	1.07	84.07	97.11	96.81	93.94	92.98

Table 1. Results on the **Multi-task benchmark**. C100, C10, F and S stand for CIFAR100, CIFAR10, Flowers, and SVHN. [†] is equivalent to Adapters with $\sigma = 8$. Methods are grouped according to the relative number of trainable parameters ($\leq 2\%$, $\in [2, 10]\%$, $\geq 10\%$)

rameters, unlike other PET approaches. Notably, the underwhelming performance of *linear probing* emphasizes the significance of altering the feature representations within the model when adapting to new tasks.

Notably, both *PHM* and *Compacter* demonstrate their effectiveness by achieving impressive performance while adjusting less than 2% of the parameters. Unlike NLP tasks where PETs have shown success with a small number of trainable parameters [14], visual tasks do not attain *full finetuning* performance with such limited parameter adjustments. Additionally, the subpar performance of *VPT* indicates that injecting tokens into the embedding space offers minimal benefit when the pre-training dataset differs from the downstream task. Remarkably, all PET methods consistently maintain similar performance rankings across all tasks, suggesting that the optimal adaptation strategy is independent on the specific downstream task.

Adapters achieve impressive results with a slightly higher number of trainable parameters (1.37M, 4.90% of the total) for $\sigma = 32$. Remarkably, *Adapters* outperform *AdaptFormer* [4] while utilizing fewer parameters (92.91% with 1.37M parameters compared to 92.27% with 2.98M parameters). This outcome highlights the superiority of adapting representations after both MSA and MLP blocks, as demonstrated in the architecture of *Adapters* (Fig. 2), over solely acting on the MLP block, as in done *AdaptFormer*.

We observe that *MiMi* significantly reduces the number of parameters by 4 times the initial size (0.40M, 1.43%)

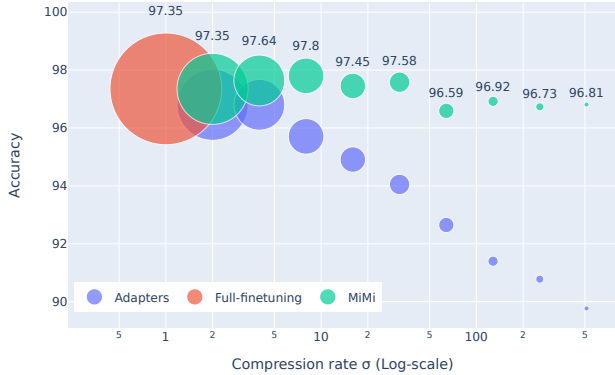


Figure 4. Comparison of top-1 accuracy versus compression rate σ on VGG-Flowers. All MiMi results originate from the same MiMi run. Adapters are trained for the exact same number of epochs as their MiMi counterparts. The size of blob markers represents the number of trainable parameters.

while outperforming all PET methods in the Multi-task benchmark. In particular, MiMi outperforms *adapters- $n_i = 47$* despite having fewer parameters, demonstrating that our iterative training procedure improves the parameter efficiency. To further emphasize the performance gap between the two approaches, we introduce Fig. 4 illustrating the performance as a function of the number of trainable parameters for VGG-Flowers (for CIFAR-100 dataset in supplementary). We observe the significant performance gap between vanilla adapters compared to adapters trained with the MiMi approach.

Furthermore, MiMi outperforms methods with similar trained parameters, in all the compression ranges. In particular, in the most challenging one (with 0.30M parameters), MiMi outperforms the closest approach, BitFit, which trains 0.34M parameters, showing a gain in average accuracy larger than 3% and 2%, for Multi-Task and DomainNet benchmarks, respectively.

Upon comparing *adapter* with uniform and proportional parameter distribution (N_i vs σ), results are in favor of allocating parameters proportionally to the layer dimension. Notably, adapters with $\sigma = 32$ outperform adapters with $N_i = 47 \forall i$ in both the Multi-task (93.53% vs 92.91%) and DomainNet (70.65% vs 69.39%) benchmarks. This observation suggests that the task-specific nature of the last layers, with higher dimensionality, necessitates greater adaptation. Furthermore, we demonstrate that reducing the size of adapters ($N_i = 23$) negatively affects performance, with a marginal drop in Multi-task (0.23%) but a more consistent decrease in DomainNet (1.01%). These findings underscore *the unsatisfactory performance obtained from training adapters in a vanilla fashion and serve as motivation for our specific training procedure.*

MiMi versus Vanilla training. Looking at the Multi-task benchmark (Fig. 3, Table 1), we observe that *MiMi* signif-

icantly reduces the number of parameters by $4\times$ (0.40M, 1.43%) while outperforming all PET methods in the Multi-task benchmark. In particular, *MiMi* outperforms *adapters- $N_i = 47$* despite having fewer parameters, demonstrating that our iterative training procedure improves the parameter efficiency. To further emphasize the performance gap between the two approaches, we introduce Fig. 4, we observe *the significant performance gap between vanilla adapters compared to adapters trained with MiMi approach.*

4.2. Ablation study

Importance score for MiMi. Next, we move on to our design choice of dimensionality reduction inside adapters throughout the adaptation cycles. We report the contribution of various components of MiMi with different setups.

- *Vanilla Adapters*: corresponds to injecting adapters with a compression rate σ .
- *Random*: we select randomly a percentage of neurons for each adapter to be removed.
- *Gradient-Based $L^1(\nabla)$* : We determine the neurons to be removed based on the L^1 norm of the gradients.
- *Local neuron selection*: We uniformly choose a percentage of neurons to be removed, independently applied to both down-sampling and up-sampling layers.
- *Global neuron selection*: The number of neurons to be removed per adapter is determined using equation 3 given ρ , considering the scaling factor if applicable. Additionally, we assess our scoring function without the inclusion of the scaling factor $n_i + m_i$. This modified version of our score is referred to as \mathcal{I}_0 .
- *MiMi*: our method as in Alg. 1.

To compare the different methods we proceed as follows. When using an iterative method, we always start from the baseline model where $\sigma = 32$. When using a non-iterative method: we start with adapters of $\sigma_0 = \sigma_{target}/(1 - \rho)$ and prune once only after the first cycle. Training continues for $C - 1$ cycles to guarantee fair comparison with iterative methods. Results are reported in Table 2.

Discussion. Table 2 summarizes the performance of local and global neuron selection for adapters. Firstly, we observe that reducing the number of parameters in vanilla adapters (higher values of σ in Fig. 4) leads to a drop in performance. Additionally, we find that using the magnitude of parameters instead of activations is advantageous. Activation-based scoring methods rely on input images, requiring batch or dataset-level statistics, which are computationally less efficient.

Secondly, global neuron selection proves to be superior to local neuron selection. The former method focuses on finetuning adapters in specific layers while completely removing other adapters, while the latter removes the same amount of neurons from each adapter’s layer. Moreover, MiMi surpasses SA by employing structured pruning (neu-

Method	Selection	Score	Iter.	Scale	σ				
					32	64	128	256	512
Vanilla	-	-	-	-	94.80	90.12	89.42	88.85	86.03
	Random	-	✓	-	95.43	95.80	95.11	95.12	
	Local (DW)	$L^1(w)$		-	95.46	95.06	94.28	93.79	
	Local	$L^1(\nabla)$	✓	-	95.46	96.17	96.11	96.42	
	Local	SA [12]		-	96.10	95.57	96.15	96.23	
	Local	SA [12]	✓	-	96.41	96.65	96.72	96.72	
Baseline	Local	GraSP [43]		-	90.62	89.71	87.22	86.66	
	Local	SNIP [21]		-	93.53	92.39	91.36	90.62	
	Global	$L^1(a)$		✓	96.10	94.28	93.80	93.25	
	Global	$L^1(a)$	✓	✓	96.13	95.15	95.77	95.72	
	Global	\mathcal{I}_0		-	94.88	95.28	95.66	95.45	
	Global	\mathcal{I}		✓	96.10	95.82	96.34	96.50	
MiMi	Global	\mathcal{I}	✓	✓	96.59	96.92	96.73	96.81	

Table 2. Performance analysis for neuron selection on VGG-Flowers. $L^1(w)$ and $L^1(a)$ denote the magnitude pruning of the parameters and the activations respectively. Local (DW) represents local pruning applied to down-sampling layers only.

ron removal) instead of unstructured pruning (weight removal), to reduce the size of adapters. Additionally, MiMi incorporates a look-ahead strategy that accounts for the impact of up-sampling layers, ensuring consistently high performance. Notably, with MiMi, adapter size can be reduced for efficient computation, unlike SA.

In the final cycles, MiMi identifies crucial adapters for adaptation, prioritizing their finetuning. This approach improves model performance by adjusting specific latent representations tailored to the downstream task while using fewer parameters. MiMi consistently outperforms both GraSP and SNIP across all σ values due to its iterative pruning approach. Pruning at initialization as done in SNIP/GraSP, before the adapters have been trained are less effective. Since the initialization is random, they are missing out on retaining potentially important weights.

Table 2 reveals that MiMi achieves superior performance compared to vanilla adapter training on the VGG-Flowers dataset, with a performance gap of 6.14%, when using $\sigma = 64$ (regardless of local/global neuron selection). Notably, this performance gap increases as we reduce the adapter size to $\sigma = 256, 512$. Furthermore, when comparing to a vanilla L^1 importance scoring, we observe the benefits of considering both down-sampling and up-sampling parameters for the adapters. This approach consistently improves performance across compression rates ranging from 0.5% to over 3%. Notably, the performance gap becomes more prominent at higher compression rates.

Finally, scaling the importance score according to equation 3 enhances the performance of the *Global* method by approximately 1% across all σ values.

Parameter allocation analysis with MiMi. In Fig. 5, we visualize the distribution of removed and remaining neurons achieved through the application of MiMi on VGG-Flowers and CIFAR-10. Notably, this illustration highlights

the contrasting outcomes between local neuron selection, which uniformly removes neurons from adapters, and the utilization of global neuron selection. Remarkably, we observe that the latter approach completely removes certain adapters from the model (evident in layers 4, 5, 7, and 8 of VGG-Flowers), while redistributing a significant number of parameters to other adapters.

Moreover, *global neuron selection* exhibits distinct adaptations for each dataset, as evidenced in Fig. 5. Notably, the distribution of removed neurons varies between CIFAR-10 and VGG-Flowers. In the case of CIFAR-10, fewer adapters are completely removed compared to VGG-Flowers. Conversely, for VGG-Flowers, only adapters at later stages are retained, suggesting that early layer representations are well-suited for this particular dataset. However, for CIFAR-10, the remaining adapters are dispersed throughout all layers of the ViT model, indicating that almost all the layers’ representations need to be finetuned. These observations highlight the adaptability and dataset-specific optimizations achieved through *global neuron selection*. To provide a more comprehensive analysis, we also present normalized plots in supplementary material.

ViT variants with MiMi. We evaluate the performance of MiMi on different ViT backbones, including Swin-S/L (~50M/~197M parameters), ViT [6] (~86M parameters), and CVT [44] (~20M parameters).

For three training cycles, we compare the three baselines: *finetuning*, *adapters*, and *MiMi*. Table 3 presents the best scores achieved in the final cycle. Remarkably, MiMi achieves comparable performance to full model finetuning, with a margin of 1.2%, 1.2%, and 1.4% for ViT-B/16, Swin-S, and CvT, respectively. This is accomplished by finetuning less than 1.5% of the parameters, including the head classifier. MiMi surpasses vanilla adapters’ performance with four times fewer parameters across all ViT backbones (ViT, Swin-T, and CvT). These experiments demonstrate the generalizability of MiMi to various ViT backbones.

Evaluating Inference Cost/Storage Footprint. In this section, we conduct a comprehensive analysis of the GFLOPS at inference time and storage footprint for PETs methods in the context of multi-task learning. Table 4 presents the findings, including the number of trainable parameters and the storage requirement in MegaBytes (MB) for saving the Swin-T model after finetuning per task T .

Storing a complete ViT for each task can impose significant burdens on storage space and computational resources. With ViTs containing millions of parameters, these storage requirements quickly accumulate in multi-task settings. However, by storing only a subset of the model’s parameters, both storage costs and computational resources for training can be significantly reduced.

Larger model components, like attention and MLP blocks, demand significant storage due to their extensive

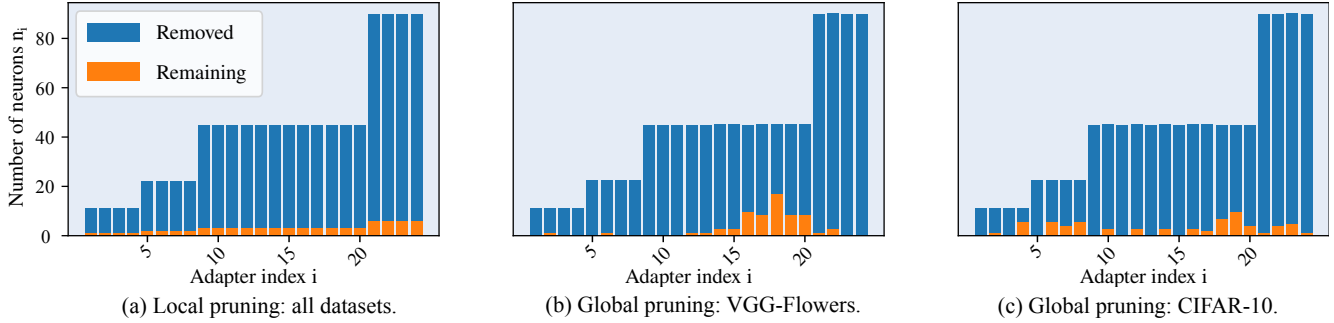


Figure 5. Layer-wise analysis of adapter’s neurons distribution at 4^{th} cycle. Bar plots represent the number of neurons n_i at each adapter i for VGG-Flowers and CIFAR-10, respectively. Global neuron selection leads to different neuron distributions depending on the dataset. Compared to VGG-Flowers, fewer adapters are completely removed on CIFAR-10.

	Method	# Params (M) ↓	Trained (%) ↓	C100	C10	V	S	Mean ↑
ViT-B-16	Finetune	85.90	100	91.22	99.01	99.32	97.68	96.81
	Adapters	0.96	0.89	89.39	98.02	97.69	94.17	94.82
	MiMi	0.62	0.54	89.86	98.09	98.75	94.94	95.41
	MiMi	0.37	0.32	89.84	98.17	98.85	95.32	<u>95.55</u>
Swin-S	Finetune	48.80	100	90.12	98.88	98.37	98.16	96.38
	Adapters	0.41	4.88	89.05	98.48	94.60	97.25	94.84
	MiMi	0.23	2.75	88.86	98.53	96.16	97.22	95.19
	MiMi	0.11	1.32	88.62	98.50	96.68	96.94	<u>95.18</u>
Swin-L	Finetune	197M	100%	95.12	99.34	99.67	98.22	98.08
	Adapters	20.1M	10.2%	94.31	99.46	99.76	97.98	<u>97.88</u>
	MiMi	10.9M	5.53%	94.78	99.44	99.51	99.77	98.38
	MiMi	6M	3.04%	92.92	99.30	99.74	97.96	97.48
CVT	Finetune	19.65	100	90.01	98.68	97.98	98.09	96.19
	Adapters	0.78	4.00	86.68	97.91	88.93	96.96	92.62
	MiMi	0.47	2.40	86.47	97.98	93.28	97.17	<u>93.73</u>
	MiMi	0.28	1.44	85.87	97.77	94.31	96.67	93.66

Table 3. Performance of our method using different ViT backbones on the Multi-task benchmark. The highest score is in bold and the second best score is underlined. C100, C10, F, and S stand for CIFAR100, CIFAR10, Flowers, and SVHN datasets.

Method	# Params (M) ↓	Storage (MB) ↓	GFLOPS ↓	Accuracy (%) ↑
Full-finetuning	27.8	111	8.72	97.35
Att-block	8.93	34.7	8.72	97.79
MLP-blocks	17.54	73.4	8.72	96.50
Full-model W/ Adapter ($\sigma = 32$)	1.37	115.3	9.06	96.27
BitFit	0.34	0.34	8.72	87.85
VPT-Deep (100 tokens)	0.32	160.1	18.40	34.88
AdaptFormer-64	0.84	1.63	9.08	90.50
SSF	0.28	0.96	8.72	95.59
Fact-TK ₃₂	0.33	1.18	10.6	87.46
Adapters ($\sigma = 32$)	1.37	4.30	9.06	96.27
Adapters ($n_i = 47$)	1.37	4.40	9.26	92.72
Adapters ($n_i = 1$)	0.30	0.47	8.74	89.77
Linear-prob	0.27	0.31	8.72	76.80
MiMi (3 cycles)	0.40	0.63	8.92	96.81
MiMi (4 cycles)	0.30	0.47	8.82	96.81

Table 4. Comparison with different PET methods on VGG-Flowers with respect to inference cost (GFLOPs).

trainable parameters. On the other hand, finetuning only the head (*Linear-prob*) is lightweight but comes at the cost of

performance. Notably, MiMi achieves a +6% performance enhancement compared to Adapters $n_i = 1$ while maintaining similar storage needs. However, MiMi offers both robust performance and reduced memory demands, positioning it as a superior alternative.

In terms of GFLOPS during inference, full fine-tuning, and similar variants, such as Att-block and MLP-blocks, achieve the lowest GFLOP values at 8.72. However, they come at the expense of a high memory footprint. On the other hand, VPT-Deep (100 tokens) stands out with the highest GFLOPS at 18.40, thanks to an increase in the embedding space for each layer to 100 tokens. This emphasizes that fewer parameters do not necessarily guarantee computational efficiency. MiMi in its 3-cycle and 4-cycle variants, achieves GFLOPS values of 8.92 and 8.82, respectively. This efficiency is attributed to our method’s ability to completely remove some adapters, effectively reducing the computational cost during inference.

5. Conclusion

In this work, we propose MiMi, a training algorithm to learn small adapters for the problem of ViT efficient finetuning. Rather than directly training adapters with few parameters, we propose to start with large adapters, and then iteratively select the more important neurons in every adapter. Our training procedure estimates the hidden dimension for each adapter, reducing the number of trainable parameters and even removing adapters if unnecessary. We empirically demonstrate the greater performance of MiMi to adapters and show that our method achieves excellent performance with low numbers of trainable parameters. Our ablation study validates the positive impact of our novel importance score to estimate the hidden dimension of each adapter.

Acknowledgements. This paper has been supported by the French National Research Agency (ANR) in the framework of its JCJC. Furthermore, this research was partially funded by Hi!PARIS Center on Data Analytics and Artificial Intelligence.

References

- [1] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bit-Fit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Dublin, Ireland, May 2022. Association for Computational Linguistics. 2, 5
- [2] Rodrigo Berriel, Stéphane Lathuilière, Moin Nabi, Tassilo Klein, Thiago Oliveira-Santos, Nicu Sebe, and Elisa Ricci. Budget-aware adapters for multi-domain learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 1
- [3] Yves Chauvin. A back-propagation algorithm with optimal use of hidden units. *Advances in neural information processing systems*, 1, 1988. 4
- [4] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition, 2022. 1, 2, 3, 5
- [5] Yann Le Cun, John S. Denker, and Sara A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*. Morgan Kaufmann, 1990. 2
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 7
- [7] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2018. 3
- [8] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [9] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2
- [10] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015. 2, 4
- [11] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2021. 1
- [12] Shwai He, Liang Ding, Daize Dong, Miao Zhang, and Dacheng Tao. Sparseadapter: An easy approach for improving the parameter-efficiency of adapters, 2022. 2, 7
- [13] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, 2017. 2
- [14] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*. PMLR, 2019. 1, 2, 3, 4, 5
- [15] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 1, 2
- [16] Shengding Hu, Zhen Zhang, Ning Ding, Yadao Wang, Yasheng Wang, Zhiyuan Liu, and Maosong Sun. Sparse structure search for parameter-efficient tuning, 2022. 2
- [17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning, 2022. 1, 2, 5
- [18] Shibo Jie and Zhi-Hong Deng. Fact: Factor-tuning for lightweight adaptation on vision transformer, 2023. 5
- [19] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 5
- [20] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). 4
- [21] Namhoon Lee, Thalayasingam Ajanthan, and Philip H. S. Torr. Snip: Single-shot network pruning based on connection sensitivity, 2018. 2, 7
- [22] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling and shifting your features: A new baseline for efficient model tuning, 2023. 5
- [23] Yuhui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco De Nadai. Efficient training of visual transformers with small datasets. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 2, 4
- [24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1, 2, 3, 4
- [25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 4
- [26] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *ACL/IJCNLP*, 2021. 1
- [27] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1
- [28] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019. 2
- [29] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network

- pruning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [2](#)
- [30] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, volume abs/1608.08710. OpenReview.net, 2017. [4](#)
- [31] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bisacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. [4](#)
- [32] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. [4](#)
- [33] Yingwei Pan, Yehao Li, Qi Cai, Yang Chen, and Ting Yao. Multi-source domain adaptation and semi-supervised domain adaptation with focus on visual domain adaptation challenge 2019, 2019. [4](#)
- [34] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online, Apr. 2021. Association for Computational Linguistics. [1](#), [2](#), [3](#)
- [35] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017. [1](#)
- [36] Sylvestre-Alvise Rebuffi, Andrea Vedaldi, and Hakan Bilen. Efficient parametrization of multi-domain deep neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. [1](#), [3](#)
- [37] Alex Renda, Jonathan Frankle, and Michael Carbin. Comparing rewinding and fine-tuning in neural network pruning. In *International Conference on Learning Representations*, 2020. [4](#)
- [38] Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. Adapterdrop: On the efficiency of adapters in transformers. In *EMNLP (1)*, 2021. [3](#)
- [39] Enzo Tartaglione, Skjalg Lepsøy, Attilio Fiandrotti, and Gianluca Francini. Learning sparse neural networks via sensitivity-driven regularization. *Advances in neural information processing systems*, 31, 2018. [2](#)
- [40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 2021. [1](#)
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. [1](#)
- [42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. [2](#)
- [43] Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow, 2020. [2](#), [7](#)
- [44] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. [7](#)
- [45] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruysen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. [4](#)
- [46] Aston Zhang, Yi Tay, SHUAI Zhang, Alvin Chan, Anh Tuan Luu, Siu Hui, and Jie Fu. Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n$ parameters. In *International Conference on Learning Representations*, 2021. [5](#)
- [47] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Michael C. Mozer, and Yoram Singer. Identity crisis: Memorization and generalization under extreme overparameterization. In *International Conference on Learning Representations*, 2020. [3](#)
- [48] Han Zhou, Xingchen Wan, Ivan Vulić, and Anna Korhonen. Autopeft: Automatic configuration search for parameter-efficient fine-tuning, 2023. [2](#)