CONSOLIDATOR: MERGEABLE ADAPTER WITH GROUPED CONNECTIONS FOR VISION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, transformers have shown strong ability as visual feature extractors, surpassing traditional convolution-based models in various scenarios. However, the success of vision transformers largely owes to their capacity of accommodating numerous parameters. As a result, new challenges for adapting a well-trained transformer to downstream tasks arise. On resource-limited devices, classic fine-tuning, which tunes and stores a full copy of parameters in the pretrained model for every downstream task, is usually impracticable for the shortage of storage space. However, few works have focused on how to efficiently and effectively transfer the knowledge in a vision transformer. Existing methods did not dive into the properties of visual features, leading to inferior performance. Moreover, some of them bring heavy inference cost though benefiting storage. To tackle these problems, we propose consolidator to achieve efficient transfer learning for vision transformers. Our consolidator modifies the pretrained model with the addition of a small set of tunable parameters to temporarily store the task-specific knowledge while freezing the backbone model during adaptation. Motivated by the success of group-wise convolution, we adopt grouped connections across the features extracted by transformer blocks to construct tunable parts in a consolidator. To further enhance the model capacity to transfer knowledge under constrained storage budget and keep inference efficient, we consolidate the parameters in two stages: 1. between adaptation and storage 2. between loading and inference. On a series of downstream visual tasks, our consolidator can reach better performance than full fine-tuning with less than 0.5% parameters stored per task, and outperform state-of-the-art parameter-efficient tuning methods.

1 INTRODUCTION

Recently, transformer architectures originated from natural language processing (NLP) (Vaswani et al., 2017) demonstrate considerable capacity in computer vision (Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021b). Vision transformers, along with traditional convolutional neural networks (CNNs) (Krizhevsky et al., 2012; He et al., 2016; Simonyan & Zisserman, 2014), are widely used as feature extractors to generate strong and general visual representations via deriving knowledge from massive images. Thanks to the abundant information in such visual representations, we can rapidly and effectively adapt the pretrained models to downstream tasks by a simple fine-tuning strategy.

Although fine-tuning can reach admirable performance, it inflicts heavy storage burdens. Since fine-tuning intensively adjusts all parameters in the pretrained model, it maintains a full copy of the model's parameters for each task when used for various downstream tasks. Therefore, although simple, fine-tuning can cause a huge storage burden, resulting in impracticality in real world scenarios, especially in resource-constrained situations, *e.g.*, embedded systems. This problem would become more severe due to the increasingly explosive scale of vision transformers.

Efforts have been made to reduce the storage overhead of full fine-tuning for transformers. For example, adapter (Houlsby et al., 2019; Karimi Mahabadi et al., 2021), prompt tuning (Li & Liang, 2021; Lester et al., 2021; Zhou et al., 2021) and LoRA (Hu et al., 2021) inject tunable parameters

and freeze the backbone during adaptation. In vision field, VPT (Jia et al., 2022) directly leverage learnable prompts for vision transformers, and NOAH (Zhang et al., 2022) searches for the optimal combinations of the three representative modules, *i.e.*, adapter, LoRA and VPT. Despite their success, existing methods suffer from two common conflicts: 1. trade-off between the inference efficiency and the adaptation performance 2. trade-off between the adaptation performance and the quantity of stored parameters. Previous works (Houlsby et al., 2019) show that introducing more tunable parameters can achieve more fruitful results. However, extra parameters can bring significantly larger computation and storage cost, resulting in low inference efficiency and more storage space. Therefore, one essential question is raised: *can we design a module that can share the same inference cost as an ordinary model while enjoying superior capacity against existing methods*?

In this paper, we propose a generic module, dubbed consolidator, to tackle the aforementioned issue. The proposed consolidator is designed as a mergeable adapter which accompanies the fully connected (FC) layer in the transformer blocks. Specifically, to enrich the model capacity under limited parameter budget, we take inspirations from the success of group-wise convolution (Howard et al., 2017; Ma et al., 2018; Liu et al., 2022) and build our consolidator as grouped connected (GC) layers. To enhance the flexibility, we further reorder channels for each group connections, following by a droppath regularizer. Benefited from the inference-time linearity of GC, channel reorder and droppath operations, the proposed consolidator can be perfectly consolidated into the original FC layer of a vision transformer, leading to no extra inference cost.

Our consolidator can be easily expanded as a multi-branch topology without breaking the linearity. Practically, we can simultaneously equip several GC layers with channel reordering for communications between different groups of feature channels. After adaptation, we can first consolidate the multi-branch GC layers into one single sparse parameter matrix and store the sparse matrix for each task. Such property can enhance the model's transferability and achieve a considerable storage reduction when the number of tasks scales up. During inference, such sparse parameter matrix can be merged into the backbone model as well, resulting in no inference cost. Thanks to the twice consolidation, the proposed consolidator can greatly promote the efficient and effective visual adaptation.

To verify the superiority of consolidator, we conduct extensive experiments and analysis on a series downstream recognition tasks. Experimental results show that our consolidator can surpass full fine-tuning by 6.96% top-1 accuracy with merely 0.35% learnable parameters per task. Compared to state-of-the-art methods, such as adapter, LoRA and NOAH, our method can consistently reach better performance while enjoying no inference cost. On other fundamental visual tasks, *i.e.*, object detection and semantic segmentation, our consolidator shows great power as well.

Overall, we summarize our contributions as follows. (i) We propose a basic module, dubbed consolidator, for efficient transfer learning. To enhance the transferability under limited tunable parameters, our consolidator is designed as a mergeable grouped connected (GC) layer with a channel reorder and a droppath regularizer. We extend the single branch to a multi-branch topology for better flexibility and transferability. (ii) We design a two-stage consolidation process by merging corresponding parameters in training-storage phase and loading-inference phase. In this way, we can maximally dig the adaptation capacity of the model under a constrained storage budget, with no extra inference cost. (iii) We conduct extensive experiments and analysis on various downstream tasks. Results show that the proposed consolidator method can consistently outperform state-of-the-art methods with fewer stored parameters but superior performance.

2 Related Works

Parameter-efficient transfer learning. In the vision field, there are successful explorations (Rebuffi et al., 2017; 2018) towards parameter-efficient transfer learning of CNNs. However, due to the disparity in parametric form of transformer and CNN, it is infeasible to apply such methods to transformer-based architectures. In the language field, a series of methods (Houlsby et al., 2019; Pfeiffer et al., 2021; Li & Liang, 2021; Lester et al., 2021; Zaken et al., 2021; Hu et al., 2021; Karimi Mahabadi et al., 2021; Liu et al., 2021a; Ding et al., 2022a) have also been proposed to efficiently transfer the knowledge of pretrained transformers to downstream language tasks. However, relative few people focused on how to adapt vision transformers efficiently, on which VPT (Jia et al., 2022) and NOAH (Zhang et al., 2022) are two explorations. VPT directly apply prompt-tuning (Li & Liang, 2021; Liu et al., 2021a) for vision transformers. NOAH first trains a large supernet with three

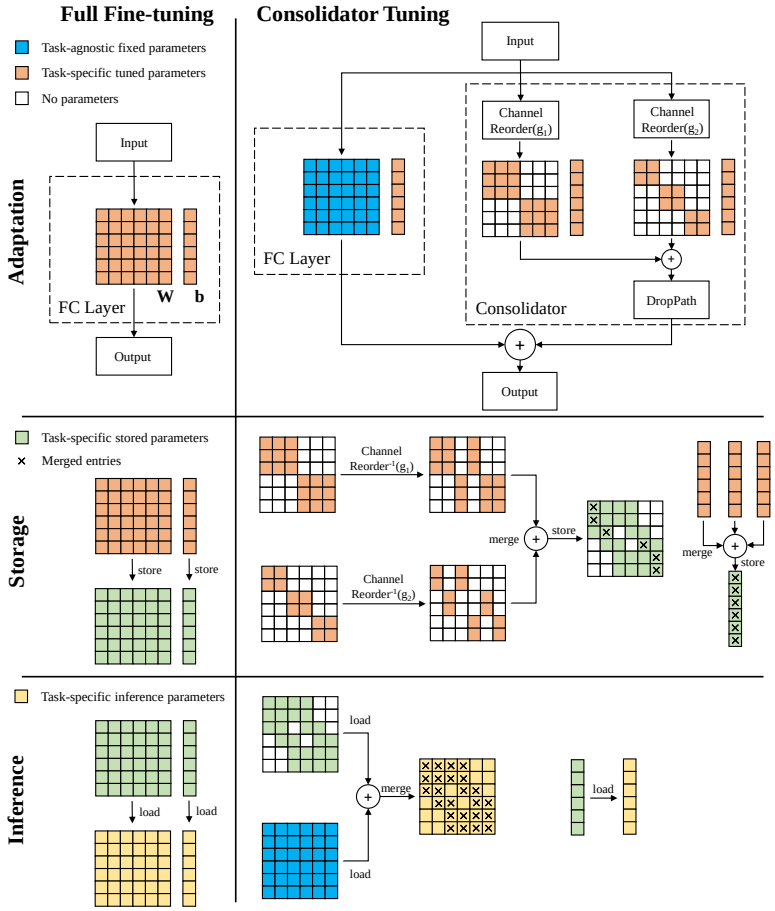

Figure 1: Consolidator tuning versus full fine-tuning. Consolidator adds tunable multi-branch grouped connected layers to the original fully connect layers. The tunable parameters are merged via addition into one single sparse matrix before storage to reduce the needed storage space. Between loading and inference, the parameters in the sparse matrix will be merged back into the original fully connected layer. Consolidator greatly enlarge the model's adaptation capacity under a constrained storage budget with no extra inference cost. **Best viewed in color.**

modules, VPT, LoRA and adapter, and then searches for the optimal configurations of each module for every transformer block using evolution algorithm (Chen et al., 2021). They also shows that classic methods such as LoRA and adapter can also lead to good performance for vision transformers. However, existing methods suffer from the two trade-offs as we discussed in Section 1, resulting in difficulties to fully dig the adaptation capacity of vision transformers efficiently. To solve the problems, we present consolidator. We design a two-stage consolidation process to perfectly balance the trade-offs, leading to efficient and effective visual adaptation.

Inference-efficient structures. Many workss (Ding et al., 2019; Guo et al., 2020; Ding et al., 2021b;c;a; 2022b) strive to design a generic convolution architecture to realize superior capacity while enjoying no inference cost. For example, RepVGG (Ding et al., 2021c) integrates extra 1x1 convolution to strengthen the main 3x3 convolution. However, all these methods are designed for CNNs, and how to effectively enlarge capacity of a vision transformer remains unclear. LoRA (Hu et al., 2021) offers a possible solution by adding parallel low-rank branches for FC layers in transformers, and receive impressive results on language tasks. In this paper, we adopt parallel GC layers to replace the functionality of the original FC layers in vision transformers, which shows strong ability for visual adaptation. Furthermore, we expand the existing one-stage training-inference consolidation to a two-stage process: 1. training-storage consolidation 2. loading-inference consolidation. Such two-stage design can maximally dig the adaptation capacity of the pretrained transformer under a constrained storage budget, with no extra inference cost.

3 METHODOLOGY

3.1 PRELIMINARIES

A typical vision transformer (Dosovitskiy et al., 2020) consists of L serial blocks. In each encoder, there are a multi-head self attention module (MHSA) and a multi-layer percepertron (MLP). Formally, a batch of input images $\mathbf{x}_{input} \in \mathbb{R}^{B\times 3\times H\times W}$ will be first reshaped into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{B\times N\times (P^2\cdot C)}$, where C is the number of channels and (P, P) is the resolution of each patch, and $N = NW/P^2$ is the number of patches. Then the patches are mapped to D channel dimensions with a linear projection. Next, a cls token is appended and we can get $\mathbf{x}_1 \in \mathbb{R}^{B\times (N+1)\times D}$. Here we use $\mathbf{x}_l \in \mathbb{R}^{B\times (N+1)\times D}$ to denote the input of l-th ($1 \le l \le L$) block. Its output $\mathbf{x}_{l+1} = \mathbf{x}'_l + \mathrm{MLP}(\mathrm{LayerNorm}(\mathbf{x}'_l))$ where $\mathbf{x}'_l = \mathbf{x}_l + \mathrm{MHSA}(\mathrm{LayerNorm}(\mathbf{x}_l))$.

The basic components in the forward process of MHSA and MLP are FC layers. Features are processed by its subsequent FC layer, and transferred to the next layer. To achieve parameter-efficient transfer, we follow LoRA (Hu et al., 2021) and formulate our consolidator for FC layer (see Fig. 1). We demonstrate that such design can realize both parameter efficiency and performance effectiveness in Section 4.

3.2 CONSOLIDATOR

For efficient transfer learning, we merely tune and store the parameters in consolidators while freezing other parameters in the pretrained model. In this subsection, we will introduce our design of consolidator, an efficient and effective module for adapting vision transformers, in detail.

Grouped connections. Inspired by the success of group convolution in extracting visual features, we therefore assume that the cross-channel information exchange is redundant in visual adaptation, and aim ti design consolidator by reducing the cross-channel connections between sequential features to minimize the number of stored parameters for downstream tasks while keeping maximum capacity. Therefore, for each FC layer, we add a concurrent module consisting of grouped connected layer. Formally, for an input $\mathbf{x} \in \mathbb{R}^D$, the output $\mathbf{x}' \in \mathbb{R}^E$ of a GC layer with group g, weight $\mathbf{W} \in \mathbb{R}^{g\times \frac{E}{g}\times \frac{D}{g}}$ and bias $\mathbf{b} \in \mathbb{R}^E$ is formulated by $\mathbf{x}' = \mathrm{GC}(\mathbf{x}) = \sum_{j=1}^{g} \mathrm{Pad}(\mathbf{W}_j \mathbf{x}_{\frac{(j-1)D}{g}:\frac{jD}{g}}, j) + \mathbf{b}$.

Here $\mathrm{Pad}(\mathbf{z}, j)$ prepends $\frac{(j-1)D}{g}$ zeros and appends $\frac{(g-j)D}{g}$ zeros for $\mathbf{z} \in \mathbb{R}^{\frac{E}{g}}$ according to another input j. In this way, the output channels in j-th group only interact with the input channels in j-th group, and thus we reduce the cross-channel connections as expected.

To flexibly reach different ratios of stored parameters, we adopt a multi-branch topology in our consolidator. There is a GC layer with weight $\mathbf{W}^{(i)} \in \mathbb{R}^{g^{(i)}\times \frac{E}{g^{(i)}}\times \frac{D}{g^{(i)}}}$ and bias $\mathbf{b}_i \in \mathbb{R}^E$ for i-th branch with group $g^{(i)}$. During adaptation, consolidator and the original FC layer take the same input and their outputs are summed up to produce the new output $\mathbf{y}$. Formally, for each FC layer with weight $\mathbf{W} \in \mathbb{R}^{E\times D}$ and bias $\mathbf{b} \in \mathbb{R}^E$, the output of thw whole layer modified by m GC branches is $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b} + \sum_{i=1}^{m}(\sum_{j=1}^{g^{(i)}} \mathrm{Pad}(\mathbf{W}_j^{(i)} \mathbf{x}_{\frac{(j-1)D}{g^{(i)}}:\frac{jD}{g^{(i)}}}, j) + \mathbf{b}^{(i)})$.

Channel reorder. To flexibly tunes the total number parameters in consolidator and enrich the exchange of information flow, we prepend a "ChannelReorder" operation to every branch in our consolidator by manually adjusting the permutation of input features along channel dimension. In general, we adopt shuffle operation (Zhang et al., 2018; Ma et al., 2018) to accomplish such purpose.

Formally, given input $\mathbf{x} \in \mathbb{R}^{*\times D}$ where "$*$" means any number of dimensions including none, we shuffle it into g groups and perform a recombination across the last dimension. Formally, we first reshape $\mathbf{x}$ into $\mathbf{x}' \in \mathbb{R}^{*\times g\times \frac{D}{g}}$, and then transpose the last two dimension and get $\mathbf{x}'' \in \mathbb{R}^{*\times \frac{D}{g}\times g}$, and then reshape $\mathbf{x}''$ into $\mathbf{x}''' \in \mathbb{R}^{*\times D}$, which is the final output. A pythonic style formulation is $\mathrm{ChannelReorder}(g, \mathbf{x}) = (\mathbf{x}.\mathrm{reshape}(*, g, \frac{D}{g})).\mathrm{transpose}(-2, -1).\mathrm{reshape}(*, D)$

We set shuffle groups $g = g_i$ in i-th branch, where g_i is the group of the corresponding GC layer. In this way, there are only few overlaps between the weight matrices of distinct branches, and the model capacity is greatly expanded each time a new branch is contained.

Stochastic depth of pretrained weight. To further enlarge the model's adaptation capacity, we append a droppath (Huang et al., 2016) layer to each branch. For small downstream datasets, dropping the consolidator path with a higher ratio p can help reduce overfitting and catastrophic forgetting, which may be beneficial to the performance. Empirically shown in Section 4.8, droppath is more effective than standard dropout (Srivastava et al., 2014) in current situation probably for the following reasons. The parameters of the whole layer degrade into the pretrained weight parameters with probability p. The frozen pretrained parameters contain domain-generic knowledge, which may help the adaptation. Overall, a model modified by consolidator will have a stochastic depth of pretrained state of parameters during each forward pass, and different consolidators will be activated for different training samples.

Two-stage consolidation. Now we have constructed all the elements of a consolidator. Formally, the output of the whole layer after being modified by consolidator is $\mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b} + \text{Droppath}(p, \sum_{i=1}^{m}(\sum_{j=1}^{g^{(i)}} \text{Pad}(\mathbf{W}_j^{(i)}\text{ChannelReorder}(g^{(i)}, \mathbf{x})_{\frac{(j-1)D}{g^{(i)}}:\frac{jD}{g^{(i)}}}, j) + \mathbf{b}^{(i)}))$.

Since the all the operations in a consolidator are all inference-time linear, we can easily consolidate the domain-specific knowledge into the domain-agnostic knowledge in pretrained backbone model, in both training-storage phase and loading-inference phase.

1. Training-storage consolidation. All we need to store in a consolidator are those $\mathbf{W}^{(i)}$ and $\mathbf{b}^{(i)}$. However, there are some parameters corresponding to the same entry, and thus they can be merged into a single one. As shown in Fig. 1, we also tune bias of the original FC layer in addition to the parameters in consolidator. It is easy to find that the duplicate biases in all consolidator branches and the original bias can be merge into a single one. And there are some duplicate entries in the weight matrix as well, so we can merge all weight matrices into one single sparse matrix. Consolidating such duplicate entries can benefit a lot for storage cost. Formally, we use $\widetilde{\mathbf{W}}$ and $\widetilde{\mathbf{b}}$ to denote the matrix that we need to store on the disk. Since channel reorder is a linear operation, we can apply a reverse operation to $\mathbf{W}^{(i)}$ to simulate the effect of the reorder applied to $\mathbf{x}$. And we have $\widetilde{\mathbf{W}} = \sum_{i=1}^{m} \text{ChannelReorder}^{-1}(g^{(i)}, \text{Compact}(\mathbf{W}^{(i)}))$. Here Compact reshapes the input matrix for the preparation of reordering its channels. It is easy to verify that $\text{ChannelReorder}^{-1}(g^{(i)}, \bullet) = \text{ChannelReorder}(\frac{D}{g^{(i)}}, \bullet)$.

2. Loading-inference consolidation. After loading the merged sparse weight matrix and merged bias matrix to memory, we can directly add them back to the weight matrix and bias matrix of the original FC layer. Formally, we use $\hat{\mathbf{W}}$ and $\hat{\mathbf{b}}$ to denote the final weight and bias of the FC layer for inference. Then $\hat{\mathbf{W}} = \mathbf{W} + \widetilde{\mathbf{W}}$, $\hat{\mathbf{b}} = \widetilde{\mathbf{b}}$. In this way, no additional inference cost is brought.

Overall, our consolidator reduces the storage space by grouped connected layers, and consolidate some duplicate entries. The training time non-linearity, *e.g.*, droppath, which turns out to be linear in inference time, effectively enriches model capacity under constrained storage budget. Finally, we can consolidate the task-specific knowledge into the backbone model via merging the inference time linear components to enjoy free, efficient and effective transfer learning.

4 Experiments

4.1 Experimental settings

Baselines. We select several state-of-the-art parameter-efficient methods as our baselines, including Full, Head, Bias (Zaken et al., 2021), Adapter (Houlsby et al., 2019), VPT (Jia et al., 2022), LoRA (Hu et al., 2021) and NOAH (Zhang et al., 2022). Note that among the state-of-the-art methods, Adapter, VPT and NOAH will bring heavy inference cost to the pretrained model which may cause troubles in resource-limited devices, while LoRA and our consolidator are more friendly for deployment with no extra inference cost brought.

VTAB-1k. We first verify the effectiveness of our method on VTAB-1k (Zhai et al., 2019) benchmark, which contains a wide range of visual domains in 19 datasets. Each dataset is divided for training set (800 images), validation set (200 images) and test set (the original set). The final adaptation accuracy is reported on test set. All models are fine-tuned on train+val set.

Table 1: Full results on the VTAB-1k (Zhai et al., 2019) benchmark. Our consolidator reaches the best overall performance against all the other competitors. Especially, Consolidator gives strongest results on 10 of 19 datasets, and surpasses the state-of-the-art search-based method NOAH without searching for the optimal architecture for each dataset and bringing any inference cost.

		Natural							Specialized				Structured								
	# params	Cifar100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLAB	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	Average
Full	100%	68.9	87.7	64.3	97.2	86.9	87.4	38.8	79.7	95.7	84.2	73.9	56.3	58.6	41.7	65.5	57.5	46.7	25.7	29.1	68.97
Head	0.04%	63.4	85.0	63.2	97.0	86.3	36.6	51.0	78.5	87.5	68.6	74.0	34.3	30.6	33.2	55.4	12.5	20.0	9.6	19.2	57.64
Bias	0.10%	72.8	87.0	59.2	97.5	85.3	59.9	51.4	78.7	91.6	72.9	69.8	61.5	55.6	32.4	55.9	66.6	40.0	15.7	25.1	65.22
VPT	0.75%	**78.8**	90.8	65.8	98.0	88.3	78.1	49.6	81.8	**96.1**	83.4	68.4	68.5	60.0	46.5	72.8	73.6	47.9	**32.9**	37.8	71.97
Adapter	0.36%	69.8	91.2	68.8	99.0	89.9	85.7	53.8	82.3	95.5	83.7	**76.1**	82.4	65.0	48.2	80.5	74.5	49.7	29.8	39.2	74.28
LoRA	0.34%	68.1	90.1	69.8	98.9	90.8	84.8	54.0	83.2	95.6	84.2	73.9	82.4	68.7	49.4	80.0	81.7	46.1	31.4	41.8	74.64
NOAH	0.52%	69.6	**92.7**	70.2	99.1	90.4	86.1	53.7	84.4	95.4	83.9	75.8	**82.8**	**68.9**	49.9	81.7	**81.8**	48.3	32.8	**44.2**	75.48
Ours	0.35%	73.9	90.6	**72.9**	**99.3**	**91.3**	**91.4**	**55.5**	**86.5**	95.5	**85.1**	75.4	81.2	67.5	**50.5**	**81.7**	78.6	**51.6**	31.9	37.2	**75.93**

Full data setting. VTAB-1k merely contains 1000 images for training per task, where the capacity of a huge vision transformer might be redundant. Therefore, to check the performance of parameter-efficient methods in a data-sufficient situation, we select 10 widely-used datasets for visual recognition in various domains and use the origianl training, validation and test splits for experiments. On average, a dataset in this setting contains a lot more images for training, leaving huge space for the model to adapt. The chosen datasets include natural pictures (Caltech101 (Fei-Fei et al., 2004), Cifar10 (Krizhevsky et al., 2009), Cifar100 (Krizhevsky et al., 2009)), fine-grained classification (CUB200 (Wah et al., 2011), OxfordFlowers (Nilsback & Zisserman, 2008), OxfordPets (Parkhi et al., 2012), StanfordDogs (Khosla et al., 2011)), textures (DTD (Cimpoi et al., 2014)), scene classification (SUN397 (Xiao et al., 2010)) and satellite images (EuroSAT (Helber et al., 2019)).

Please refer to appendix for more details.

4.2 Main results

We first choose a ViT-B (Dosovitskiy et al., 2020) with 86M parameters, which is supervised trained on IN-21k (Deng et al., 2009), as our target model.

VTAB-1k Tab. 1 presents the full results on VTAB-1k benchmark. Overall, our consolidator is the best parameter-efficient method. On 10 of 19 datasets, consolidator achieves the best top-1 accuracy. Notably, consolidator surpasses a search-based state-of-the-art method NOAH with no need to exhaustively search for the module configurations, which is quite time-consuming in training.

Full data setting Tab. 2 presents the full results on full data setting. Overall, our consolidator still performs best. An interesting observation is that the rank of full fine-tuning rises as the training data increase. None of the parameter-efficient methods can reach comparable performance with full tine-tuning other than our consolidator within 0.5% parameters stored. In contrast, Adapter, LoRA, NOAH and consolidator can reach 5%-6% higher accuracy on VTAB-1k than full fine-tuning.

4.3 Results for self-supervised vision transformer

Then we use a self-supervised trained vision transformer on IN-1k as our target model. We choose MoCo v3 (He et al., 2020) as the pretraining method. Seen from Tab. 3, on both VTAB-1k and full data setting, consolidator consistently reaches highest average top-1 accuracy. LoRA falls behind others in this setting, showing incompatibility with MoCo v3 learned representations. Adapter and NOAH still keep competitive performance, though worse than consolidator.

4.4 Results for more pretrained models

To further verify the generalization ability of consolidator, we conduct extensive experiments in Tab. 4 based on full data setting. First, we apply consolidator to supervised learned visual repre-

Table 2: Full results on data-sufficient scenarios. It is more challenging to take full adavantage of a large amount of data within a fairly small number of stored parameters. In this setting, full fine-tuning is a strong competitor and only consolidator can reach comparable or better performance than full fine-tuning with merely around 0.5% parameters stored.

		Full data										
	# params	Caltech101	Cifar10	Cifar100	CUB	DTD	Flowers102	Pets	Dogs	Sun397	EuroSAT	Average
Full	100%	90.9	99.1	**92.6**	**87.7**	**75.6**	98.9	**93.9**	84.8	70.9	**98.6**	89.30
Head	0.10%	90.0	88.8	71.6	83.5	71.3	98.2	90.5	78.4	67.0	95.6	83.49
Bias	0.22%	90.2	99.0	91.9	86.4	73.5	98.5	92.6	85.4	71.0	97.7	88.62
LoRA	0.53%	90.1	99.0	91.8	85.4	71.9	98.1	92.1	85.6	71.3	97.2	88.25
Adapter	0.55%	91.0	99.0	92.2	86.3	72.6	98.7	92.5	85.9	71.9	97.4	88.75
Ours	0.50%	**91.4**	**99.1**	92.5	87.0	74.5	**99.0**	93.2	**86.4**	**72.1**	97.9	**89.31**

Table 3: Adaptation results for a MoCo v3 ViT-B. For self-supervised vision transformers, consolidator still performs best on both VTAB-1k and full data setting.

	VTAB-1k		Full data	
	# params	Average	# params	Average
Full	100%	69.55	100%	86.24
Head	0.04%	59.62	0.10%	75.42
Bias	0.10%	69.15	0.22%	80.96
LoRA	0.34%	72.73	0.53%	79.70
Adapter	0.36%	73.22	0.55%	86.15
NOAH	0.42%	73.55	——	——
Ours	0.35%	**74.29**	0.50%	**86.41**

Table 4: Adaptation performance for more pretrained models in full data setting. For all the models, consolidator can reach better performance than full fine-tuning within a very small set of parameters which is needed to be stored. Generally, it is more difficult to reach better performance than full fine-tuning when the model capacity is insufficient, *i.e.*, the model does not accomodate enough parameters (ViT-L v.s. ViT-S).

		Supervised Learning			MAE
Architecture		ViT-S	ViT-L	Swin-B	ViT-B
Total Parameters		22M	303M	87M	86M
Full	Average	89.10	90.07	91.00	82.77
	# params	100%	100%	100%	100%
Ours	Average	**89.12**	**90.52**	**91.28**	**83.32**
	# params	**5.06%**	**0.33%**	**0.77%**	**1.78%**

sentation with larger (ViT-L) or smaller (ViT-S) parameter amounts than standard ViT-B. Compared with full fine-tuning, we achieve a comparable performance for ViT-S with 5.06% parameters. For ViT-L, merely 0.33% parameters can lead to 0.45 higher than full fine-tuning. Then we conduct experiment on Swin-B, a hierarchical architecture using shifted windows to introduce locality for better recognition. We observe a 0.28 improvement while storing only 0.77% parameters. Finally, we conduct experiment on ViT-B pretrained by a generative SSL method, MAE, as a comparison with the former contrastive SSL method MoCo v3. Our consolidator still earns consistent improvement, storing 1.78% parameters onto the disk while enjoying 0.55 higher accuracy.

Generally, we can conclude that it is more difficult to reach better performance than full fine-tuning when the model capacity is insufficient, *i.e.*, the model does not accomodate enough parameters (ViT-L v.s. ViT-S) to acquire full information from massive training data and adapt itself to the new domain.

4.5 Adaptation across varying scales of stored parameters

Next, we seek to find the principle between downstream accuracy and the number of stored parameters for all parameter-efficient methods with flexible parameter scale, based on full data setting. Results are shown in Fig. 2 left. LoRA, adapter and consolidator reach better performance as the number of parameters increases. In various parameter scales (from 0.5% to 10%), our consolidator consistently outperforms all competitors by a clear margin.

4.6 Adaptation across varying data sampling ratios

Fig. 2 right shows adaptation results corresponding to varying sampling rates of datasets, based on full data setting. For all sampling rates, consolidator keeps best adaptation ability. In addition, as

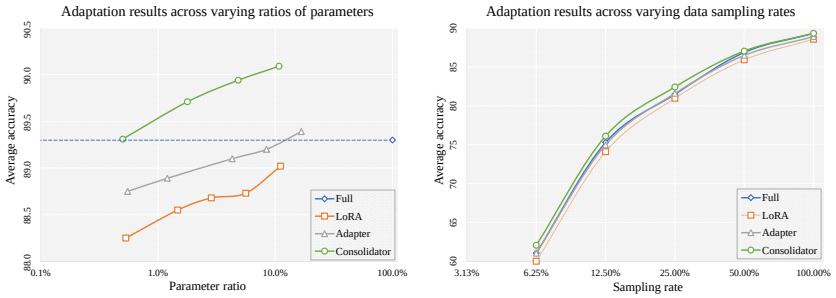

Figure 2: Left: the Adaptation results corresponding to varying ratios of stored parameters. Clearly, consolidator consistently outperforms LoRA, adapter and full fine-tuning by a significant margin across a wide range of parameter scales from 0.5% to 10%. And each of the three methods will reach higher accuracy if we increase its storage budget and tuning more parameters. Right: adaptation results corresponding to varying sampling rates of datasets. Consolidator reaches the best performance for all sampling rates. As the sampling rate decreases, full fine-tuning slightly falls off its advantage over adapter and LoRA, which is consistent with previous observations on VTAB-1k and full data setting.

Table 5: Performance on downstream object detection and semantic segmentation tasks. Compared with tuning detection/segmentation head only, consolidator reaches much more better mAP/mIoU with negligible parameters increase, and surpasses the performance of tuning all parameters in both backbone and task head as well.

Method	Backbone	Object Detection			Semantic Segmentation		
		Framework	# params	mAP	Framework	# params	mIoU
Full	Swin-B	Faster-RCNN	100%	85.85	UperNet	100%	83.24
Head			16.74%	84.89		28.48%	82.31
Consolidator			**17.22%**	**86.64**		**28.89%**	**83.96**

the sampling rate decreases, full fine-tuning slightly falls off its advantage over Adapter and LoRA, which is consistent with previous observations on VTAB-1k and full data setting.

4.7 Results on object detection and semantic segmentation.

We further verify the effectiveness on downstream object detection and semantic segmentation tasks. We adopt Swin-Base as the backbone which can provide hierarchical features. Experiments are done on PASCAL VOC 07+12 (Everingham et al., 2010) for detection and PASCAL VOC 12 (Everingham et al., 2010) for segmentation. We adopt Faster R-CNN (Ren et al., 2015) and UperNet (Xiao et al., 2018) separately for detection and segmentation framework. Seen from Tab. 5, our consolidator significantly outperforms full training and detection/segmentation head training with a small amount of parameters stored, showing great potential for broader usage.

4.8 Ablation studies

We conduct controlled experiments to better identify the effect of individual components in our module design. We report tuned parameters, stored parameters and accuracy separately in Tab. 6. Experiments in this section are conducted on 5 of the 10 datasets with different domains: Caltech101, DTD, OxfordFlowers, StanfordDogs and EuroSAT.

Droppath *v.s.* Dropout. We first investigate our choice of Droppath. Droppath with a consolidator of ($g^{(1)} = 96$, $g^{(2)} = 192$) as the base model. As shown in Tab. 6, compared with dropout and no drop-type layer, droppath can obtain 0.44 an 0.48 performance improvement, respectively, well demonstrating the effectiveness of encouraging stochastic depth for consolidator.

ChannelReorder. The effect of ChannelReorder operation mainly lies in separating the entries in different branches to reduce the repetitive ones. In a multi-branch case like ($g^{(1)} = 96, g^{(2)} = 192$), ChannelReoder brings 0.38 accuracy impovement. Furthermore, it is helpful even if there is only one branch like ($g^{(1)} = 384$), where ChannelReorder still slightly raises the accuracy by 0.1.

Table 6: Effect of individual designs for a consolidator module.

Branches	Original Bias	Extra Bias	Channel Reorder	Dropout	Droppath	Tuned Params	stored Params	Accuracy	ΔAccuracy
96, 192	✓	✓	✓	✗	✓	1.95%	1.75%	87.36	-0.00
96, 192	✓	✓	✓	✓	✗	1.95%	1.75%	86.92	-0.44
96, 192	✓	✓	✓	✗	✗	1.95%	1.75%	86.88	-0.48
96, 192	✓	✓	✗	✗	✓	1.95%	1.24%	86.98	-0.38
96, 192	✗	✓	✓	✗	✓	1.85%	1.75%	87.26	-0.10
96, 192	✓	✗	✓	✗	✓	1.76%	1.75%	86.92	-0.44
96, 192	✗	✗	✓	✗	✓	1.66%	1.65%	86.82	-0.54
384	✓	✓	✓	✗	✓	0.56%	0.47%	87.08	-0.00
384	✓	✓	✗	✗	✓	0.56%	0.47%	86.98	-0.10
384, 384	✓	✓	✓	✗	✓	0.92%	0.47%	86.96	-0.12
384	✗	✗	✓	✗	✓	0.37%	0.37%	86.84	-0.00
384 (unst)	✗	✗	**N/A**	✗	✓	0.37%	0.37%	86.70	-0.14

Duplication of bias and weight. Based on the consolidator with ($g^{(1)} = 96, g^{(2)} = 192$), we can see that duplicating bias is relatively effective. Compared with tuning the original bias and only tuning two extra bias parameters, tuning all three bias parameters can lead to 0.44 and 0.1 performance improvement, respectively, though they share the same storage cost. Additionally, we also compare tuning original bias *v.s.* not tuning bias, and the former only has a slight 0.1 accuracy advantage, which further verifies the effectiveness owes mostly to the dedicate consolidation parts design instead of simple bias tuning. Besides bias, we test on a consolidator with ($g^{(1)} = 96$) to investigate the effect of integrating duplicate branches to duplicate the weight. However, this kind of consolidation does not bring any improvement.

Structured *v.s.* Unstructured. One potential limitation of consolidator may be that $g^{(i)}$ can not be selected arbitrarily for it must be a factor of the channels. This may cause trouble when fairly few parameters, *e.g*. 0.0001%, are required to be tuned. A supplemental method is to adopt unstructured sparsity instead of current structured block-wise sparsity in consolidator branches to flexibly control the number of parameters. We simulate the situation using a consolidator with ($g^{(1)} = 384$) for comparison with the unstructured one. Seen from Tab. 6, unstructured branch whose tunable parameter number equals to that of a branch with $g = 384$ faces slight accuracy drop, 0.14. In summary, unstructured consolidator can be a sub-optimal choice when needed, with slight performance drop and more training cost due to the unstructured sparse matrix is unfriendly to the hardware.

5 Conclusions

We propose consolidator, a novel method to achieve both parameter- and inference-efficient visual transfer learning for transformers. Consolidator adds a few tunable, mergeable modules along each fully connected layer in the pretrained model and keep most of the original parameters freezed during adaptation. We creatively design two-stage consolidation to dramatically boost the model capacity under a given storage budget. The duplicate entries in multi-branches of consolidator will be merged into a single matrix and stored onto disk. Finally, we consolidate the task-specific knowledge in consolidator into the tasks-agnostic knowledge in pretrained backbone model, without any extra inference cost introduced. On various vision tasks, consolidator outperforms all state-of-the-art competitors significantly, showing strong scalability and generalization ability. We hope to motivate the community to tap more effective and efficient methods for visual transfer learning.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Minghao Chen, Houwen Peng, Jianlong Fu, and Haibin Ling. Autoformer: Searching transformers for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12270–12280, October 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*, 2022a.

Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1911–1920, 2019.

Xiaohan Ding, Tianxiang Hao, Jianchao Tan, Ji Liu, Jungong Han, Yuchen Guo, and Guiguang Ding. Resrep: Lossless cnn pruning via decoupling remembering and forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4510–4520, 2021a.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Diverse branch block: Building a convolution as an inception-like unit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10886–10895, 2021b.

Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13733–13742, 2021c.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11963–11975, 2022b.

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.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Shuxuan Guo, Jose M Alvarez, and Mathieu Salzmann. Expandnets: Linear over-parameterization to train compact convolutional networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

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*, pp. 2790–2799. PMLR, 2019.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pp. 646–661. Springer, 2016.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

Junguang Jiang, Yang Shu, Jianmin Wang, and Mingsheng Long. Transferability in deep learning: A survey. *arXiv preprint arXiv:2201.05867*, 2022.

Bo Fu Junguang Jiang, Baixu Chen and Mingsheng Long. Transfer-learning-library. `https://github.com/thuml/Transfer-Learning-Library`, 2020.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34, 2021.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*. Citeseer, June 2011.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021a.

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*, pp. 10012–10022, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

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*, pp. 487–503, 2021.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8119–8127, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

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*, pp. 10347–10357. PMLR, 2021.

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.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 418–434, 2018.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, 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.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021.

A Experiment Settings

A.1 Image recognition

A.1.1 Datasets

VTAB-1k. VTAB-1k (Zhai et al., 2019) consists of 19 visual datasets in three groups: Natural, Specialized and Structured, containing images collecting from a wide range of visual domains. Each dataset is divided for training set (800 images), validation set (200 images) and test set (the original set). The final adaptation accuracy is reported on test set. All models are fine-tuned on the full 1000 labeled images, *i.e.*, train+val set. The validation set is used for tuning some hyperparameters.

Full data. We select 10 widely-used visual datasets with sufficient data to further verify the effectiveness of the parameter-efficient methods. The datasets are shown in following part:

Caltech101 (Fei-Fei et al., 2004) contains pictures of natural objects belonging to 101 classes, with 3060 images for training and 6084 images for test.

CIFAR-10/100 (Krizhevsky et al., 2009) contains pictures of natural objects belonging to 10/100 classes, with 50000 images for training and 10000 images for test.

Caltech-UCSD Birds-200-2011 (CUB) (Wah et al., 2011) contains pictures of birds belonging to 200 fine-grained bird classes, with 5994 images for training and 5794 images for test.

Describable Textures Dataset (DTD) (Cimpoi et al., 2014) contains texture images in the wild belonging to 47 human-centric classes, with 5640 images in total which are equally split into the training, validation and test set.

OxfordFlowers102 (Nilsback & Zisserman, 2008) contains pictures of flowers belonging to 102 fine-grained flower classes, with 1020 images for training, 1020 images for validation and 6149 images for test.

Oxford-IIITPets (Parkhi et al., 2012) contains pictures of pets belonging to 37 fine-grained pet classes, with 3680 images for training and 3669 images for test.

StanfordDogs (Khosla et al., 2011) contains pictures of dogs belonging to 120 fine-grained dog classes, with 12000 images for training and 8580 images for test.

Scene UNderstanding (SUN397) (Xiao et al., 2010) contains pictures of scenes belonging to 397 classes, with 76128 images for training, 10875 images for validation and 21750 images for test.

EuroSAT (Helber et al., 2019) contains pictures based on Sentinel-2 satellite images belonging to 10 classes, with 27000 images in total.

Some of the datasets do not contain a validation split, and we will manually select $\tilde{1}0\%$ random images from training set as the validation set.

To better utilize the massive data in various domains, we follow the practice in (Junguang Jiang & Long, 2020; Jiang et al., 2022) to make preparation and divide data splits. For data augmentation, we adopt a standard pipeline. In training, we do a random resize crop to 224×224, random horizontal flip and normalization for each input image. In test, we do a resize to 256×256, center crop to 224×224 and normalization for each input image.

A.1.2 TRAINING HYPERPARAMETERS.

On **VTAB-1k**, we follow the hyperparameters in VPT (Jia et al., 2022) for full fine-tuning, Head, Bias and VPT, and mainly follow the hyperparameters in NOAH (Zhang et al., 2022) for adapter, LoRA, NOAH and consolidator. The detailed setting for hyperparameters can be found in Tabs. 7 and 8.

On **full data setting**, we do a quick grid search to choose a proper set of training hyperparameters based on the performance of full fine-tuning for every well-trained visual representation. All training hyperparameters are shown in Tab. 8. Following the practice in (Jiang et al., 2022; Junguang Jiang & Long, 2020), we reduce the learning rate of backbone parameters to 0.1x that of head parameters. We then adopt the same hyperparameters for all parameter-efficient variants as well as our consolidator after the hyperparameters have been chosen according to full fine-tuning results.

All experiments are conducted on a single NVIDIA RTX 3090 with 24G memory.

A.1.3 METHODS.

Full: Ordinary fine-tuning. It tunes all parameters and stores all parameters to disk for every downstream tasks, leading to heavy storage cost.

Head: Also know as linear probing. It only tunes the classification head and freezes other parameters.

Bias (Zaken et al., 2021): Bias only tunes the biases and freezes all the weights in a pretrained model.

Adapter (Houlsby et al., 2019): Adapter inserts a sequence of tunable parameters, including a down projection, a non-linearity (here we use GELU (Hendrycks & Gimpel, 2016)) and a up projection layer, into each encoder block. The adapters are serially connected with the backbone layers.

LoRA (Hu et al., 2021): LoRA adds a concurrent branch containing low rank weight matrices for efficient parameter update. A LoRA module contains serial down projection and up projection layer without non-linearity, and thus it can be merged into the backbone parameters before inference.

VPT (Li & Liang, 2021; Jia et al., 2022): VPT appends tunable virtual tokens to the inputs of transformer blocks, which will participate in the calculation of the subsequent blocks along with the actual tokens.

NOAH (Zhang et al., 2022): NOAH first trains a large supernet with all three modules, VPT, LoRA and adapter, and then searches for the optimal configurations of each module for each layer using the NAS algorithm introduced by AutoFormer (Chen et al., 2021). In the end, NOAH retrains the best subnet candidates to produce the final result.

Implementation details. For **VTAB-1k**, we follows the same implementations for LoRA and adapter with NOAH (Zhang et al., 2022). For NOAH, we enlarge its subnet retraining learning rate range from {1e-3} to {1e-2, 1e-3} as in Tab. 8. Because we found that higher learning rate is helpful for the adaptation of self-supervised models like MoCo v3 transformers.

For **Full data setting**, the detailed implementations are shown as follows. For LoRA (Hu et al., 2021), we follow its original implementation to do low-rank re-parameterization and tune merely two weight matrices $\mathbf{W}_q$ and $\mathbf{W}_v$, which generate the attention matrices $\mathbf{Q}$ and $\mathbf{V}$, in every transformer encoder block. For bias, we tune all the biases (including the parameters outside encoder blocks) in the model. For adapter, we follow its original implementation to additionally tune the parameters of LayerNorm (Ba et al., 2016) and adds the tunable adapter modules to the end of each MHSA and MLP before calculating the residual connections. For consolidator, we follow the practice of adapter to tune the LayerNorm as well and only adds consolidator to the linear layers in MHSA and MLP of each transformer encoder block. We sweep the drop ratio in $\{0.0, 0.2, 0.5, 0.8\}$. We choose training hyperparameters according to the performance of full-finetuning for each visual representation. The hyperparameters are finally configured as in Tab. 7. When applying consolidator to Swin-B, we skip the first stage of transformer encoder blocks that contain relatively small part of parameters (0.6% of total parameters) compared with others, and merely inserts consolidator for linear layers in MHSA and MLP of blocks of last 3 stages.

Table 7: The method-related hyperparameters.

	model	Setting	LoRA	Adapter	Consolidator
Supervised, MoCo v3	ViT-B	VTAB-1k	rank=10	hidden=9	($g^{(1)}$=384)
		Full data	rank=8	hidden=16	($g^{(1)}$=384)
Supervised	ViT-S	Full data	——	——	($g^{(1)}$=48, $g^{(2)}$=64, $g^{(3)}$=96)
Supervised	ViT-L	Full data	——	——	($g^{(1)}$=512)
Supervised	Swin-B	Full data	——	——	($g^{(1)}$=256)
MAE	ViT-B	Full data	——	——	($g^{(1)}$=96, $g^{(2)}$=192)

Table 8: The training hyperparameters.

	VTAB-1k		Full data				
	Supervised	MoCo v3	Supervised			MAE	MoCo v3
model	ViT-B	ViT-B	ViT-S	ViT-L	Swin-B	ViT-B	ViT-B
pretraining dataset	IN-21k	IN-1k	IN-21k	IN-21k	IN-21k	IN-1k	IN-1k
optimizer	adamw	adamw	sgd	sgd	sgd	sgd	sgd
warmup epochs	10	10	5	5	5	5	5
epochs	100	100	40	40	40	40	40
batch size	64	64	128	32	64	64	64
lr	1e-2, 1e-3	1e-2, 1e-3	1e-3	1e-3	1e-3	1e-2	1e-2
wd	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4

A.2 Object detection

On downstream object detection task, we adopt Faster R-CNN (Ren et al., 2015) framework with FPN (Lin et al., 2017) to verify the effectiveness of consolidator on Pascal VOC 07+12 dataset (Everingham et al., 2010). We use a Swin-B pretrained on IN-21k as the backbone model. The consolidator setting is the same as the setting of Swin-B in Tab. 7. For hyperparameters, we adopt AdamW as optimizer with a learning rate of 1e-4 and weight decay of 5e-2, and train for 8 epochs in total. The learning rate is decayed by a factor of 10 after 6 epochs. The first 300 iterations are trained with a warmup ratio of 1e-3 for learning rate. The data augmentation is the same as the default strategy in mmdetection (Chen et al., 2019).

A.3 Semantic segmentation

On downstream semantic segmentation task, we adopt UperNet (Xiao et al., 2018) to verify the effectiveness of consolidator on Pascal VOC 12 dataset (Everingham et al., 2010). We use a Swin-B pretrained on IN-21k as the backbone model. The consolidator setting is the same as the setting of Swin-B in Tab. 7. For hyperparameters, we adopt AdamW as optimizer with a learning rate of 6e-5 and weight decay of 1e-2, and train for 20000 iterations in total. The learning rate is decayed by a polynomial scheduler with a power of 1.0. The first 200 iterations are trained with a warmup ratio of 1e-6 for learning rate. The data augmentation is the same as the default strategy in mmsegmentation (Contributors, 2020).

B Inference Cost of Classic Parameter-efficient Transfer Learning Methods

Some of the state-of-the-art parameter-efficient tuning methods, *e.g.* adapter (Houlsby et al., 2019) and prompt-based tuning (Li & Liang, 2021), introduce non-negligible extra cost in inference period. In contrast, consolidator shares totally identical inference cost with the original architecture. We quantitatively show such cost in Tab. 9. As the number of tuned parameters increases, adapter and prefix tuning lead to more inference latency and GPU memory usage, which is unfavorable for the model deployment, especially on resource-constrained devices such as mobile or embedded systems.

Table 9: Inference latency (ms/image) and GPU memory (MB) usage of ViT-B (Dosovitskiy et al., 2020) with a mini-batch size of 64. The inference latency is averaged over 100 batches of running. All the data are measured on the same NVIDIA RTX 3090 and Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz.

	# params	Latency	ΔLatency	GPU Memory	ΔGPU Memory
Full fine-tuning / Consolidator	100% / any	1.73	+0%	11776	+0%
Adapter(hidden=384)	16.67%	2.06	+19%	13664	+16%
Adapter(hidden=24)	1.20%	1.88	+9%	12730	+8%
VPT(length=100)	1.17%	2.72	+52%	18970	+61%
VPT(length=10)	0.21%	1.83	+6%	12528	+6%

Table 10: Sensitivity test on consolidator's hyperparameters. Given a particular target storage budget, we may have several choices in selecting the branches and groups. As seen below, such choices make little influence on the final results. The performance of consolidator is relatively stable under a given storage budget.

# params	0.75%	1.01%	4.88%	6.42%
branches$_{more}$	(256,768)	(256,384,768)	(48,64,96)	(24,48)
acc$_{more}$	87.02	87.04	87.18	87.28
branches$_{less}$	(192)	(128)	(32,64)	(16)
acc$_{less}$	86.98	87.00	87.14	87.36
Δacc	0.04	0.04	0.04	-0.08

C SENSITIVITY OF GROUPS AND BRANCHES IN CONSOLIDATOR

Given a particular target storage budget, we may have several choices in selecting the branches and groups. As seen in Tab. 10, such choices make little influence on the final results. The performance of consolidator is relatively stable under a given storage budget. When the budget increases, the performance of consolidator increases as well. Such monotonical property is helpful for real-world applications, making it easy to tune hyperparameters and rapidly find an optimal candidate under a given storage budget.