Structured Model Probing: Empowering Efficient Transfer Learning by Structured Regularization

Zhi-Fan Wu Chaojie Mao Xue Wang Jianwen Jiang Yiliang Lv Rong Jin
Alibaba Group

Abstract

*Despite encouraging results from recent developments in transfer learning for adapting pre-trained model to downstream tasks, the performance of model probing is still lagging behind the state-of-the-art parameter efficient tuning methods. Our investigation reveals that existing model probing methods perform well for the easy case when the source domain (where models are pre-trained) and the adapted domain are similar, but fail for the difficult case when the two domains are significantly different. Simply incorporating features extracted from multiple layers and increasing complexity of the probing model can mitigate the gap in the difficult case, but degrades the performance in the easy case. To address this challenge, we propose structured model probing (***SMP***) that is able to deliver good performance for both cases through* **structured regularization***. The regularization performs feature selection leveraging model structure as a prior, and controls the complexity of the probing model through the weights of selected structures. This enables us to construct a simple adaptation model, with a small number of selected features and a linear prediction model, for the easy case; and to automatically increase the complexity of adaptation model, with a large number of selected features and a non-linear model, for the difficult case. Our extensive empirical studies show that SMP significantly outperforms the state-of-the-art methods for parameter efficient tuning, and at the same time, still maintains the advantage of computational efficiency for probing-based methods.*

1. Introduction

The parameters of models have grown drastically with the rapid development of deep learning [5, 37], posing challenges for adapting pre-trained models to downstream tasks. The widely used transfer strategy, fully fine-tuning, becomes infeasible due to excessive computation and storage overhead. Recently, parameter efficient tuning [19, 23] has gained attention for reducing storage costs and improving tuning performance with few trainable parameters when limited data are available for downstream tasks. Despite the progress, parameter efficient tuning methods still face significant computational overhead in training due to the need to perform forward and backward passes through the pre-trained model at each training step.

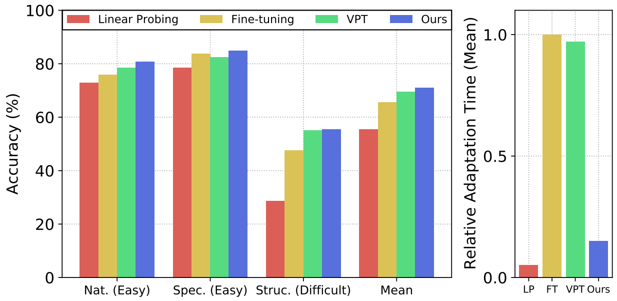


Figure 1. Performance comparison on different dataset groups (left) and mean relative adaptation time (right) on VTAB-1k benchmark.

Model probing, which leverages features extracted from the pre-trained model to perform adaptation, is promising due to its efficiency and simplicity: The training procedure is efficient since the pre-trained model is fixed and only performs one forward propagation to extract features, and probing model is decoupled from pre-trained model that makes it convenient to deploy small sized task-specific downstream models. However, it is a nuisance that probing methods are usually outperformed by state-of-the-art parameter efficient tuning methods. This is true even after several recent improvements made for the probing methods [16, 45]. In this paper, we aim to improve the performance of model probing significantly without sacrificing its simplicity and efficiency.

We first investigate when model probing methods perform worse than the tuning-based methods on widely used visual task adaptation benchmark VTAB-1k [44] as shown in Figure 1. Target domains are categorized as *easy* and *difficult* cases based on their similarity to the source domain where the pre-trained model is trained on (see Section 3.1 for details). We find simple linear probing yields comparable performance as the two tuning-based approaches (fine-tuning and VPT [23]) for easy cases, whereas a significant

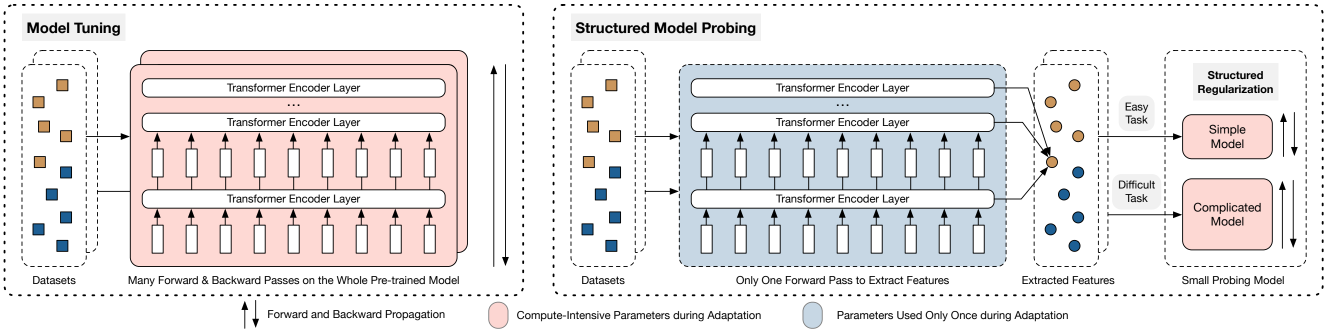

Model Tuning
Transformer Encoder Layer
Transformer Encoder Layer
Transformer Encoder Layer
Datasets
Many Forward & Backward Passes on the Whole Pre-trained Model
Structured Model Probing
Transformer Encoder Layer
Transformer Encoder Layer
Transformer Encoder Layer
Datasets
Only One Forward Pass to Extract Features
Extracted Features
Easy Task
Difficult Task
Structured Regularization
Simple Model
Complicated Model
Small Probing Model
Forward and Backward Propagation
Compute-Intensive Parameters during Adaptation
Parameters Used Only Once during Adaptation

Figure 2. Adaptation pipelines of Model Tuning (left) and the proposed Structured Model Probing (right).

performance gap is observed between linear probing and tuning-based methods for difficult cases. This suggests that model probing exhibits different behavior for various target domains, and performance gap arises for target domains that dissimilar to the source domain.

This is not surprising that we expect a more complicated model than a linear probing to compensate the difference between the source domain and the target domain. In general, we can increase model complexity in two ways, either by introducing more features extracted from the pre-trained model, or by introducing a non-linear model for prediction. A *main challenge* is how to develop a unified framework of model probing that is able to handle both the easy and difficult adaptations effectively. This is because, according to our study in Section 3, for easy cases, a complicated probing model (with more features and non-linear probing) performs significantly *worse* than a simple linear probing.

To this end, we propose structured model probing, or **SMP** for short, that effectively address both easy and difficult cases of domain adaptation. The key to our unified framework is **structured regularization**. It performs feature selection leveraging the structure of the pre-trained model as a prior, and controls the complexity of the probing model through the weights assigned to selected features. Thus, for the easy case, this regularization constructs a simple linear model with a few selected features; and for the difficult case, it automatically introduces more features from the pre-trained model and more non-linearity in the prediction module based on the value of loss function, as shown in Figure 2. Our empirical studies show that the proposed structured model probing performs significantly better than state-of-the-art tuning-based approaches, while maintaining the advantage of computational efficiency.

Our main contributions are as follows:

1. We observe and demonstrate the conflicting behaviors of model probing: a complicated probing model with more features and non-linear transformation can improve the performance for difficult cases of domain adaptation, but it yields worse performance than linear probing for easy cases of domain adaptation.
2. We propose structured model probing (SMP) based on structured regularization. This regularization enables us to construct simple probing model for easy tasks, and automatically increase model complexity for difficult tasks, to simultaneously address both the easy and difficult cases of domain adaptation.
3. We conduct extensive empirical studies on various adaptation tasks. The proposed SMP method shows superior experimental results compared to state-of-the-art probing and tuning methods, and requires significantly lower training costs compared to tuning methods.

2. Related Work

Transfer learning has been extensively studied in vision domain [34, 48]. It aims to improve the performance on target domain by leveraging the related information contained in source domain. Transfer learning based on pre-trained model [8, 20, 21] has received significant attention due to its simplicity and good empirical performance [23, 45].

Model tuning adapts the pre-trained model to new tasks via updating model parameters. Fully fine-tuning is a widely used strategies due to its efficacy and simplicity [9, 25, 44]. However, it suffers from huge storage and computation overhead due to the ever-expanding size of large-scale pre-trained model [37, 40]. Heuristic approaches have been proposed to improve fine-tuning performance [3, 6, 17, 18, 27]. Parameter efficient tuning [19] aims to reduce the storage requirement of transferred model by updating a few parameters attached to the pre-trained model [23, 47], like Prompt Tuning [29, 30] or Adapter [22]. While originally studied in natural language processing tasks due to the dominance of large-scale pre-trained language models [5, 14], there is growing interest in applying these methods to computer vision tasks [7, 23, 31, 41, 46, 47]. Though parameter efficient tuning reduces storage cost significantly, the training cost remains high due to the forward and backward propagation need to pass through the entire pre-trained model [11].

Model probing utilizes frozen extracted features for adapting the pre-trained model to new tasks, thus achieving efficient adaptation. The most popular method, linear probing, suffers from performance degradation [25], which limits its application in transfer learning [20, 21]. Recent methods [16, 36, 45] have been proposed to improve probing performance by leveraging side network [38, 45] or incorporating intermediate representations [2, 16, 32], but there is still a significant performance gap compared to tuning methods. This motivates us to investigate the limitation of probing methods, and come up with a better solution.

3. Understanding the Limitation of Model Probing for Transfer Learning

In this section, we analyze the performance of probing methods across various downstream domains to understand their limitations, motivated by the observation in Figure 1. We assess feature informativeness for different tasks and investigate the necessity of non-linear transformation.

3.1. Analysis Setup

Pre-trained model and source data. The ViT-B/16 [15] is used as the backbone because of the extensive usage of transformer-based models. The model is pre-trained on ImageNet-21k [12] in a supervised learning manner.

Target data. We use the VTAB-1k benchmark, consisting of 19 datasets that cover natural images (natural), images captured by specialist equipment (specialized), and images generated from simulated environments (structured), as our target data for transfer learning. We follow the setting in [44] and use 1000 samples as the training set per task to evaluate adaptation with limited data as suggested in [44].

Domain similarity. Several previous works have proposed metrics to measure the similarity between source domain and target domain [1, 10, 13, 16, 33]. Here, we adopt the domain similarity measure defined in [16]:

$$\textit{Domain similarity} = \text{Acc}_{\text{Linear}} - \text{Acc}_{\text{Scratch}}. \tag{1}$$

A higher domain similarity suggests that the features extracted from a pre-trained model can be utilized for the target domain, leading to performance gains via a linear classifier compared to training from scratch. Conversely, a lower domain similarity indicates that training from scratch on the target domain may be more effective than using extracted features. Thus, the domain similarity is evaluated based on the benefit gained by a linear classifier compared to training from scratch. We also report the *label-feature correlation* [13] in Suppl. C.8, which exhibits a high Spearman rank correlation (0.854) with the domain similarity scores.

Domain similarity also reflects how difficult to adapt from source domain to target domain. As shown in Figure 4, tasks in natural and specialized groups have higher domain similarity to source domain where the pre-trained model is trained on, so we mark them as easy cases. And tasks in structured groups have lower domain similarity to source domain, thus they are labeled as difficult cases in Figure 1.

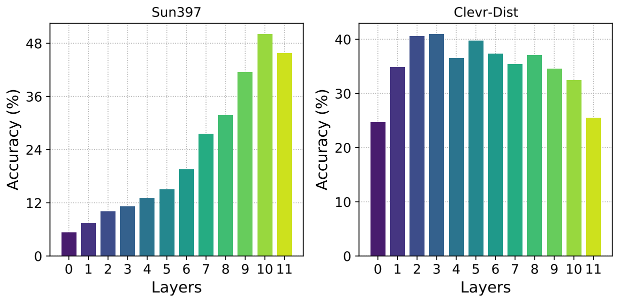


Figure 3. Performance of features extracted from different layers on Sun397 (easy) and Clevr-Dist (difficult).

3.2. Informativeness of Extracted Features

To quantify the informativeness of features extracted from different layers for downstream tasks, we use leave-one-out k-NN testing as the surrogate performance criterion:

$$\text{Informativeness} = \frac{\sum_{i=1}^{m} \sum_{x_j \in N(x_i)} \mathbb{1}\{y_i = y_j\}}{km}, \tag{2}$$

where m is the number of samples, x is the extracted features of a sample, $N(x_i)$ represents k nearest neighbors of x_i. We simply set $k = 1$, which results in leave-one-out 1-NN testing. We choose k-NN, instead of linear classifier, for informativeness measurement because k-NN is a non-parametric classifier and thus avoids training additional parameters, making the measurement more robust.

In Figure 3, we plot the informativeness of features extracted from different layers on two tasks from different domains. For Sun397, an easy domain adaption task, we observe the best performance achieved by features extracted from deep layers. This explains the good performance of linear probing for easy domain adaption tasks, as observed in Figure 1. While for Clevr-Dist, a difficult task with target domain significantly different from the source domain, we observe that no layer has outstanding performance. The results echo the findings in [16].

Furthermore, we conduct experiments on gradually incorporating features extracted from different layers of the pre-trained model, demonstrating the different behaviors on easy and difficult tasks, which can be found in Section 5.4. The results show for easy tasks, more features may lead to overfitting, but for difficult tasks, performance improves with more features. Therefore, when the target domain is very different from the source domain, we need to provide probing model with diverse features extracted from multiple layers of the pre-trained model.

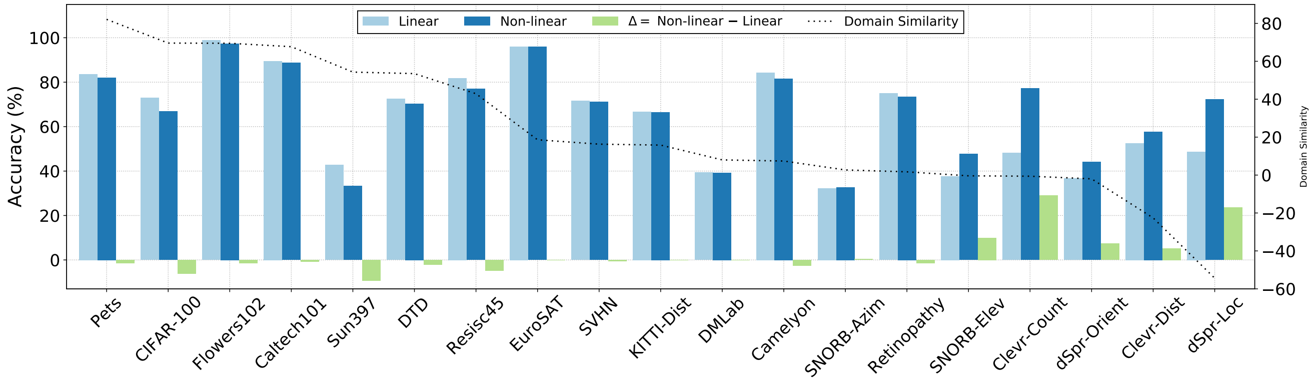


Figure 4. Performance of extracted features with linear/ non-linear transformation on VTAB-1k. Datasets are ordered from easy tasks (left) to difficult tasks (right) according to domain similarity.

3.3. Effects of Non-linear Transformation

Since there is no significantly informative layer for difficult domain adaption tasks, a linear combination may not be enough for exploiting the information in extracted features. We further examine the efficacy of non-linear combination on extracted features, as neural networks achieve superior performance through non-linear feature transformation.

We employ an multiple-layer perceptron (MLP) for non-linear transformation and compare its performance with linear transformation. Features extracted from different structures of the pre-trained model are concatenated together as input (See Section 4.1 for details).

In Figure 4, we clearly observe that adding non-linearity to feature transformation can significantly boost the performance for the difficult domain adaption tasks (i.e. the datasets on right end of Figure 4). This phenomenon indicates that simply performing linear transformation on extracted features can not exploit the full potential of representations extracted from the pre-trained model. On the other hand, we notice that the performance on easy tasks is not improved, and is even *hurt* with the introduction of non-linearity. We speculate that this is because the linear transformation can already make full use of extracted information, while non-linearity increases model capacity, which eventually leads to overfitting to the training data.

Discussion. Our investigation reveals that the optimal probing model differs depending on the domain of downstream task. Incorporating diverse features from multiple layers of the pre-trained model and performing non-linear transformation on them can significantly improve the performance on difficult cases. However, this yields worse results on easy tasks. On the contrary, simply performing linear combination on features from deep layers can achieve satisfactory performance for easy adaptation tasks, but not for difficult ones. The conflicting behaviors of model probing motivates us to develop a unified framework that can solve the adaptation problems for both easy and difficult adaptation tasks.

4. Structured Model Probing

In this section, we propose structured model probing, a unified probing method which can deal with both easy and difficult cases of adaptation tasks. We first present how to thoroughly extract features in a structured way, then introduce our probing model with structured regularization.

4.1. Structured Feature Extraction

Incorporating diverse features extracted from the pre-trained model can improve the performance on difficult adaptation tasks for model probing, as shown in Section 3.2. To provide diverse features for the probing model, we extract features from all structures in a Vision Transformer (ViT). These structures include features before self-attention, features after self-attention, and features after Feed Forward Network (FFN), for each transformer layer. We also include tokenized image inputs and pre-logits. We use $V = \{v_i\}_i^{|V|}$ represents all structures in ViT, and $|V|$ is the total number of structures. For an input sample, we denote each set of tokens extracted from structure v_i as $X_i \in \mathbb{R}^{t\times c}$, where t is the dimension of tokens and c is the dimension of channels.

To reduce feature redundancy while preserving diversity, we apply 1D average pooling along different dimensions on tokens X_i, resulting in two types of aggregation. To maintain channel information, we aggregate X_i along the token dimension, which produces x_i^t. And to preserve spatial information, we aggregate X_i along the channel dimension, resulting in x_i^c. Finally, we concatenate all structured features into a single vector, $x = [x_1^t, x_1^c, ..., x_{|V|}^t, x_{|V|}^c]$. We denote s_i as the index set of elements in x, and $S = \{s_i | i = 1...|s|\}$.

4.2. Probing with Structured Regularization

Our structured regularization contains two components: *structured sparsity regularizer* and *structured non-linearity regularizer*. We introduce them as follows.

4.2.1 Structured Feature Selection

As shown in Section 3.2, the informativeness of extracted features varies with tasks, and redundant features lead to performance degradation on some tasks. Thus, we propose to perform feature selection during the training stage of the probing model to prevent overfitting. A popular way is to use group lasso [16, 43] by constructing groups on feature dimension over the weight matrix of a linear classifier:

$$\min_{\theta} L(\theta) + \lambda \sum_{i=1}^{d} ||\theta_i||_2, \text{ where } ||\theta_i||_2 = \sqrt{\sum_{j=1}^{n} \theta_{ij}^2}. \quad (3)$$

We use $\theta \in \mathbb{R}^{d \times n}$ to represent weight matrix, where d is the feature dimension and n is the number of classes. However, due to the high dimensionality of features, the final achieved sparsity is hardly the optimal result. It is worth noting that the structure of pre-trained model provides a natural way to divide features as groups, and features extracted from the same structure have higher relevance to each other. Thus, we leverage the structure of the pre-trained model, to achieve the structured sparsity with a **structured sparsity regularizer**:

$$\min_{\theta} L(\theta) + \lambda \sum_{s \in S} \frac{1}{|s|} ||\theta_s||_2, \text{ where } ||\theta_s||_2 = \sqrt{\sum_{i \in s} \sum_{j=1}^{n} \theta_{ij}^2}, \quad (4)$$

$s \in S$ is the index set of a specific structure in feature as described in Section 4.1, and $|s|$ represents its dimension. Compared with constructing groups by features, in our regularizer, we construct groups by leveraging the structure of the pre-trained model. This formalization can significantly reduce the combinations of selection and thus prevent overfitting. With such an advantage, our method achieves superior performance with a small number of samples under high-dimensional features.

4.2.2 Structured Non-linear Transformation

Our discovery in Section 3.3 shows non-linear transformation can facilitate transfer performance for difficult tasks, while could reduce performance for easy adaptation tasks, mostly due to over-fitting. This means that we need a model that allows flexibly incorporating non-linearity to a linear model. We propose a structured non-linearity regularizer that nicely integrates feature selection (i.e. choosing the subset of informative features from the pre-trained model) with model selection (i.e. choosing either a linear model or non-linear model for prediction) inspired by hierarchical sparse modeling [28, 42]. Specifically, we construct our model by the combination of a linear model and a neural network:

$$f(x) = \theta^\top x + f_W(x), \quad (5)$$

where $\theta \in \mathbb{R}^{d \times n}$ is a linear model, and f_W represents an MLP with model weights W. We use $W^{(1)}$ to represent the input layer of the MLP, and the parameters of rest layers are denoted as $W^{(2:)}$.

We utilize the feature selection result $||\theta||_2$, the ℓ_2 norm of weights assigned to the selected features, to regularize the complexity of the non-linear model:

- When a small number of features are good enough for prediction, we expect a small value for $||\theta||_2$. By using $||\theta||_2$ to control the complexity of non-linear model, we expect the resulting model to be mostly linear.
- With increasing $||\theta||_2$, more diverse structures from the pre-trained model are needed for prediction. By using $||\theta||_2$ to regularize the size of non-linear model, we expect the resulting model to be more non-linear.

A way to achieve such control is to add a constraint to the objective function in Eq.(4) as follows:

$$\begin{aligned} &\min_{\theta, W} L(\theta, W) + \lambda \sum_{s \in S} ||\theta_s||_2 \\ \text{subject to } &||W_s^{(1)}||_2 \le M_1 ||\theta_s||_2, \ s = 1...|S|, \\ &||W^{(2:)}||_2 \le M_2 ||\theta||_2. \end{aligned} \quad (6)$$

There are two constraints in Eq.(6): the first constraint is used to maintain the input feature sparsity of the neural network, and the second term is used to control the non-linearity through the linear model. To achieve efficient implementation, we convert the constraint in Eq.(6) into a regularization term, which we call **structured non-linearity regularizer**:

$$\begin{aligned} \Omega(\theta, W) = &\sum_{s \in S} \max\{||W_s^{(1)}||_2 - M_1 ||\theta_s||_2, 0\} \\ &+ \max\{||W^{(2:)}||_2 - M_2 ||\theta||_2, 0\}. \end{aligned} \quad (7)$$

This regularization term means that, when $||W||_2 > M||\theta||_2$, W receives penalty until it achieves similar complexity as $M||\theta||$. Then our final objective function is:

$$\min_{\theta, W} L(\theta, W) + \lambda_1 \sum_{s \in S} ||\theta_s||_2 + \lambda_2 \Omega(\theta, W). \quad (8)$$

We use λ to control the penalty strength, and simply set $\lambda_2 = 0.1$ works well in our experiments.

Loss function. Since our model $f(x)$ consists of both linear and non-linear parts, we decouple these two components in loss function. This modification enables the structured sparsity regularizer to control the complexity of $||\theta||$ without the interference from the non-linear part:

$$L(\theta, W) = \text{CE}(\theta^\top x, y) + \text{CE}(\text{sg}(\theta^\top x) + f_W(x), y), \quad (9)$$

where y is the one-hot label vector of x, CE is cross-entropy loss and sg means stop-gradient.

Table 1. Median test accuracy (%) over 3 seeds on the VTAB-1k benchmark using ViT-B/16 pretrained on ImageNet-21k. * indicates results are obtained from [23] and † indicates results are obtained from [16].

	Natural								Specialized					Structured									
	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Mean (Natural)	Camelyon	EuroSAT	Resisc45	Retinopathy	Mean (Specialized)	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Elev	Mean (Structured)	Mean (All)
Tuning Methods																							
Scratch†	7.6	19.1	13.1	29.6	6.7	19.4	2.3	14.0	71.0	71.0	29.3	72.0	60.8	31.6	52.5	27.2	39.1	66.1	29.7	11.7	24.1	35.3	32.8
Fine-tuning*	68.9	87.7	64.3	97.2	86.9	87.4	38.8	75.9	79.7	95.7	84.2	73.9	83.7	56.3	58.6	41.7	65.5	57.5	46.7	25.7	29.1	47.6	65.6
Bias*	72.8	87.0	59.2	97.5	85.3	59.9	51.4	73.3	78.7	91.6	72.9	69.8	78.3	61.5	55.6	32.4	55.9	66.6	40.0	15.7	25.1	44.1	62.0
VPT*	78.8	90.8	65.8	98.0	88.3	78.1	49.6	78.5	81.8	96.1	83.4	68.4	82.4	68.5	**60.0**	**46.5**	**72.8**	**73.6**	**47.9**	32.9	37.8	55.0	69.4
Probing Methods																							
Linear	77.2	86.8	66.6	99.1	88.8	35.7	55.1	72.7	78.4	89.6	72.2	73.7	78.5	31.0	29.9	35.2	54.9	11.7	27.7	14.4	23.7	28.6	55.4
Side-Tuning*	60.7	60.8	53.6	95.5	66.7	34.9	35.3	58.2	58.5	87.7	65.2	61.0	68.1	27.6	22.6	31.3	51.7	8.2	14.4	9.8	21.8	23.4	45.6
Head2Toe	75.3	**90.9**	**75.0**	**99.5**	86.1	**83.5**	50.9	80.2	84.4	95.7	**84.4**	74.3	84.7	51.7	59.4	44.0	65.6	47.1	40.3	32.5	41.1	47.7	67.5
Linear$_{Struc.}$	72.9	89.5	72.5	98.9	83.6	71.6	42.7	76.0	84.3	96.0	81.8	**75.0**	84.3	48.2	52.5	39.4	66.7	48.7	36.7	32.3	37.7	45.3	64.8
Linear$_{Struc.}$ w/ FSR	78.2	90.1	73.4	99.3	88.1	71.7	45.4	78.0	83.9	96.1	81.9	74.1	84.0	48.5	52.8	39.6	64.4	51.4	38.6	32.9	37.0	45.7	65.7
Linear$_{Struc.}$ w/ SSR	79.0	90.2	74.1	99.1	**90.4**	73.4	54.2	80.1	84.6	**96.3**	83.8	74.4	**84.8**	48.8	52.9	39.8	66.9	48.7	38.9	32.5	37.9	45.8	66.6
MLP$_{Struc.}$	66.8	88.7	70.3	97.4	82.1	71.1	33.3	72.8	81.6	95.9	77.0	73.5	82.0	77.2	57.7	39.3	66.5	72.3	44.1	32.7	47.7	54.7	67.1
SMP w/o SNR	71.3	89.6	72.3	99.0	85.6	72.7	45.0	76.5	84.2	95.9	81.9	72.8	83.7	77.4	57.1	40.8	67.1	72.4	44.2	32.6	48.4	55.0	69.0
SMP	**79.3**	**90.9**	74.9	99.3	**90.4**	75.0	**55.3**	**80.7**	**84.8**	**96.3**	83.1	**75.0**	**84.8**	**77.5**	58.0	40.8	67.5	72.5	44.5	**33.0**	**49.0**	**55.4**	**70.9**

5. Experiments

In this section, we evaluate the proposed structured model probing method on multiple datasets across various tasks, and provide detailed analysis of the proposed method.

Implementation details. We evaluate SMP on the popular vision backbone, ViT-B/16 [15] pre-trained on ImageNet-21k [12], which is widely used across relative research works [16, 23]. We provide details of probing model and hyperparameter selection in the supplementary material.

Compared methods. We compare SMP with other commonly used and state-of-the-art tuning (Fine-tuning, Bias [6] and VPT [23]) and probing methods (Linear Probing, Side-Tuning [45] and Head2Toe [16]). All hyperparameters of baseline methods are carefully tuned, and further information is provided in supplementary material.

5.1. Experiments on VTAB-1k

Experimental results on the VTAB-1k benchmark (Table 1) indicate that SMP outperforms other methods, achieving a mean performance of 70.9%, 3.4% higher than Head2Toe and 1.5% higher than VPT. SMP also significantly outperforms Linear Probing, Bias, and Side-Tuning. SMP achieves optimal results in all three dataset groups. Especially in Structured group, SMP attains a mean accuracy of 55.4%, surpassing other probing methods by a significant margin. This demonstrates the incorporation of non-linear transformation is important for difficult adaptation tasks.

Ablation studies. We provide results of baseline methods that contribute to SMP, to validate the efficacy of structured regularization and the contribution of each component.

Linear$_{Struc.}$ represents extracting features by structured feature extraction proposed in Section 4.1, then training a linear classifier. Compare to Linear Probing, it gets 9% average performance gain, indicating that solely extracting features from the last layer is inadequate for model probing.

We further compare feature sparsity regularization in Eq.(3) (represents by Linear$_{Struc.}$+ FSR) to the proposed structured sparsity regularizer (represents by Linear$_{Struc.}$+ SSR). Structured sparsity regularizer outperforms feature sparsity regularization, indicating that feature selection can mitigate the overfitting caused by high dimensions, and leveraging model structure as a prior can bring extra benefits.

We apply non-linear transformation on extracted features, represents by MLP$_{Struc.}$. We find the performance on structured group gets a significant improvement over Linear$_{Struc.}$, demonstrating the importance of non-linear transformation on difficult adaptation tasks. However, the performance on easy tasks, i.e., natural and specialized groups, is degraded, suggesting the risk of overfitting from non-linear transformation. We solve this problem by leveraging structured non-linearity regularizer (SNR), which can automatically control the non-linearity through the norms of selected structures.

5.2. Experiments on Few-shot Learning

In this section, we evaluate our method in few-shot learning scenario, where the number of training samples varies from 1 to 16 shots per class following existing studies. Five fine-grained recognition tasks are chosen, i.e., Food101 [4], CUB-200-2011 [39], Stanford Cars [26], Stanford Dogs [24]

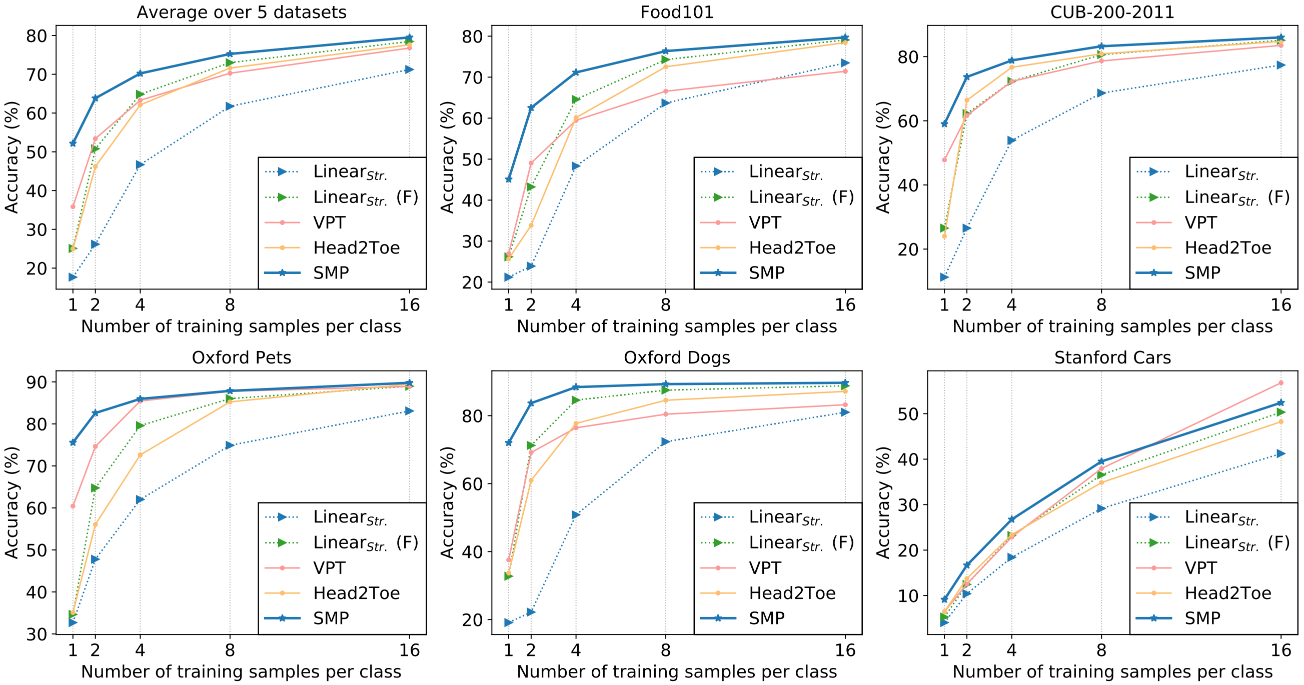


Figure 5. Performance of few-shot learning on fine-grained visual recognition tasks. Linear$_{Str.}$ (F) indicates Linear$_{Struc.}$ w/ FSR.

and Oxford Pets [35]. We compared SMP with two strong baselines, Linear$_{Struc.}$ and Linear$_{Struc.}$ w/ FSR, and two prominent methods, VPT (tuning) and Head2Toe (probing).

Experimental results presented in Figure 5 show that SMP consistently outperforms other methods in most cases, demonstrating its effectiveness in few-shot scenario, and its ability to improve performance with increasing training data. Notably, feature sparsity regularization performs worse than SMP, highlighting the importance of leveraging pre-trained model structure as prior knowledge in low-shot scenario.

5.3. Experiments on More Transfer Scenarios

Larger downstream datasets. We provide average results on 9 full-size downstream datasets in Table 2. It can be observed even in the larger data regime, SMP still achieves competitive performance compared with tuning methods. Full results are presented in Suppl. B.1.

Different pre-trained models. We present VTAB-1k results on ViT-B/16 and ViT-L/14 pretrained by CLIP in Suppl. B.2. Despite the usage of stronger pre-trained models, SMP continuously outperforms baseline methods due to the incorporation of diverse features and non-linear transformation, as well as its flexible framework.

Different architectures. SMP is a versatile method that can be applied to convolutional neural networks as well. In Suppl. B.3, we present the VTAB-1k results obtained from ImageNet pre-trained ResNet-50 models, where SMP outperforms other methods and achieves superior performance.

Table 2. Average accuracy (%) on 9 full-size downstream datasets.

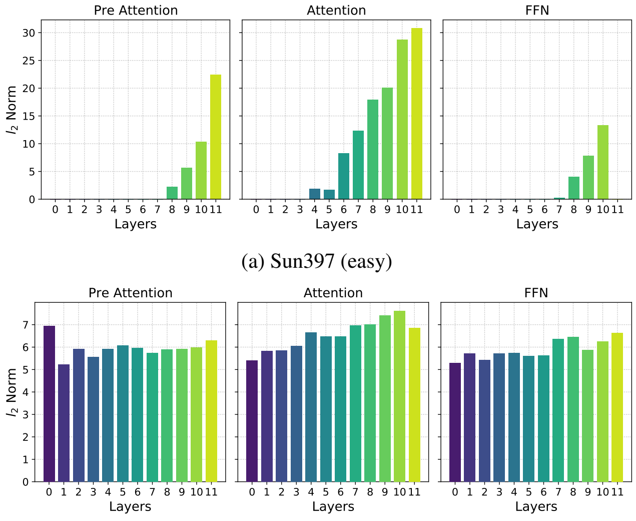
	Fine-tuning	VPT	Linear	SMP
Average Accuracy	88.4	**88.7**	81.7	**88.7**

5.4. Analysis and Discussion

In this section, we conduct various experiments to analyze the components of the proposed SMP method. Full results of this section are presented in Suppl. C.

Visualization of group norms. We visualize the ℓ_2 norm of each structure on different tasks in Figure 6, which helps us to understand the behavior of the proposed structured regularization. For Sun397, an easy adaptation task, we can observe the structures from deep layers have larger norms in Figure 6a. This validates our analysis in Section 3.2, that for easy adaptation tasks, features extracted from deep layers contain more information. For Clevr-Dist, a difficult adaptation task, the structures have similar norms overall, which means for difficult tasks, information contained in intermediate layers is also important, and diverse features can improve the performance, as we find in Section 3.2. This motivates the proposed structured non-linearity regularization, utilizing ℓ_2 norm to control the complexity of non-linear model.

Retraining probing model by gradually incorporating structures. We retrain a probing model by gradually incorporating structures selected based on the magnitude of their ℓ_2 norm, to validate their contribution to the prediction. Results are presented in Figure 7. For Sun397, an easy adaptation task, we can see in Figure 7a that there are redun-



(a) Sun397 (easy)

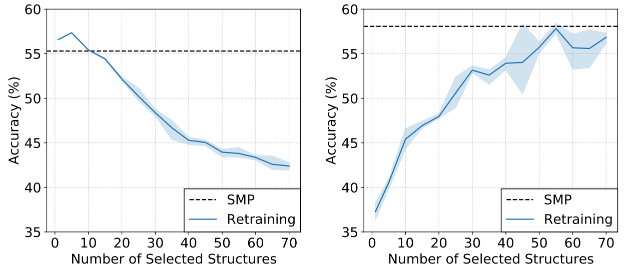
(b) Clevr-Dist (difficult)

Figure 6. ℓ_2 norm of structure groups on different tasks.

dant structures. And with a few selected structures, we can achieve comparable results compared with the original SMP, justifying the zero norm of some structured observed after performing structure regularization in Figure 6a. However, for Clevr-Dist, a difficult adaptation task, comparable performance is achieved until selecting 55 structures, suggesting intermediate structures also contribute to the final prediction. This demonstrates the necessity of proposed structured regularizer, which resolves the contradiction between easy and difficult domain adaptation tasks.

Impact of feature aggregation designs. The primary objective of performing token and channel aggregation is to minimize the redundancy in the extracted raw features, while maintaining feature diversity in the compact aggregated representation. To achieve this, two types of aggregation are designed to preserve channel and spatial information, respectively. We provide ablation studies in Suppl. C.5. Overall, channel information (aggregating along tokens) is essential for all datasets, while spatial information (aggregating along channels) is beneficial for tasks that are sensitive to spatial location, e.g., dSpr-Loc and sNORB-Azim.

Impact of different structures. To increase the feature diversity, we incorporate features extracted from various structures, coupled with a structured sparse regularizer to conduct model fitting and feature selection simultaneously. This flexible framework enables us to avoid the costly manual selection of candidate structures. We provide analysis of structures in Suppl. C.6. The results show that incorporating all candidate structured features yields best mean results, which demonstrates the efficacy of our method.



(a) Sun397 (easy)

(b) Clevr-Dist (difficult)

Figure 7. Retraining performance by gradually incorporating structures with larger ℓ_2 norm obtained by SMP.

Table 3. Cost comparison (Average results on VTAB-1k). * indicates relative results compared with Fine-tuning.

	Linear	FT	VPT	SMP
Adaptation time (rel.)*	0.05	1	0.97	0.15
Extra Inference FLOPs (G)	0	0	11.01	0.01
Extra parameters (M)	0.04	85.83	0.64	6.91

Cost of SMP. We compare the cost of SMP with baseline methods, including relative adaptation time, extra inference FLOPs, and extra parameters. The average results on VTAB-1k are shown in Table 3. SMP only needs 15% adaptation time on average compared with fine-tuning. On the contrary, the adaptation time of VPT is 97% of fine-tuning. This results in a significantly lower adaptation cost of SMP compared to tuning methods. SMP is also very efficient during inference time, and only introduces 0.01G inference FLOPs compared with fine-tuning. It can be noticed that VPT introduces extra 11.01G inference FLOPs, which is far more than SMP. SMP requires 6.91M extra parameters on average, which is the 8% of fine-tuning. Though the requirement of extra parameters is higher than VPT, we can leverage the retraining method proposed in Section 5.4, which can further reduce the extra parameters. Overall, SMP is a highly efficient method with low adaptation cost, fast inference, and fewer extra parameters, while maintaining superior performance compared to tuning-based methods.

6. Conclusion

In this paper, we present structured model probing, an effective yet efficient probing method for transfer learning. Our investigation reveals that model probing behaves differently for easy and difficult adaptation tasks. To mitigate this conflicting behavior, we propose structured model probing, a method that is able to achieve good performance in both cases. With the proposed structured regularization, we can construct a simple adaptation model for easy adaptation cases, and increase the complexity of the adaptation model for difficult cases. Our method outperforms state-of-the-art tuning methods while maintaining computational efficiency. Our work highlights the potential of model probing, which motivates researchers to better exploit the use large-scale pre-trained model and apply model probing to diverse tasks.

References

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6429–6438, 2019. 3

[2] Alessandro Achille, Aditya Golatkar, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. LQF: linear quadratic fine-tuning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15729–15739, 2021. 3

[3] Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. Better finetuning by reducing representational collapse. In *9th International Conference on Learning Representations (ICLR)*, 2021. 2

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, pages 446–461, 2014. 6

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. 1, 2

[6] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. 2, 6

[7] Shoufa Chen, Chongjian GE, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020. 2

[9] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9620–9629, 2021. 2

[10] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4109–4118, 2018. 3

[11] Mostafa Dehghani, Yi Tay, Anurag Arnab, Lucas Beyer, and Ashish Vaswani. The efficiency misnomer. In *10th International Conference on Learning Representations (ICLR)*, 2022. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 3, 6

[13] Aditya Deshpande, Alessandro Achille, Avinash Ravichandran, Hao Li, Luca Zancato, Charless C. Fowlkes, Rahul Bhotika, Stefano Soatto, and Pietro Perona. A linearized framework and a new benchmark for model selection for fine-tuning. *arXiv:2102.00084*, 2021. 3

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019. 2

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations (ICLR)*, 2021. 3, 6

[16] Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C. Mozer. Head2Toe: Utilizing intermediate representations for better transfer learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 6009–6033, 2022. 1, 3, 5, 6

[17] Weifeng Ge and Yizhou Yu. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10–19, 2017. 2

[18] Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogério Schmidt Feris. SpotTune: Transfer learning through adaptive fine-tuning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4805–4814, 2019. 2

[19] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. In *10th International Conference on Learning Representations (ICLR)*, 2022. 1, 2

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 2, 3

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022. 2, 3

[22] 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 *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2790–2799, 2019. 2

[23] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim.

Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 6

[24] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition (CVPR Workshop)*, 2011. 6

[25] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2661–2671, 2019. 2, 3

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshop (ICCV Workshop)*, pages 554–561, 2013. 6

[27] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *10th International Conference on Learning Representations (ICLR)*, 2022. 2

[28] Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22:127:1–127:29, 2021. 5

[29] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3045–3059, 2021. 2

[30] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4582–4597, 2021. 2

[31] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5196–5205, 2022. 2

[32] Fangzhou Mu, Yingyu Liang, and Yin Li. Gradients as features for deep representation learning. In *8th International Conference on Learning Representations (ICLR)*, 2020. 3

[33] Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 7294–7305, 2020. 3

[34] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 2

[35] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3498–3505, 2012. 7

[36] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2227–2237, 2018. 3

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1, 2

[38] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. LST: Ladder side-tuning for parameter and memory efficient transfer learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[39] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011. 6

[40] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *10th International Conference on Learning Representations (ICLR)*, 2022. 2

[41] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Learning to prompt for continual learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 139–149, 2022. 2

[42] Xiaohan Yan and Jacob Bien. Hierarchical Sparse Modeling: A choice of two group lasso formulations. *Statistical Science*, 32(4):531 – 560. 5

[43] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1): 49–67, 2006. 5

[44] 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:1910.04867*, 2019. 1, 2, 3

[45] Jeffrey O. Zhang, Alexander Sax, Amir Zamir, Leonidas J. Guibas, and Jitendra Malik. Side-tuning: A baseline for network adaptation via additive side networks. In *European Conference on Computer Vision (ECCV)*, pages 698–714, 2020. 1, 2, 3, 6

[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16795–16804, 2022. 2

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[48] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021. 2