Institute of Media, Information, and Network
SHANGHAI JIAO TONG UNIVERSITY

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Adapting Shortcut with Normalizing Flow: An Efficient Tuning Framework for Visual Recognition

WED-PM-344

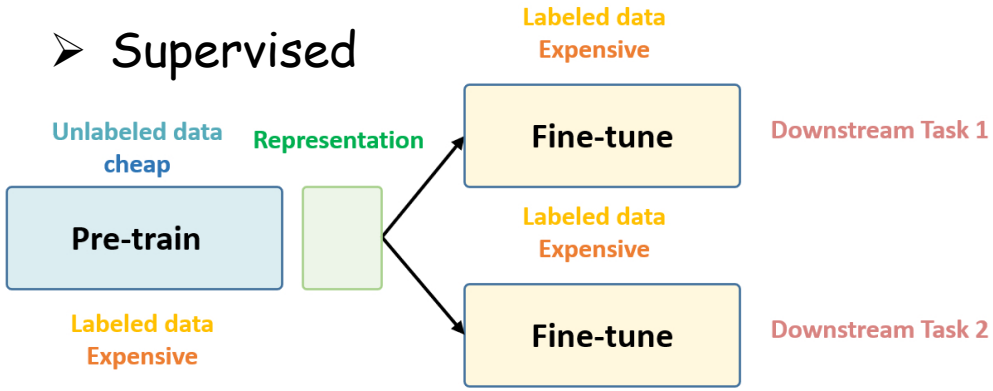
Yaoming Wang[1], Bowen shi[1], Xiaopeng Zhang[2], Jin Li[1], Yuchen Liu[1], Wenrui Dai[1], Chenglin Li[1], Hongkai Xiong[1], Tian Qi[2]

[1]Shanghai Jiao Tong University, China; [2] Huawei Cloud, China

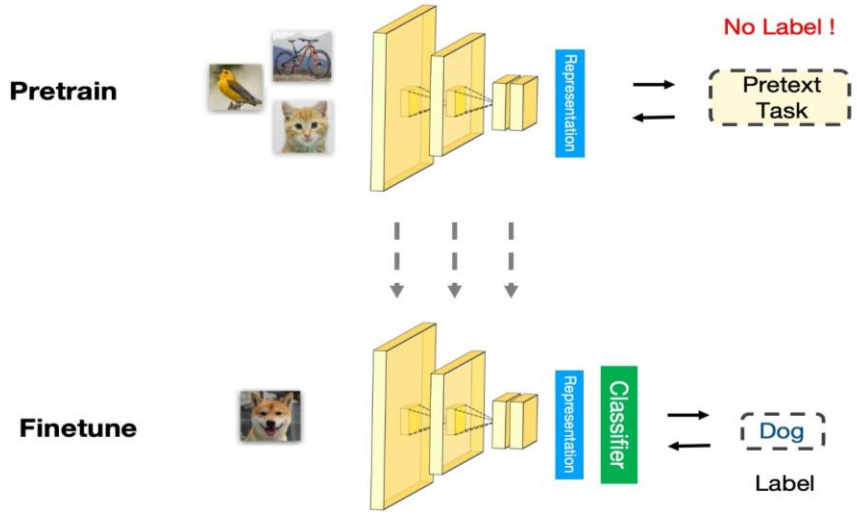
Background

Pretrain-then-finetune

- Supervised

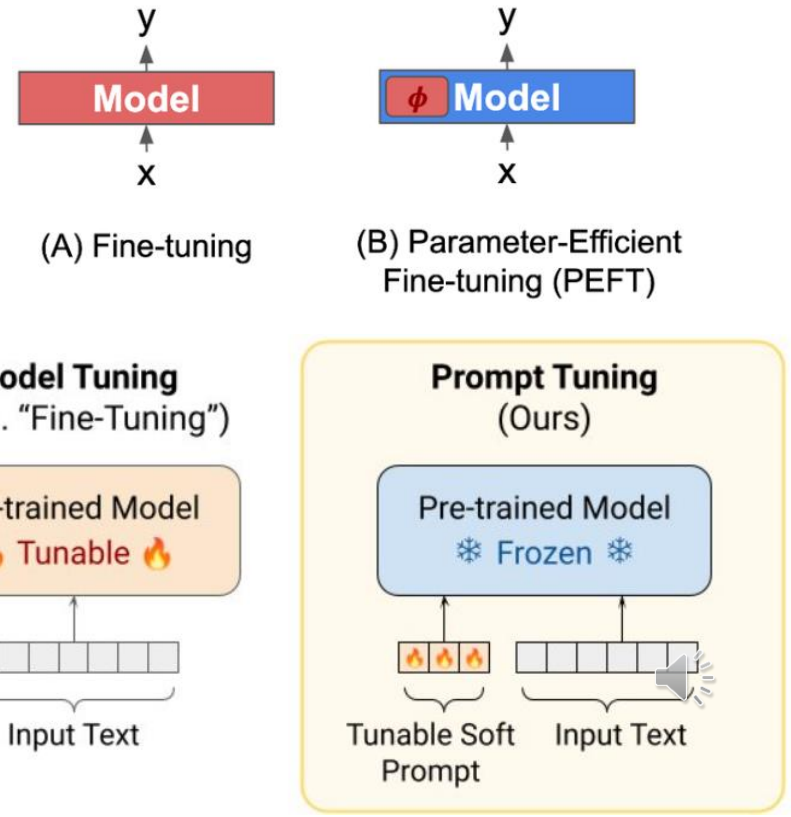


- Self-supervised



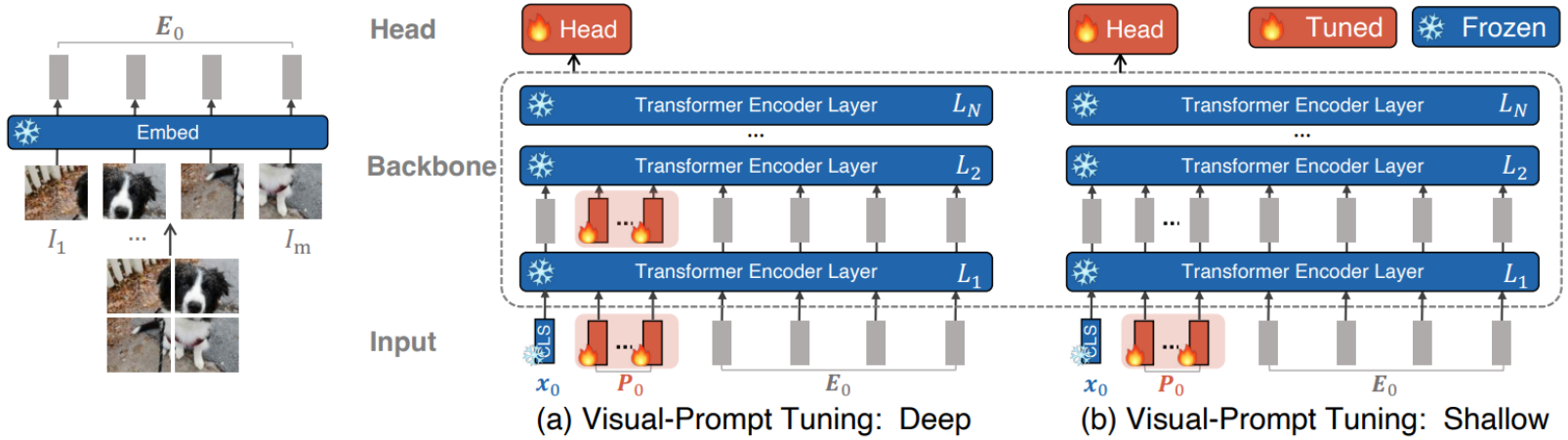
Parameter-efficient fine-tuning

- **freeze** most model parameters
- **update** a small number of model parameters

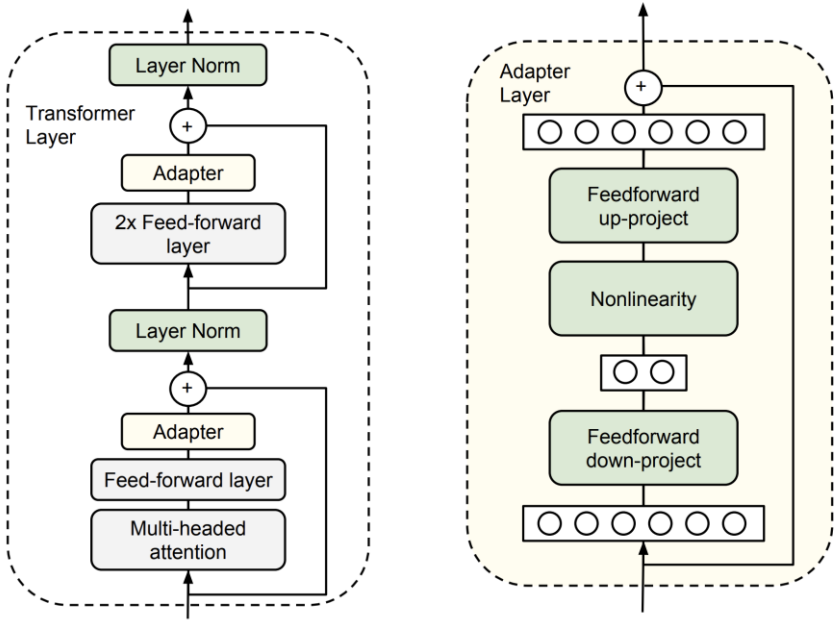


Related Work

Visual Prompt Tuning



Adapter

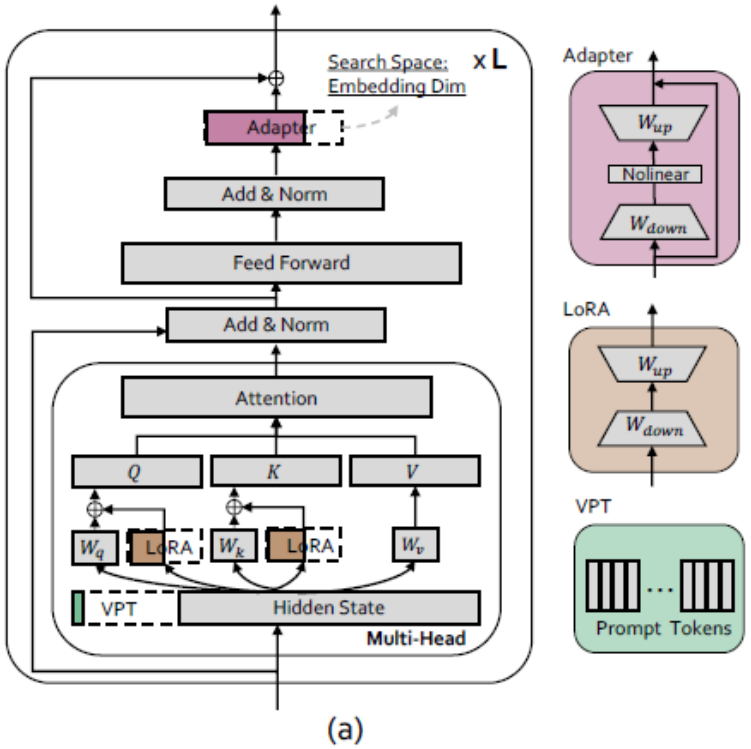


Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In International Conference on Machine Learning, pages 2790–2799. PMLR, 2019.

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.

Related Work

NOAH: Neural Prompt Search



Crossover

(a) Inputs:

1. Two random NOAH subnet architectures with 12 blocks

VPT | Adapter | LoRA

Block size = 12 | 12 | 12

A: {{10, 5, 5, 5, 10, 50, 0,..., 0},{5, 5, 10, 0, 0, 0,..., 0},{5, 5, 5, 0..., 0}}

↳ Embedding dimension

B: {{5, 10, 5, 50, 5, 50, 0,..., 0},{5, 5, 5, 0,..., 0},{5, 5, 5, 10..., 0}}

(b) Implementation:

{ A_{VPT}, $A_{Adapter}$, A_{LoRA} } { B_{VPT}, $B_{Adapter}$, B_{LoRA} }

A_{VPT} $B_{Adapter}$ A_{LoRA}

Randomly select a prompt module from A or B

Mutation

(a) Inputs:

1. One random NOAH subnet architecture with 12 blocks

A: {{10, 5, 5, 5, 10, 50 , 0,...,0},{5, 5, 10, 0, 0, 0,...,0},{5, 5, 5, 0...,0}}

2. Search space
   - Embedding dimension: {5, 10, 50, 100}
   - Depth number: {0, 3, 6, 9, 12}

(b) Implementation:

Depth number = 6 | 3 | 3

A: {{10, 5, 5, 5, 10, 50 , 0,...,0},{5, 5, 10, 0, 0, 0,...,0},{5, 5, 5, 0...,0}}

New depth = 3 | New depth = 6 | New depth = 3

{{10, 5, 5, 0, 0, 0,..., 0}, { 5, 5, 10, 5, 10, 5,..., 0},{5, 5, 5, 0,...,0}}

Set to zero | Sample from the search space | Unchanged

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

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Adapting Shortcut with Normalizing Flows

Preliminary: Shortcut Connection

$$x_{l+1} = h(x_l) + \mathcal{F}(x_l, W_l). \quad y = x_l + \sum_{i=l}^{L} \mathcal{F}(x_i, W_i).$$

K-Lipschitz Constraint

Since the model parameters are pretrained on large datasets, it implies an implicit Lipschitz constraint:

$$||\mathcal{F}(x_l^2, W_l) - \mathcal{F}(x_l^1, W_l)|| \leq K_l ||x_l^2 - x_l^1||$$

Some layer that are unconstrainted will propagate error when transferring to downstream datasets.

In **parameter efficient fine-tuning**, the model needs to fit the unseen downstream dataset when the most of the model parameters **are frozen**.

Adapting shortcut

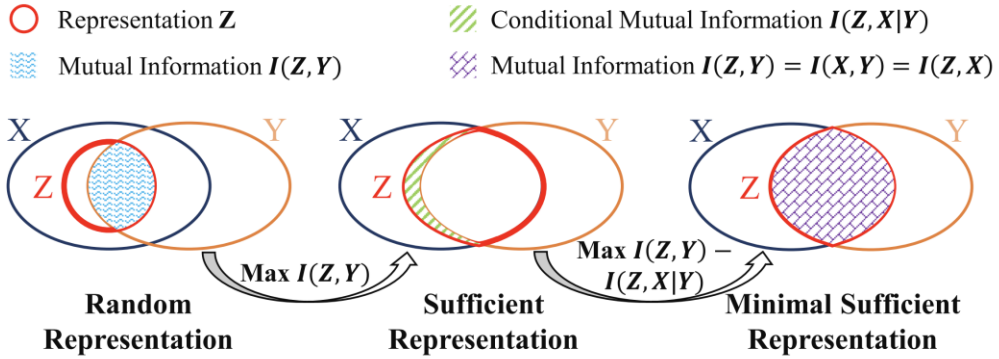
$$||y^2 - y^1|| \leq \left(\prod_{i=l}^{L} (K_i^{\phi} + K_i) \right) ||x_l^2 - x_l^1||$$

Table 6. Experiments on adapting on shortcut or residual. Here SNF-s is the abbreviation of SNF-shallow and SNF-d is the abbreviation of SNF-deep.

Method	Caltech101		Cifar100	
	SNF-s	SNF-d	SNF-s	SNF-d
on shortcut	**93.5**	**94.0**	**84.3**	**84.0**
on residual	90.8	92.2	83.2	83.7

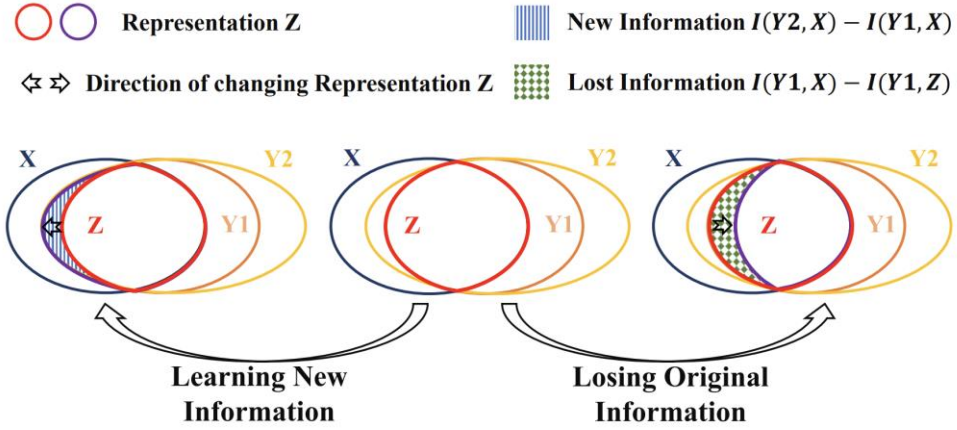
Adapting Shortcut with Normalizing Flows

Assumption 1. *For pretrained model, the learned feature z is the minimal sufficient representation for learning information of label y from input x as $I(x, y) = I(x, z) = I(y, z)$. The intuitive diagram is illustrated in Fig. 2.*



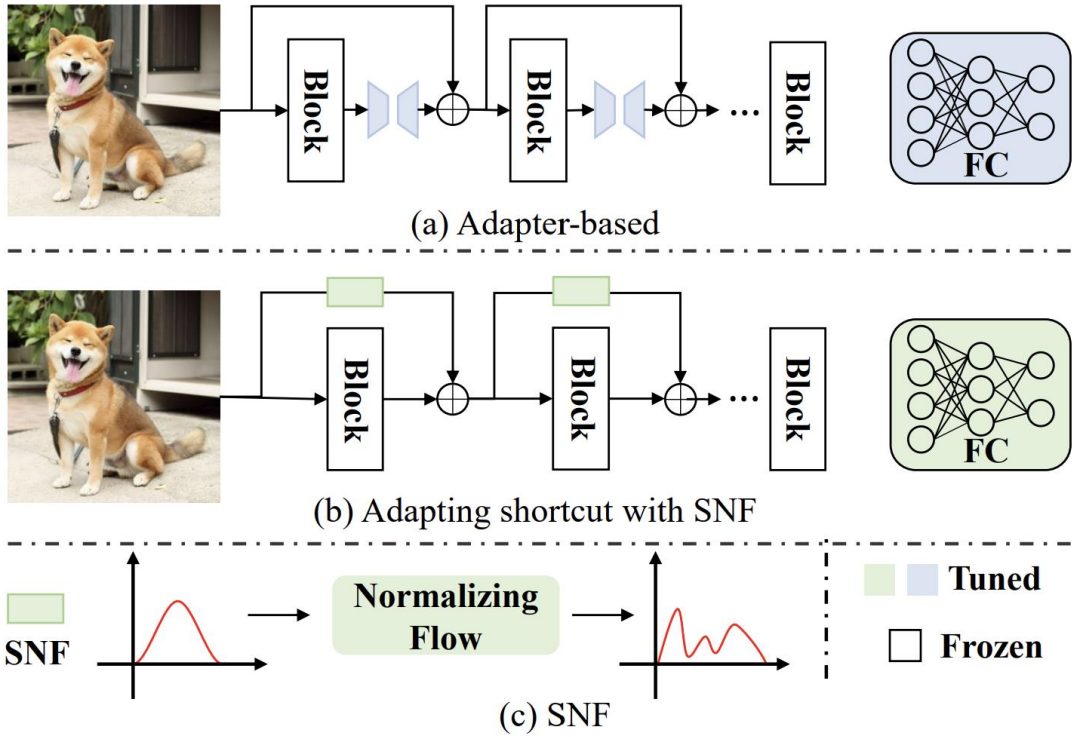
Assumption 2. *For downstream datasets, the label information y_2 is assumed to contain the label information y_1 of the pretrained large dataset as $y_1 \subseteq y_2$. And the mutual information $I(x, y_1) \leq I(x, y_2)$ for input x.*

Proposition 1. *For downstream datasets, we can hardly learn new information about the input when most features are frozen, then if we learn representation z_{new} in a way that information is lost, i.e., $I(z_{new}, x) < I(z, x)$, the representation will also lose information about the label y_2 as $I(z_{new}, y_2) < I(z, y_2)$.*



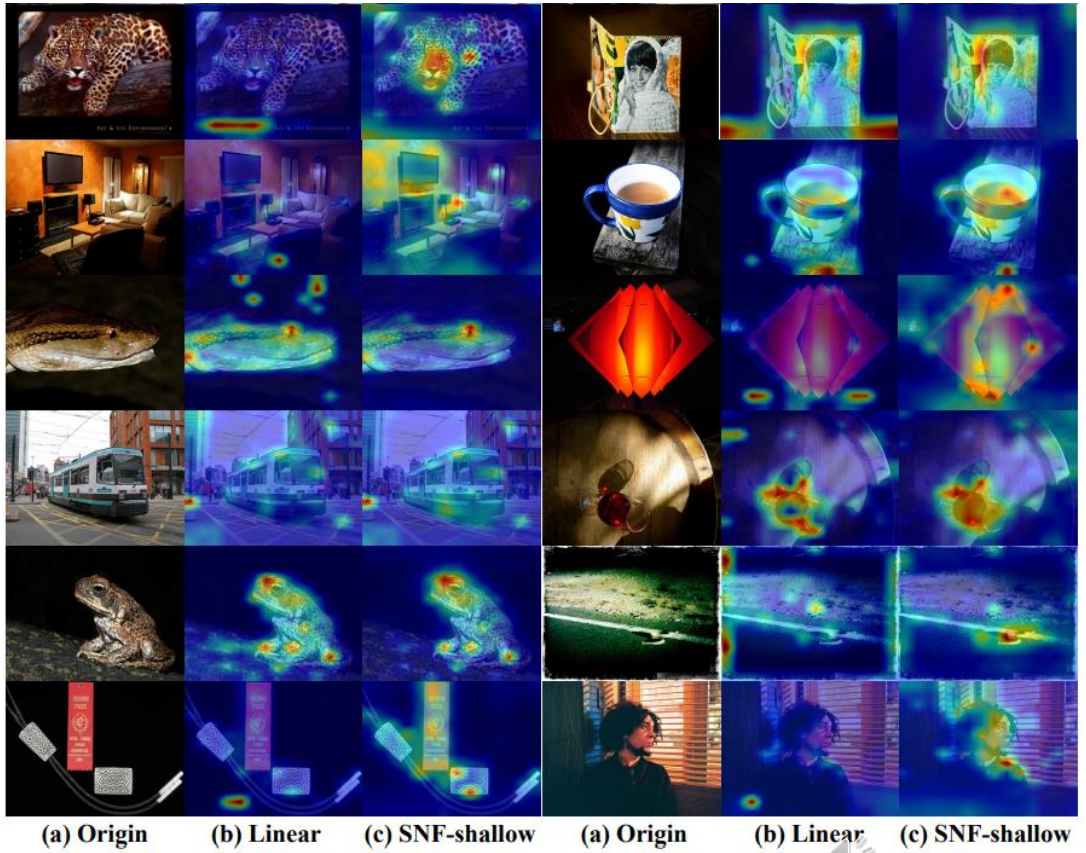
Proposition 2. *For $z_1, z_2 \in \mathbb{R}^N$, the mutual information $I(z_1, z_2)$ equals to $I(z_1, z_1)$ when the mapping $z_2 = f_\psi(z_1), f_\psi : \mathbb{R}^N \to \mathbb{R}^N$ is invertible and smooth.*

Adapting Shortcut with Normalizing Flows



$$z^J = f_J \circ \cdots \circ f_2 \circ f_1(z^0) \qquad \ln q(z^J) = \ln q(z^0) - \sum_{j=1}^{J} \ln \left| \det \frac{\partial f_j}{\partial z_{j-1}} \right|$$

$$f(z) = z + \lambda \cdot h(\gamma^T \cdot z + \beta) \qquad \ln \left| \det \frac{\partial f}{\partial z} \right| = \ln \left| I + \lambda^T \cdot h'(\gamma^T \cdot z + \beta) \cdot \gamma \right|$$



Experiments

Preliminary: Dataset Introduction

Dataset	Description	#Classes	Train	Val	Test
Fine-grained visual recognition tasks (FGVC)					
CUB-200-2011 [28]	Fine-grained bird species recognition	200	5,994	-	5,794
NABirds [26]	Fine-grained bird species recognition	555	23,929	-	24,633
Oxford Flowers [23]	Fine-grained flower species recognition	102	2,040	-	6,149
Stanford Dogs [15]	Fine-grained dog species recognition	120	12,000	-	8,580
Stanford Cars [7]	Fine-grained car recognition	196	8,144	-	8,041
Visual Task Adaptation Benchmark (VTAB-1k)					
Cifar100 [17]	Natural	100	800/1000	200	10,000
Caltech101 [6]		102			6,084
DTD [4]		47			1,880
Oxford-Flowers102 [22]		102			6,149
Oxford-Pets [24]		37			3,669
SVHN [21]		10			26,032
Sun397 [30]		397			21,750
Patch Camelyon [27]	Specialized	2	800/1000	200	32,768
EuroSAT [9]		10			5,400
Resisc45 [3]		45			6,300
Retinopathy [14]		5			42,670
Clevr/count [13]	Structured	8	800/1000	200	15,000
Clevr/distance [13]		6			15,000
DMLab [1]		6			22,735
KITTI-Dist [8]		4			711
dSprites/location [20]		16			73,728
dSprites/orientation [20]		16			73,728
SmallNORB/azimuth [18]		18			12,150
SmallNORB/elevation [18]		9			12,150

Few-shot Learning					
Food-101 [2]	Daily fine-grained food recognition	101	(1/2/4/8/16)*(#Classes)	-	25,250
Stanford Cars [16]	Daily fine-grained car recognition	196		-	8,041
Oxford-Flowers102 [22]	Daily fine-grained flower species recognition	102		-	6,149
FGVC-Aircraft [19]	Daily fine-grained Aircraft species recognition	100		-	3,333
Oxford-Pets [24]	Daily fine-grained pet species recognition	37		-	3,669
Domain Generalization					
ImageNet-V2 [25]	Variants of ImageNet with domain shifts	1000	-	-	10,000
ImageNet-Sketch [29]		1000	-	-	50,889
ImageNet-A [11]		1000	-	-	7,500
ImageNet-R [10],		1000	-	-	30,000
Other Visual Recognition Tasks					
ImageNet [5]	Other general visual recognition	1,000	16*(#Classes)	50,000	150,000
Cifar-100 [17]		100	50,000	-	10,000

Experiments

Table 2. Per-task fine-tuning results on VTAB-1k benchmark. The backbone is ViT-B/16, and we ignore the linear layer when calculate the amount of learnable parameters. Bold represents the best performance, underlined represents the second best performance.

Methods	#Params(M)	Natural							Specialized				Structured								Average
		Cifar100	Caltech101	DTD	Flower102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	sNORB-Ele	
Traditional Fine-tuning																					
Full	85.8	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	65.57
Linear	0	63.4	85.0	63.2	97.0	86.3	36.6	51.0	78.5	87.5	68.5	74.0	34.3	30.6	33.2	55.4	12.5	20.0	9.6	19.2	52.94
Other approaches																					
Bias [2]	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	62.1
VPT-shallow [16]	0.07	77.7	86.9	62.6	97.5	87.3	74.5	51.2	78.2	92.0	75.6	72.9	50.5	58.6	40.5	67.1	68.7	36.1	20.2	34.1	64.9
VPT-deep [16]	0.60	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	69.4
LoRA [15]	0.25	67.1	91.4	69.4	98.8	90.4	85.3	54.0	84.9	95.3	84.4	73.6	**82.9**	**69.2**	49.8	78.5	75.7	47.1	31.0	44.0	72.3
NOAH [42]	0.39	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**	73.2
Ours																					
SNF-shallow	**0.036**	**84.3**	93.5	**72.7**	**99.3**	**91.3**	89.5	54.3	**85.7**	96.2	**85.5**	74.1	81.1	61.0	48.9	82.3	75.4	**49.3**	31.1	41.7	73.5
SNF-deep	0.25	84.0	**94.0**	**72.7**	**99.3**	**91.3**	**90.3**	**54.9**	**87.2**	**97.3**	**85.5**	74.5	82.3	63.8	49.8	**82.5**	75.8	49.2	31.4	42.1	**74.1**

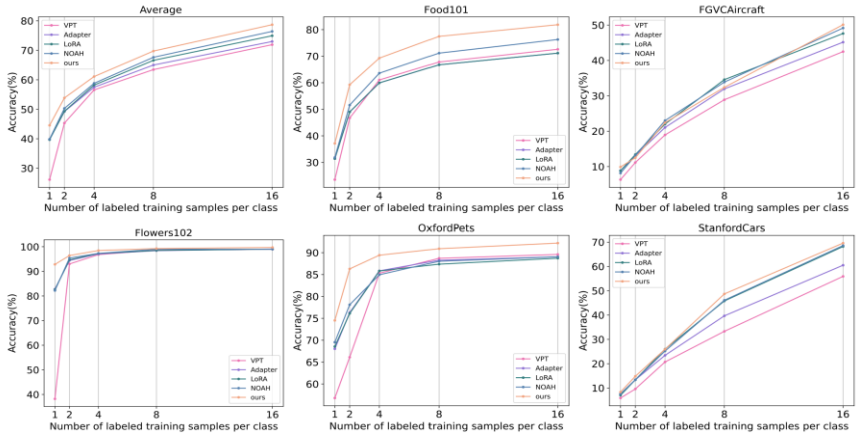


Figure 4. Results of few-shot learning on five daily visual recognition datasets (Best viewed in color).

Table 3. Results on domain generalization, the source domain is ImageNet-1k. The backbone is ViT-B/16.

Method	Source	$\rightarrow \mathcal{D}_{V2}$	$\rightarrow \mathcal{D}_{Sketch}$	$\rightarrow \mathcal{D}_{A}$	$\rightarrow \mathcal{D}_{R}$
Adapter [14]	70.5	59.1	16.4	5.5	22.1
VPT [16]	70.5	58.0	18.3	4.6	23.2
LoRA [15]	70.8	59.3	20.0	6.9	23.3
NOAH [42]	71.5	66.1	24.8	11.9	28.5
Ours	**78.5**	**66.4**	**27.0**	**12.2**	**30.4**

Experiments

Table 1. Per-task fine-tuning results on FGVC datasets. The backbone is ViT-B/16, and we ignore the linear layer when calculate the amount of learnable parameters. Bold represents the best performance, underlined represents the second best performance.

Methods	Params(M)	CUB-200-2011	NABirds	Oxford Flowers	Stanford Dogs	Stanford Cars	Mean
Linear probing	0	85.3	75.9	97.9	86.2	51.3	79.32
Full-tuning	85.8	87.3	82.7	98.8	89.4	84.5	88.54
Sidetune [41]	-	84.7	75.8	96.9	85.8	48.6	78.35
Adapter [14]	0.23	87.1	84.3	98.5	89.8	68.6	85.67
VPT-shallow [16]	0.08	86.7	78.8	98.4	**90.7**	68.7	84.62
VPT-deep [16]	0.66	88.5	84.2	99.0	90.2	83.6	89.11
Results of our methods							
SNF-shallow	**0.036**	90.0	86.7	99.6	89.3	83.3	89.78
SNF-deep	0.25	**90.2**	**87.4**	**99.7**	89.5	**86.9**	**90.74**

Table 4. Results on different backbones.SNF-s† and SNF-d† are the abbreviation of SNF-shallow and SNF-deep, respectively.

Methods	ViT-B			ViT-L			Swin-B			ResNet-50			ResNet-101		
	DTD	EuroSAT	#Params(M)	DTD	EuroSAT	#Params(M)	DTD	EuroSAT	#Params(M)	DTD	EuroSAT	#Params(M)	DTD	EuroSAT	#Params(M)
Linear	63.2	87.5	0	67.7	94.6	0	77.1	94.0	0	60.5	88.4	0	58.7	87.8	0
Full	64.3	95.7	85.8	68.5	95.7	306	76.5	96.6	87.1	61.1	**95.9**	24.5	60.3	**95.6**	42.6
SNF-s†	**72.7**	96.2	0.036	73.5	96.6	0.10	77.4	96.8	0.05	62.8	95.6	0.03	62.5	**95.6**	0.06
SNF-d‡	**72.7**	**97.3**	0.25	**74.7**	**97.0**	0.69	**78.1**	**97.3**	0.34	**63.5**	95.7	0.18	**63.4**	95.1	0.43

Ablation

Table 5. Ablation study on the length of flow model.

Methods	Params(M)	NABirds	Stanford Cars	Mean
1-layer	**0.036**	86.7	83.3	85.00
3-layer	0.14	87.2	84.9	86.05
5-layer	0.25	**87.4**	86.9	87.15
7-layer	0.36	**87.4**	**87.5**	**87.45**

Table 6. Experiments on adapting on shortcut or residual. Here SNF-s is the abbreviation of SNF-shallow and SNF-d is the abbreviation of SNF-deep.

Method	Caltech101		Cifar100	
	SNF-s	SNF-d	SNF-s	SNF-d
on shortcut	**93.5**	**94.0**	**84.3**	**84.0**
on residual	90.8	92.2	83.2	83.7

Table 7. Experiments on bottleneck. The backbone is ViT-B/16.

Method	Caltech101	Cifar100	#Params
SNF-shallow	93.5	**84.3**	**0.036**
SNF-deep	**94.0**	84.0	0.25
Bottleneck on shortcut	90.9	79.6	0.29

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Institute of Media, Information, and Network

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Q & A

Many Thanks !