Gradient-based Parameter Selection for Efficient Fine-Tuning

Zhi Zhang[1,2]*, Qizhe Zhang[2]*, Zijun Gao[2], Renrui Zhang[3,4], Ekaterina Shutova[1], Shiji Zhou[5], Shanghang Zhang[2]†
[1]ILLC, University of Amsterdam,
[2]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University,
[3]MMLAB, CUHK, [4]Shanghai AI Laboratory
[5]Department of Automation, Tsinghua University,
{z.zhang, e.shutova}@uva.nl, {theia, shanghang}@pku.edu.cn

Parameter-efficient Fine-tuning (PEFT)

- **Challenging**

  Given the increasing size of the pre-trained models, fine-tuning all the parameters in the model is memory-intensive and data-inefficient, when fine-tuining multiple downstream tasks.

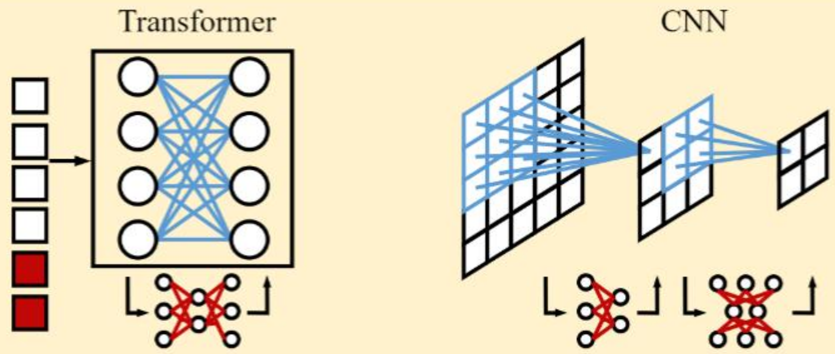
- **PEFT**

  Aims to fine-tune a minimal number of parameters to fit downstream tasks while keeps most of the parameters frozen.

Existing Methods and Limitations

- **Current typical methods**

  Adapter, LoRA, VPT.



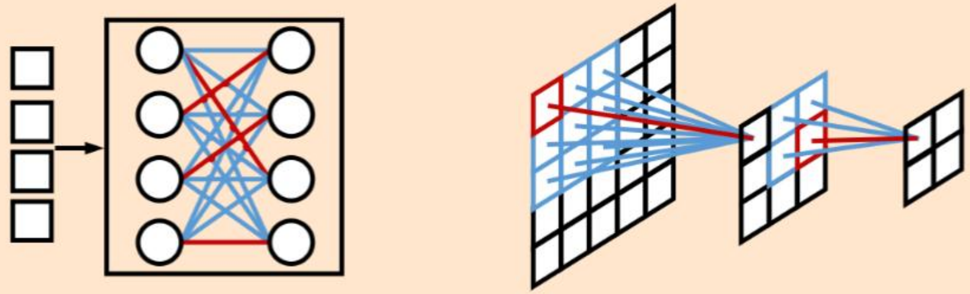
(a) Existing representative PEFT methods

- **Limitation**
  - Introducing additional learnable parameters into the backbone.
    - Disrupting the original architecture.
    - Increasing computational costs during training and/or inference stages.
  - Lacking generalizability across various model architectures.

Our method -- Overview

Overview:

- Selecte parameters from the original model
- Finetune the selected parameters and keep the remaining parameters fixed.



(b) Gradient-based Parameter Selection (Ours)

- Comparison:

Method	Mean Acc.	Params. (%)	Model Agnostic	No extra Train param.	No extra Infer params.	Task Adaptive
Full [43]	70.36	100	✓	✓	✓	✗
Linear [43]	58.48	0.08	✓	✓	✓	✗
Bias [92]	67.54	0.20	✓	✓	✓	✗
Adapter [36]	60.04	0.35	✗	✗	✗	✗
VPT [43]	73.53	0.76	✗	✗	✗	✗
LoRA [38]	75.16	0.90	✗	✗	✓	✗
SSF [58]	76.77	0.32	✗	✗	✓	✗
GPS (ours)	78.64	0.36	✓	✓	✓	✓

How to select parameters: **Two aspects**

- Importance for downstream tasks

  Gredient value: parameters with the largest gradient value indicate the fastest changes in the loss function along the gradient direction.

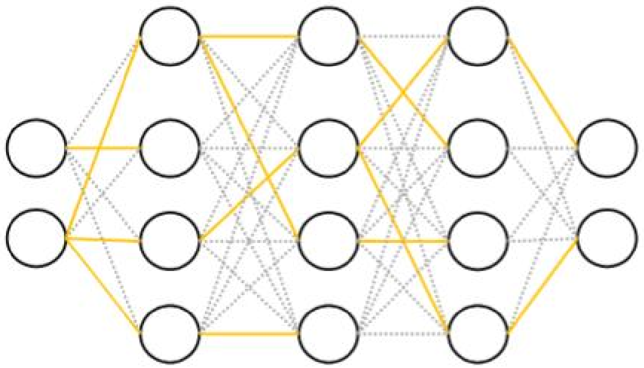
- Involving all neurons

  Every neuron in the network should be involved, as it can potentially adjust all neurons' states to better fit a task during finetuing stage.

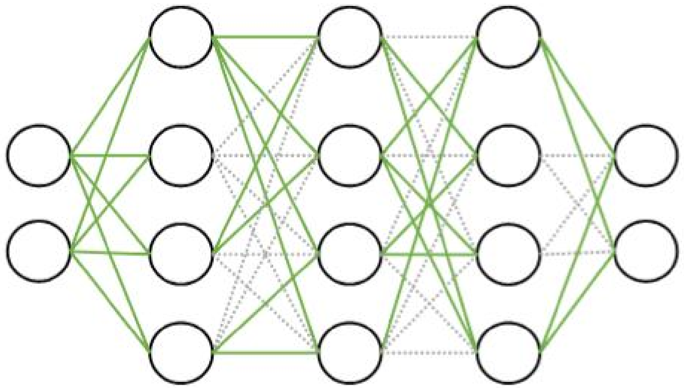
How to select parameters： **Combination**

Combination：

For certain task, we first calculate the gradient for all model parameters. Then, for each neuron in the network, we select the top-K connections (parameters) with the highest gradient value (modulus) among all input connections to that neuron.

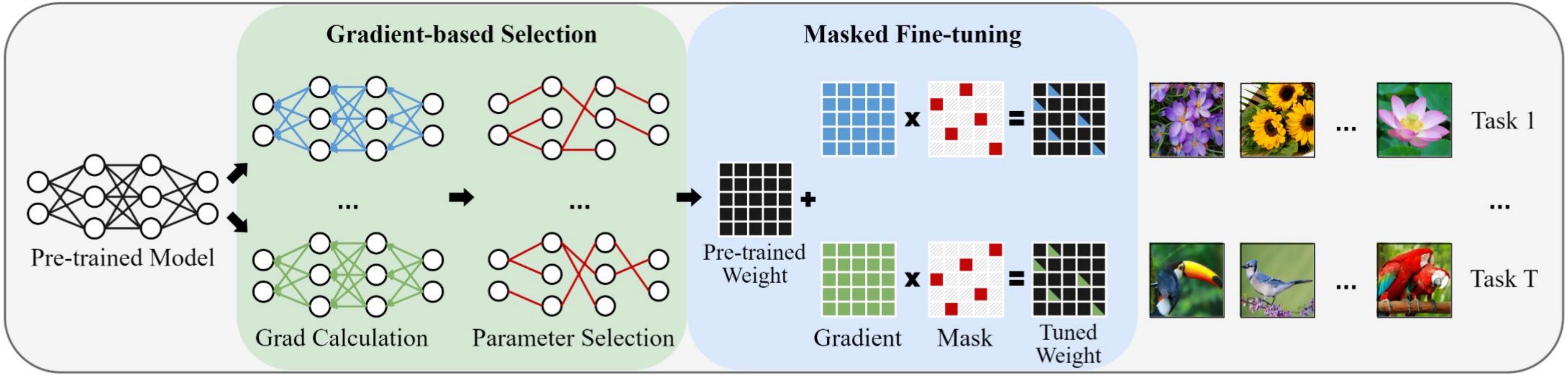


(a) One input connection



(b) Two input connections

Gradient-based Parameter (GPS) Selection for PEFT



Overview

- Parameter selection
- Masked fine-tuning

Experiments--Image Classification (FGVC)

Dataset	CUB -2011	NA- Brids	Oxford Flowers	Stan. Dogs	Stan. Cars	Mean Acc.	Params. (%)
Full [43]	87.3	82.7	98.8	89.4	84.5	88.54	100.00
Linear [43]	85.3	75.9	97.9	86.2	51.3	79.32	0.21
Bias [92]	88.4	84.2	98.8	91.2	79.4	88.40	0.33
Adapter [36]	87.1	84.3	98.5	89.8	68.6	85.66	0.48
LoRA [38]	85.6	79.8	98.9	87.6	72.0	84.78	0.90
VPT-Shallow [43]	86.7	78.8	98.4	90.7	68.7	84.62	0.29
VPT-Deep [43]	88.5	84.2	99.0	90.2	83.6	89.11	0.99
SSF [58]	89.5	85.7	99.6	89.6	89.2	90.72	0.45
SPT-Adapter [30]	89.1	83.3	99.2	91.1	86.2	89.78	0.47
SPT-LoRA [30]	88.6	83.4	99.5	91.4	87.3	90.04	0.60
GPS (Ours)	**89.9**	**86.7**	**99.7**	**92.2**	**90.4**	**91.78**	0.77

Table 2. Performance comparisons on FGVC with ViT-B/16 models pre-trained on ImageNet-21K.

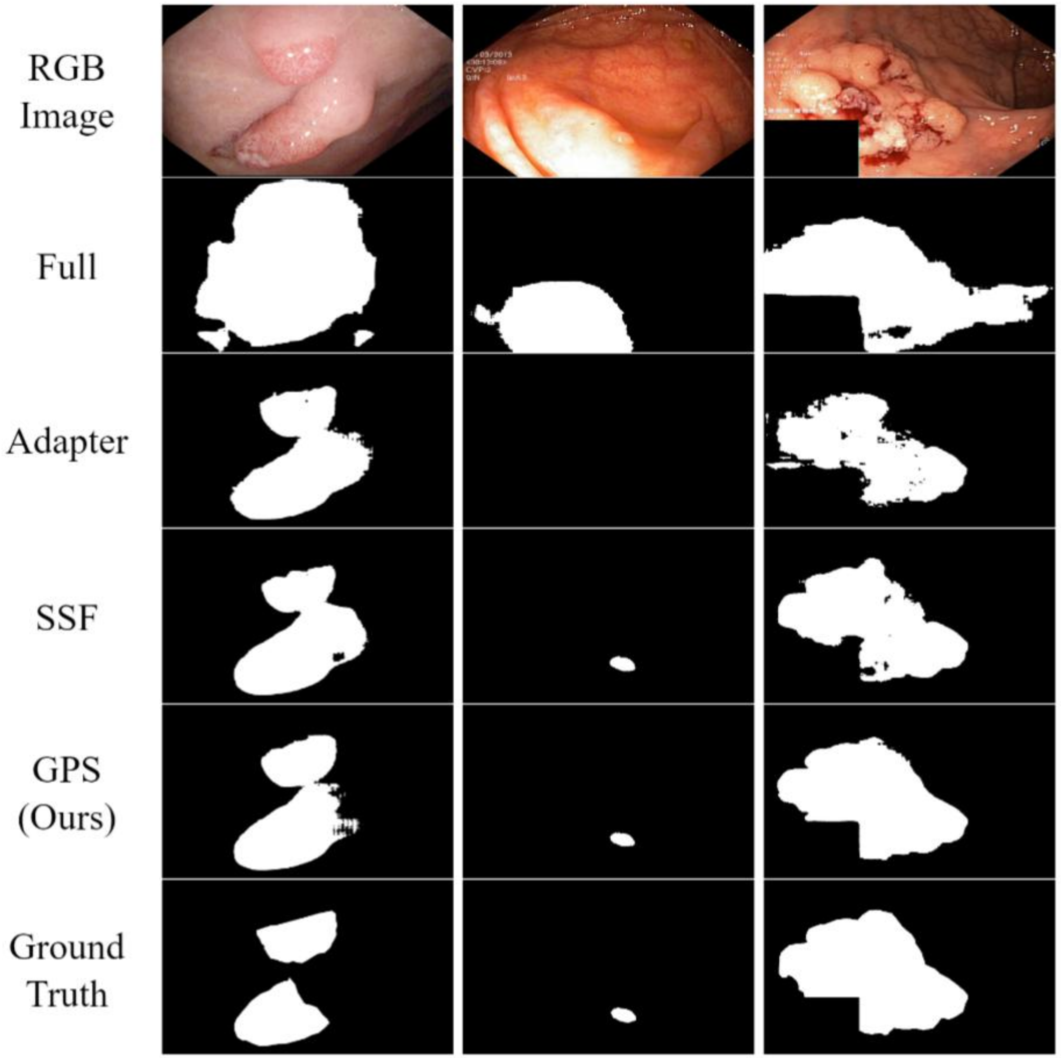
Experiments--Image Classification (VTAB)

Method \ Dataset	Natural							Specialized				Structured								VTAB	
	CIFAR-100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Patch Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr/count	Clevr/distance	DMLab	KITTI/distance	dSprites/loc	dSprites/ori	SmallNORB/azi	SmallNORB/ele	Mean Acc.	Mean Params. (%)
Full [43]	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	100.00
Linear [43]	63.4	85.0	64.3	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	53.00	0.05
Bias [92]	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	62.05	0.16
Adapter [36]	74.1	86.1	63.2	97.7	87.0	34.6	50.8	76.3	88.0	73.1	70.5	45.7	37.4	31.2	53.2	30.3	25.4	13.8	22.1	55.82	0.31
LoRA [38]	68.1	91.4	69.8	99.0	90.5	86.4	53.1	85.1	95.8	84.7	74.2	83.0	66.9	50.4	81.4	80.2	46.6	32.2	41.1	72.63	0.90
VPT-Shallow [43]	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	64.85	0.13
VPT-Deep [43]	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	69.43	0.70
SSF [58]	69.0	92.6	75.1	**99.4**	**91.8**	90.2	52.9	87.4	95.9	**87.4**	75.5	75.9	62.3	53.3	80.6	77.3	54.9	29.5	37.9	73.10	0.28
SPT-ADAPTER [30]	72.9	93.2	72.5	99.3	91.4	88.8	**55.8**	86.2	96.1	85.5	75.5	83.0	**68.0**	51.9	81.2	51.9	31.7	**41.2**	**61.4**	73.03	0.44
SPT-LoRA [30]	73.5	93.3	72.5	99.3	91.5	87.9	55.5	85.7	**96.2**	85.9	75.9	**84.4**	67.6	52.5	82.0	81.0	51.1	30.2	41.3	74.07	0.63
GPS (Ours)	**81.1**	**94.2**	**75.8**	**99.4**	91.7	**91.6**	52.4	**87.9**	**96.2**	86.5	**76.5**	79.9	62.6	**55.0**	**82.4**	**84.0**	**55.4**	29.7	46.1	**75.18**	0.25

Table 3. Performance comparisons on VTAB-1k with ViT-B/16 models pre-trained on ImageNet-21K.

Experiments--Semantic Segmentation (Polyp)

Method	mDice (↑)	mIoU (↑)	Params. (M)
Full [43]	71.1	55.7	93.8
Linear [43]	71.6	46.6	4.06
Bias [92]	86.5	69.1	4.16
Adapter [6]	84.8	66.7	4.12
SSF [58]	87.3	71.7	4.26
GPS (Ours)	**88.1**	**72.5**	4.22



Experiments--Different Architectures

Method \ Dataset	CUB-200 -2011	NABrids	Oxford Flowers	Stanford Dogs	Stanford Cars	Mean Acc.	Mean Params. (M)	Mean Params. (%)
ViT-B/16 + Full	87.3	82.7	98.8	89.4	84.5	88.54	85.98	100.00
ViT-B/16 + Linear	85.3	75.9	97.9	86.2	51.3	79.32	0.18	0.21
ViT-B/16 + SSF	89.5	85.7	99.6	89.6	89.2	90.72	0.39	0.45
ViT-B/16 + GPS (Ours)	**89.9**	**86.7**	**99.7**	**92.2**	**90.4**	**91.78**	0.66	0.77
Swin-B + Full	90.7	**89.8**	99.5	88.9	**93.2**	92.42	86.98	100.00
Swin-B + Linear	90.6	86.8	99.2	88.3	74.6	87.90	0.24	0.28
Swin-B + SSF	90.5	88.4	**99.7**	88.7	90.4	91.54	0.49	0.56
Swin-B + GPS (Ours)	**90.8**	88.9	**99.7**	**92.7**	90.7	**92.56**	0.83	0.95
ConvNeXt-B + Full	**91.2**	**90.4**	99.6	89.9	**94.1**	93.04	87.81	100.00
ConvNeXt-B + Linear	90.6	86.9	99.3	89.7	73.5	88.00	0.24	0.28
ConvNeXt-B + SSF	90.8	89.0	**99.7**	90.4	92.5	92.48	0.50	0.56
ConvNeXt-B + GPS (Ours)	91.0	89.6	**99.7**	**93.7**	92.6	**93.32**	0.79	0.90

Table 9. Performance comparisons on FGVC benchmark with different model architectures.

Thank you for your attention!!!