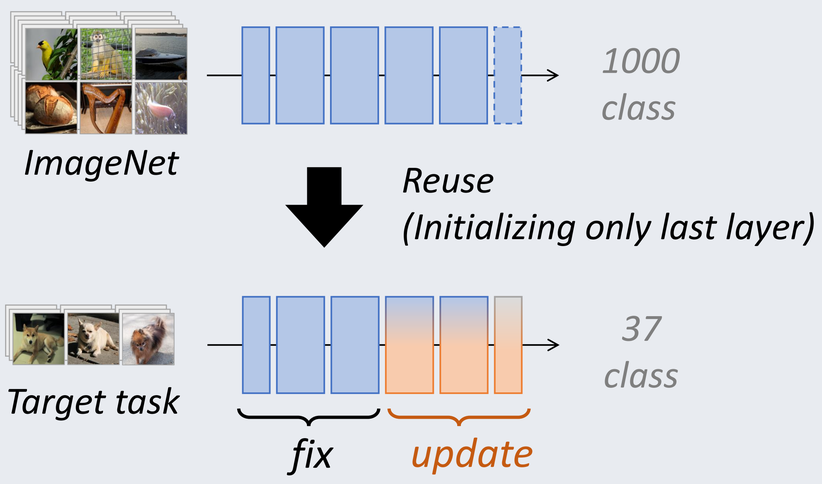
A Simple Finetuning Strategy Based on Bias-Variance Ratios of Layer-Wise Gradients

Mao Tomita[1], Ikuro Sato[2,1], Rei Kawakami[1], Nakamasa Inoue[1], Satoshi Ikehata[3,1], Masayuki Tanaka[1]

[1] Institute of Science Tokyo, Japan, [2] Denso IT Laboratory, Japan, [3] National Institute of Informatics, Japan

1. Background

Finetuning is a common deep learning technique for adapting a pretrained model to a new, related tasks.



2. Which layers should we finetune?

- Performance strongly depends on the choice of tuning layers.
- Which layers should be tuned depends on the target task.

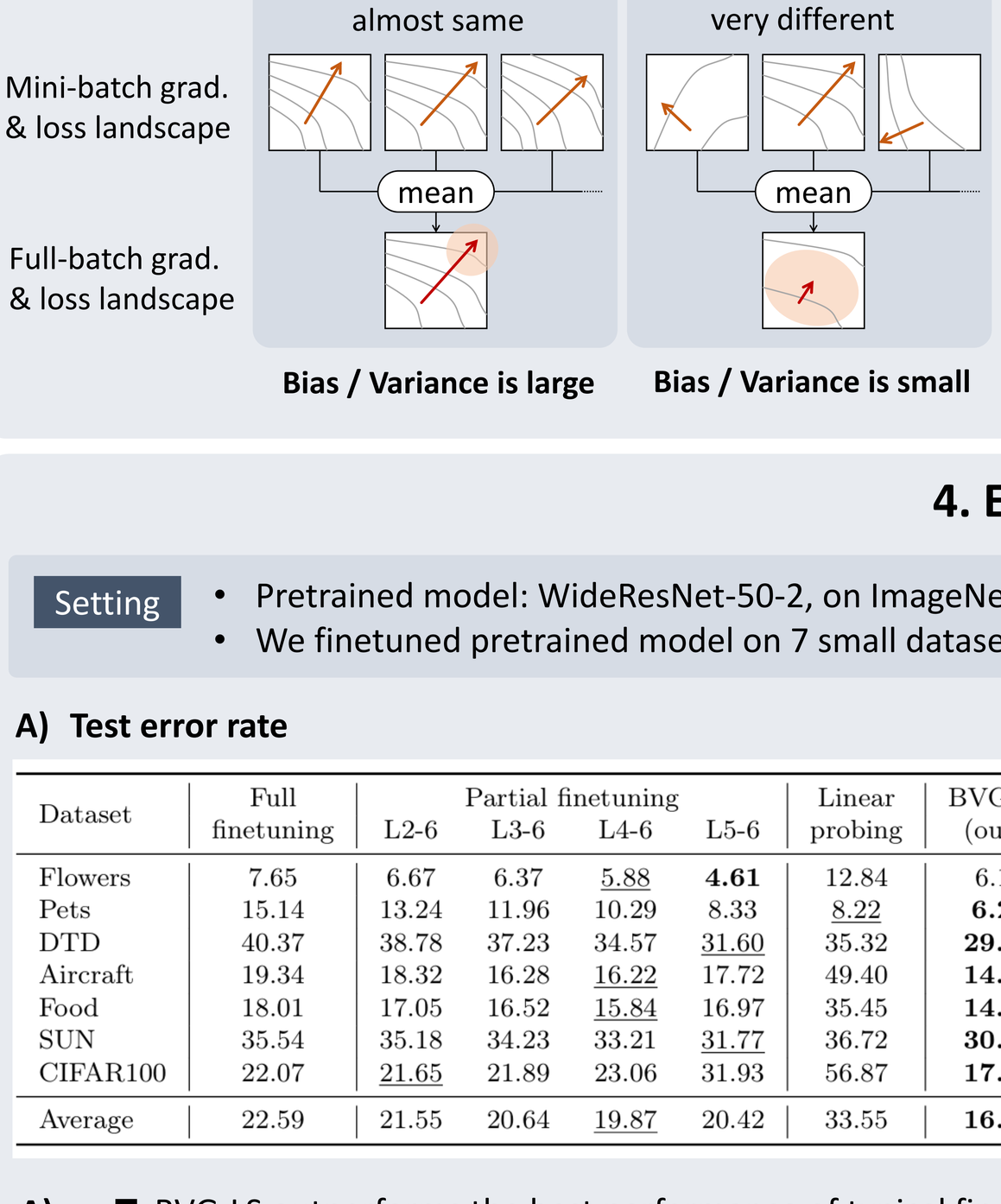
(Test error rate)

Method / Task	(A) Tuning all layers **Full finetuning**	(B) Tuning last few layers **Partial finetuning**	(C) Tuning only last layer **Linear probing**
Pet	15.1 %	10.3 %	**8.2 %**
Food	18.0 %	**15.8 %**	35.5 %
CIFAR100	**22.1 %**	23.1 %	56.9 %

Our aim To propose a simple finetuning strategy that is independent of the target task by identifying the layers that should be tuned based on some score.

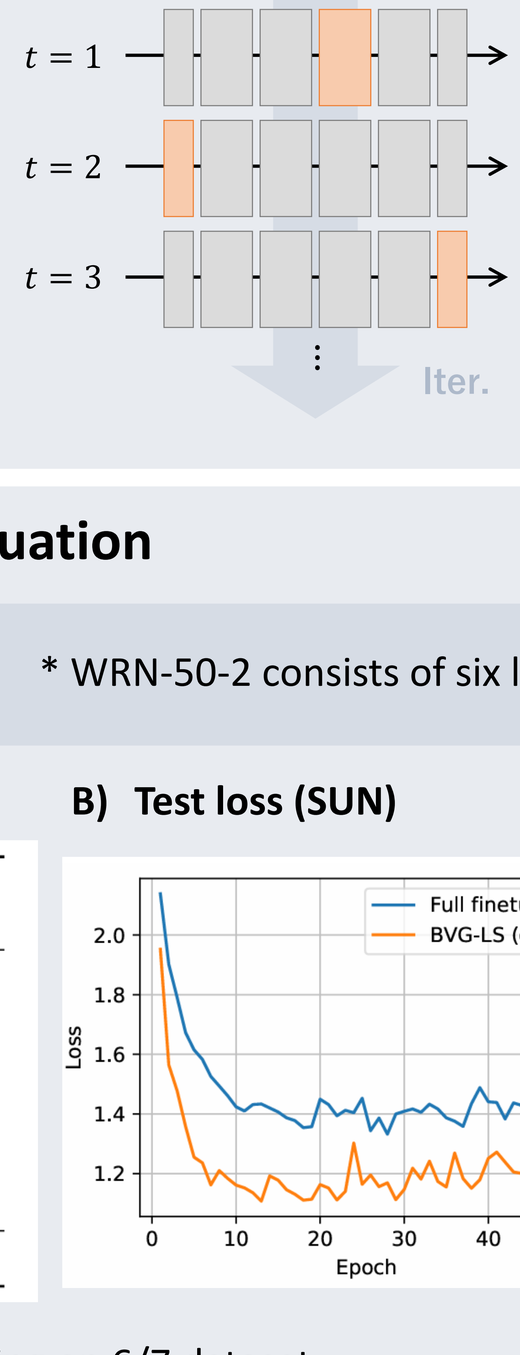
3. Proposed method < BVG-LS: Bias-Variance Guided Layer Selection >

Idea We focus on bias-variance ratio of minibatch gradients



In each iteration, we update only the layer with the highest value of bias-variance ratio $r^{(l,t)}$.

Layer with the highest $r^{(l,t)}$ (Update) Other layers (No Update)



$g_i^{(l,t)}$: minibatch gradient of l- th layer

$$b_i^{(l,t)} = \alpha b_i^{(l,t-1)} + (1-\alpha) g_i^{(l,t)}$$

$$v_i^{(l,t)} = \alpha v_i^{(l,t-1)} + (1-\alpha)\left(g_i^{(l,t)}\right)^2$$

$$r^{(l,t)} = \frac{\sum_i \left(b_i^{(l,t)}\right)^2}{\sum_i \left(v_i^{(l,t)} - \left(b_i^{(l,t)}\right)^2\right)}$$

Bias / Variance

4. Evaluation

Setting
- Pretrained model: WideResNet-50-2, on ImageNet
- We finetuned pretrained model on 7 small dataset

* WRN-50-2 consists of six layers:

A) Test error rate

Dataset	Full finetuning	Partial finetuning L2-6	L3-6	L4-6	L5-6	Linear probing	BVG-LS (ours)
Flowers	7.65	6.67	6.37	5.88	**4.61**	12.84	6.18
Pets	15.14	13.24	11.96	10.29	8.33	8.22	**6.26**
DTD	40.37	38.78	37.23	34.57	31.60	35.32	**29.20**
Aircraft	19.34	18.32	16.28	16.22	17.72	49.40	**14.60**
Food	18.01	17.05	16.52	15.84	16.97	35.45	**14.19**
SUN	35.54	35.18	34.23	33.21	31.77	36.72	**30.23**
CIFAR100	22.07	21.65	21.89	23.06	31.93	56.87	**17.25**
Average	22.59	21.55	20.64	19.87	20.42	33.55	**16.84**

B) Test loss (SUN)

C) Layer selection frequencies

A)
- BVG-LS outperforms the best performance of typical finetuning on 6/7 dataset.
- BVG-LS significantly outperforms full finetuning on all 7/7 datasets, although both methods do update all layers in the end.

B)
- Despite fewer updates per layer, BVG-LS undergoes faster loss descent and achieves smaller test loss than full finetuning .

C)
- The last layer undergoes dominant updates in early stage.
- Surprisingly, the selection frequency of the input layer is quite high (roughly, the second highest).

This work is supported by DENSO IT LAB Recognition and Learning Algorithm Collaborative Research Chair (Science Tokyo.)