PACE: marrying generalization in PArameter-efficient fine-tuning with Consistency rEgularization

Australian National University

Yao Ni†, Shan Zhang‡,†, Piotr Koniusz§,†

†The Australian National University ‡The University of Adelaide §Data61♥CSIRO

firstname.lastname@anu.edu.au

THE UNIVERSITY of ADELAIDE

CSIRO

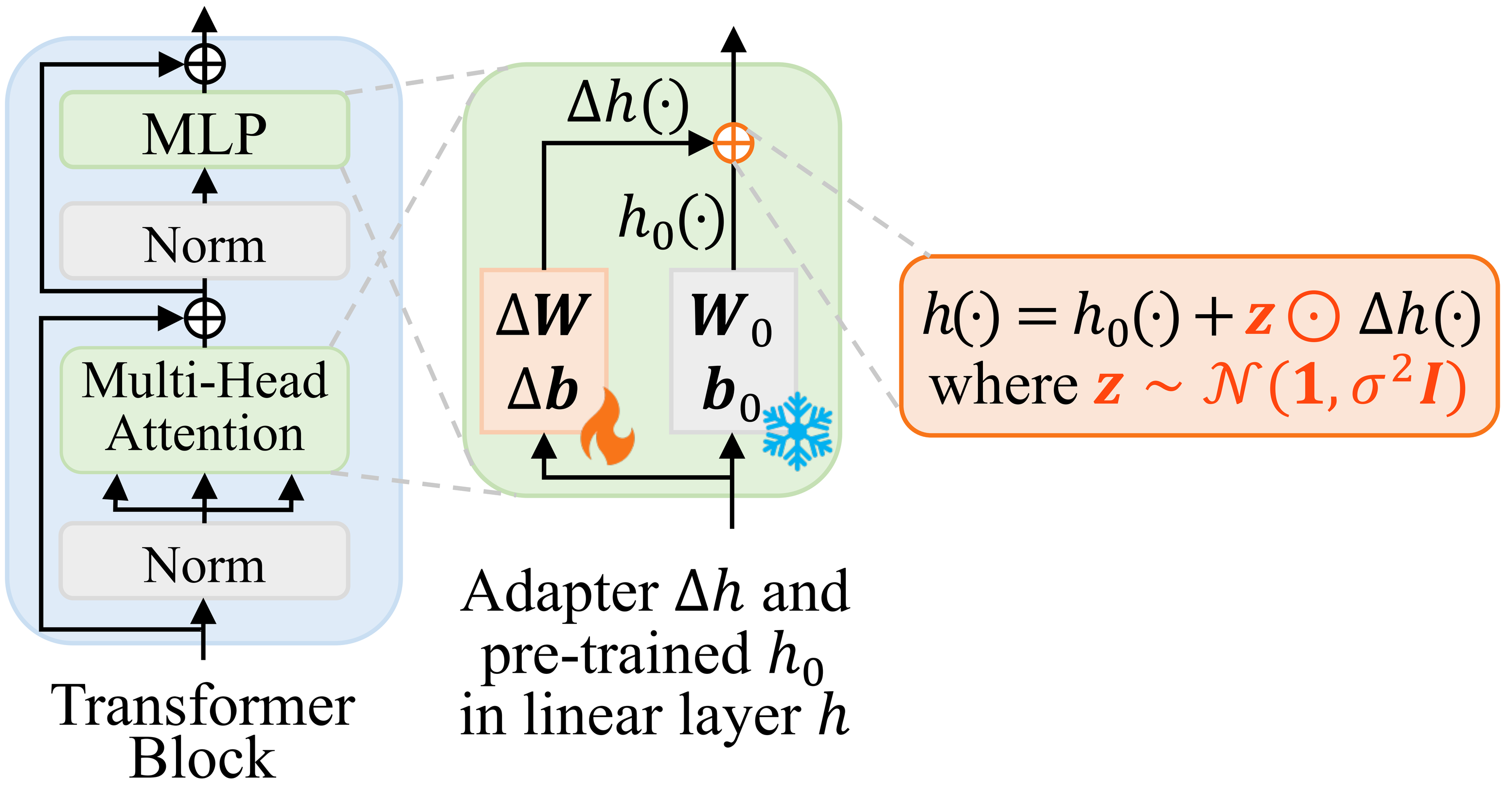
Goal

Background: Pre-trained transformers are growing larger. While Parameter-Efficient Fine-Tuning (PEFT) improves performance by tuning a small subset of parameters, it struggles with limited generalization and suffers from forgetting pre-trained knowledge.

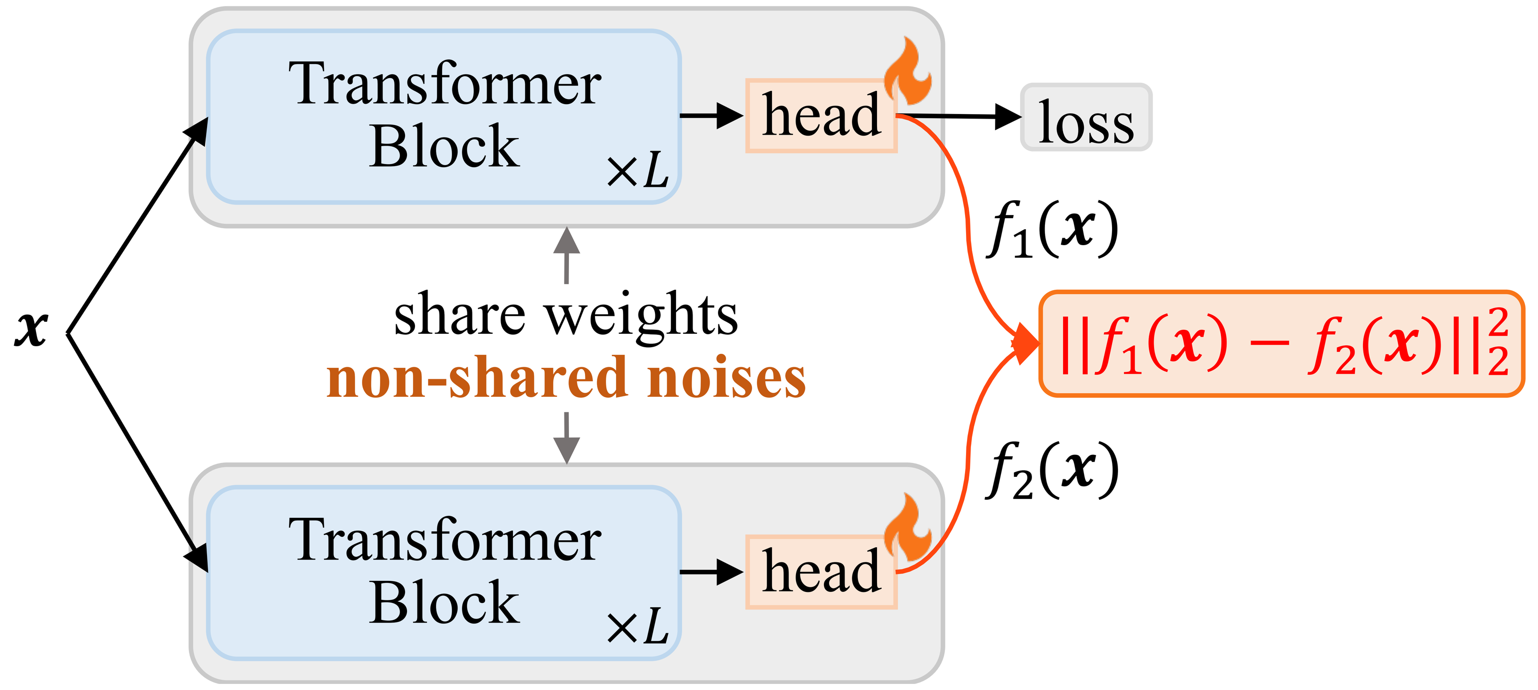
Goal: Improve generalization & retain pre-trained knowledge.

Pipeline

Step 1: Add multiplicative noise to adapter features.



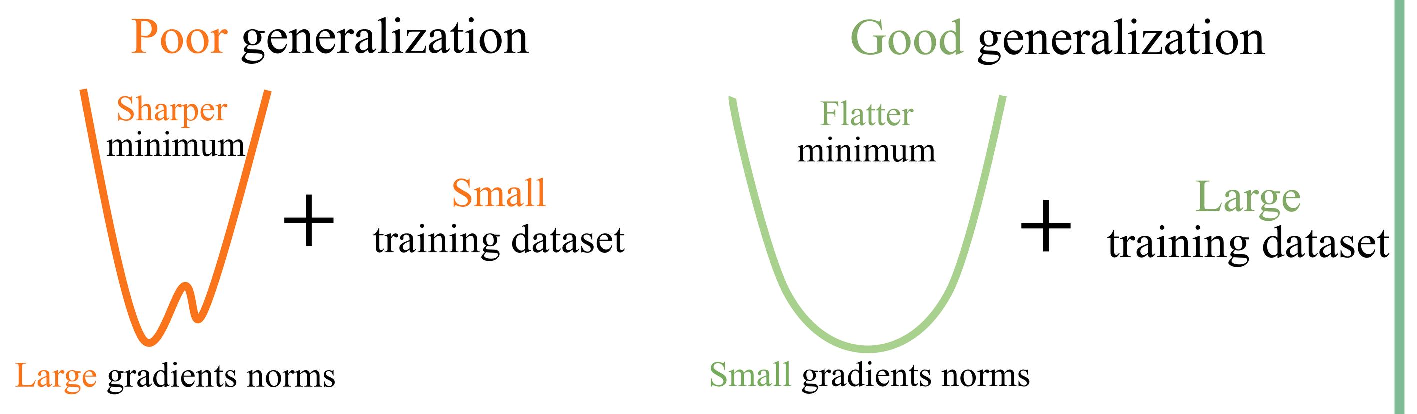
Step 2. Enforce consistency regularization across perturbations.



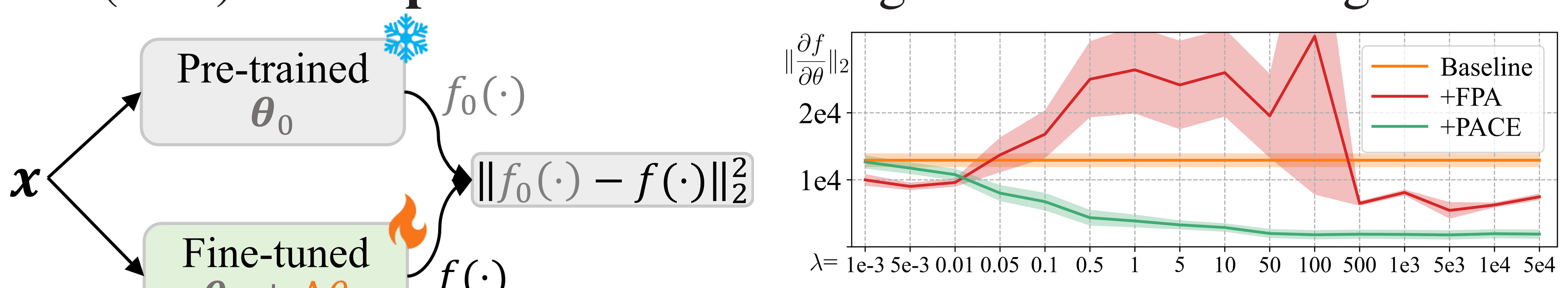
W_0, b_0: pre-trained linear layer weights; $\Delta W, \Delta b$: adapter weights; z: noise, $\mathcal{N}$: Gaussian distribution; x: sample; L: number of blocks; σ^2: noise variance; f_1, f_2: fine-tuned models with different noises.

Method

Thm. 1: Smaller norms of gradients & larger datasets improve generalization on unseen data.



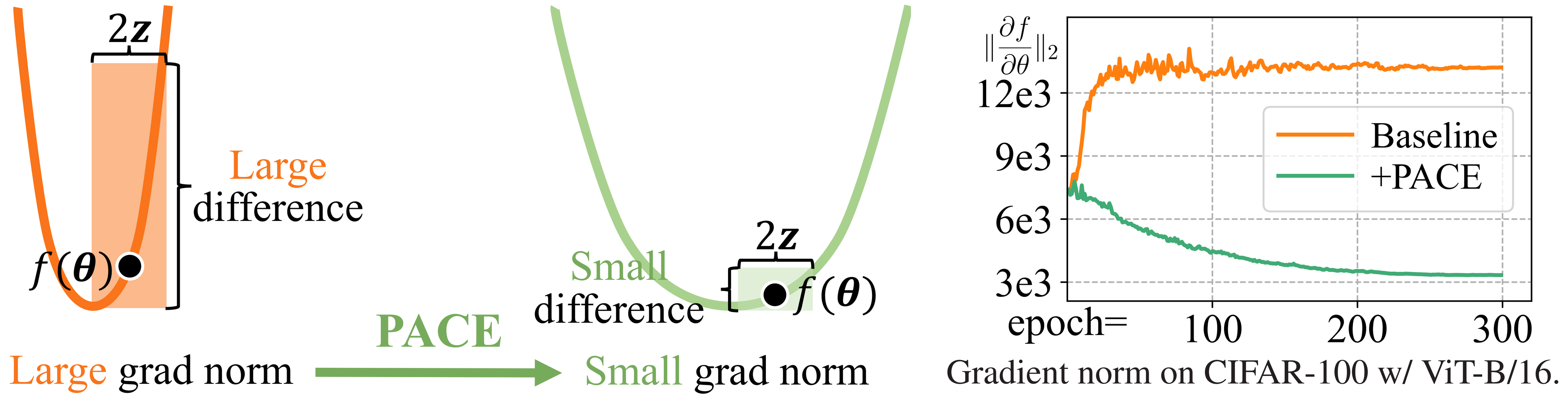
Prop. 1: For small dataset, aligning fine-tuned model with pre-trained one (FPA) retains **pre-trained** knowledge but cannot reduce gradients.



Grad. norms & reg. strength λ (CIFAR-100, ViT-B/16)

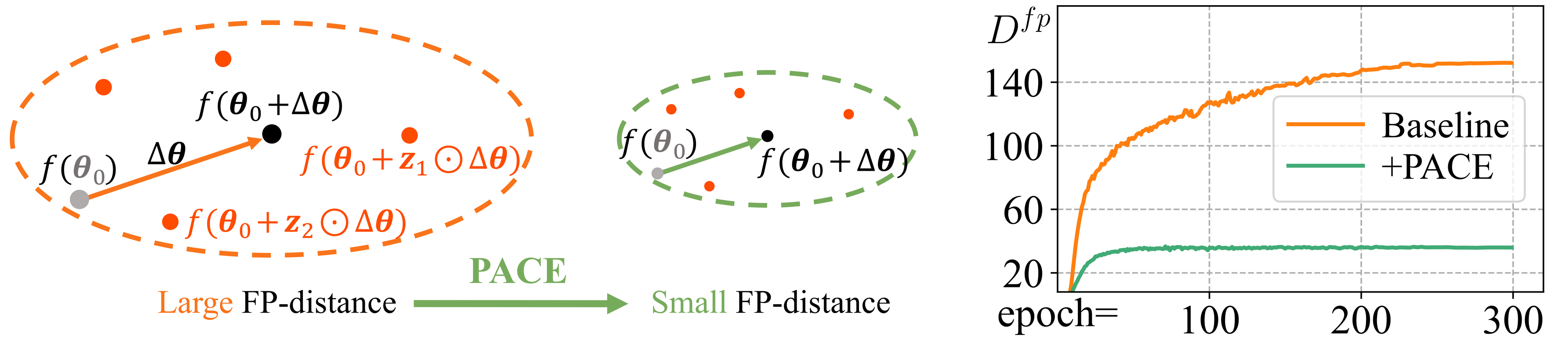
Solution: PACE perturbs adapter features & enforces consistency.

Thm. 2: PACE regularizes gradient norms.



Gradient norm on CIFAR-100 w/ ViT-B/16.

Thm. 3: PACE reduces distance between fine-tuned & pre-trained.



FP-distance: between fine-tuned & pre-trained models; θ_0: pre-trained weights; $\Delta\theta$: adapters weights; $z_1, z_2 \sim \mathcal{N}(\mathbf{1}, \sigma^2 I)$.

Distance between fine-tuned and pre-trained models (D^{fp}) on CIFAR-100 w/ ViT-B/16.

Experiments

Results on VTAB-1K with ViT-B/16. Mean Acc. is the average of group mean values.

Method	Natural							Specialized				Structured								
	Cifar100	Caltech101	DTD	Flowers102	Pets	SVHN	Sun397	Camelyon	EuroSAT	Resisc45	Retinopathy	Clevr-Count	Clevr-Dist	DMLab	KITTI-Dist	dSpr-Loc	dSpr-Ori	sNORB-Azim	NsORB-Ele	Mean Acc.
Full	68.9	87.7	64.3	97.3	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	68.9
Linear	64.4	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	57.6
VPT-Deep	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	72.0
Adapter	69.2	90.1	68.0	98.8	89.9	82.8	54.3	84.0	94.9	81.9	75.5	80.9	65.3	48.6	78.3	74.8	48.5	29.9	41.6	73.9
AdaptFormer	70.8	91.2	70.5	99.1	90.9	86.6	54.8	83.0	95.8	84.4	76.3	81.9	64.3	49.3	80.3	76.3	45.7	31.7	41.1	74.7
LoRA	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	74.5
NOAH	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	74.2
RepAdapter	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	76.1
RLRR	75.6	92.4	72.9	99.3	91.5	89.8	57.0	86.8	95.2	85.3	75.9	79.7	64.2	53.9	82.1	83.9	53.7	33.4	43.6	76.7
GLoRA	76.4	92.9	74.6	**99.6**	**92.5**	91.5	57.8	87.3	**96.8**	88.0	76.0	83.1	67.3	54.5	**86.2**	83.8	52.9	37.0	41.4	78.0
Baseline	74.9	93.3	72.0	99.4	91.0	91.5	54.8	83.2	95.7	86.9	74.2	83.0	70.5	51.9	81.4	77.9	51.7	33.6	44.4	76.4
+PACE	**79.0**	**94.2**	73.6	99.4	92.4	**93.7**	**58.0**	**87.4**	96.4	**89.3**	**77.1**	**84.9**	**70.9**	**54.9**	84.3	**84.7**	**57.3**	**39.3**	**44.8**	**79.0**

Results on FGVC with ViT-B/16.
* denotes using augmented ViT by AugReg.

Method	CUB -2011	NA-Birds	Oxford Flowers	Stan. Dogs	Stan. Cars	Mean Acc.
Full	87.3	82.7	98.8	89.4	84.5	85.9
Linear	85.3	75.9	97.9	86.2	51.3	79.3
VPT	88.5	84.2	99.0	90.2	83.6	89.1
LoRA	88.3	85.6	99.2	91.0	83.2	89.5
SSF*	89.5	85.7	99.6	89.6	89.2	90.7
ARC*	89.3	85.7	**99.7**	89.1	**89.5**	90.7
RLRR*	89.8	85.3	99.6	90.0	90.4	91.0
$\text{LoRA}_{\text{mul}}+\text{VPT}_{\text{add}}$	88.9	87.1	99.4	91.2	87.5	90.8
+PACE	**89.8**	**87.3**	99.5	**92.2**	88.8	**91.5**

Results on domain adaptation with ViT-B/16 pretrained on ImageNet-21K.

Method	Source	Target				Mean
	ImageNet	-Sketch	-V2	-A	-R	Acc.
Full	63.9	18.5	52.5	3.2	21.2	31.8
Linear	67.9	14.4	60.8	9.4	25.6	35.6
Adapter	70.5	16.4	59.1	5.5	22.1	34.7
VPT	70.5	18.3	58.0	4.6	23.2	34.7
LoRA	70.8	20.0	59.3	6.9	23.3	36.0
NOAH	71.5	24.8	66.1	11.9	28.5	40.5
GLoRA	78.3	30.6	67.5	13.3	31.0	44.1
$\text{LoRA}_{\text{mul}}+\text{VPT}_{\text{add}}$	78.3	30.6	68.5	14.1	32.5	44.8
+PACE	**79.0**	**31.8**	**69.4**	**16.3**	**35.2**	**46.3**

Results for GLUE w/ $\text{RoBERTa}_{\text{base}}$. Matthew's correlation for COLA, Pearson correlation for STSB, and accuracy for others.

Method	COLA	STSB	MRPC	RTE	QNLI	SST2	Avg.
Full	63.6	91.2	90.2	78.7	92.8	94.8	85.2
BitFit	62.0	90.8	**92.7**	81.5	91.8	93.7	85.4
Adapt	62.6	90.3	88.4	75.9	93.0	94.7	84.2
VeRA	65.6	90.7	89.5	78.7	91.8	94.6	85.2
LoRA	63.4	91.5	89.7	86.6	93.3	95.1	86.6
+PACE	**66.2**	**92.0**	91.4	**86.9**	**93.6**	**95.6**	**87.6**

Results for GSM-8K using Phi-3-mini-4k-instruct.

Method	Accuracy
Pre-trained	62.01
Full	73.16
LoRA	75.66
+PACE	**78.77**

Yao Ni Seeking PostDoc Job. Scan his CV.

Yao Ni CV

PhD opportunities.

Github Code