NEURAL INFORMATION PROCESSING SYSTEMS

Georgia State University

NEC Laboratories America

UCONN

VB-LoRA: Extreme Parameter Efficient Fine-Tuning with Vector Banks

Yang Li, Georgia State University

Shaobo Han, NEC Laboratories America

Shihao Ji, University of Connecticut

Try VB-LoRA

```
import peft
peft.VBLoRAConfig(...)
```

Background

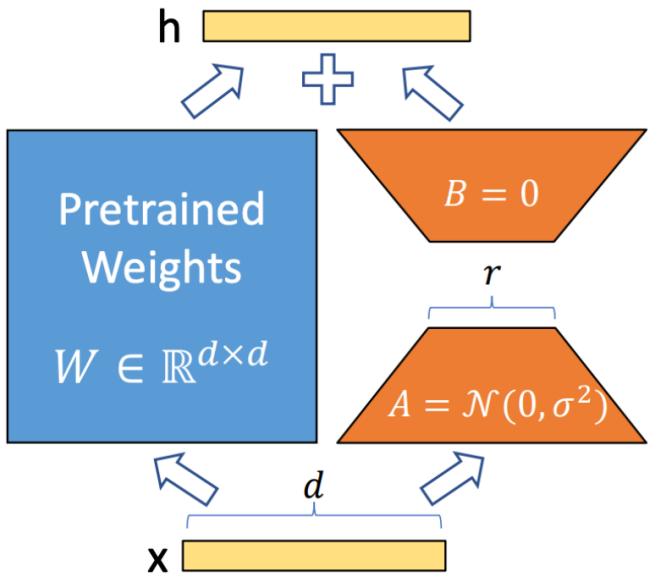
- Large Models
  - Llama3, Claude, BERT, etc.

- Fine-tuning large models can lead to even better performance on specific downstream tasks.

- Parameter-Efficient Fine-Tuning (PEFT)

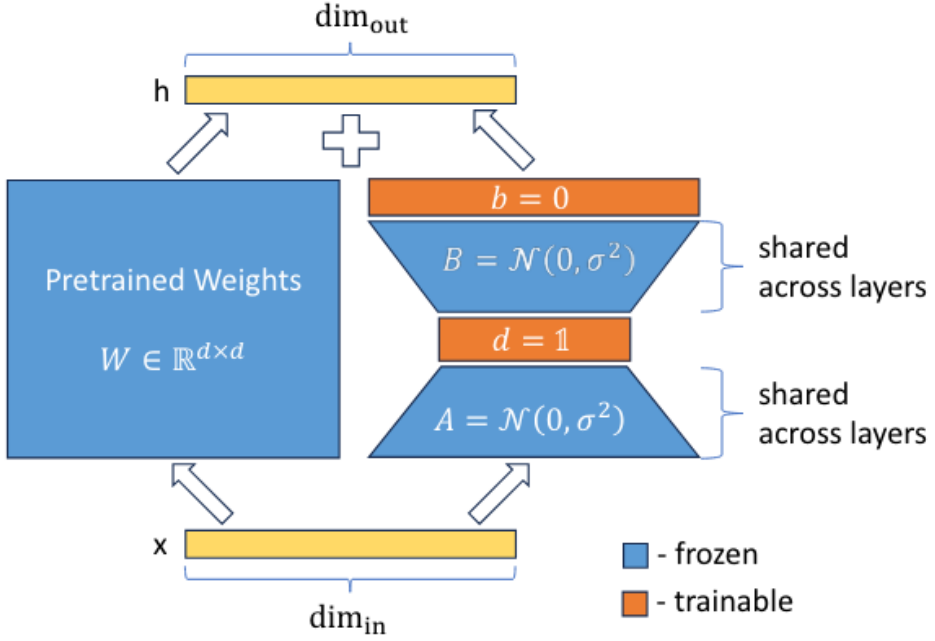
Prior Work

The figures are sourced from the original paper.



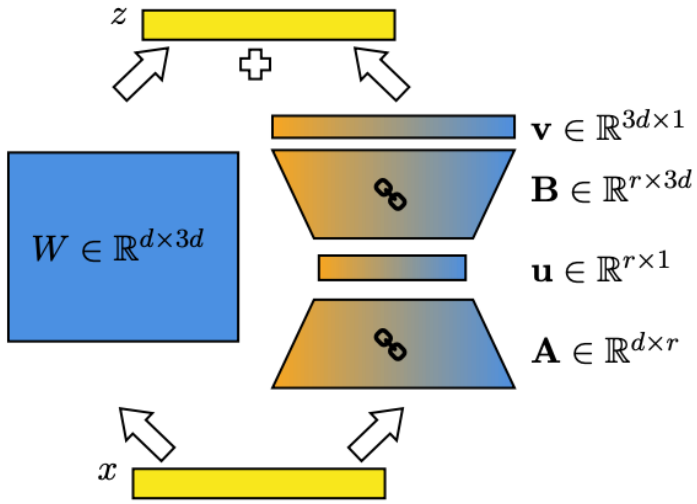
LoRA

$\Delta W = AB$



VeRA

$\Delta W = AdBb$

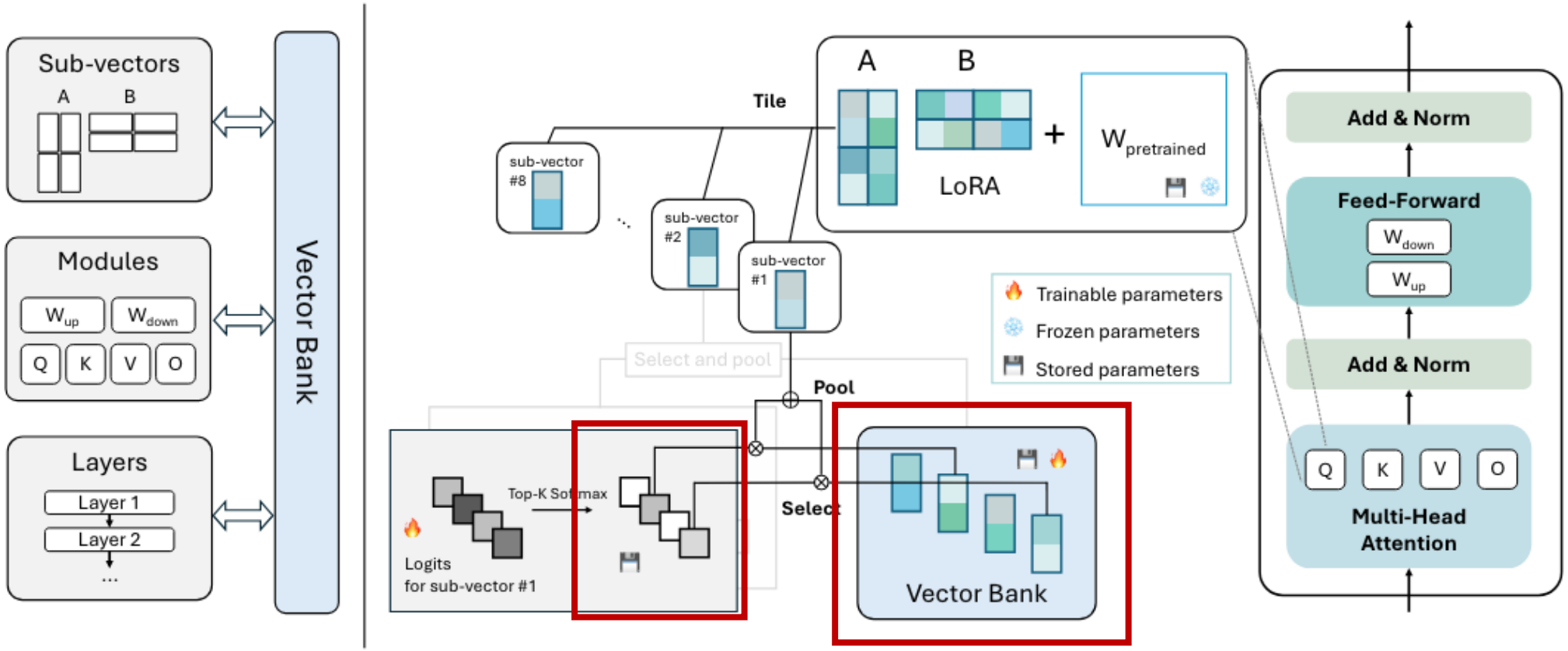


Tied-LoRA

$\Delta W = AuBv$

VB-LoRA

Divide *LoRA matrices into sub-vectors*



A differentiable ***top-k admixture module (TKAM)*** *to form sub-vectors*

Vector bank *is* ***shared*** *across sub-vectors, modules, and layers*

Natural Language Understanding

Method	# Params	SST-2	MRPC	CoLA	QNLI	RTE	STS-B	Avg.
LoRA$_{qv}$	0.786M	96.2±0.5	90.2±1.0	68.2±1.9	94.8±0.3	85.2±1.1	92.3±0.5	87.8
VeRA$_{qv}$	0.061M	**96.1±0.1**	90.9±0.7	68.0±0.8	94.4±0.2	85.9±0.7	91.7±0.8	87.8
Tied-LoRA$_{qv}$	0.066M	94.8±0.6	89.7±1.0	64.7±1.2	94.1±0.1	81.2±0.1	90.8±0.3	85.9
VB-LoRA$_{qv}$	**0.024M**	**96.1±0.2**	**91.4±0.6**	**68.3±0.7**	**94.7±0.5**	**86.6±1.3**	**91.8±0.1**	**88.2**
VeRA$_{all}$	0.258M	96.6±0.5	90.9±0.8	68.5±1.4	94.4±0.4	85.9±1.2	**92.2±0.2**	88.1
Tied-LoRA$_{all}$	0.239M	94.8±0.3	90.0±0.4	66.8±0.1	94.1±0.1	82.3±2.0	91.6±0.2	86.6
VB-LoRA$_{all}$	**0.033M**	**96.3±0.2**	**91.9±0.9**	**69.3±1.5**	**94.4±0.2**	**87.4±0.7**	91.8±0.2	**88.5**

Roberta-large / GLUE benchmark

Instruct Tuning

Model	Method	# Params	Score
Llama2-7B	w/o FT	-	4.79
	LoRA	159.9M	5.63
	VB-LoRA	**0.8M (0.5%)**	**5.71**
Llama2-13B	Full-FT	-	5.38
	LoRA	250.3M	6.13
	VB-LoRA	**1.1M (0.4%)**	**6.31**

Llama2 7B and 13B / Cleaned Alpaca dataset

Take Away

- Divide-and-share paradigm
- Extreme parameter efficiency
- Try our method on Hugging Face PEFT

```
import peft
peft.VBLoRAConfig(...)
```

Paper

Code

Doc on HF