SONY

MambaPEFT:
Exploring Parameter-Efficient Fine-Tuning for Mamba

Masakazu Yoshimura*, Teruaki Hayashi*, and Yota Maeda *Equally contributed

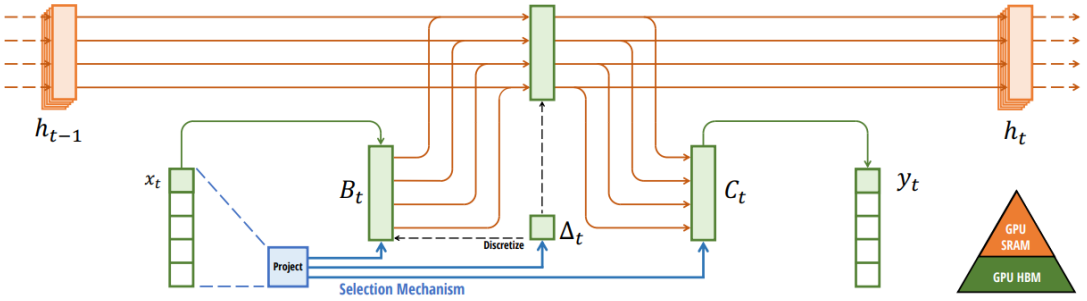
(Sony Group Corporation)

Introduction

Mamba is successful in various areas

However, unlike Transformer, it still lacks parameter-efficient fine-tuning(PEFT) methods to adapt the large pre-trained model to diverse tasks.

PEFT is crucial for Mamba ecosystem



Mamba

A model that consist of the strengths of state space models (SSM) / RNNs and Transformers.

- Efficient like SSMs / RNNs
  - The computational cost scales linearly with sequence length, not quadratic.
- High accuracy like Transformers
  - The proposed selective scan works similarly to Attention.

Our contribution

- Investigate the feasibility of existing PEFT methods for Mamba
- Re-design and propose PEFT methods specific to Mamba with extensive experimentation
- Further performance improvements with HybridPEFT, in which the optimal combination of multiple PEFT methods and their hyper-parameters are searched

Overview

Investigate, improve, and propose 20 variations of 7 PEFT methods

Partial tuning - Update selected params

A, D, causal_conv1d, cls_embed, bias, …

Additive method - Add parameters

Prompt-tuning, Adapter,

Affix-tuning, **Additional-scan** ← New Mamba-specific PEFT

Reparameterization method

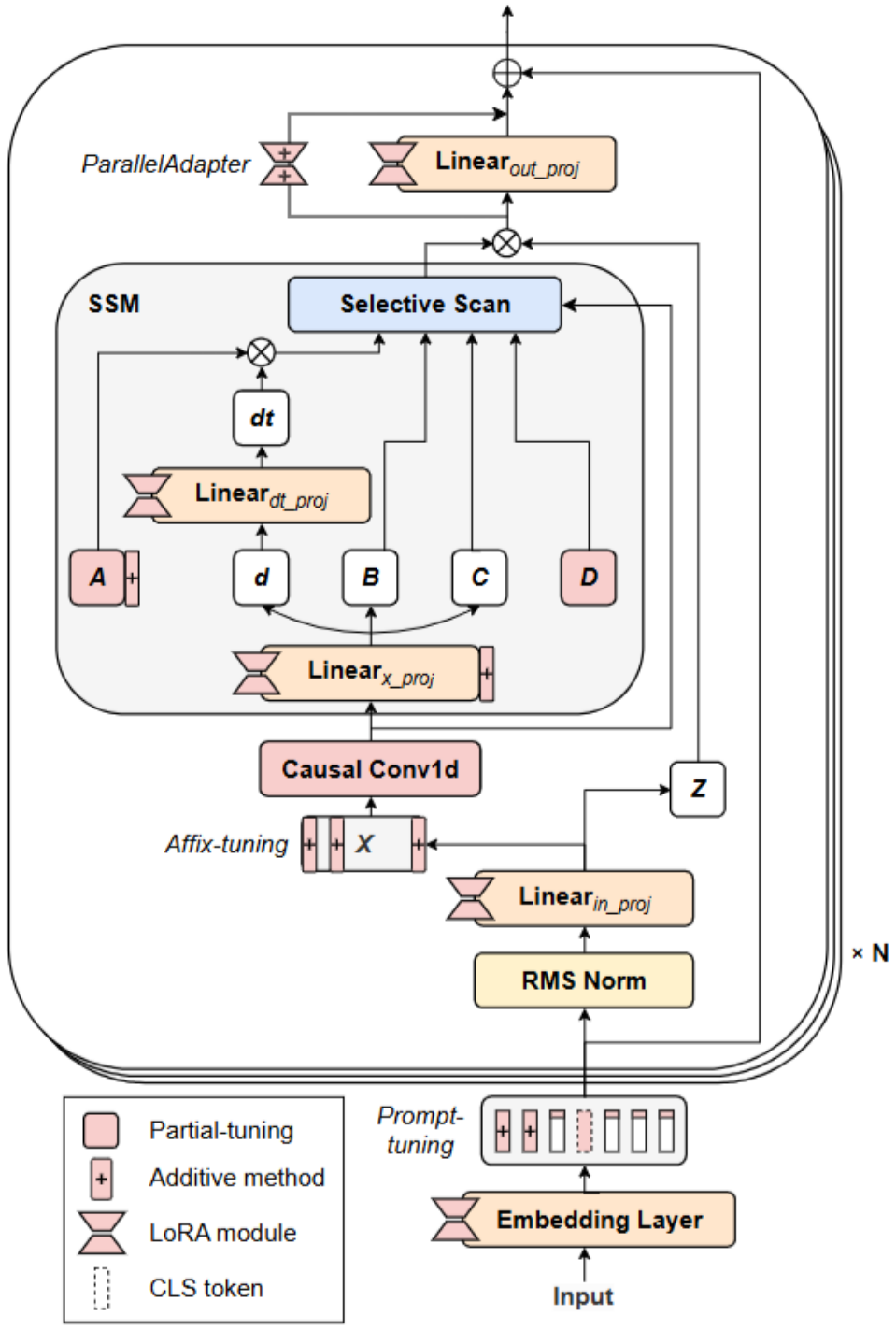
LoRA

Where to tune? (in_proj, out_proj, …)

LoRA$_p$ (partial-LoRA)

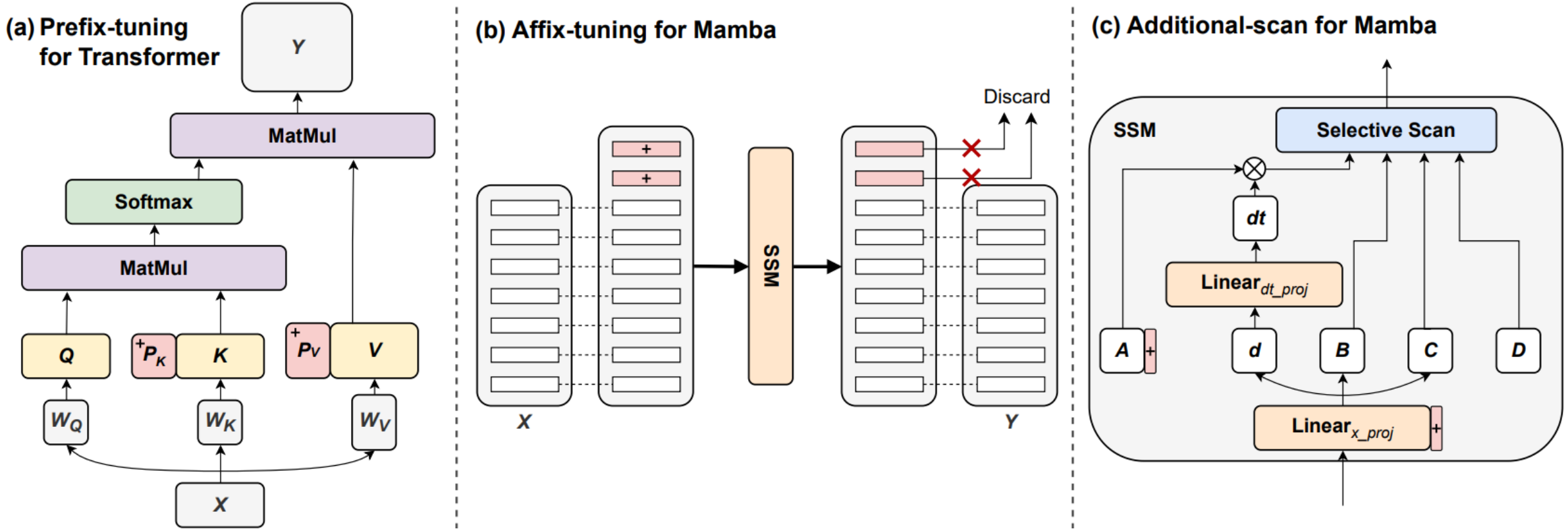
Investigate more finely about where to tune

Apply LoRA on the partial weight with respect to outputs (X, Z, B, C, …)



Proposed PEFT methods specific to Mamba

Affix-tuning and Additional-scan



Affix-tuning

Insert tokens at arbitrary position before SSM and discard after output

Additional-scan

Add learnable dimensions to the hidden state in SSM

Evaluation and Findings

Vtab-1K (vision, 1K training data)

Model	Method	#Params (K)	Avg.
ViT-S	Scratch	21,704	26.20
	Full	21,704	53.47
	Linear Probing	9	51.74
	FacT-TK	16	66.96
	LoRA	628	68.68
	Adaptformer	333	68.97
	SPT-LoRA	414	69.38
	Adapter+	122	69.87
Vim-S	Scratch	25,450	25.42
	Full	25,450	47.08
	Linear Probing	9	52.75
	Conv1d-tuning	156	69.09
	Prompt-tuning (w/o proj)	12	56.77
	Prompt-tuning	307	62.54
	Affix-tuning (w/o proj)	230	65.04
	Affix-tuning	117,000	70.29
	Additional-scan	672	68.65
	ParallelAdapter	663	70.96
	LoRA(out_proj)	2,663	71.12
	LoRA(in_proj)	1,483	71.25
	LoRA$_p$(X)	1,778	71.52
	Hybrid (w/ proj)	117,236	**72.05**
	Hybrid (w/o proj)	1,044	71.80

Commonsense (language, 170K training data)

Model	Method	#Params(%)	Avg.
Pythia 160M	Full	100	42.0
	LoRA	0.72	41.6
Mamba 130M	Full	100	43.8
	SLL LoRA	1.45	42.7
	Additional-scan	0.51	42.7
	Affix-tuning (w/o proj)	0.17	40.6
	Affix-tuning	64.64	43.2
	LoRA(in_proj)	2.23	42.8
	LoRA$_p$(X)	2.67	**43.7**
Pythia 1.4B	Full	100	50.5
	LoRA	0.44	50.5
Mamba 1.4B	Full	100	53.0
	SLL LoRA	4.64	52.7
	Additional-scan	0.26	53.5
	Affix-tuning (w/o proj)	0.09	**53.9**
	LoRA(in_proj)	1.13	52.6
	LoRA$_p$(X)	1.36	53.7

(1) Mamba benefits from PEFT more than Transformers

Larger improvement from full fine-tuning

(2) LoRA$_p$(X) is effective with limited data

(3) Additional-scan is effective with large data

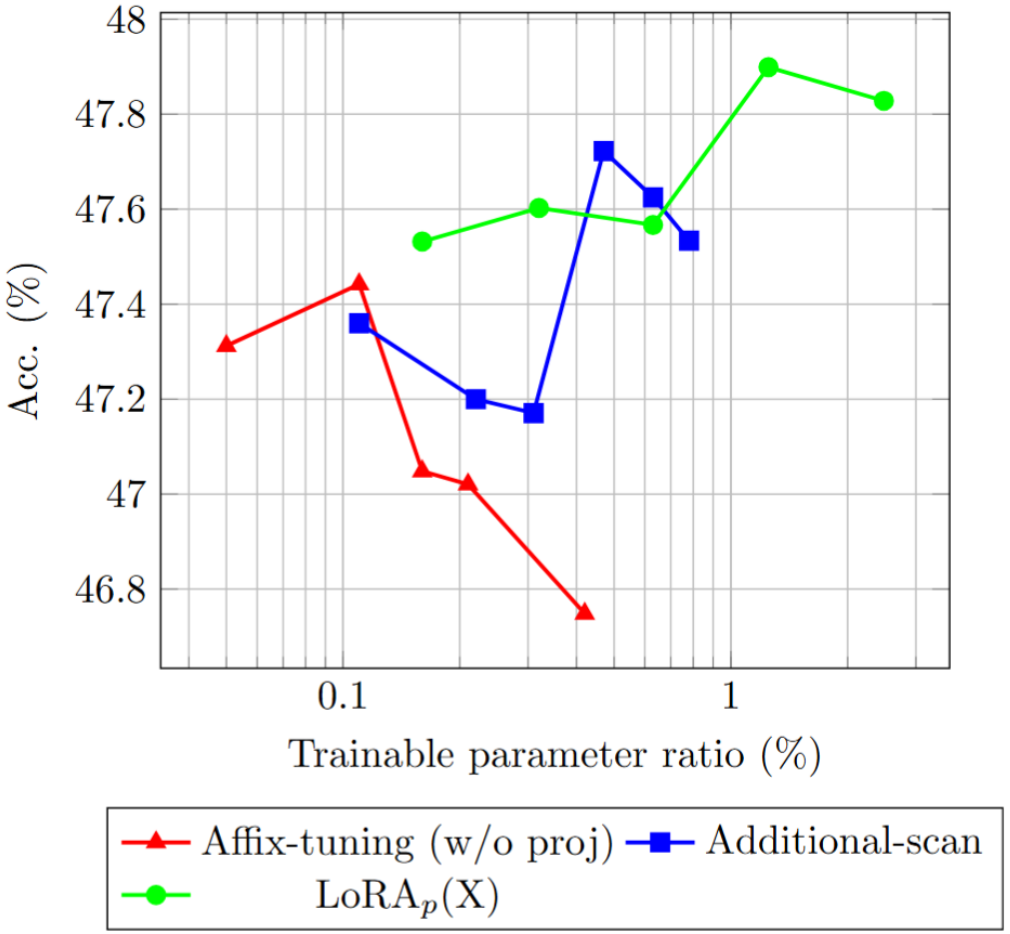
(4) Affix-tuning is effective for large Mamba models

(5) Performance improvements with HybridPEFT

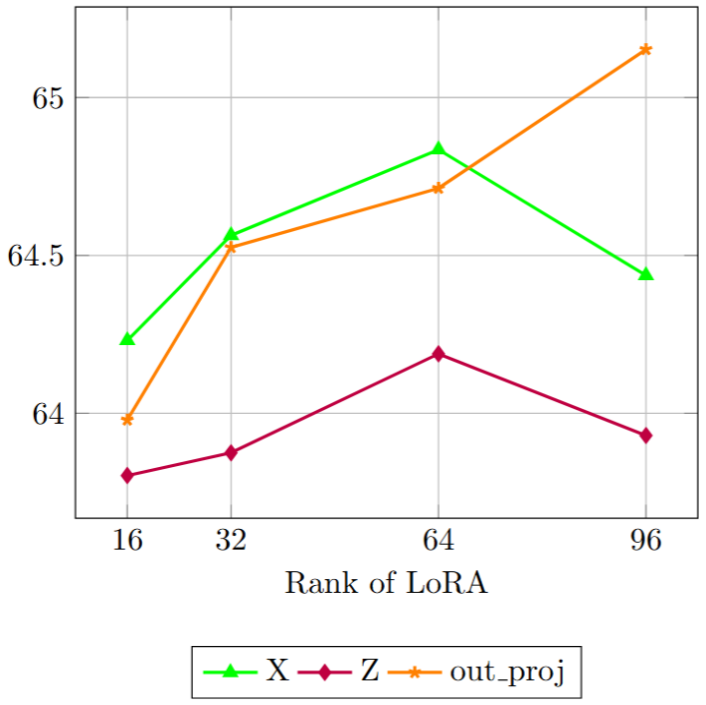
Evaluation and Findings

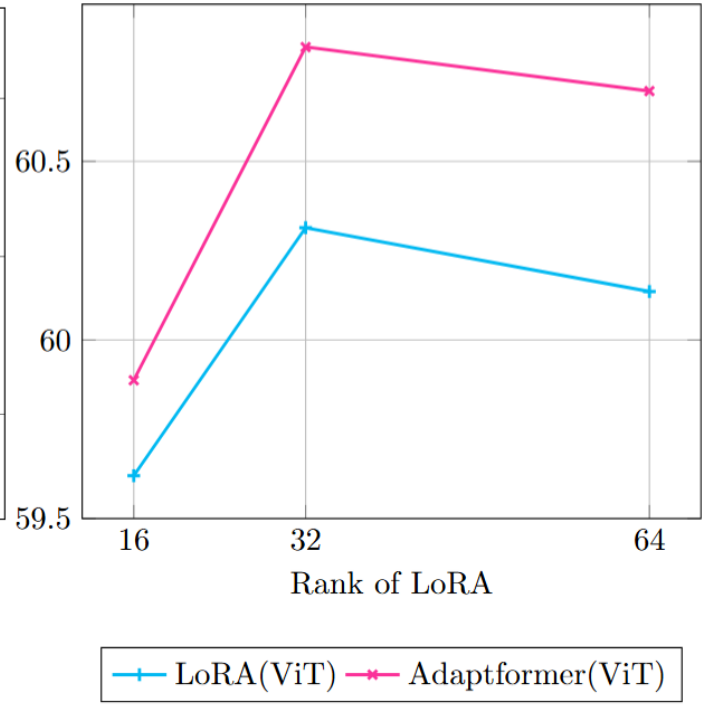
(6) We should choose a suitable PEFT method depending on the computational budget

No method is always superior



(7) PEFT for Mamba can be improved by adding more parameters





More findings and experiments can be found in the paper

- Unlike Transformer, LoRA is better applied to a specific module.
- Simply combining high performance PEFTs to create a hybrid PEFT will degrade performance.

Conclusion

Conclusion

- Investigate the feasibility of existing PEFT methods for Mamba
- Re-design and propose PEFT methods specific to Mamba with extensive experimentation
- Further performance improvements with HybridPEFT, in which the optimal combination of multiple PEFT methods and their hyper-parameters are searched

Future Direction

- Based on our findings regarding the optimal tuning locations and the optimal number of parameters, future works can focus on the application or algorithm of the PEFT for Mamba.

The code is open sourced!