S²FT: Efficient, Scalable and Generalizable LLM Fine-tuning by Structured Sparsity

Xinyu Yang[1], Jixuan Leng[1], Geyang Guo[2], Jiawei Zhao[3], Ryumei Nakada[4], Linjun Zhang[4], Huaxiu Yao[5], Beidi Chen[1]
[1] Carnegie Mellon University, [2] Georgia Tech, [3] Caltech, [4] Rutgers, [5] UNC-Chapel Hill

Introduction

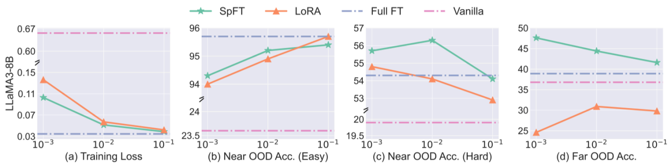
Why using S²FT instead of Full FT or LoRA?

	High Quality		Efficient Training		Scalable Serving		
	ID	OOD	Time	Memory	Fusion	Switch	Parallelism
Full FT	✔✔	✔	✘	✘	✘	✘	✘
LoRA	✔	✘	✔	✔	✔	✔	✔
S²FT	✔	✔✔	✔✔	✔✔	✔✔	✔✔	✔✔

Structured Sparse Fine-Tuning (S²FT), is a family of PEFT methods for LLMs that achieves high quality, efficient training, and scalable serving simultaneously. Compared to LoRA, S²FT offers several key advantages: **(i) High Quality:** enhanced generalization ability on both commonsense and arithmetic reasoning with 4.6% and 1.3% average improvements, **(ii) Efficient Training:** 10% reduced training time and memory, **(iii) Scalable Serving:** effective fusion, fast switch, and efficient parallelism when serving multiple adapters. These features are particularly valuable for the large-scale, real-world deployment of foundation models in various domains.

Observation

Sparse FT demonstrate better generalization ability.

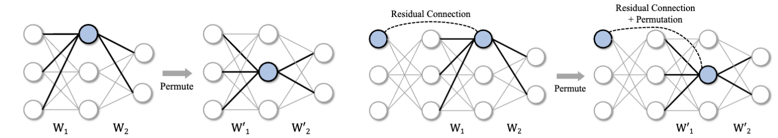


The counterintuitive observation that selecting channels with the smallest activations leads to improved performance further support this finding.

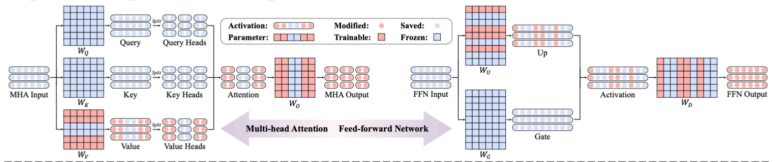
Task	S²FT-R	S²FT-W		S²FT-A		S²FT-S		S²FT-G	
		Large	Small	Large	Small	Large	Small	Large	Small
Knowledge	86.6	85.9 (-0.7)	85.3 (-1.3)	84.7 (-1.9)	87.3 (+0.7)	85.1 (-1.5)	87.2 (+0.6)	85.4 (-1.2)	86.2 (-0.4)
Arithmetic	79.6	78.4 (-1.2)	78.4 (-1.2)	77.1 (-2.5)	80.0 (+0.4)	76.8 (-2.8)	79.8 (+0.2)	77.8 (-1.8)	79.5 (-0.1)

Method

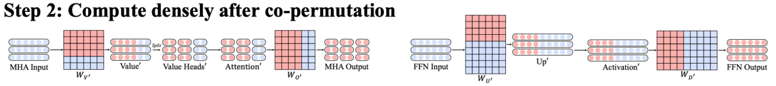
Discover Coupled Structures in LLMs.



Step 1: Select sparsely with coupled structures



Step 2: Compute densely after co-permutation



Experimental Results

a) High Quality on **Commonsense Reasoning**:

Method	#Param	BoolQ	PIQA	SIQA	HellaSwag	Wino	ARC-e	ARC-c	OBQA	Avg. ↑
Full FT	100	73.9	86.2	79.1	93.1	85.8	88.1	78.2	84.0	83.6
LoRA	0.70	70.8	85.2	79.7	92.5	84.9	88.9	78.7	84.4	82.5
DoRA	0.71	74.6	89.3	79.9	95.5	85.6	90.5	80.4	85.8	85.2
S²FT	0.70	75.0	89.0	80.7	96.5	88.0	92.5	83.4	87.8	86.6

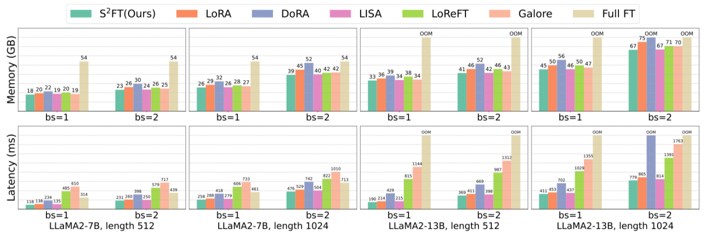
b) High Quality on **Arithmetic Reasoning**:

Method	#Param	MultiArith	GSM8K	AddSub	AQuA	SingleEq	SVAMP	MAWPS	Avg. ↑
Full FT	100	99.2	62.0	93.9	26.8	96.7	74.0	91.2	77.7
LoRA	0.70	99.5	61.6	92.7	25.6	96.3	73.8	90.8	77.2
DoRA	0.71	98.8	62.7	92.2	26.8	96.9	74.0	91.2	77.5
S²FT	0.70	99.7	65.8	93.7	31.5	97.8	76.0	92.4	79.6

c) High Quality on **Instruction-Following** :

	Method	Writing	Roleplay	Reasoning	Code	Math	Extraction	STEM	Humanities	Avg.
Mistral-7B	Vanilla	5.25	3.20	4.50	1.60	2.70	6.50	6.17	4.65	4.32
	Full FT	5.50	4.45	5.45	2.50	3.25	5.78	4.75	5.45	4.64
	LoRA	5.30	4.40	4.65	2.35	3.30	5.50	5.55	4.30	4.41
	Galore	5.05	5.27	4.45	1.70	2.50	5.21	5.52	5.20	4.36
	LISA	6.84	3.65	5.45	2.20	2.75	5.65	5.95	6.35	4.85
	Ours	**6.95**	4.40	**5.50**	**2.70**	**3.55**	5.95	**6.35**	**6.75**	**5.27**
LLaMA2-7B	Vanilla	2.75	4.40	2.80	1.55	1.80	3.20	5.25	4.60	3.29
	Full FT	5.55	6.45	3.60	1.75	2.00	4.70	6.45	7.50	4.75
	LoRA	6.30	5.65	4.05	1.60	1.45	4.17	6.20	6.20	4.45
	Galore	5.60	6.40	3.20	1.25	1.95	5.05	6.57	7.00	4.63
	LISA	6.55	**6.90**	3.45	1.60	**2.16**	4.50	6.75	7.65	4.94
	Ours	**6.75**	6.60	**4.15**	**1.65**	1.85	4.75	**7.45**	**8.38**	**5.20**

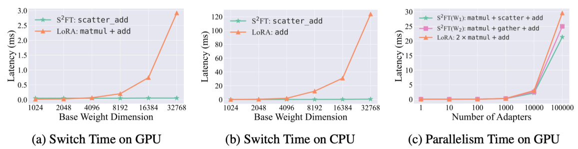
d) Efficient Training with **varying sequence lengths and batch sizes** :



e) Scalable Serving through **effective adapter fusion**:

Task	LoRA			S²FT		
	Adapter 1	Adapter 2	Fused	Adapter 1	Adapter 2	Fused
Commonsense	83.1	32.1	79.8 (-3.3)	86.6	42.3	84.0 (-2.6)
Arithmetic	12.0	77.2	71.6 (-5.6)	12.8	79.6	75.3 (-4.3)

f) Scalable Serving through **fast switch and efficient parallelism**:



Available sources:

- Code: https://github.com/Infini-AI-Lab/S2FT
- Blog: https://infini-ai-lab.github.io/S2FT-Page/
- Paper: https://arxiv.org/abs/2412.06289