Dec 2-7, 2025

RefLoRA: Refactored Low-Rank Adaptation for Efficient Finetuning of Large Models

Yilang Zhang

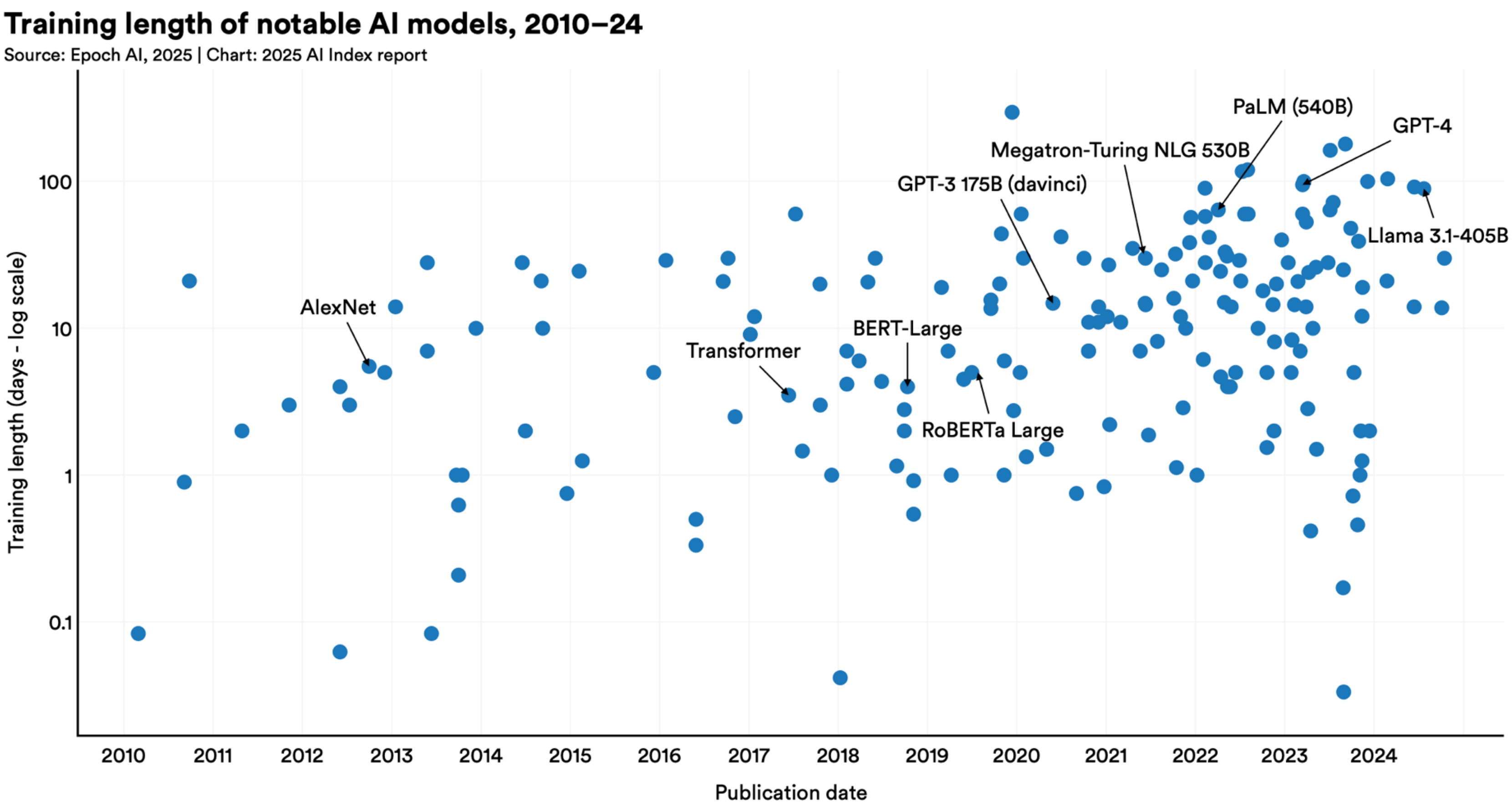
Dept. of ECE, University of Minnesota

Aknowledgement: *Bingcong Li, Georgios B. Giannakis*

NSF grants 2220292, 2212318, 2312547, and 2332547

Challenge in finetuning

- **Challenge:** rapid growth of model size vs. prohibitive computational overhead



https://hai.stanford.edu/ai-index/2025-ai-index-report

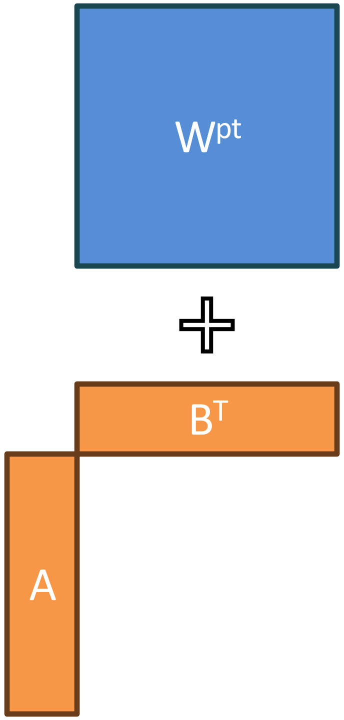
Low-rank adaptation

- Low-rank adaptation (LoRA)

  - Consider a general weight matrix $\mathbf{W} = \mathbf{W}^{\text{pt}} + \mathbf{W}^{\text{ft}} \in \mathbb{R}^{m \times n}$

  - Low-rank "adapters" $\mathbf{W}^{\text{ft}} = \mathbf{A}\mathbf{B}^\top, \quad \mathbf{A} \in \mathbb{R}^{m \times r}, \mathbf{B} \in \mathbb{R}^{n \times r} \ (r \ll m, n)$

$$\min_{\mathbf{A},\mathbf{B}} \ell(\mathbf{W}^{\text{pt}} + \mathbf{A}\mathbf{B}^\top)$$



  - Learnable parameters $\mathcal{O}((m+n)r) \ll \mathcal{O}(mn)$

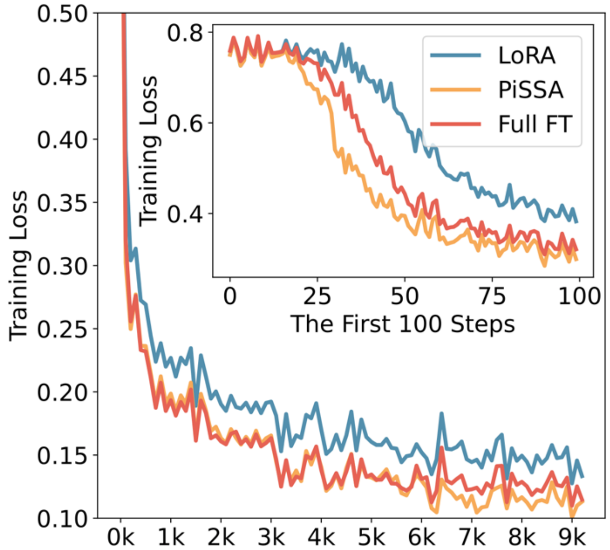
  - Initialize $\mathbf{A}_0 \sim \mathcal{N}(0, \sigma^2), \ \mathbf{B}_0 = \mathbf{0}$ ➤ $\mathbf{W}_0 = \mathbf{W}^{\text{pt}}$

  - Update
$$\mathbf{A}_{t+1} = \mathbf{A}_t - \eta \nabla \ell(\mathbf{W}_t)\mathbf{B}_t := \mathbf{A}_t + \Delta\mathbf{A}_t$$
$$\mathbf{B}_{t+1} = \mathbf{B}_t - \eta \nabla \ell(\mathbf{W}_t)^\top \mathbf{A}_t := \mathbf{B}_t + \Delta\mathbf{B}_t$$

  - ✓ Markedly reduced (up to 100x) computational cost

E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," in *ICLR*, 2022.

Limitations of LoRA



- Unbalanced update and slow convergence

$$\Delta \mathbf{A}_0 = -\eta \nabla \ell(\mathbf{W}_0)\mathbf{B}_0 := \mathbf{0},\ \Delta \mathbf{B}_0 = -\eta \nabla \ell(\mathbf{W}_0)^\top \mathbf{A}_0$$

  - Improved initialization: PiSSA [Meng et al'24], LoRA-GA [Wang et al'24], …

- Nonunique factorization and inconsistent update
  - Define $\Delta \mathbf{W}_t := \mathbf{W}_{t+1} - \mathbf{W}_t = \mathbf{A}_{t+1}\mathbf{B}_{t+1}^\top - \mathbf{A}_t\mathbf{B}_t^\top$
  - Generally, $\tilde{\mathbf{A}}_t\tilde{\mathbf{B}}_t^\top = \mathbf{A}_t\mathbf{B}_t^\top \quad \not\Rightarrow \quad \Delta\tilde{\mathbf{W}}_t = \Delta \mathbf{W}_t$
  - **Ex.** $\tilde{\mathbf{A}}_t = c\mathbf{A}_t, \tilde{\mathbf{B}}_t = c^{-1}\mathbf{B}_t,\ c \neq 0, 1$
    - $$\begin{aligned}\Delta \mathbf{W}_t &= \mathbf{A}_t\Delta\mathbf{B}_t^\top + \Delta\mathbf{A}_t\mathbf{B}_t^\top + \Delta\mathbf{A}_t\Delta\mathbf{B}_t^\top \\ \Delta\tilde{\mathbf{W}}_t &= c^2\mathbf{A}_t\Delta\mathbf{B}_t^\top + c^{-2}\Delta\mathbf{A}_t\mathbf{B}_t^\top + \Delta\mathbf{A}_t\Delta\mathbf{B}_t^\top\end{aligned}$$
  - Invariance-aware optimizers: LoRA-RITE [Yen et al'25], Lora-Pro [Wang et al'25], …

Goal: Identify among all factorizations the optimal one that minimizes the loss per step

F. Meng, Z. Wang, and M. Zhang, "PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models," in *NeurIPS*, 2024.

Parameter symmetries in LoRA

Structure of equivalent low-rank factorizations

Lemma 1 (ours). *Assuming* $\operatorname{rank}(\mathbf{A}_t) = \operatorname{rank}(\mathbf{B}_t) = r$, *it holds that*

$$\{(\tilde{\mathbf{A}}_t, \tilde{\mathbf{B}}_t) \mid \tilde{\mathbf{A}}_t \tilde{\mathbf{B}}_t^\top = \mathbf{A}_t \mathbf{B}_t^\top\} = \{(\mathbf{A}_t \mathbf{P}_t, \mathbf{B}_t \mathbf{P}_t^{-\top}) \mid \mathbf{P}_t \in \mathbb{R}^{r \times r} \text{ invertible}\}.$$

In addition, if $\mathbf{P}_t$ *is orthogonal, it further holds*

$$\Delta \tilde{\mathbf{W}}_t = \Delta \mathbf{W}_t.$$

- $\mathbf{P}_t \in \mathrm{GL}(r)$ characterizes all alternative factorizations
- When $\tilde{\mathbf{A}}_t = \mathbf{A}_t \mathbf{P}_t$ and $\tilde{\mathbf{B}}_t = \mathbf{B}_t \mathbf{P}_t^{-\top}$,

$$\Delta \mathbf{W}_t = \mathbf{A}_t \Delta \mathbf{B}_t^\top + \Delta \mathbf{A}_t \mathbf{B}_t^\top + \Delta \mathbf{A}_t \Delta \mathbf{B}_t^\top$$
$$\Delta \tilde{\mathbf{W}}_t = \mathbf{A}_t \mathbf{P}_t \mathbf{P}_t^\top \Delta \mathbf{B}_t^\top + \Delta \mathbf{A}_t (\mathbf{P}_t \mathbf{P}_t^\top)^{-1} \mathbf{B}_t^\top + \Delta \mathbf{A}_t \Delta \mathbf{B}_t^\top$$

- $\Delta \tilde{\mathbf{W}}_t$ determined by SPD matrix $\mathbf{S}_t := \mathbf{P}_t \mathbf{P}_t^\top \in \mathbb{S}_{++}^r$
- Let $\mathbf{P}_t \overset{\text{svd}}{=} \mathbf{U}_t \mathbf{\Sigma}_t \mathbf{V}_t^\top$; arbitrary orthogonal matrix $\mathbf{V}_t$ not affecting $\mathbf{S}_t = \mathbf{U}_t \mathbf{\Sigma}_t^2 \mathbf{U}_t^\top$

Goal: identify the optimal $\mathbf{S}_t \in \mathbb{S}_{++}^r$ that minimizes $\ell(\mathbf{W}_t + \Delta \tilde{\mathbf{W}}_t(\mathbf{S}_t))$ for each t

Y. Zhang, B. Li, and G. B. Giannakis, "RefLoRA: Refactored Low-Rank Adaptation for Efficient Fine-Tuning of Large Models," in *NeurIPS*, 2025.

Refactoring via upper bound minimization

- Refactored low-rank adaptation (RefLoRA)

Proposition 2 (ours). *If ℓ is L-Lipschitz smooth, it follows that*

$$\ell(\mathbf{W}_t+\Delta\tilde{\mathbf{W}}_t(\mathbf{S}_t)) \leq \frac{L\eta^2}{2}\|\nabla\ell(\mathbf{W}_t)\|_2^2\Big(\|\mathbf{A}_t\mathbf{S}_t^{\frac{1}{2}}\|_{\mathrm{F}}^2+\|\mathbf{B}_t\mathbf{S}_t^{-\frac{1}{2}}\|_{\mathrm{F}}^2-\frac{1}{L\eta}\Big)^2+\mathcal{O}(L\eta^3)+\mathrm{Const.}$$

- Minimize upper bound

$$\mathbf{S}_t^* = \operatorname*{arg\,min}_{\mathbf{S}_t\in\mathbb{S}_{++}^r} \Big(\|\mathbf{A}_t\mathbf{S}_t^{\frac{1}{2}}\|_{\mathrm{F}}^2 + \|\mathbf{B}_t\mathbf{S}_t^{-\frac{1}{2}}\|_{\mathrm{F}}^2 - \frac{1}{L\eta}\Big)^2$$

Theorem 3 (ours). *Define* $\tilde{\mathbf{S}}_t = (\mathbf{A}_t^\top\mathbf{A}_t)^{-\frac{1}{2}}[(\mathbf{A}_t^\top\mathbf{A}_t)^{\frac{1}{2}}\mathbf{B}_t^\top\mathbf{B}_t(\mathbf{A}_t^\top\mathbf{A}_t)^{\frac{1}{2}}]^{\frac{1}{2}}(\mathbf{A}_t^\top\mathbf{A}_t)^{-\frac{1}{2}}$ *and* $\tilde{C}_t = \|\mathbf{A}_t\check{\mathbf{S}}_t^{1/2}\|_{\mathrm{F}}^2 + \|\mathbf{B}_t\check{\mathbf{S}}_t^{-1/2}\|_{\mathrm{F}}^2$ *. It holds that*

$$\mathbf{S}_t^* \begin{cases} = \tilde{\mathbf{S}}_t & , \text{ if } \eta \geq \frac{1}{\tilde{C}_t L} \\ \ni \Big[(\tilde{C}_t L\eta)^{-1} \pm \sqrt{(\tilde{C}_t L\eta)^{-2}-1}\Big]\tilde{\mathbf{S}}_t & , \text{ otherwise} \end{cases} .$$

- "Refactor" via $\tilde{\mathbf{A}}_t = \mathbf{A}_t\mathbf{S}_t^{*1/2}$, $\tilde{\mathbf{B}}_t = \mathbf{B}_t\mathbf{S}_t^{*-1/2}$, and update via e.g., Adam(W)

Y. Zhang, B. Li, and G. B. Giannakis, "RefLoRA: Refactored Low-Rank Adaptation for Efficient Fine-Tuning of Large Models," in *NeurIPS*, 2025.

Computational overhead

- Overhead comparison

Method	Time	Space
LoRA forward/backward	$\Omega(mn)$	$\Omega(mn)$
LoRA-Pro [55]	$\mathcal{O}(m^2r + (m+n+r)r^2)$	$\mathcal{O}(m^2 + (m+n+r)r)$
LoRA-RITE [61]	$\mathcal{O}((m+n+r)r^2)$	$\mathcal{O}((m+n+r)r)$
RefLoRA (Thm. 3)	$\mathcal{O}((m+n+r)r^2)$	$\mathcal{O}(r^2)$
RefLoRA-S (Thm. 4)	$\mathcal{O}((m+n)r)$	$\mathcal{O}(1)$

  - Efficiency in both time and memory

- Simplified refactoring (RefLoRA-S)

  - Constrain $\mathbf{S}_t = s_t\mathbf{I}_r,\ s_t \in \mathbb{R}_{++}$
  - $\tilde{\mathbf{A}}_t = \sqrt{s_t}\mathbf{A}_t,\ \tilde{\mathbf{B}}_t = \frac{1}{\sqrt{s_t}}\mathbf{B}_t$

Theorem 4 (ours). *For RefLoRA-S, it holds that*

$$s_t^* = \begin{cases} \frac{\|\mathbf{B}_t\|_\mathrm{F}}{\|\mathbf{A}_t\|_\mathrm{F}} & ,\ \text{if } \eta \geq \frac{1}{2\|\mathbf{B}_t\|_\mathrm{F}\|\mathbf{A}_t\|_\mathrm{F}L} \\ \frac{\frac{1}{L\eta} \pm \sqrt{\frac{1}{L^2\eta^2} - 4\|\mathbf{A}_t\|_\mathrm{F}^2\|\mathbf{B}_t\|_\mathrm{F}^2}}{2\|\mathbf{A}_t\|_\mathrm{F}^2} & ,\ \text{otherwise} \end{cases}.$$

RefLoRA properties

- Balanced Gram matrices $\tilde{\mathbf{A}}_t^\top \tilde{\mathbf{A}}_t = \tilde{\mathbf{B}}_t^\top \tilde{\mathbf{B}}_t$
  - Balanced Frobenius norm $\|\tilde{\mathbf{A}}_t\|_{\mathrm{F}}^2 = \|\tilde{\mathbf{B}}_t\|_{\mathrm{F}}^2$
  - Inherently satisfied by SVD-based initialization such as PiSSA
  - Maximal loss reduction

Theorem 5 (ours). *The solution* $\mathbf{S}_t = \tilde{\mathbf{S}}_t$ *minimizes the lower bound*

$$0 \geq \underbrace{\langle \nabla_{\tilde{\mathbf{A}}_t} \ell(\mathbf{W}_t), \Delta\tilde{\mathbf{A}}_t \rangle_{\mathrm{F}} + \langle \nabla_{\tilde{\mathbf{B}}_t} \ell(\mathbf{W}_t), \Delta\tilde{\mathbf{B}}_t \rangle_{\mathrm{F}}}_{\approx \Delta\ell(\mathbf{W}_t)} \geq -\eta \|\nabla \ell(\mathbf{W}_t)\|_2^2 (\|\mathbf{A}_t \mathbf{S}_t^{1/2}\|_{\mathrm{F}}^2 + \|\mathbf{B}_t \mathbf{S}_t^{-1/2}\|_{\mathrm{F}}^2).$$

- Consistent weight updates

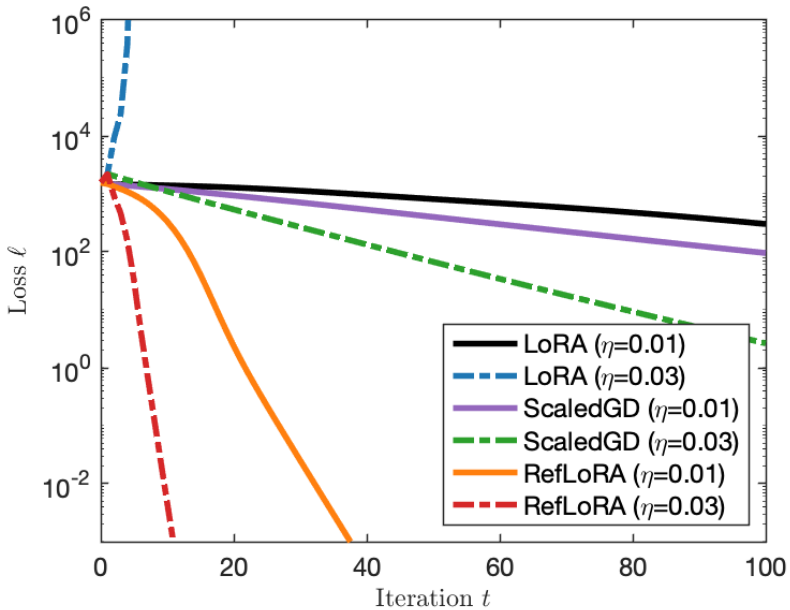
Theorem 6 (ours). *For any* $\mathbf{A}'_t \mathbf{B}'^\top_t = \mathbf{A}_t \mathbf{B}_t^\top$, *let* $\Delta\tilde{\mathbf{W}}'_t$ *and* $\Delta\tilde{\mathbf{W}}_t$ *be their weight updates with RefLoRA. It always holds*

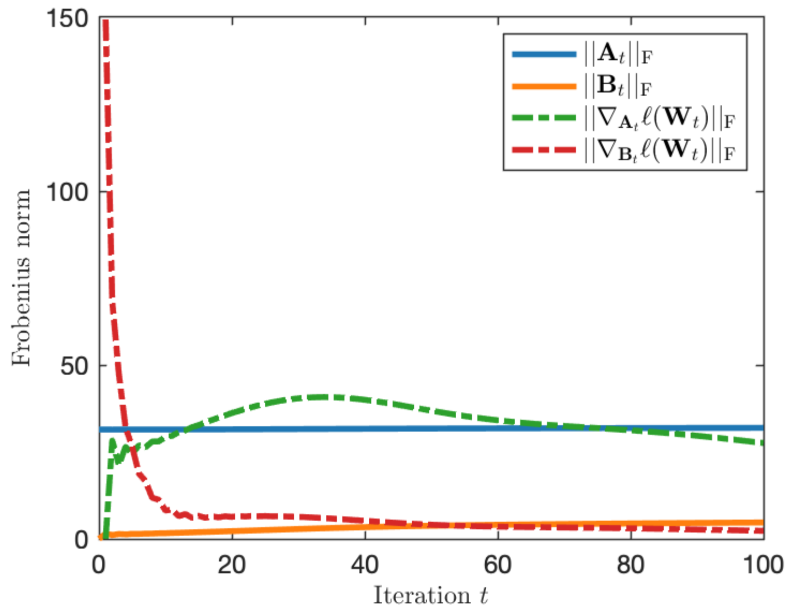
$$\Delta\tilde{\mathbf{W}}'_t = \Delta\tilde{\mathbf{W}}_t.$$

Toy test

- Matrix factorization $$\min_{\mathbf{A},\mathbf{B}} \frac{1}{2}\|\mathbf{Y}-\mathbf{A}\mathbf{B}^\top\|_F^2$$

  - Can be viewed as LoRA on a single-layer model, with whitened inputs
  - ScaledGD: tailored for low-rank matrix factorization; popular among LoRA variants





- Left: RefLoRA demonstrates stable and faster convergence
- Right: LoRA slows down due to the unbalanced update

Natural language understanding (NLU)

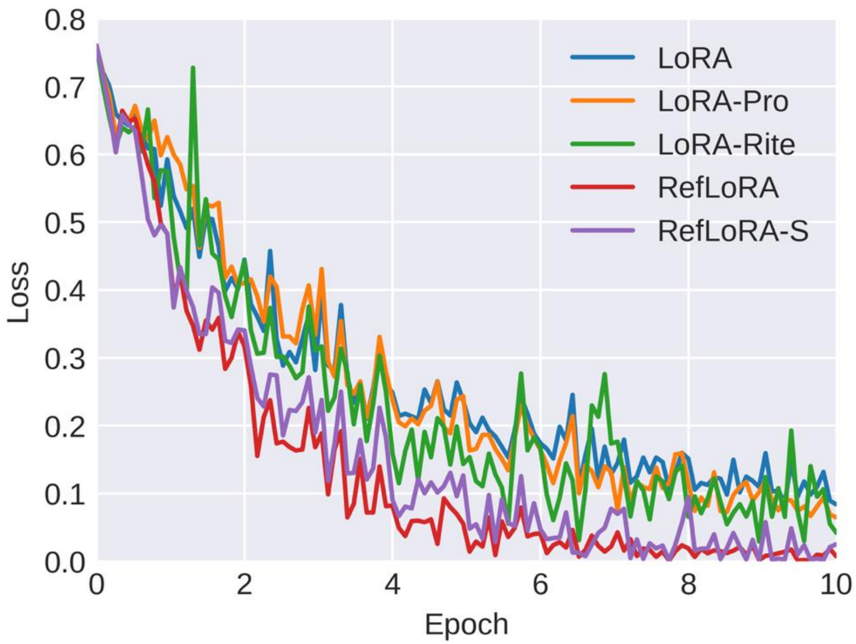
- GLUE benchmark [Wang et al'19] using DeBERTaV3-base [He et al'23]
- 5 random runs with $r = 8$

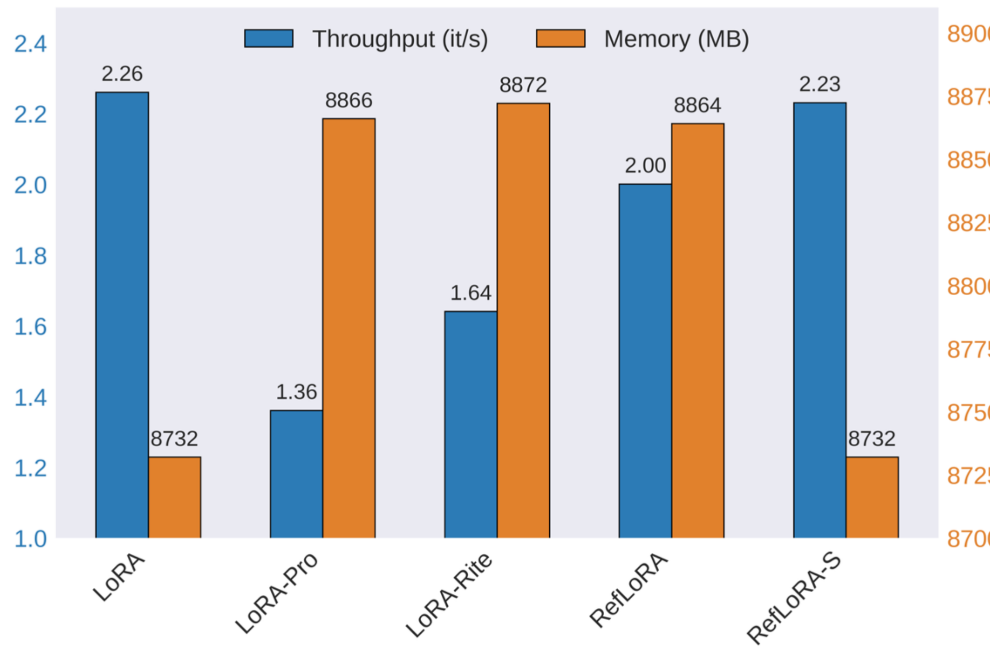
Method	Params	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	All
		Mcc	Acc	Acc	Corr	Acc/F1	M/Mm	Acc	Acc	Avg
Full FT	184M	69.19	95.63	89.46	91.60	92.40/89.80	89.90/90.12	94.03	83.75	88.25
BitFit	0.1M	66.96	94.84	87.75	91.35	88.41/84.95	89.37/89.91	92.24	78.70	86.20
HAdapter	1.22M	68.64	95.53	89.95	91.48	91.91/89.27	90.13/90.17	94.11	84.48	88.28
PAdapter	1.18M	68.77	95.61	89.46	91.54	92.04/89.40	90.33/90.39	94.29	85.20	88.41
LoRA	1.33M	69.82	94.95	89.95	91.60	91.99/89.38	90.65/90.69	93.87	85.20	88.50
DoRA	1.33M	70.85	95.79	90.93	91.79	92.07/-	90.29/-	94.10	86.04	88.98
AdaLoRA	1.27M	71.45	**96.10**	90.69	91.84	92.23/89.74	**90.76/90.79**	**94.55**	88.09	89.46
LoRA-Pro	1.33M	71.36	95.76	90.20	91.92	92.19/89.60	90.23/90.19	94.29	85.56	88.94
LoRA-RITE	1.33M	69.55	95.41	90.93	91.79	92.02/89.42	90.22/90.33	94.42	85.20	88.69
RefLoRA	1.33M	**71.73**	95.99	**91.42**	92.03	92.28/89.70	90.23/90.41	94.40	**88.09**	**89.52**
RefLoRA-S	1.33M	70.66	95.76	90.44	**92.21**	**92.43/89.89**	90.13/90.17	94.16	87.73	89.19

- Outperforms SOTA methods on 5 out of 8 datasets; comparable on the rest 3
- Best average performance; marginal drop (0.33%) with lightweight RefLoRA-S

https://gluebenchmark.com

Convergence and overhead tests





- Faster and stabler convergence
- Higher throughput and reduced memory usage than SOTA approaches
- RefLoRA-S is efficient and scalable as LoRA

Commonsense reasoning

- Commonsense-170k [Hu et al'23] with LLaMA series [Touvron et al'23; Grattafiori et al'24]

	r	Method	Params	BoolQ	PIQA	SIQA	HS	WG	ARCe	ARCc	OBQA	Avg
ChatGPT-3.5-turbo			-	73.1	85.4	68.5	78.5	66.1	89.8	79.9	74.8	77.0
LLaMA-7B	32	LoRA	0.83%	66.42	80.03	77.84	82.88	81.85	79.92	63.40	77.20	76.19
		PrecLoRA	0.83%	68.96	80.95	77.43	81.54	80.27	78.83	64.16	79.20	76.42
		NoRA+	0.83%	69.85	81.83	77.38	82.09	80.03	79.67	64.25	78.60	76.71
		DoRA	0.84%	69.7	83.4	78.6	87.2	81.0	81.9	66.2	79.2	78.4
		LoRA-RITE	0.84%	69.82	82.75	78.55	84.72	81.69	82.15	66.23	81.40	78.54
		RefLoRA	0.83%	69.60	82.48	79.53	88.25	82.56	81.57	66.64	80.20	**78.85**
		RefLoRA-S	0.83%	70.18	82.48	78.15	87.41	82.08	81.52	65.36	81.60	78.60
	16	DoRA	0.43%	70.0	82.6	79.7	83.2	80.6	80.6	65.4	77.6	77.5
		RefLoRA	0.41%	69.66	82.43	79.43	87.38	81.22	80.68	65.44	78.60	**78.11**
		RefLoRA-S	0.41%	67.65	81.50	79.07	88.28	81.77	81.23	64.59	78.60	77.84
LLaMA2-7B	32	LoRA	0.83%	69.8	79.9	79.5	83.6	82.6	79.8	64.7	81.0	77.6
		PrecLoRA	0.83%	71.47	81.50	78.81	85.97	80.43	81.14	66.55	81.00	78.36
		NoRA+	0.83%	70.52	81.94	79.07	87.66	82.24	82.70	67.06	80.20	78.92
		DoRA	0.84%	71.8	83.7	76.0	89.1	82.6	83.7	68.2	82.4	79.7
		LoRA-RITE	0.84%	71.04	82.43	79.79	89.12	84.53	83.88	68.77	81.20	80.10
		RefLoRA	0.83%	72.54	83.79	80.04	86.94	84.85	86.36	71.50	80.20	**80.78**
		RefLoRA-S	0.83%	73.36	83.84	80.76	90.02	82.48	84.55	67.92	82.60	80.69
	16	DoRA	0.43%	72.0	83.1	79.9	89.1	83.0	84.5	71.0	81.2	**80.5**
		RefLoRA	0.41%	71.38	82.43	80.35	90.49	83.43	84.05	69.28	82.00	80.43
		RefLoRA-S	0.41%	72.08	83.03	80.45	85.89	83.27	84.30	69.88	82.00	80.11
LLaMA3-8B	32	LoRA	0.70%	70.8	85.2	79.9	91.7	84.3	84.2	71.2	79.0	80.8
		PrecLoRA	0.70%	70.73	85.80	78.86	91.87	83.66	85.10	71.08	82.40	81.19
		NoRA+	0.70%	71.16	85.10	79.48	92.22	83.35	85.86	72.27	83.20	81.58
		DoRA	0.71%	74.6	89.3	79.9	95.5	85.6	90.5	80.4	85.8	85.2
		LoRA-RITE	0.84%	74.19	89.44	81.52	95.44	86.74	90.45	80.12	86.60	85.56
		RefLoRA	0.70%	75.35	88.74	80.91	95.71	86.66	90.49	80.20	87.40	85.68
		RefLoRA-S	0.70%	75.50	89.72	81.11	95.59	87.29	90.99	79.78	86.00	**85.75**
	16	DoRA	0.35%	74.5	88.8	80.3	95.5	84.7	90.1	79.1	87.2	85.0
		RefLoRA	0.35%	75.26	88.79	81.37	95.85	85.64	90.11	80.55	86.60	**85.52**
		RefLoRA-S	0.35%	74.92	89.01	80.60	95.75	85.24	90.45	80.89	86.40	85.41

- Highest average accuracies in 5 out of 6 setups

Z. Hu, et al., "LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models," in *NeurIPS*, 2024.

Subject-driven image generation

- DreamBooth [Ruiz et al'23] using Stable Diffusion v1.4 [Rombach et al'21] with $r = 4$

Loss	LoRA	LoRA-Pro	LoRA-RITE	RefLoRA
Avg$\pm$std	0.100 ± 0.015	0.099 ± 0.015	0.095 ± 0.016	$\mathbf{0.086} \pm 0.017$

Prompt:
"a dog eating nachos"

- Clearer details and better object fidelity

N. Ruiz, et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," in *CVPR*, 2023.

Concluding remarks

- LoRA suffers from slow convergence, and inconsistent update
- Non-unique factorization characterized by an SPD matrix
- RefLoRA(-S) optimizing loss upper bound
- Extensive numerical tests on various tasks and models
- Faster convergence and consistent update with minimum overhead

Thank You!