Numerical Algorithm-driven Generative Model Design and Adaptation

Hongteng Xu

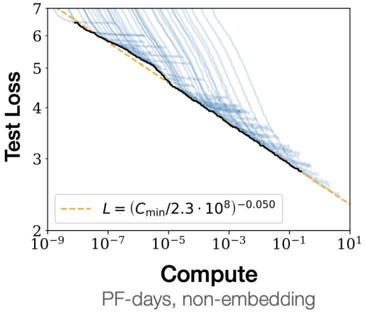
GSAI, RUC

Aug. 10, 2024

中国人民大學
RENMIN UNIVERSITY OF CHINA

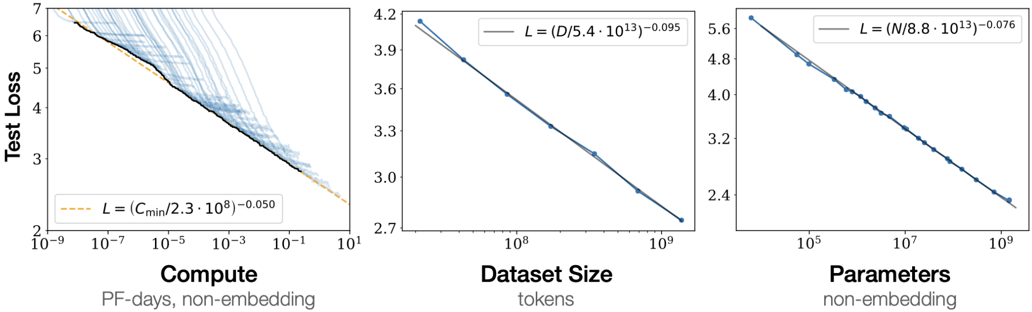
高瓴人工智能学院
Gaoling School of Artificial Intelligence

Demands under "Scaling Laws" of AIGC



Kaplan, Jared, et al. Scaling laws for neural language models. arXiv:2001.08361 (2020).

Demands under "Scaling Laws" of AIGC

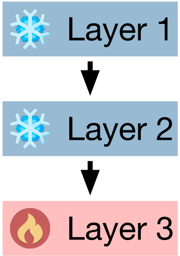


Kaplan, Jared, et al. Scaling laws for neural language models. arXiv:2001.08361 (2020).

- **New model architectures with lower complexity**
- **Parameter-efficient fine-tuning (PEFT) strategies**

Parameter-efficient Fine-tuning Methods

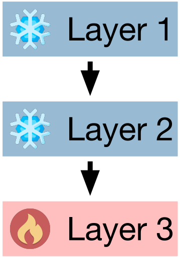
Partial Frozen



Sung, Yi-Lin, et al. "Training neural networks with fixed sparse masks." NeurIPS 2021.

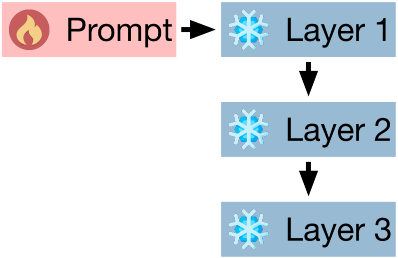
Parameter-efficient Fine-tuning Methods

Partial Frozen



Sung, Yi-Lin, et al. "Training neural networks with fixed sparse masks." NeurIPS 2021.

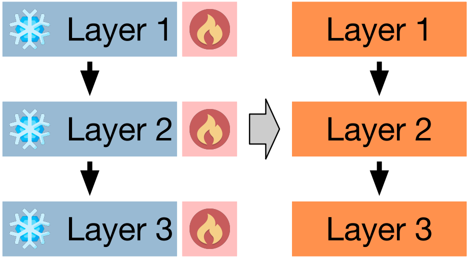
Soft prompt fine-tuning



Li, Xiang Lisa, and Percy Liang. "Prefix-Tuning: Optimizing Continuous Prompts for Generation." ACL 2021.

LoRA: The Mainstream Adapter-based Strategy

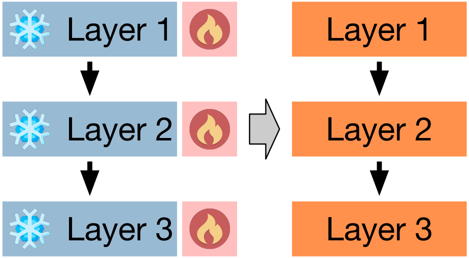
Adapter-based fine-tuning

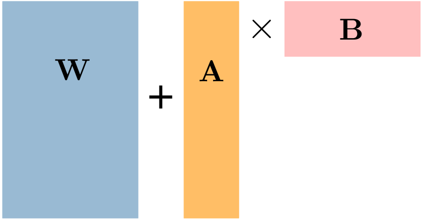


LoRA: The Mainstream Adapter-based Strategy

Adapter-based fine-tuning

Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." ICLR 2022.

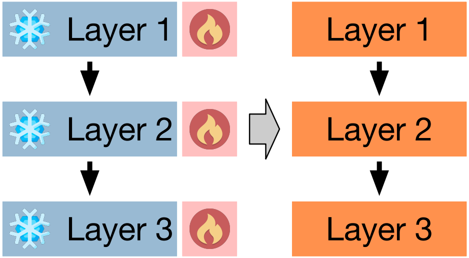


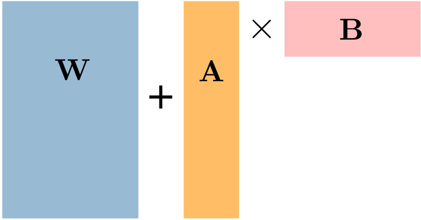


LoRA: The Mainstream Adapter-based Strategy

Adapter-based fine-tuning

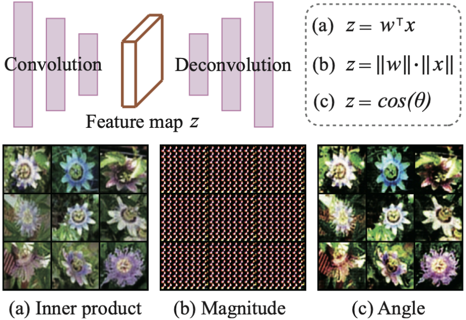
Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." ICLR 2022.





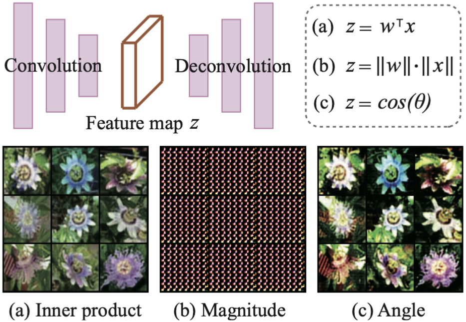
- **Structure Adjustment:** adjust matrices' ranks
- **Initialization Improvement:** mainly based on the SVD of weight matrices
- **Parameter Quantization:** lower bits, sparser matrices

OFT: Another Potential Solution

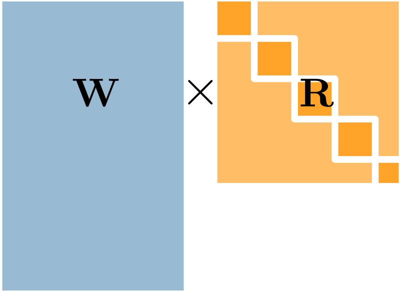


Qiu, Zeju, et al. "Controlling text-to-image diffusion by orthogonal finetuning." NeurIPS 2023.

OFT: Another Potential Solution

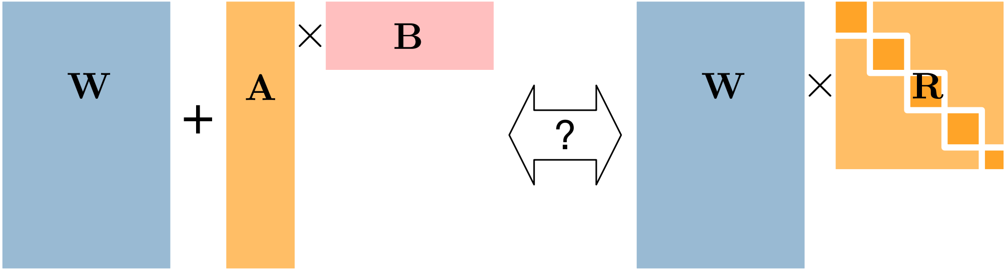


Qiu, Zeju, et al. "Controlling text-to-image diffusion by orthogonal finetuning." NeurIPS 2023.



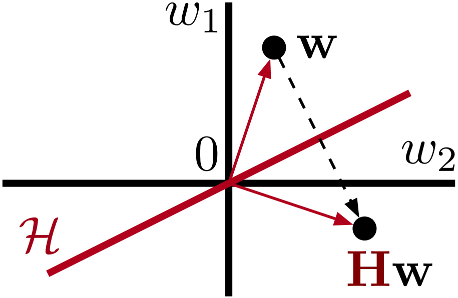
Angular information matters
→ Orthogonal fine-tuning (OFT)

Is There A Bridge between LoRA and OFT?



Householder Reflection: A Simple Orthogonal Transform

Alston Householder

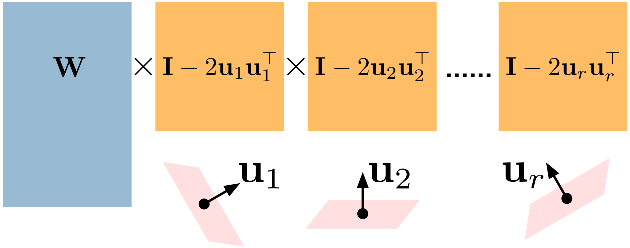


$$\boldsymbol{H} = \boldsymbol{I} - 2\boldsymbol{u}\boldsymbol{u}^\top, \quad \boldsymbol{u} \in \mathbb{S}^{d-1} \qquad (1)$$

Householder Reflection Adaptation (HRA)

- Implement OFT by a chain of Householder reflections

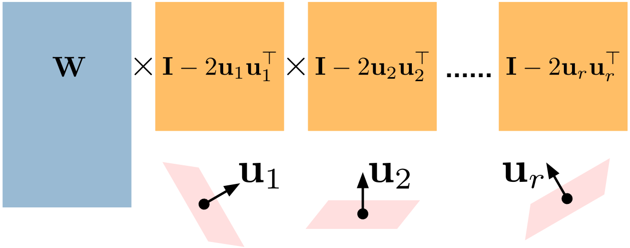
$$\boldsymbol{z} = \boldsymbol{W}\underbrace{\left(\prod_{i=1}^{r}\boldsymbol{H}_i\right)}_{\boldsymbol{H}^{(r)}}\boldsymbol{x} = \boldsymbol{W}\left(\prod_{i=1}^{r}(\boldsymbol{I} - 2\boldsymbol{u}_i\boldsymbol{u}_i^\top)\right)\boldsymbol{x}, \text{ with } \{\boldsymbol{u}_i \in \mathbb{S}^{d-1}\}_{i=1}^{r}. \tag{2}$$



Householder Reflection Adaptation (HRA)

- Implement OFT by a chain of Householder reflections

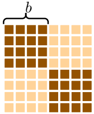
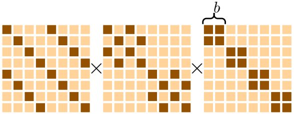
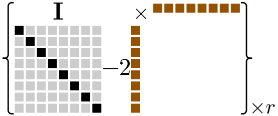
$$\boldsymbol{z} = \boldsymbol{W}\underbrace{\left(\prod_{i=1}^{r}\boldsymbol{H}_i\right)}_{\boldsymbol{H}^{(r)}}\boldsymbol{x} = \boldsymbol{W}\left(\prod_{i=1}^{r}(\boldsymbol{I} - 2\boldsymbol{u}_i\boldsymbol{u}_i^{\top})\right)\boldsymbol{x}, \text{ with } \{\boldsymbol{u}_i \in \mathbb{S}^{d-1}\}_{i=1}^{r}. \quad (2)$$



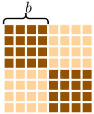
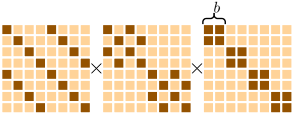
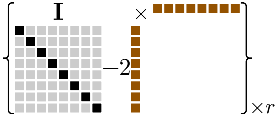
- Implement $\boldsymbol{W}\boldsymbol{H}^{(r)}\boldsymbol{x}$ with low complexity ($\mathcal{O}(d(r + d_{\text{out}}))$ for $\boldsymbol{W} \in \mathbb{R}^{d_{\text{out}}\times d}$)

$$1)\ \boldsymbol{x}^{(j+1)} = \boldsymbol{x}^{(j)} - 2\langle \boldsymbol{u}_{r-j}, \boldsymbol{x}^{(j)}\rangle \boldsymbol{u}_{r-j}, \text{ for } j = 0, ..., r-1. \quad 2)\ \boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}^{(r)}. \quad (3)$$

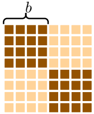
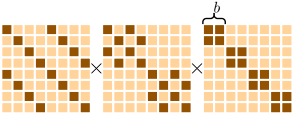
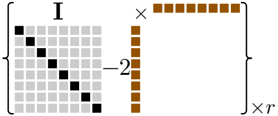
Comparisons with Existing OFTs

Method	OFT	BOFT	Our HRA
Implementation	$\boldsymbol{R}^{(b)} = \mathrm{diag}(\{\boldsymbol{R}_i\}_{i=1}^{d/b})$	$\boldsymbol{B}^{(m,b)} = \prod_{i=1}^{m} \boldsymbol{B}_i^{(b)}$	$\boldsymbol{H}^{(r)} = \prod_{i=1}^{r} \boldsymbol{I} - 2\boldsymbol{u}_i\boldsymbol{u}_i^{\top}$
Illustration			

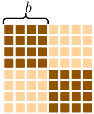
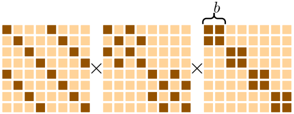
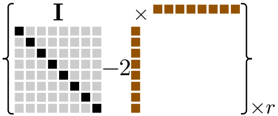
Comparisons with Existing OFTs

Method	OFT	BOFT	Our HRA
Implementation	$\boldsymbol{R}^{(b)}=\text{diag}(\{\boldsymbol{R}_i\}_{i=1}^{d/b})$	$\boldsymbol{B}^{(m,b)}=\prod_{i=1}^{m}\boldsymbol{B}_i^{(b)}$	$\boldsymbol{H}^{(r)}=\prod_{i=1}^{r}\boldsymbol{I}-2\boldsymbol{u}_i\boldsymbol{u}_i^{\top}$
Illustration			
#Parameters	$\frac{d(b-1)}{2}\sim db$	$\frac{dm(b-1)}{2}\sim dmb$	rd
Complexity	$\mathcal{O}(d(b^2+b+d_{\text{out}}))$	$\mathcal{O}(d((b^2+b)m+d_{\text{out}}))\sim$ $\mathcal{O}(d((b^2+d)m+d_{\text{out}}))$	$\mathcal{O}(d(r+d_{\text{out}}))$
	Cayley Transform: $\boldsymbol{R}=(\boldsymbol{I}-\boldsymbol{A})(\boldsymbol{I}+\boldsymbol{A})^{-1}$		

Comparisons with Existing OFTs

Method	OFT	BOFT	Our HRA
Implementation	$\boldsymbol{R}^{(b)} = \mathrm{diag}(\{\boldsymbol{R}_i\}_{i=1}^{d/b})$	$\boldsymbol{B}^{(m,b)} = \prod_{i=1}^{m} \boldsymbol{B}_i^{(b)}$	$\boldsymbol{H}^{(r)} = \prod_{i=1}^{r} \boldsymbol{I} - 2\boldsymbol{u}_i\boldsymbol{u}_i^\top$
Illustration			
#Parameters	$\frac{d(b-1)}{2} \sim db$	$\frac{dm(b-1)}{2} \sim dmb$	rd
Complexity	$\mathcal{O}(d(b^2 + b + d_{\text{out}}))$	$\mathcal{O}(d((b^2 + b)m + d_{\text{out}})) \sim$ $\mathcal{O}(d((b^2 + d)m + d_{\text{out}}))$	$\mathcal{O}(d(r + d_{\text{out}}))$
	Cayley Transform: $\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{I} + \boldsymbol{A})^{-1}$		
Cover $\mathbb{O}_{d\times d}$	$b = d$	$\{\boldsymbol{B}_i^{(m=\log d, b=2)}\}_{i=1}^{d-1}$	$\{\boldsymbol{u}_i\}_{i=1}^{d-1}$

Comparisons with Existing OFTs

Method	OFT	BOFT	Our HRA
Implementation	$\boldsymbol{R}^{(b)} = \text{diag}(\{\boldsymbol{R}_i\}_{i=1}^{d/b})$	$\boldsymbol{B}^{(m,b)} = \prod_{i=1}^{m} \boldsymbol{B}_i^{(b)}$	$\boldsymbol{H}^{(r)} = \prod_{i=1}^{r} \boldsymbol{I} - 2\boldsymbol{u}_i\boldsymbol{u}_i^{\top}$
Illustration			
#Parameters	$\frac{d(b-1)}{2} \sim db$	$\frac{dm(b-1)}{2} \sim dmb$	rd
Complexity	$\mathcal{O}(d(b^2 + b + d_{\text{out}}))$	$\mathcal{O}(d((b^2 + b)m + d_{\text{out}})) \sim$ $\mathcal{O}(d((b^2 + d)m + d_{\text{out}}))$	$\mathcal{O}(d(r + d_{\text{out}}))$
	Cayley Transform: $\boldsymbol{R} = (\boldsymbol{I} - \boldsymbol{A})(\boldsymbol{I} + \boldsymbol{A})^{-1}$		
Cover $\mathbb{O}_{d\times d}$	$b = d$	$\{\boldsymbol{B}_i^{(m=\log d, b=2)}\}_{i=1}^{d-1}$	$\{\boldsymbol{u}_i\}_{i=1}^{d-1}$

Have potentials to be more efficient in practice.

Connections to LoRA: HRA is An Adaptive LoRA

- Reformulation of the HR chain:

$$\boldsymbol{H}^{(r)} = \prod_{i=1}^{r} (\boldsymbol{I} - 2\boldsymbol{u}_i\boldsymbol{u}_i^\top) = \boldsymbol{I} + \boldsymbol{U}_r\boldsymbol{\Gamma}_r\boldsymbol{U}_r^\top, \tag{4}$$

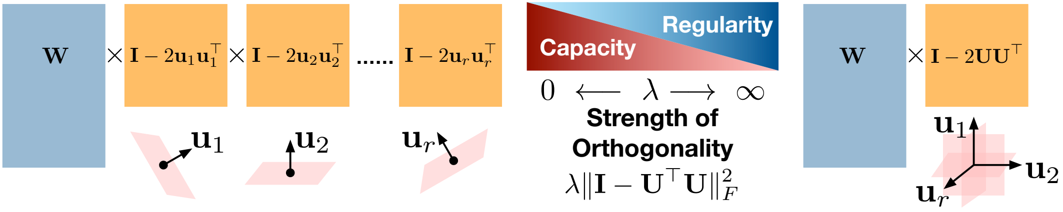
- $\boldsymbol{\Gamma}_r = [\gamma_{ij}] \in \mathbb{R}^{r\times r}$ is a upper-triangular matrix, and its upper-riangular element is

$$\gamma_{ij} = \begin{cases} -2 & i = j \\ (-2)^{j-i+1} \prod_{i=1}^{j-1} \langle \boldsymbol{u}_i, \boldsymbol{u}_{i+1} \rangle & i < j. \end{cases} \tag{5}$$

- HRA is equivalent to an adaptive LoRA, making $Range(\boldsymbol{W})$ unchanged.

$$\boldsymbol{W}\boldsymbol{H}^{(r)} = \boldsymbol{W} + \underbrace{\boldsymbol{W}\boldsymbol{U}_r\boldsymbol{\Gamma}_r}_{\boldsymbol{A}_{\boldsymbol{W},\boldsymbol{U}}}\boldsymbol{U}_r^\top. \tag{6}$$

Orthogonality: The Key of Balancing Expressiveness and Regularity



$$\min_{\{\boldsymbol{U}_r^{(l)}\}_{l=1}^L} \mathsf{Loss}(\mathcal{D}; \{\boldsymbol{U}_r^{(l)}\}_{l=1}^L) + \underbrace{\lambda \sum_{l=1}^{L} \|\boldsymbol{I}_r - (\boldsymbol{U}_r^{(l)})^\top \boldsymbol{U}_r^{(l)}\|_F^2}_{\text{Orthogonal regularizer}}, \tag{7}$$

- $\lambda \in [0, \infty)$: Normalization
- $\lambda = \infty$: (Modified) Gram-Schmidt Orthogonalization

Experiments: NLP Tasks

Table: Results (%) of various methods on GLUE development set.

Method	#Param	MNLI	SST-2	CoLA	QQP	QNLI	RTE	MRPC	STS-B	All
Full Fine-tune	184M	89.90	95.63	69.19	**92.40**	94.03	83.75	89.46	91.60	88.25
BitFit	0.10M	89.37	94.84	66.96	88.41	92.24	78.70	87.75	91.35	86.20
H-Adapter	1.22M	90.13	95.53	68.64	91.91	94.11	84.48	89.95	91.48	88.28
P-Adapter	1.18M	90.33	95.61	68.77	92.04	94.29	85.20	89.46	91.54	88.41
LoRA$_{r=8}$	1.33M	90.65	94.95	69.82	91.99	93.87	85.20	89.95	91.60	88.50
AdaLoRA	1.27M	**90.76**	96.10	71.45	<u>92.23</u>	<u>94.55</u>	88.09	90.69	91.84	89.46
OFT$_{b=16}$	0.79M	90.33	96.33	**73.91**	92.10	94.07	87.36	92.16	<u>91.91</u>	89.77
BOFT$_{b=8}^{m=2}$	0.75M	90.25	96.44	72.95	92.10	94.23	<u>88.81</u>	92.40	**91.92**	89.89
HRA$_{r=8,\lambda=0}$	0.66M	<u>90.70</u>	<u>96.45</u>	<u>73.70</u>	91.29	**94.66**	88.45	<u>93.69</u>	91.86	**90.10**
HRA$_{r=8,\lambda=10^{-5}}$	0.66M	90.43	**96.79**	71.91	91.02	94.44	**89.53**	**94.10**	91.74	<u>90.00</u>
HRA$_{r=8,\lambda=\infty}$	0.66M	90.52	95.87	70.71	90.71	94.12	87.00	92.59	91.54	89.13

Experiments: Controllable Text-to-Image Generation

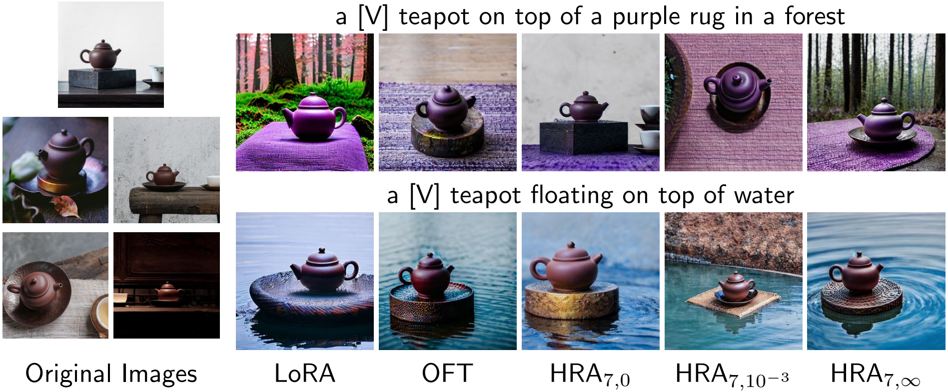
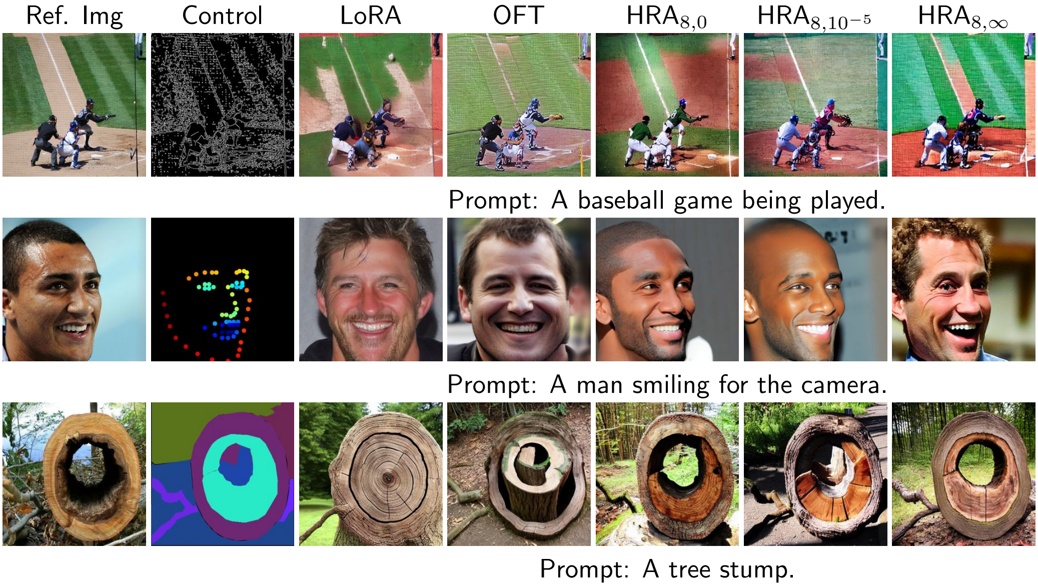


Figure: Qualitative results on subject-driven generation.

Experiments: Controllable Text-to-Image Generation



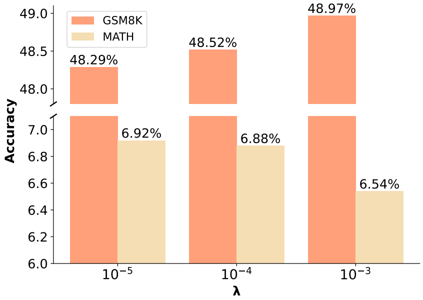
Experiments: Controllable Text-to-Image Generation

Table: Results of various methods on subject-driven generation and controllable generation.

Method	#Param (M)	Subject-driven generation				#Param (M)	Controllable generation					
		Image fidelity		Text fidelity	Diversity		C2I		S2I			L2F
		DINO↑	CLIP-I↑	CLIP-T↑	LPIPS↑		IoU↑	F1↑	mIoU↑	mAcc↑	aAcc↑	Error↓
Real Images	-	0.764	0.890	-	0.562	-	-	-	-	-	-	-
DreamBooth	859.52	0.614	0.778	0.239	0.737	859.52	0.049	0.093	7.72	14.40	33.61	146.19
ControlNet	-	-	-	-	-	361.30	0.189	0.317	20.88	30.91	61.42	7.61
T2I-Adapter	-	-	-	-	-	77.00	0.078	0.143	16.38	26.31	51.63	23.75
LoRA	0.8	0.613	0.765	0.237	0.744	1.25	0.168	0.286	22.98	35.52	58.03	7.68
COFT $_{b=4}$	23.3	0.630	0.783	0.235	0.744	26.40	0.195	0.325	26.92	40.08	62.96	6.92
OFT$_{b=4}$	23.3	0.632	0.785	0.237	0.746	26.40	0.193	0.323	27.06	40.09	62.42	7.07
BOFT$^{m=4}_{r=8}$	-	-	-	-	-	20.76	-	-	28.83	41.24	<u>67.74</u>	5.67
HRA$_{r=8,\lambda=0}$	0.69	**0.670**	**0.803**	0.238	0.758	0.89	**0.213**	**0.350**	**29.45**	**42.02**	66.83	<u>5.56</u>
HRA$_{r=8,\lambda=10^{-3}}$	0.69	<u>0.661</u>	<u>0.799</u>	<u>0.255</u>	<u>0.760</u>	0.89	<u>0.205</u>	<u>0.339</u>	<u>29.27</u>	<u>40.89</u>	**67.86**	**5.46**
HRA$_{r=8,\lambda=\infty}$	0.69	0.651	0.794	**0.274**	**0.778**	0.89	0.201	0.334	28.15	40.22	64.95	11.11

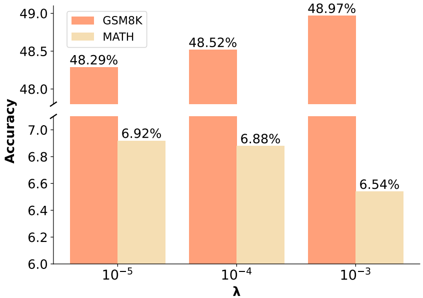
Experiments: Mathematical Reasoning

Method	Param. Ratio	GSM8K	MATH
LLaMA2-7B	-	14.6	2.5
LoRA$_{r=32}$	0.25%	50.2	7.8
OFT$_{b=16}$	0.13%	50.1	8.4
BOFT$^{m=2}_{b=8}$	0.12%	50.6	8.6
PiSSA	4.75%	53.1	7.4
HRA$_{r=8,\lambda=0}$	0.03%	47.1	6.6
HRA$_{r=16,\lambda=0}$	0.06%	52.1	8.1
HRA$_{r=32,\lambda=0}$	0.12%	$\underline{55.8}$	9.0
HRA$_{r=32,\lambda=\infty}$	0.12%	52.8	$\underline{9.2}$
HRA$_{r=32,\lambda=10^{-4}}$	0.12%	**56.3**	**9.3**



Experiments: Mathematical Reasoning

Method	Param. Ratio	GSM8K	MATH
LLaMA2-7B	-	14.6	2.5
LoRA$_{r=32}$	0.25%	50.2	7.8
OFT$_{b=16}$	0.13%	50.1	8.4
BOFT$^{m=2}_{b=8}$	0.12%	50.6	8.6
PiSSA	4.75%	53.1	7.4
HRA$_{r=8,\lambda=0}$	0.03%	47.1	6.6
HRA$_{r=16,\lambda=0}$	0.06%	52.1	8.1
HRA$_{r=32,\lambda=0}$	0.12%	<u>55.8</u>	9.0
HRA$_{r=32,\lambda=\infty}$	0.12%	52.8	<u>9.2</u>
HRA$_{r=32,\lambda=10^{-4}}$	0.12%	**56.3**	**9.3**



Method	Param. Ratio	ARC	HellaSwag	MMLU	Winogrande	HumanEval
LLaMA2-7B	-	49.74	58.90	45.92	74.11	12.80
LoRA$_{r=16}$	0.12%	48.81	56.89	40.60	71.27	11.59
HRA$_{r=32,\lambda=10^{-4}}$	0.12%	49.57	57.72	41.20	73.32	13.41

Discussion: From Model Adaptation to Model Design

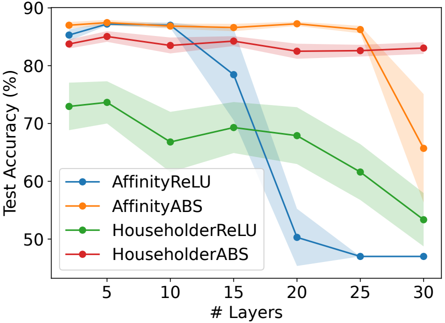
- Householder reflection $\boldsymbol{H}\boldsymbol{x} = \boldsymbol{x} - 2\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{x}$ can work as a (restricted) ResNet.
- $|\boldsymbol{x}|$ can be treated as an adaptive Householder reflection transform.

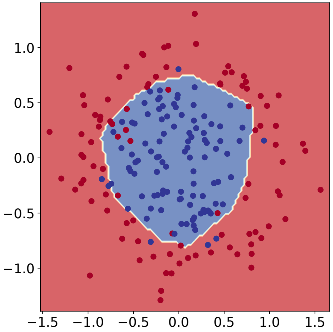
Discussion: From Model Adaptation to Model Design

- Householder reflection $\boldsymbol{H}\boldsymbol{x} = \boldsymbol{x} - 2\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{x}$ can work as a (restricted) ResNet.
- $|\boldsymbol{x}|$ can be treated as an adaptive Householder reflection transform.
- **Stacking HRs leads to a Lipschitz-1 Network.**
  - Support deeper networks without other tricks (LayerNorm, ResNet connection, ...)
  - Robust to attack because of its Lipschitz-1 property.

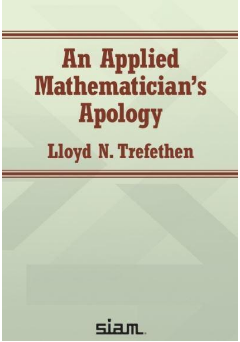
Discussion: From Model Adaptation to Model Design

- Householder reflection $\boldsymbol{H}\boldsymbol{x} = \boldsymbol{x} - 2\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{x}$ can work as a (restricted) ResNet.
- $|\boldsymbol{x}|$ can be treated as an adaptive Householder reflection transform.
- **Stacking HRs leads to a Lipschitz-1 Network.**
  - Support deeper networks without other tricks (LayerNorm, ResNet connection, ...)
  - Robust to attack because of its Lipschitz-1 property.





Roles of “Traditional” Algorithms in The Era of AIGC



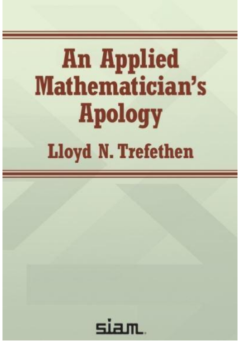
Discrete Algorithms (Computer Science)

- Sorting, Searching, Discrete Optimization, ...

Continuous Algorithms (Scientific Computing)

- SVD, Eigenproblem Solver, Equation Solver, ...

Roles of "Traditional" Algorithms in The Era of AIGC



Discrete Algorithms (Computer Science)

- Sorting, Searching, Discrete Optimization, ...

Continuous Algorithms (Scientific Computing)

- SVD, Eigenproblem Solver, Equation Solver, ...

- Ironically, everyone admits the significance of algorithms, while few of them fully exploit the power of algorithms in their research:(
- "Traditional" algorithm matters!

Summary

Householder Reflection Adaptation

- Bridge the gap between low-rank and orthogonal adaptation
- An typical and effective application of algorithm to model adaptation
- Have potentials to model design

Resources

- Paper: `https://arxiv.org/pdf/2405.17484`
- Code: `https://github.com/DaShenZi721/HRA`
- Email: hongtengxu@ruc.edu.cn

Summary

Householder Reflection Adaptation

- Bridge the gap between low-rank and orthogonal adaptation
- An typical and effective application of algorithm to model adaptation
- Have potentials to model design

Resources

- Paper: `https://arxiv.org/pdf/2405.17484`
- Code: `https://github.com/DaShenZi721/HRA`
- Email: hongtengxu@ruc.edu.cn

Thanks! Q & A