



中國人民大學

RENMIN UNIVERSITY OF CHINA



高瓴人工智能学院

Gaoling School of Artificial Intelligence

Bridging The Gap between Low-rank and Orthogonal Adaptation via Householder Reflection Adaptation

Shen Yuan, Haotian Liu, Hongteng Xu

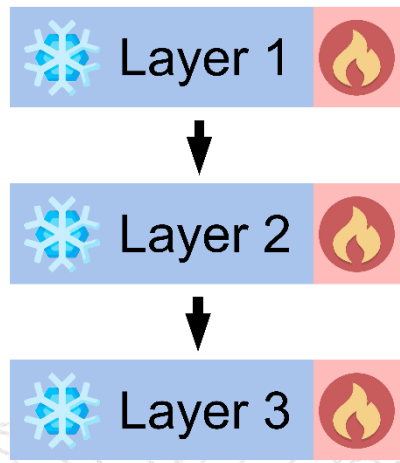
Gaoling School of Artificial Intelligence, Renmin University of China

{shenyuan721, haotianliu, hongtengxu}@ruc.edu.cn

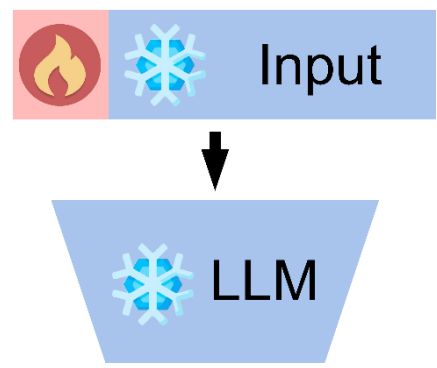
NeurIPS 2024 [Spotlight]

Parameter-efficient fine-tuning (PEFT)

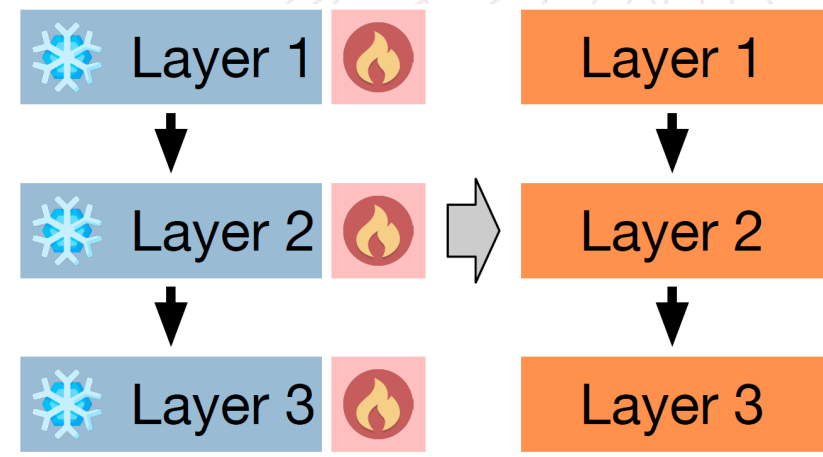
➤ Model fine-tuning



➤ Soft prompt fine-tuning



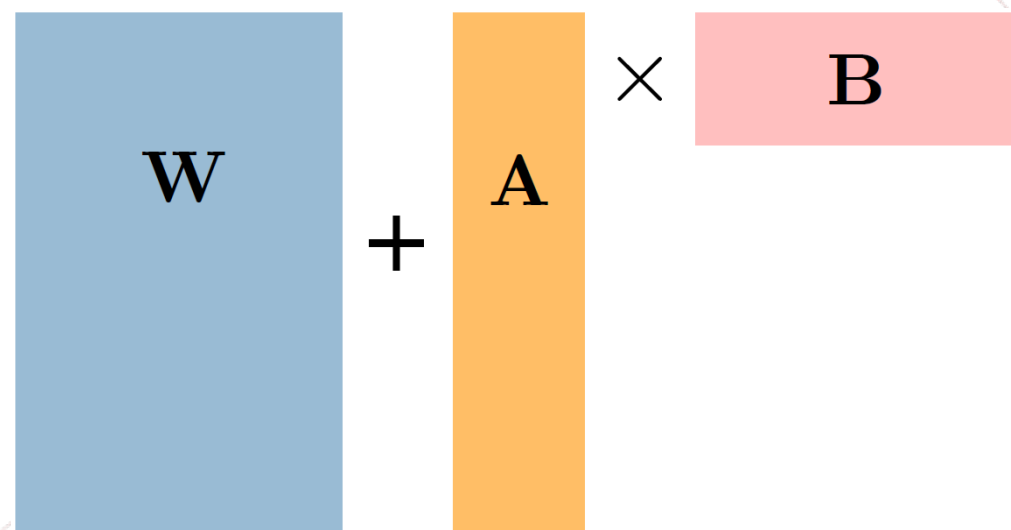
➤ Adapter-based fine-tuning





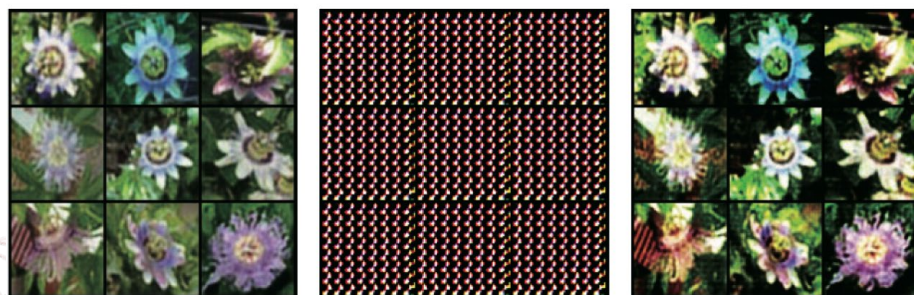
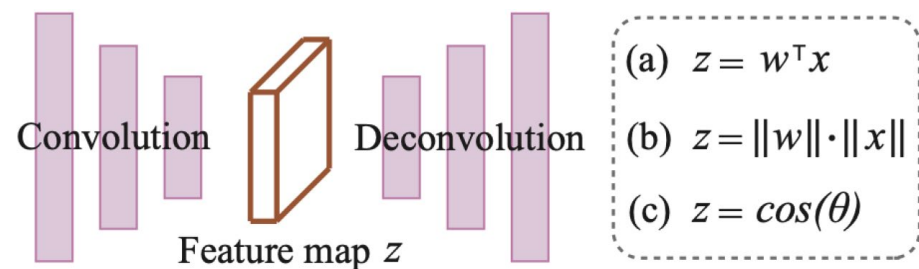
LoRA: The mainstream model adaptation method

- LoRA hypothesizes that the weights have a low “intrinsic rank” during adaptation.



OFT: The preservation of pretrained knowledge

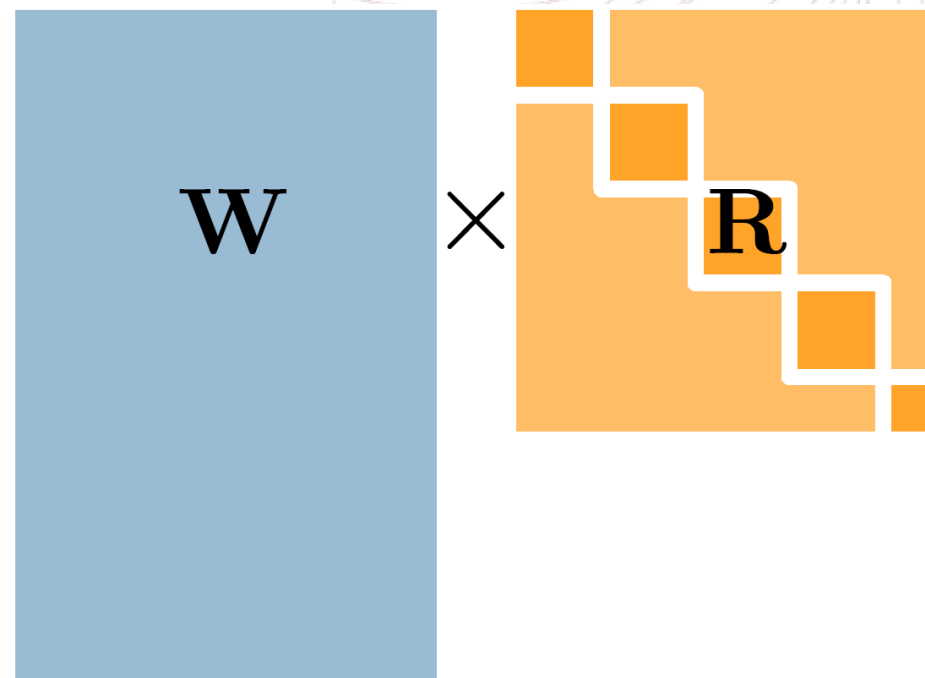
- OFT emphasizes the retention of pre-trained knowledge during adaptation.



(a) Inner product

(b) Magnitude

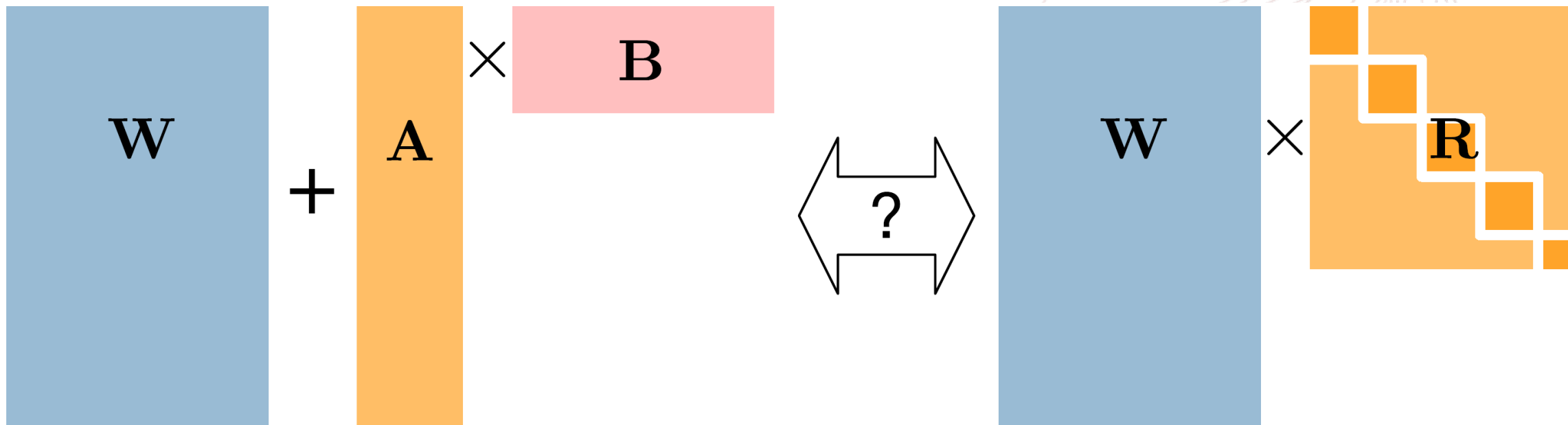
(c) Angle



Angular information matters

→ Orthogonal fine-tuning (OFT)

Is there a bridge between LoRA and OPT?

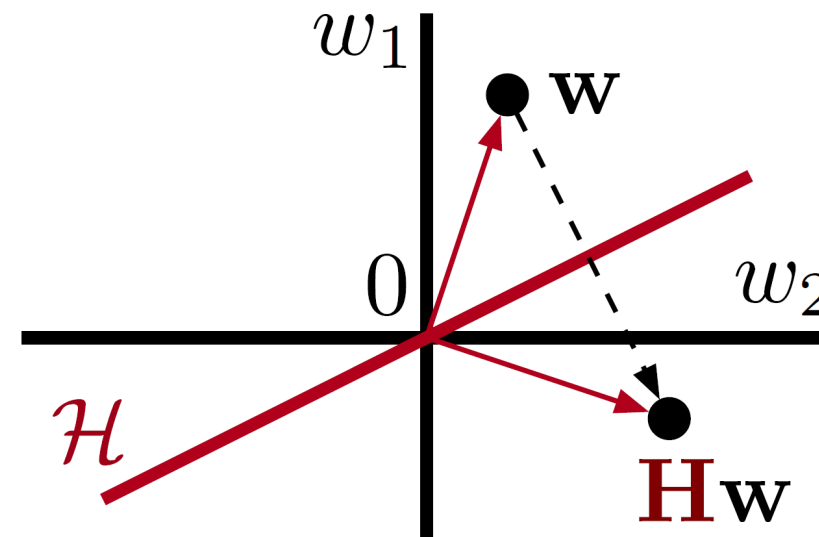




Householder Reflection: A simple orthogonal transform



Alston Householder

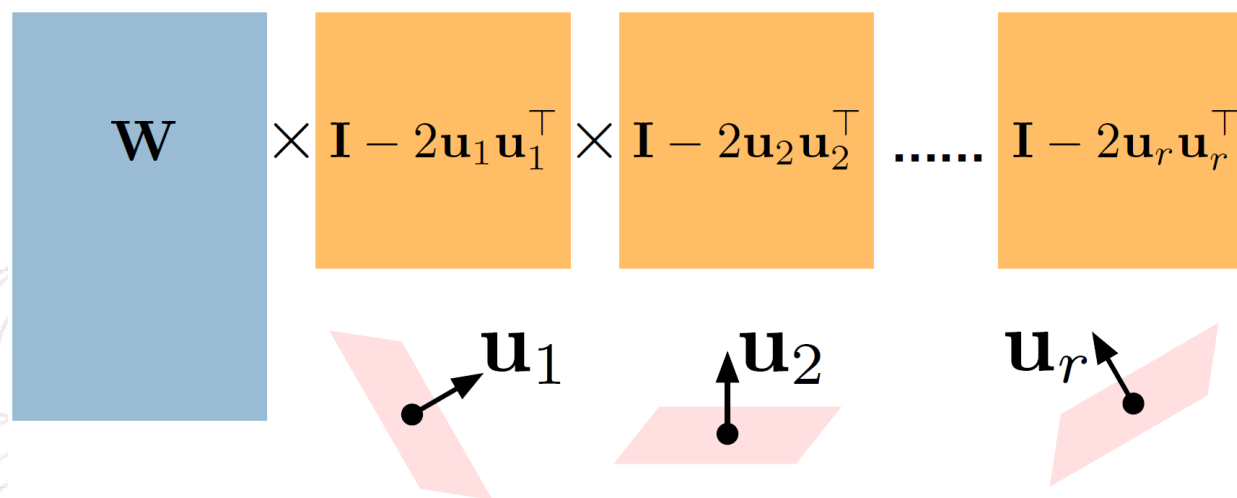


$$H = I - 2uu^T, \quad u \in \mathbb{S}^{d-1}$$

Householder Reflection Adaptation (HRA)

➤ Implement OFT by a chain of Householder reflections

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



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

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



Connections to LoRA: HRA is an adaptive LoRA

- Reformulation of the HR chain:

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

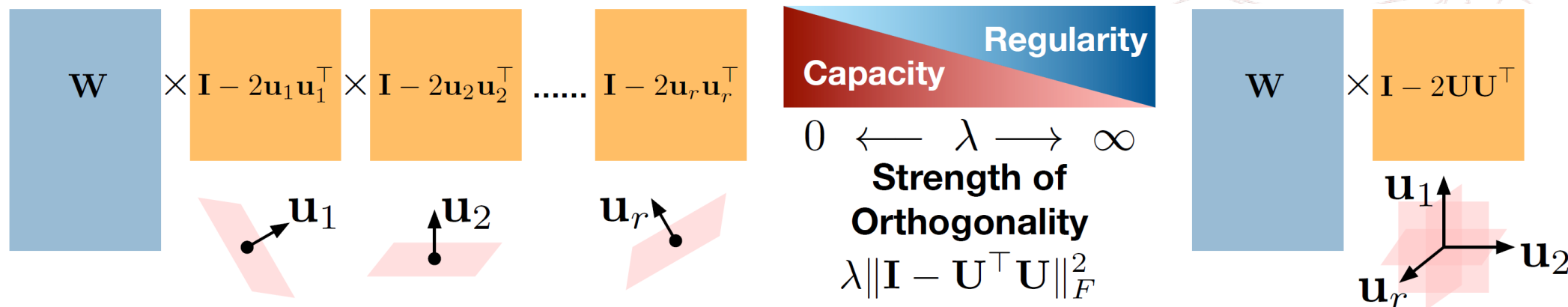
- $\mathbf{\Gamma}_r = [\gamma_{ij}] \in \mathbb{R}^{r \times r}$ is a upper-triangular matrix, and its upper-triangular element is

$$\mathbf{\Gamma}_1 = -2, \quad \mathbf{\Gamma}_r = \begin{bmatrix} \mathbf{\Gamma}_{r-1} & -2\mathbf{\Gamma}_{r-1}\mathbf{U}_{r-1}^\top\mathbf{u}_r \\ \mathbf{0}_{r-1}^\top & -2 \end{bmatrix}.$$

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

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

Orthogonality: The key of balancing expressiveness and regularity



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

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

Experiments: NLP tasks

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

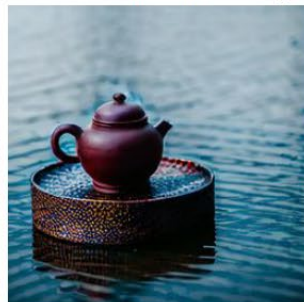
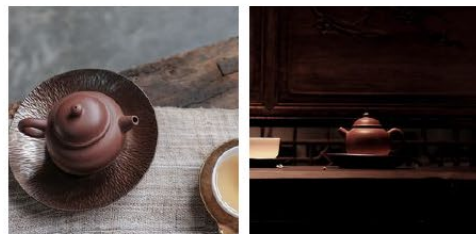
Method	#Param (M)	MNLI	SST-2	CoLA	QQP	QNLI	RTE	MRPC	STS-B	All
Full Fine-tune	184	89.90	95.63	69.19	92.40	94.03	83.75	89.46	91.60	88.25
BitFit	0.10	89.37	94.84	66.96	88.41	92.24	78.70	87.75	91.35	86.20
H-Adapter	1.22	90.13	95.53	68.64	91.91	94.11	84.48	89.95	91.48	88.28
P-Adapter	1.18	90.33	95.61	68.77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

Experiments: Controllable text-to-image generation

a [V] teapot on top of a purple rug in a forest



a [V] teapot floating on top of water



Original Images

LoRA

OFT

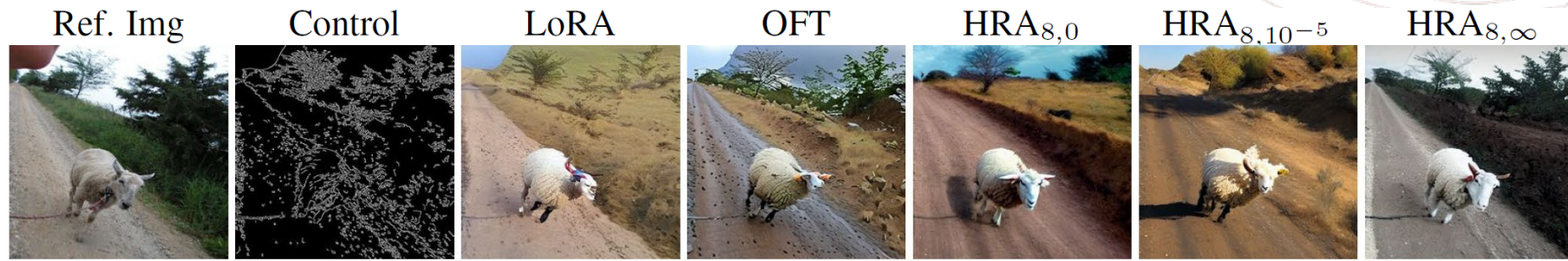
HRA_{7,0}

HRA_{7,10⁻³}

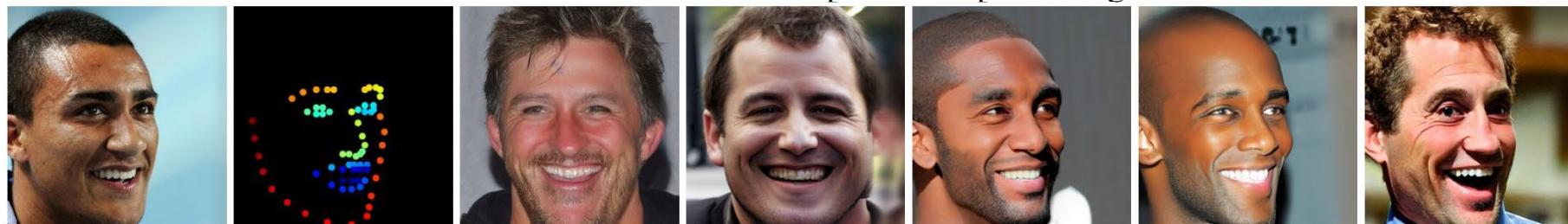
HRA_{7,∞}

Figure: Qualitative results on subject-driven generation.

Experiments: Controllable text-to-image generation



Prompt: A sheep crossing a dirt road.



Prompt: A man smiling for the camera.



Prompt: A tree stump.

Figure: Qualitative results on controllable generation.



Thank you for listening!

The code is available at <https://github.com/DaShenZi721/HRA> and [PEFT](#)!

HRA Public

DaShenZi721 update hyperparameters dd44068 · 4 days ago 16 Commits

assets update 5 months ago

generation update 5 months ago

llama modify requirements.txt 6 months ago

nlu update hyperparameters 4 days ago

README.md update 2 weeks ago

Bridging The Gap between Low-rank and Orthogonal Adaptation via Householder Reflection Adaptation

Capacity Regularity

Strength of Orthogonality $\lambda \|I - U^T U\|_F^2$

arXiv

Hugging Face

Search models, datasets, users...

Models Datasets Spaces Posts Docs Pricing

• PEFT

Search documentation Ctrl+K

MAIN EN 16,276

Polytropon

P-tuning

Prefix tuning

Prompt tuning

Layernorm tuning

VeRA

FourierFT

VB-LoRA

HRA

UTILITIES

Model merge

Helpers

Hotswapping adapters

You are viewing *main* version, which requires [installation from source](#). If you'd like regular pip install, checkout the latest stable version ([v0.13.0](#)).

Bridging The Gap between Low-rank and Orthogonal Adaptation via Householder Reflection Adaptation (HRA)

HRAConfig

HRAModel

Bridging The Gap between Low-rank and Orthogonal Adaptation via Householder Reflection Adaptation (HRA)

HRA is a simple but effective adapter-based fine-tuning method by leveraging Householder reflections. This method harnesses the advantages of both strategies, reducing parameters and computation costs while penalizing the loss of pre-training knowledge. It consistently achieves better performance with fewer trainable parameters and outperforms state-of-the-art adapters across different models, including large language models (LLMs) and conditional image generators.

The abstract from the paper is:

"While following different technical routes, both low-rank and orthogonal adaptation techniques can efficiently adapt large-scale pre-training models in specific tasks or domains based on a small piece of trainable parameters. In this study, we bridge the gap between these