NEURIPS 2025 SPOTLIGHT

Uni-LoRA: One Vector is All You Need !

Kaiyang Li, Shaobo Han, Qing Su, Wei Li, Zhipeng Cai, Shihao Ji

UCONN NEC GEORGIA STATE

HIGHLIGHTS

- **One Vecter to Rule 'All'**

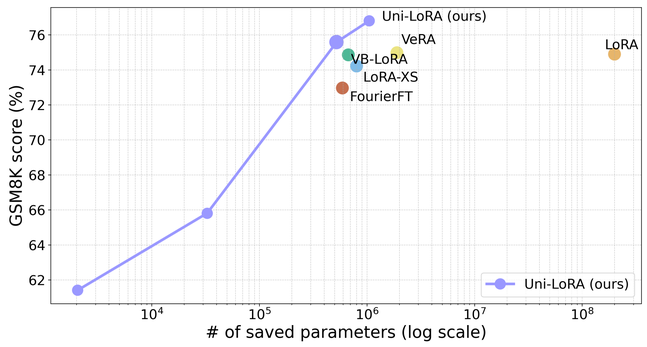
  All major LoRA variants (e.g., Tied-LoRA, VeRA, LoRA-XS, VB-LoRA) are unified as *structured low-dimensional projections* of the full LoRA space.

- **Minimal Representation**

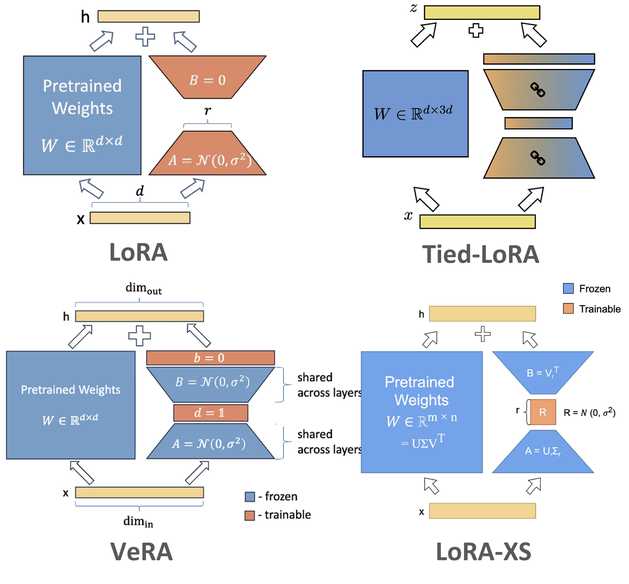
  We compress LoRA into *one tiny vector + a seed*, *enabling* full-model LoRA reconstruction.

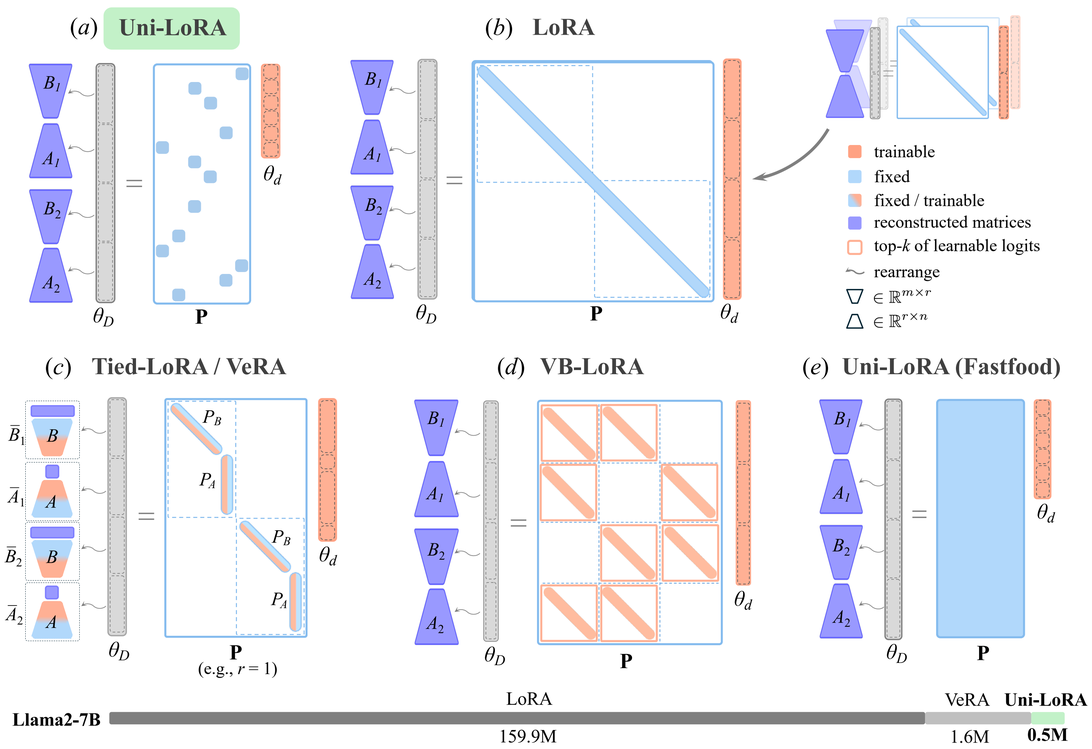
- **Extremely Parameter-Efficient**

  On Gemma-7B, we reach vanilla LoRA performance with **0.52M parameters** (only **0.006%** of the full model, **0.26%** of LoRA)



LoRA Lineup





Core Algorithm

1. Flatten all LoRA matrices into

$$\theta_D = \mathrm{Concat}\big(\mathrm{vec}_{\mathrm{row}}(B_1),\ \mathrm{vec}_{\mathrm{row}}(A_1), \ldots\big)$$

2. Optimize $\theta_d \in \mathbb{R}^d$ and project it back via

$$\theta_D = P\,\theta_d, \qquad d \ll D$$

3. Different LoRA variants use different projection matrices ***P***.

The Making of P

1. Each row of ***P*** selects exactly one nonzero entry at random.
2. For each column *j*, the selected rows are normalized by $1/\sqrt{n_j}$, where n_j is the number of rows assigned to that column.
3. In effect, we randomly group LoRA parameters and tie all parameters within each group.

The G-U-I Principles

- **Globality** ⟶ Share parameters **across layers** to eliminate redundancy and maximize reuse.
- **Uniformity** ⟶ Evenly distribute original dimensions into subspaces so information is **load-balanced.**
- **Isometry** ⟶ Preserve distances and geometry in the parameter space.

$$\|P(x-y)\| = \|x-y\|, \quad \forall x, y \in \mathbb{R}^d$$

Parameter Efficiency & Complexity

- Store only one vector θ_d (and one random seed).
- Projection time and space: $\mathcal{O}(D)$

  Gaussian $\mathcal{O}(Dd)$, Fastfood $\mathcal{O}(D \log d)$

Method	Learnable Projection	Globality	Uniformity	Isometry
VeRA	✗	✗	✗	✗
Tied-LoRA	✓	✗	✗	✗
VB-LoRA	✓	✓	✓	✗
LoRA-XS	✗	✗	✓	✓
Uni-LoRA (Fastfood)	✗	✓	✓	✓
Uni-LoRA (ours)	✗	✓	✓	✓

Results

- **Natural Language Understanding**

  On the GLUE benchmark, Uni-LoRA achieves SOTA performance while using far fewer trainable parameters.

Method	# Params	SST-2	MRPC	CoLA	QNLI	RTE	STS-B	Avg.
LoRA	0.786M	96.2±0.5	90.2±1.0	68.2±1.9	94.8±0.3	85.2±1.1	92.3±0.5	87.8
VeRA	0.061M	96.1±0.1	90.9±0.7	68.0±0.8	94.4±0.2	85.9±0.7	91.7±0.8	87.8
Tied-LoRA	0.066M	94.8±0.6	89.7±1.0	64.7±1.2	94.1±0.1	81.2±0.1	90.8±0.3	85.9
VB-LoRA	0.162M	96.1±0.2	**91.4±0.6**	68.3±0.7	**94.7±0.5**	86.6±1.3	91.8±0.1	88.2
LoRA-XS	0.025M	95.9±0.3	90.7±0.4	67.0±1.2	93.9±0.4	**88.1±0.3**	92.0±0.7	87.9
FourierFT	0.048M	96.0±0.2	90.9±0.3	67.1±1.4	94.4±0.2	87.4±1.6	92.0±0.4	88.0
Uni-LoRA	**0.023M**	**96.3±0.2**	91.3±0.6	**68.5±1.1**	94.6±0.4	86.6±1.6	**92.1±0.1**	**88.3**

- **Instruction Tuning**

  Uni-LoRA achieves higher scores than LoRA while using only **0.3%** (Llama2-7B) and **0.4%** (Llama2-13B) of the parameters.

Model	Method	# Params	Score
Llama2-7B	w/o FT	-	1.11
	LoRA	159.9M	3.23
	VB-LoRA	83M	3.46
	Uni-LoRA	**0.52M**	**3.56**
Llama2-13B	w/o FT	-	1.06
	LoRA	250.3M	4.13
	VB-LoRA	256M	4.33
	Uni-LoRA	**1.0M**	**4.43**

- **Mathematical Reasoning**

  Uni-LoRA surpasses all baselines on Gemma-7B and matches LoRA on Mistral-7B with just 0.3% of the parameters

Model	Method	# Params	GSM8K	MATH
Mistral-7B	LoRA	168M	67.70	19.68
	VeRA	1.39M	68.69	**18.81**
	LoRA-XS	0.92M	68.01	17.86
	VB-LoRA	93M	**69.22**	17.90
	Uni-LoRA	**0.52M**	68.54	18.18
Gemma-7B	LoRA	200M	74.90	31.28
	VeRA	1.90M	79.98	28.84
	LoRA-XS	0.80M	74.22	27.62
	VB-LoRA	113M	74.86	28.90
	Uni-LoRA	**0.52M**	**75.59**	**28.94**

Paper Code LinkedIn