Slaying the HyDRA: Parameter-Efficient Hyper Networks with Low-Displacement Rank Adaptation

Xiangyu Chen, Ye Wang, Matthew Brand, Pu (Perry) Wang, Jing Liu, Toshiaki Koike-Akino

Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA

Novelty

- A new **PEFT** method called **HyDRA** — an integration of hyper networks and LDR adaptation — is introduced to generalize LoRA weight updates with block-wise LDR matrices by sampling parameters from a trainable parameter pool.
- We provide a new mechanism which adjusts the size of parameter pool, providing more flexibility to balance between model size and expressiveness.
- We demonstrate that our HyDRA framework offers a high flexibility and improvement in parameter efficiency for some benchmark experiments.

LDR (Low-Displacement Rank) Matrices

- Structured matrices $\mathbf{W}$ are low rank under **displacement operator**:

$$\nabla_{\mathbf{A},\mathbf{B}}(\mathbf{W}) = \mathbf{AW} - \mathbf{WB} = \mathbf{GH}, \tag{1}$$

for $\mathbf{W} \in \mathbb{R}^{m\times n}, \mathbf{A} \in \mathbb{R}^{m\times m}, \mathbf{B} \in \mathbb{R}^{n\times n}, \mathbf{G} \in \mathbb{R}^{m\times r}, \mathbf{H} \in \mathbb{R}^{r\times n}$.

- LoRA is a *special case* of LDR family.

Circulant *Toeplitz* *Hankel* *Vandermonde*

$$\begin{bmatrix} \mathbf{c}_0 & \mathbf{c}_{-1} & \cdots & \mathbf{c}_{-(n-1)} \\ \mathbf{c}_{-(n-1)} & \mathbf{c}_0 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \mathbf{t}_{-1} \\ \mathbf{c}_{-1} & \cdots & \mathbf{c}_{-(n-1)} & \mathbf{c}_0 \end{bmatrix} \begin{bmatrix} \mathbf{t}_0 & \mathbf{t}_{-1} & \cdots & \mathbf{t}_{-(n-1)} \\ \mathbf{t}_1 & \mathbf{t}_0 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \mathbf{t}_{-1} \\ \mathbf{t}_{n-1} & \cdots & \mathbf{t}_1 & \mathbf{t}_0 \end{bmatrix} \begin{bmatrix} h_0 & \cdots & h_{n-2} & h_{n-1} \\ h_1 & \vdots & h_{n-1} & h_n \\ h_{n-2} & \vdots & \vdots & \vdots \\ h_{n-1} & h_n & \cdots & h_{2n-2} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{v}_0 & \cdots & \mathbf{v}_0^{n-1} \\ 1 & \mathbf{v}_1 & \cdots & \mathbf{v}_1^{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \mathbf{v}_{n-1} & \cdots & \mathbf{v}_{n-1}^{n-1} \end{bmatrix}$$

LDR Advantage in Memory and Complexity

- **Parameter efficient** in linear order $\mathcal{O}[rn]$.
- **Superfast operation** in sub-quadratic order $\mathcal{O}[rn \cdot \mathrm{polylog}(n)]$.

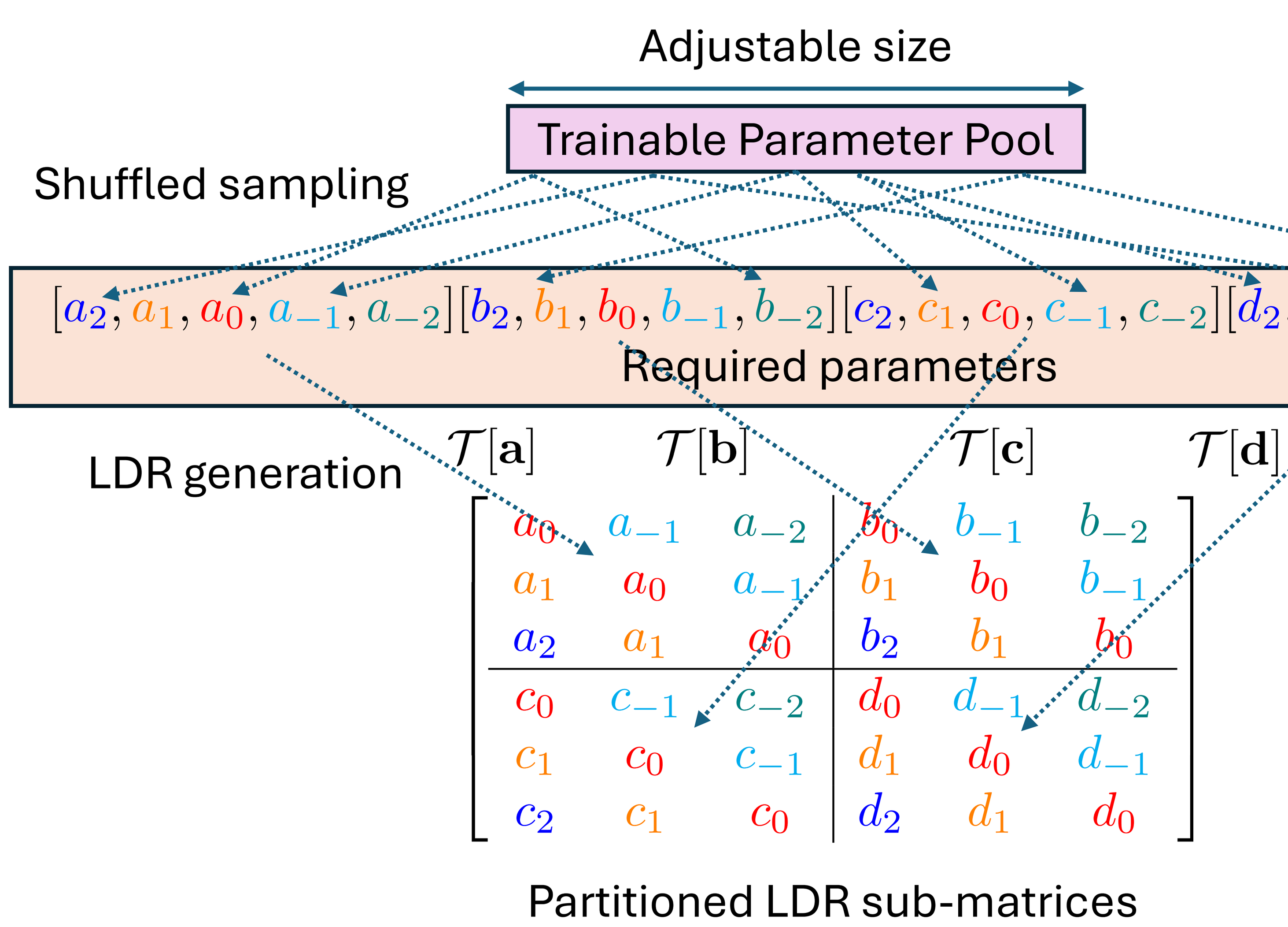
Table: Variants of LDR matrices with displacement operators ($\mathbf{Z}_f = [\mathbf{e}_2, \ldots, \mathbf{e}_n, f\mathbf{e}_1]$)

Structure Matrix $\mathbf{W}$	$\mathbf{A}$	$\mathbf{B}$	Memory	Complexity
Low-Rank	$\mathbf{I}$	$\mathbf{0}$	$2rn$	$\mathcal{O}[rn]$
Circulant	$\mathbf{Z}_1$	$\mathbf{I}$	$2rn$	$\mathcal{O}[rn\log(n)]$
Toeplitz-like	$\mathbf{Z}_1$	$\mathbf{Z}_{-1}$	$2rn$	$\mathcal{O}[rn\log(n)]$
Hankel-like	$\mathbf{Z}_1$	$\mathbf{Z}_0^\intercal$	$2rn$	$\mathcal{O}[rn\log(n)]$
Vandermonde-like	$\mathrm{diag}(\mathbf{v})$	$\mathbf{Z}_0$	$2rn+n$	$\mathcal{O}[rn\log^2(n)]$
Cauchy-like	$\mathrm{diag}(\mathbf{v})$	$\mathrm{diag}(\mathbf{u})$	$2n(r+1)$	$\mathcal{O}[rn\log^2(n)]$

HyDRA

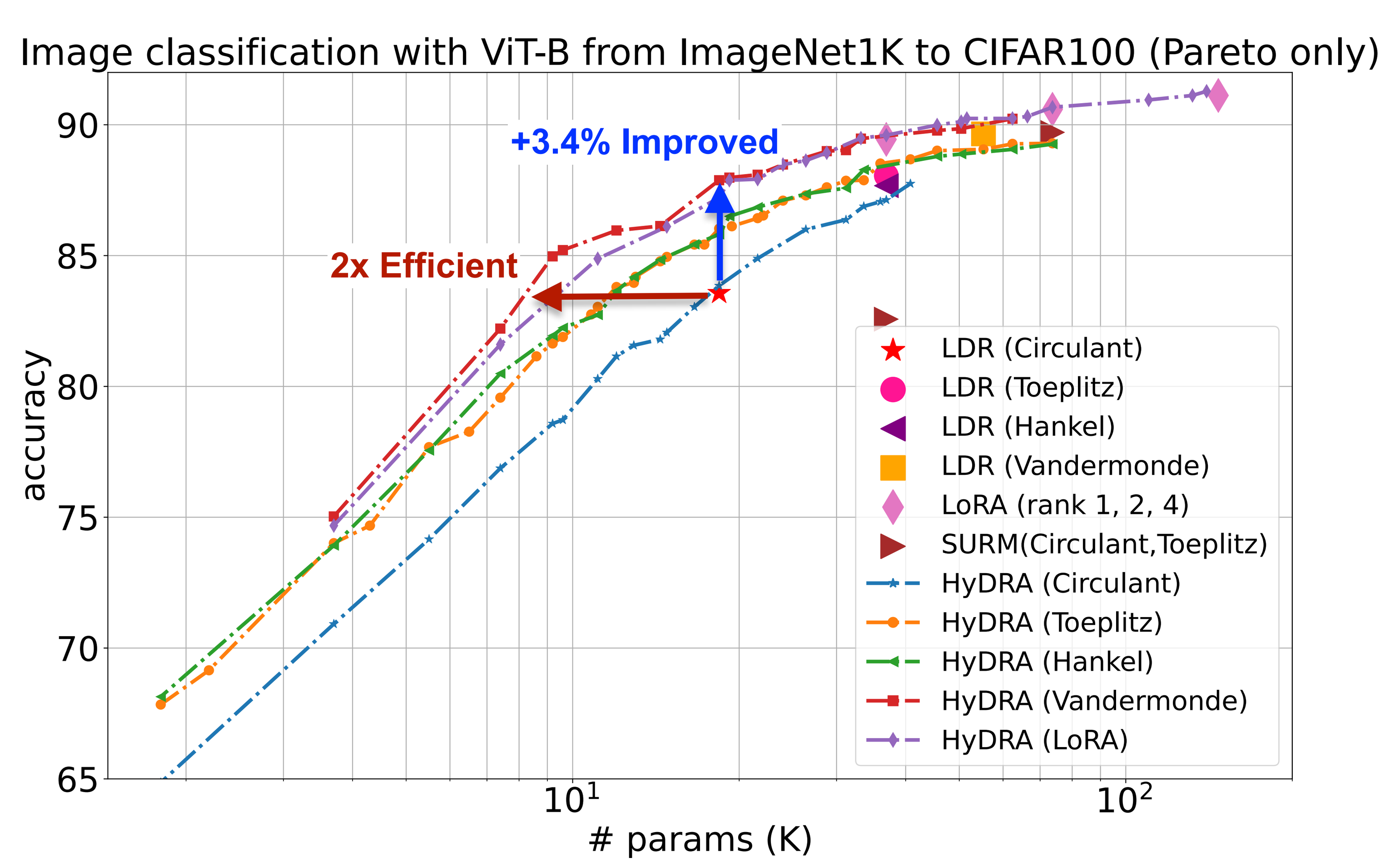
Hypernet low-Displacement Rank Adaptation (HyDRA):

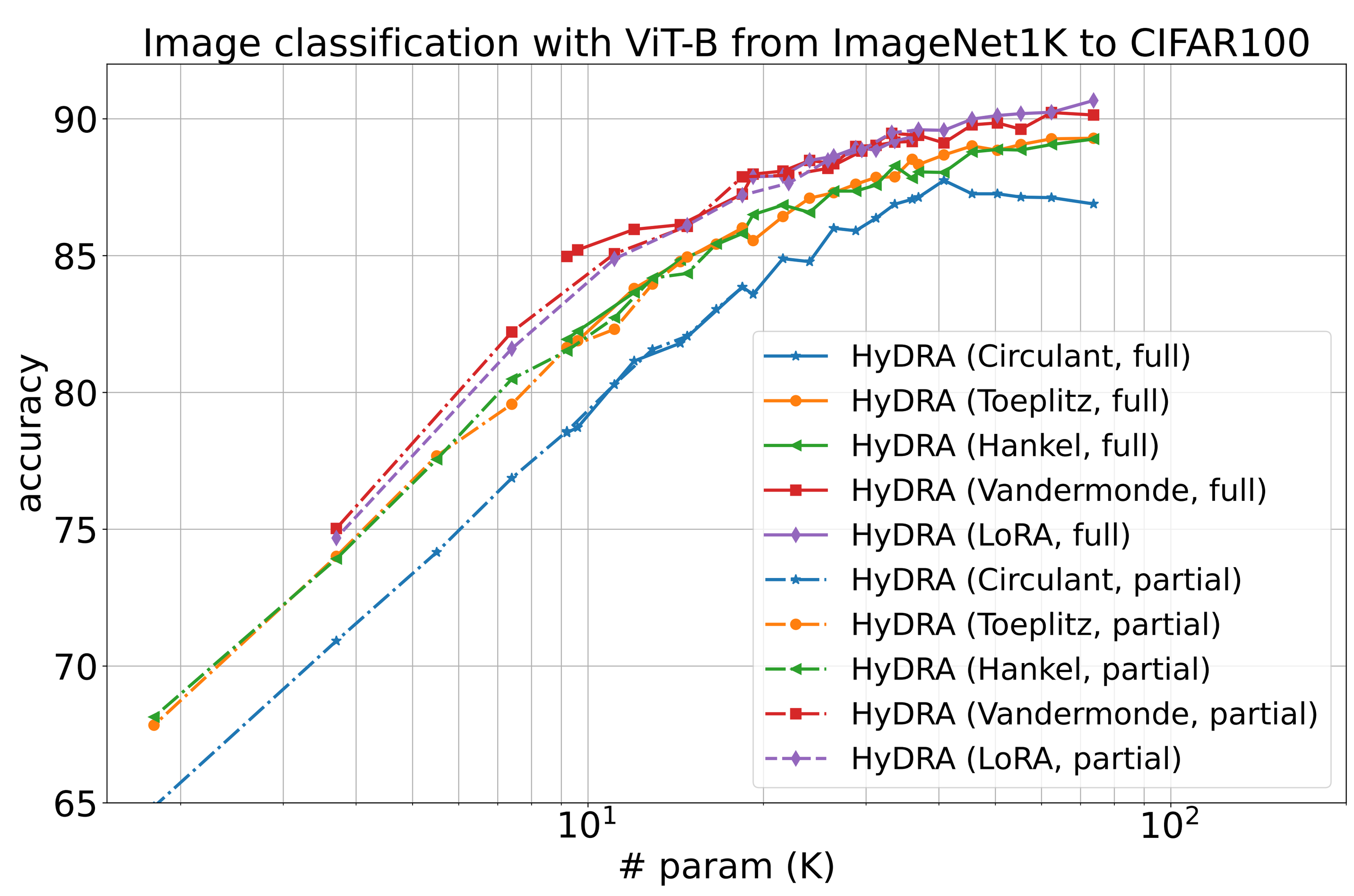
- **Hyper Network**: shuffled sampling from a parameter pool
- Flexible parameter scaling: size-unbounded parameter pool
- **LDR Partition**: Split LDR matrices to increase expressibity
- Partially trainable parameters to reduce memory



Experiments

- Transfer benchmark from ImageNet1K to CIFAR100 for ViT-Base/16 with 87M parameters.





Block Partition with LDR Sub-Matrices for HyDRA

Block partition of LDR sub-matrices still enables superfast operations: e.g., for 2×2 split

$$\Delta\mathbf{W}\mathbf{x} = \begin{bmatrix} \Delta\mathbf{W}_0 & \Delta\mathbf{W}_1 \\ \Delta\mathbf{W}_2 & \Delta\mathbf{W}_3 \end{bmatrix} \begin{bmatrix} \mathbf{x}_0 \\ \mathbf{x}_1 \end{bmatrix} = \begin{bmatrix} \Delta\mathbf{W}_0\mathbf{x}_0 + \Delta\mathbf{W}_1\mathbf{x}_1 \\ \Delta\mathbf{W}_2\mathbf{x}_0 + \Delta\mathbf{W}_3\mathbf{x}_1 \end{bmatrix}, \tag{2}$$

where $\Delta\mathbf{W}_i \in \mathbb{R}^{\frac{m}{k}\times\frac{n}{k}}$ is the ith LDR matrix for $i \in \{0,1,2,3\}$, and $\mathbf{x}_i \in \mathbb{R}^{\frac{n}{k}\times 1}$.
For instance, when $\Delta\mathbf{W}_i$ is a Toeplitz-like matrix having $\mathbf{G}_i = [\mathbf{g}_{i,1}, \ldots, \mathbf{g}_{i,r}]$ and $\mathbf{H}_i^\intercal = [\mathbf{h}_{i,1}, \ldots, \mathbf{h}_{i,r}]$, we can express as follows:

$$\Delta\mathbf{W}_i\mathbf{x}_j = \sum_{l=1}^{r} \mathsf{Krylov}(\mathbf{Z}_1, \mathbf{g}_{i,l}) \cdot \mathsf{Krylov}(\mathbf{Z}_{-1}, \mathbf{h}_{i,l})\mathbf{x}_j \tag{3}$$

where Kryolov operator is computed by **FFT/IFFT** as follows:

$$\mathsf{Krylov}(\mathbf{Z}_1, \mathbf{v})\mathbf{u} = \mathbf{ifft}(\mathbf{fft}(\mathbf{v}) \circ \mathbf{fft}(\mathbf{u})), \qquad \mathsf{Krylov}(\mathbf{Z}_{-1}, \mathbf{v})\mathbf{u} = \bar{\eta} \circ \mathbf{ifft}(\mathbf{fft}(\bar{\eta} \circ \mathbf{v}) \circ \mathbf{fft}(\bar{\eta} \circ \mathbf{u})), \tag{4}$$

where $\bar{\eta} = [1, \eta, \eta^2, \ldots, \eta^{n-1}]^\intercal$, and $\eta = \exp(i\frac{\pi}{n})$ which can be further simplified with diagonalization.

Email: koike@merl.com; Technical Report: `https://www.merl.com/publications/docs/TR2024-157.pdf`