Structured Unrestricted-Rank Matrices for Parameter Efficient Fine-tuning

Arijit Sehanobish*, Avinava Dubey*, Krzysztof Choromanski*, Somnath Basu Roy Chowdhury*,
Deepali Jain, Vikas Sindhwani, Snigdha Chaturvedi

Google DeepMind · Google Research · Columbia University · UNC NLP

Motivation

- Large Transformer models used in various applications in NLP, Vision, Speech etc.
- Fine-tuning these large models for downstream tasks becomes resource-intensive.
- Parameter Efficient Fine-tuning (PEFT) methods have emerged as an attractive method to adapt these models.
- Most PEFT methods leverage low rank matrices.

Low Displacement Rank Matrices

∇ has rank r

- $\boxed{\nabla_{\mathbf{A},\mathbf{B}}(\mathbf{M})} := \mathbf{AM} - \mathbf{MB}$
- Ex : Circulant, Toeplitz, etc.
- *Low rank matrices* are a subset of this framework (by choosing suitable A and B)

Examples of Structured Matrices

$$\begin{bmatrix} c_0 & c_{n-1} & \cdots & c_2 & c_1 \\ c_1 & c_0 & c_{n-1} & \cdots & c_2 \\ \vdots & c_1 & c_0 & \ddots & \vdots \\ c_{m-2} & \vdots & \ddots & \ddots & \vdots \\ c_{m-1} & c_{m-2} & \cdots & \cdots & c_m \end{bmatrix}$$

(a) Circulant

$$\begin{bmatrix} a_0 & a_{-1} & \cdots & \cdots & a_{-(n-1)} \\ a_1 & a_0 & a_{-1} & \cdots & \vdots \\ a_2 & a_1 & a_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & a_{-1} \\ a_{m-1} & \cdots & \cdots & a_1 & a_0 \end{bmatrix}$$

(b) Toeplitz

$$\begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}$$

(c) Kronecker

(d) $$\mathbf{W}(\mathbf{G},\mathbf{H}) = \sum_{i=1}^{r} \mathbf{Z}_1(\mathbf{g}_i)\mathbf{Z}_{-1}(\mathbf{h}_i)$$

where $$\mathbf{Z}_f(\mathbf{v}) = \begin{bmatrix} v_0 & fv_{n-1} & \cdots & fv_1 \\ v_1 & v_0 & \cdots & fv_2 \\ \vdots & \vdots & \vdots & fv_{n-1} \\ v_{n-1} & \cdots & v_1 & v_0 \end{bmatrix}$$

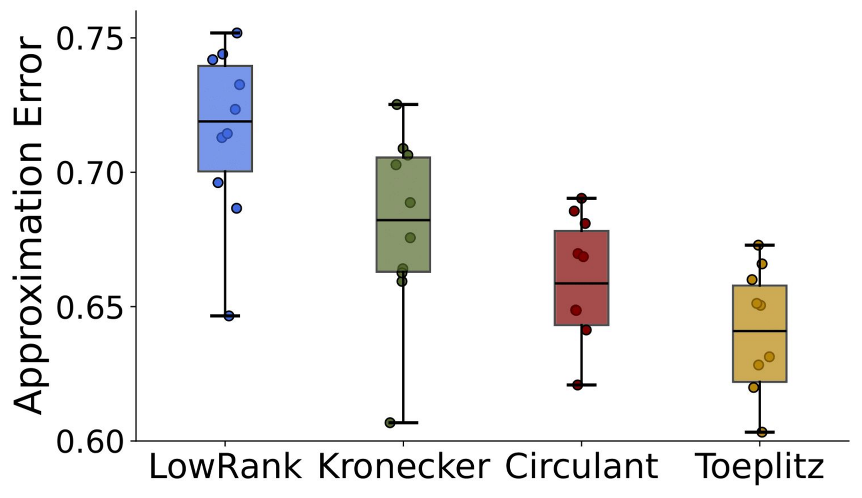
We call these matrices *Structured Unrestricted Rank Matrices* (SURM)

*Equal contribution

Main Research Question

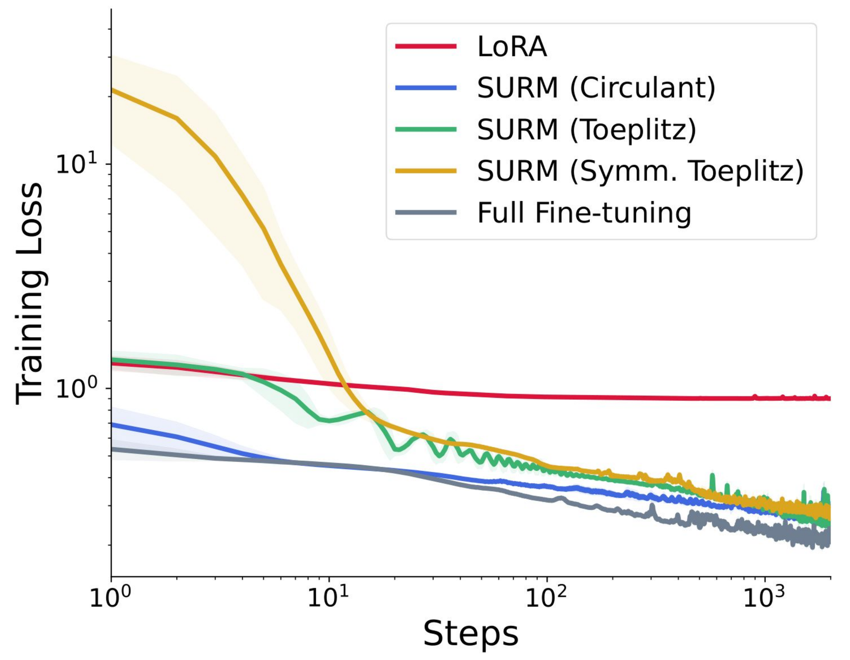
[RQ] Are there other classes of matrices that can be used in lieu of low rank ones, which perform better under the same parameter budget?

Approximation Qualities of SURMs

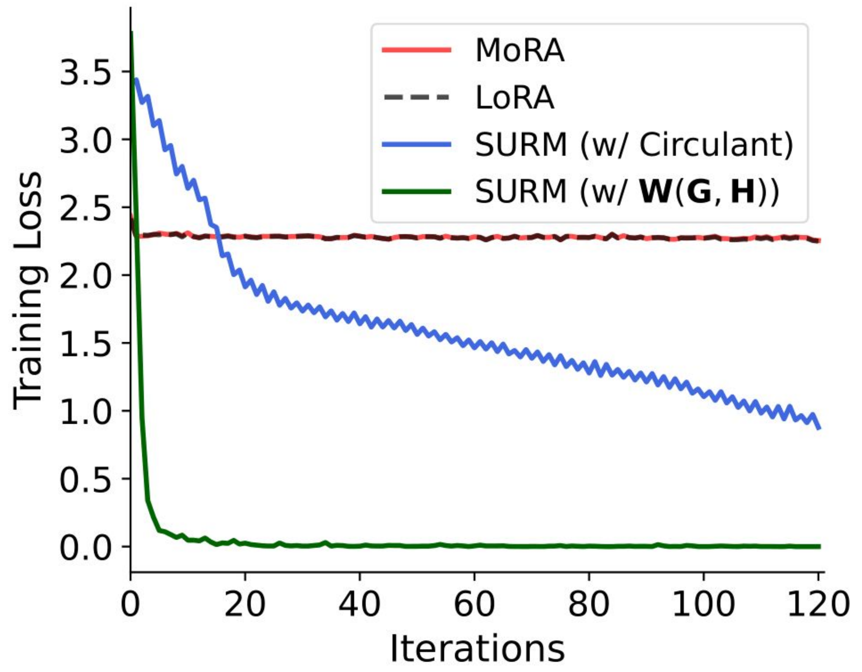


- SURMs show better approximation quality than low-rank matrices.
- Circulant and Toeplitz perform similarly to the more general **W(G,H)**.

Low Rank Matrices struggle to fit the data



Fitting a pinwheel dataset with a simple neural network with one hidden layer and varying the type of the hidden layer.



Fitting a UUID dataset with Llama-2 7B to investigate if high ranks are needed to learn OOD tasks.

Fine-tuning on Vision Datasets

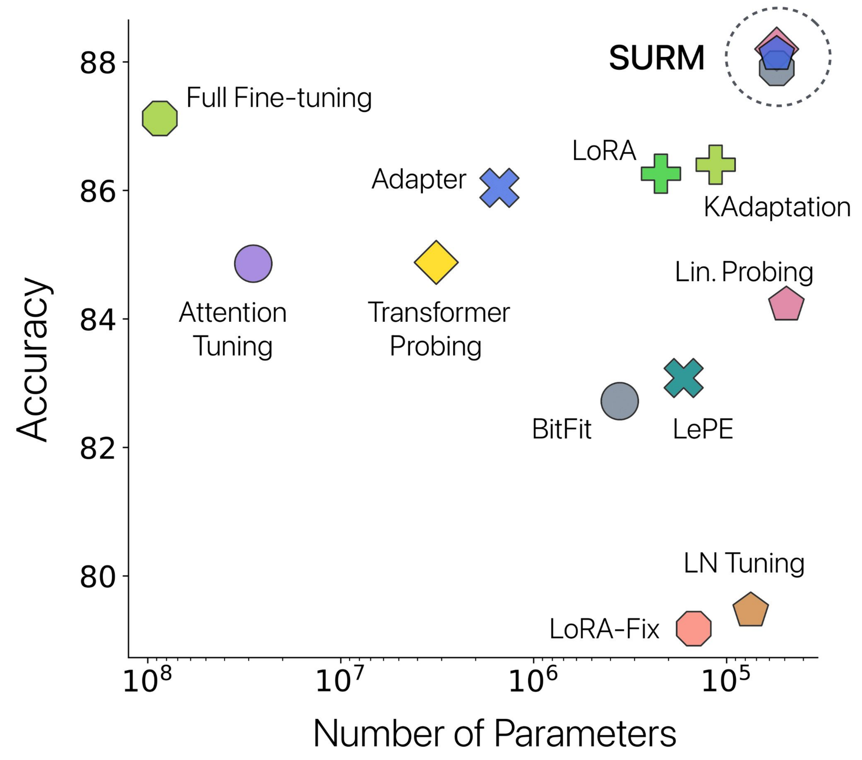


Image Classification

- **Small Data Regimes**. SURMs obtain strong performance on small scale datasets: CIFAR-10, CIFAR-100, DTD, etc.
- **Large Data regimes.** SURM match full fine-tuning performance on ImageNet and INaturalist using only 0.06% parameters.

Low Resource Training. Circulant is the most performant variant and can match the full fine tuning results with only a small fraction of data

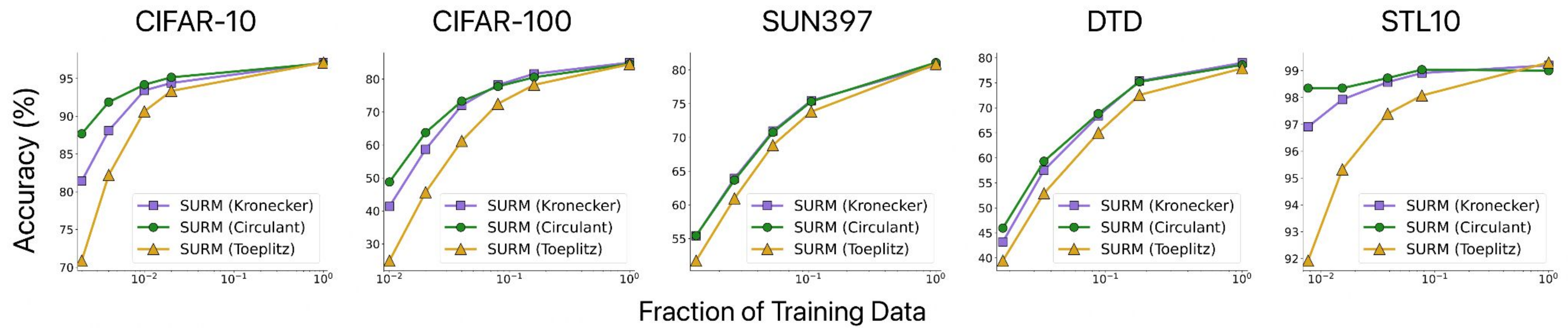
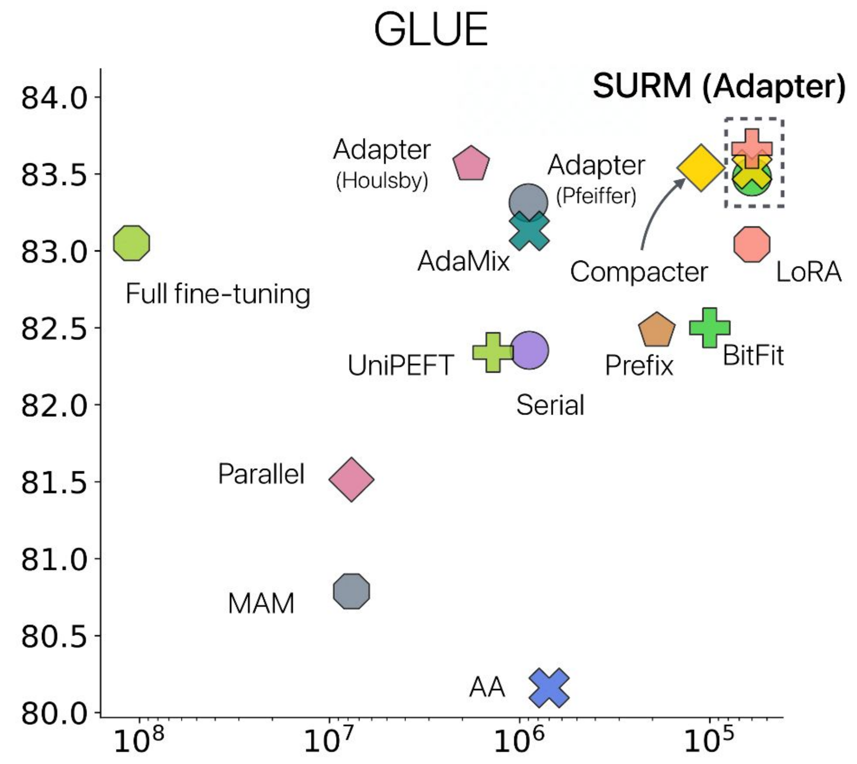


Image Segmentation. SURMs integrated in SURM compare favorably with specialized architectures developed for medical imaging on Synapse multi-organ segmentation dataset.

Fine-tuning on NLP Datasets



Results on GLUE Dataset

- SURM-Adapters outperform many strong baselines while using very few parameters.
- SURM-LoRA outperforms the baseline LoRA with the same parameter budget.

Structured Unrestricted-Rank Matrices for Parameter Efficient Fine-tuning

Arijit Sehanobish*, Avinava Dubey*, Krzysztof Choromanski*, Somnath Basu Roy Chowdhury*, Deepali Jain, Vikas Sindhwani, Snigdha Chaturvedi

Google DeepMind · Google Research · Columbia University · UNC NLP

Motivation

- Large Transformer models used in various applications in NLP, Vision, Speech etc.
- Fine-tuning these large models for downstream tasks becomes resource-intensive.
- Parameter Efficient Fine-tuning (PEFT) methods have emerged as an attractive method to adapt these models.
- Most PEFT methods leverage low rank matrices.

Low Displacement Rank Matrices

- M has displacement rank r if ∇ has rank r.
- $$\nabla_{\mathbf{A},\mathbf{B}}(\mathbf{M}) := \mathbf{AM} - \mathbf{MB}$$
- Ex : Circulant, Toeplitz etc
- *Low rank matrices* are a subset of this framework (by choosing suitable A and B)

Matrices Explored in our Work

$$\begin{bmatrix} c_0 & c_{n-1} & \cdots & c_2 & c_1 \\ c_1 & c_0 & c_{n-1} & \cdots & c_2 \\ \vdots & c_1 & c_0 & \ddots & \vdots \\ c_{m-2} & \vdots & \ddots & \ddots & \vdots \\ c_{m-1} & c_{m-2} & \cdots & \cdots & c_m \end{bmatrix}$$

(a) Circulant

$$\begin{bmatrix} a_0 & a_{-1} & \cdots & \cdots & a_{-(n-1)} \\ a_1 & a_0 & a_{-1} & \cdots & \vdots \\ a_2 & a_1 & a_0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & a_{-1} \\ a_{m-1} & \cdots & \cdots & a_1 & a_0 \end{bmatrix}$$

(b) Toeplitz

$$\begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}$$

(c) Kronecker

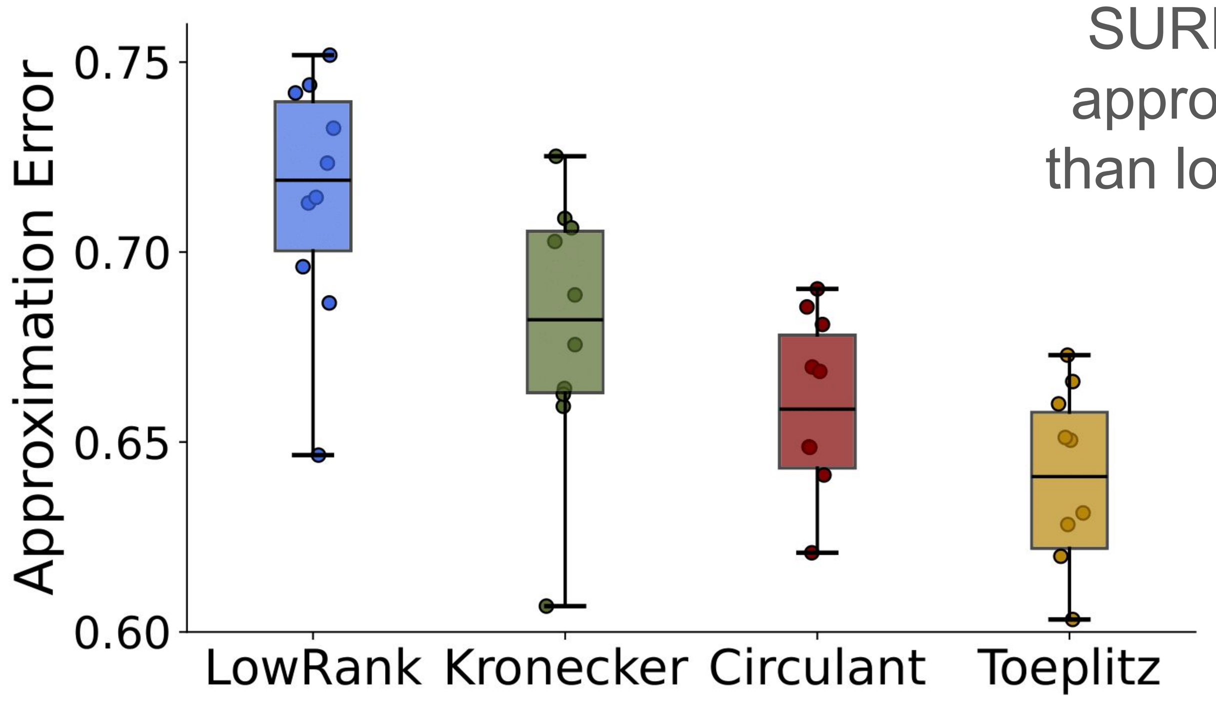
(d) $\mathbf{W}(\mathbf{G},\mathbf{H}) = \sum^{r} \mathbf{Z}_1(\mathbf{g}_i)\mathbf{Z}_{-1}(\mathbf{h}_i)$

where $\mathbf{Z}_f(\mathbf{v}) = \begin{bmatrix} v_0 & fv_{n-1} & \cdots & fv_1 \\ v_1 & v_0 & \cdots & fv_2 \\ \vdots & \vdots & \vdots & fv_{n-1} \\ v_{n-1} & \cdots & v_1 & v_0 \end{bmatrix}$

We call these matrices *Structured Unrestricted Rank Matrices* (**SURM**)

* equal contribution

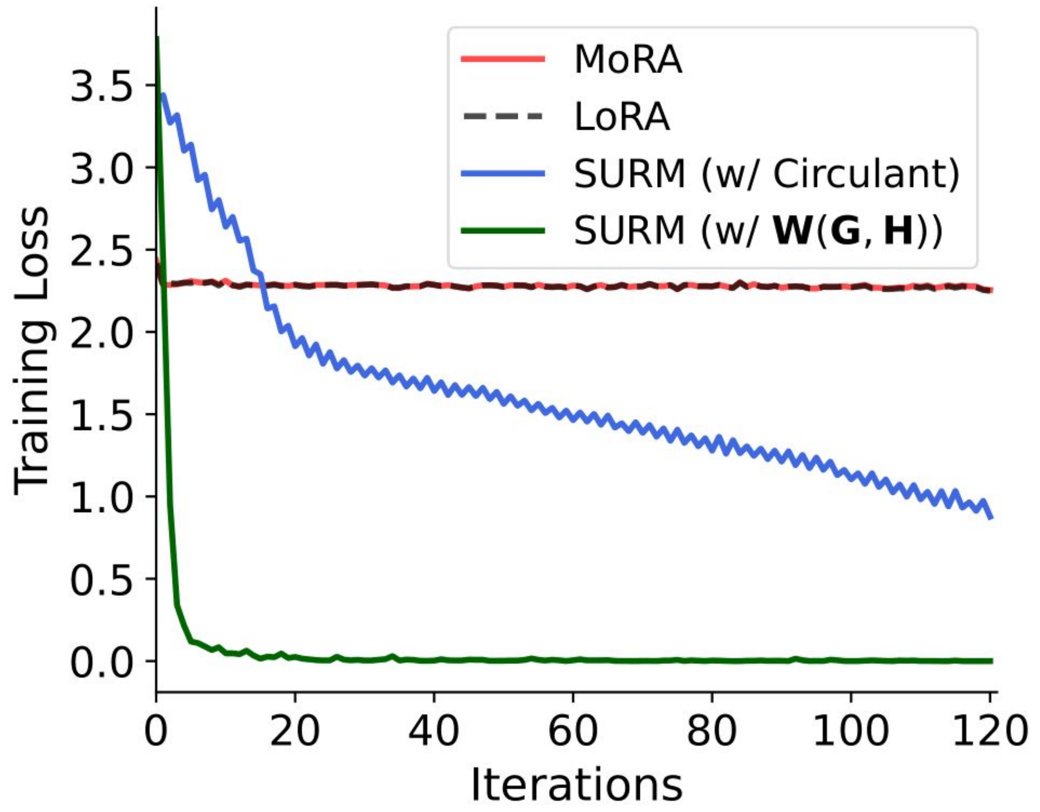
Approximation Qualities of SURMs



SURMs show better approximation quality than low-rank matrices.

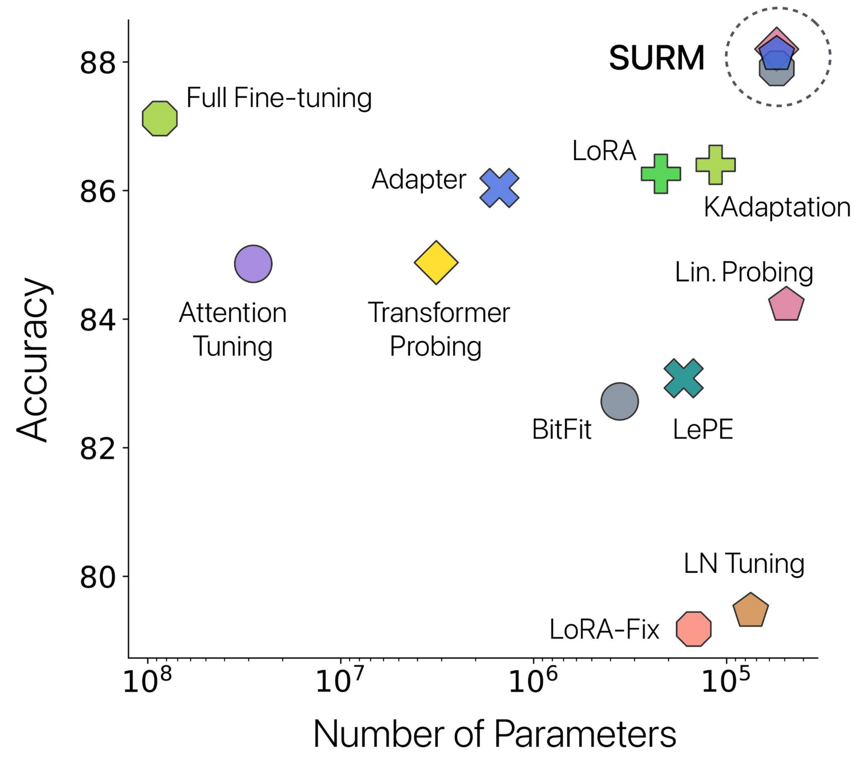
- SURMs show better approximation quality than low-rank matrices.
- Circulant and Toeplitz perform similarly to the more general **W(G,H)**.

Low Rank Matrices struggle to fit the data



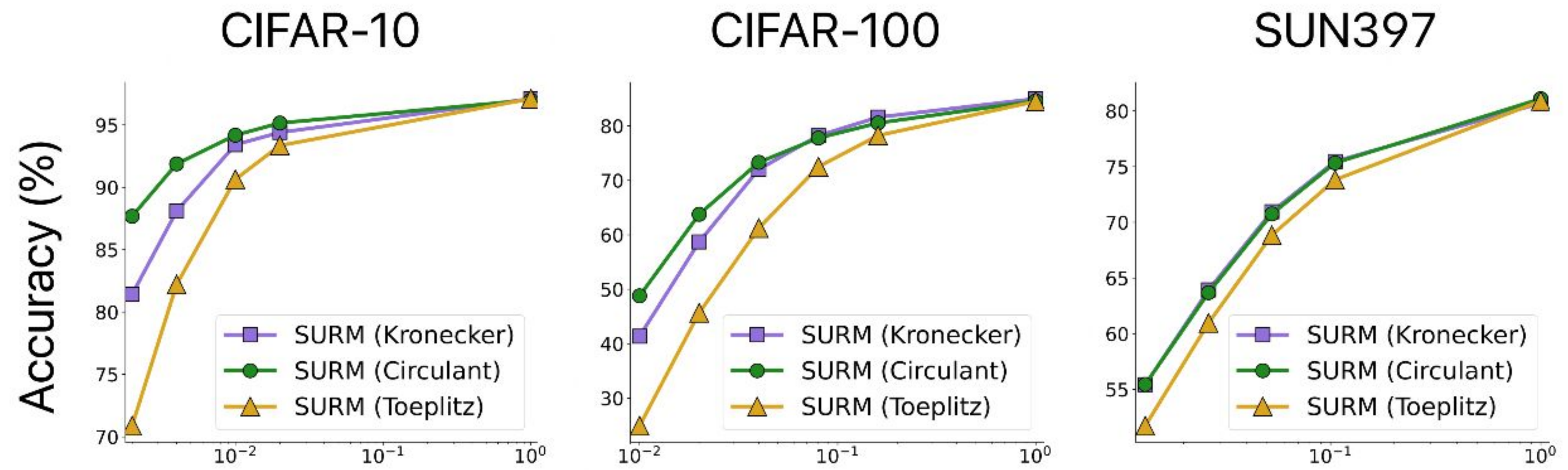
- Fitting a UUID dataset with Llama-2-7b to investigate if high ranks are needed to learn OOD tasks.

Vision Results



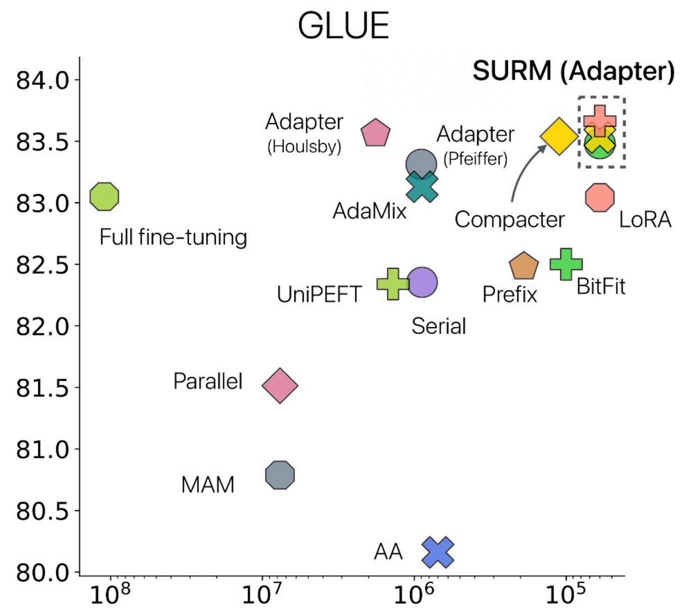
- **Image Classification**
  - **Small Data Regimes**. SURMs obtain strong performance on small scale datasets: CiFAR-10, CiFAR-100, DTD, etc.
  - **Large Data regimes.** SURM match performance to full fine tuning on ImageNet and INaturalist using only 0.06% parameters.

Low Resource Training. Circulant is the most performant variant and can match the full fine tuning results with only a small fraction of data



➢ **Image Segmentation** : SURMs integrated in SURM compare favorably with specialized architectures developed for medical imaging on Synapse multi-organ segmentation dataset.

NLP Results



Results on GLUE Dataset :

➢ SURM-Adapters outperform many strong baselines while using very few parameters.

➢ SURM (integrated into LoRA) outperforms the baseline LoRA, under the same parameter budget.