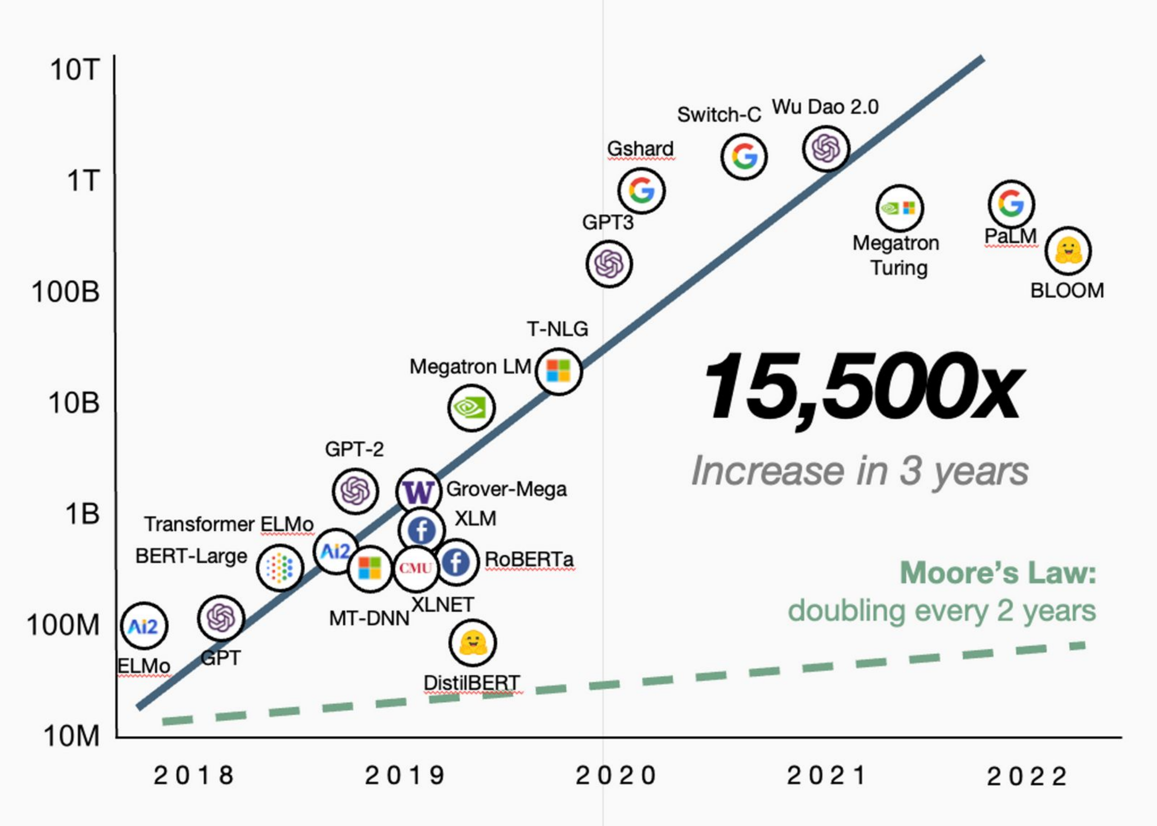
Structured Unrestricted-Rank Matrices for Parameter Efficient Fine-tuning

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

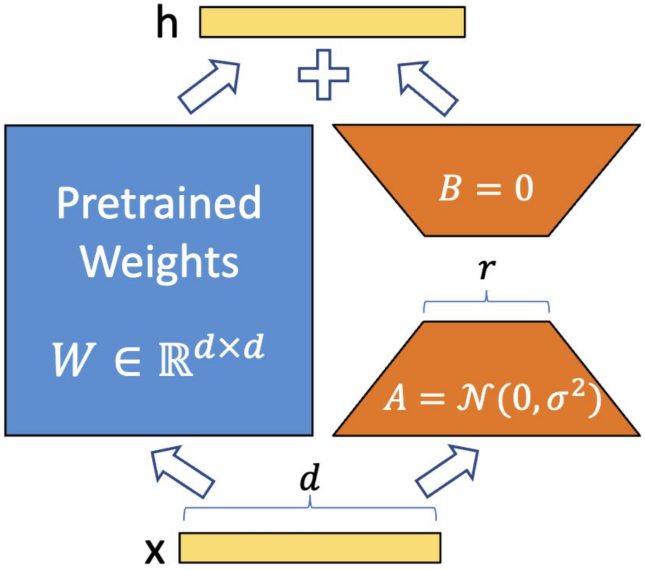
** Equal Contribution*

Correspondence to arijit.sehanobish1@gmail.com

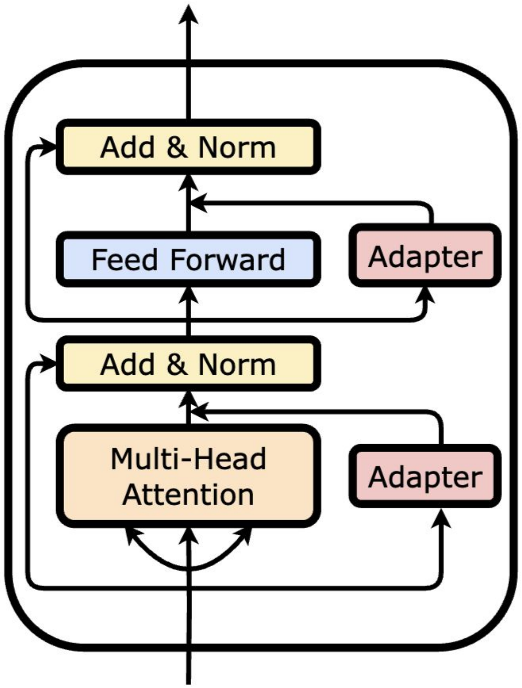
Motivation


10T
1T
100B
10B
1B
100M
10M
2018
2019
2020
2021
2022
ELMo
GPT
BERT-Large
Transformer ELMo
GPT-2
MT-DNN
XLNET
Grover-Mega
XLM
RoBERTa
DistilBERT
Megatron LM
T-NLG
GPT3
Gshard
Switch-C
Wu Dao 2.0
Megatron Turing
PaLM
BLOOM
15,500x
Increase in 3 years
Moore's Law:
doubling every 2 years

Popular PEFT Methods



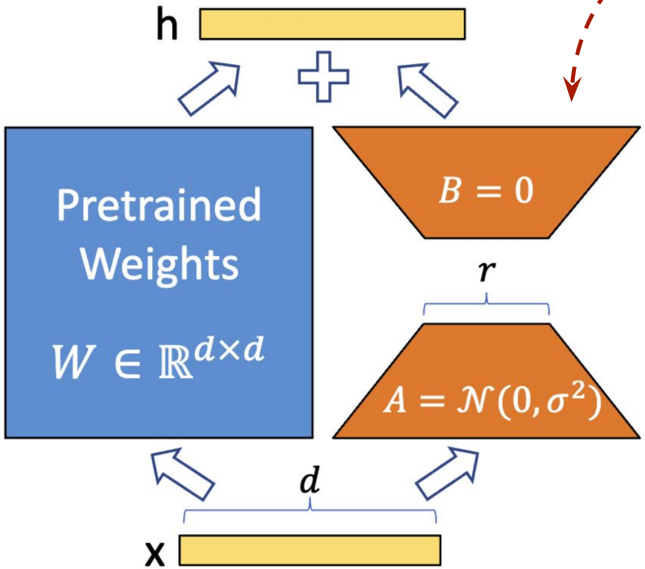
LoRA



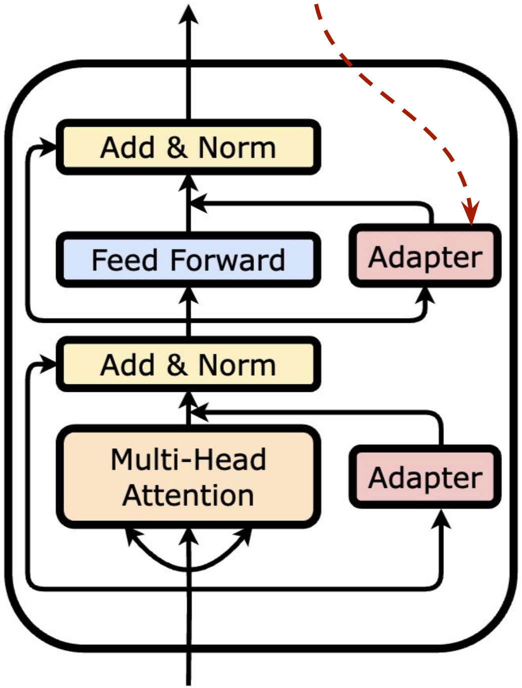
Adapters

Popular PEFT Methods

Low-Rank parameterization



LoRA



Adapters

Can we use other matrices which are not necessarily low rank in PEFT?

- *Structured Matrix* : a generic term for a m×n matrix **A** that can be represented by fewer than mn parameters. They reduce both space and time complexity when performing matrix multiplications.
- Simple example of such a matrix is a *low-rank matrix*.
- In this work, we focus on *structured matrices* that are not necessarily low rank, which we refer to as *Structured Unrestricted-Rank Matrices* (SURM). They support sub-quadratic matrix-vector multiplications.
    - **Low Displacement Rank Matrices**
    - **Kronecker Product of matrices**
- These matrices can be easily plugged in LoRA and Adapters in lieu of low-rank matrices.

Low Displacement Rank Matrices

A matrix **M** is said to have low displacement rank *r* if the displacement operator ∇ has rank *r*.

$$\nabla_{\mathbf{A},\mathbf{B}}(\mathbf{M}) := \mathbf{AM} - \mathbf{MB}$$

Low Displacement Rank Matrices

A matrix **M** is said to have low displacement rank *r* if the displacement operator ∇ has rank *r*.

$$\nabla_{\mathbf{A},\mathbf{B}}(\mathbf{M}) := \mathbf{AM} - \mathbf{MB}$$

- Choosing suitable **A**, **B** one can get a rich class of matrices including circulant, Toeplitz matrices, products and inverses of Toeplitz matrices and *low rank matrices*.

Low Displacement Rank Matrices

A matrix **M** is said to have low displacement rank *r* if the displacement operator ∇ has rank *r*.

$$\nabla_{\mathbf{A},\mathbf{B}}(\mathbf{M}) := \mathbf{AM} - \mathbf{MB}$$

- Choosing suitable **A**, **B** one can get a rich class of matrices including circulant, Toeplitz matrices, products and inverses of Toeplitz matrices and *low rank matrices*.
- Thus this framework is *strictly more general* than what is considered in the literature.

Types of Matrices studied in this 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

Types of Matrices studied in this 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

Types of Matrices studied in this 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

Types of Matrices studied in this 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

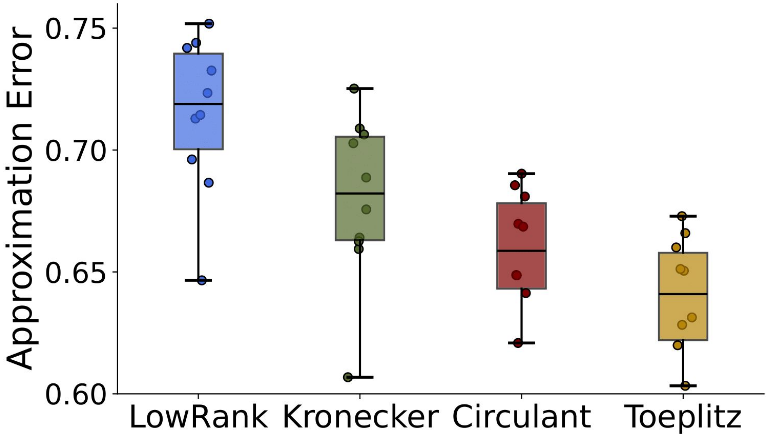
$$\begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}$$

(c) Kronecker

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

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

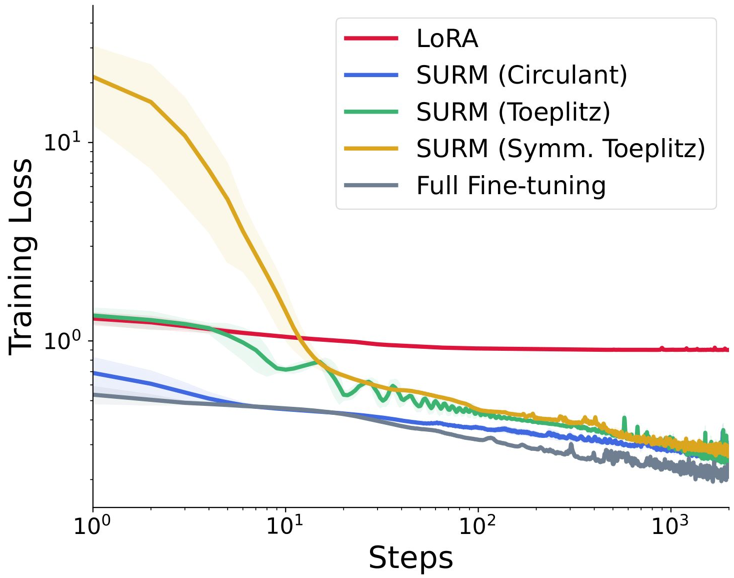
SURMs are good approximators



- SURMs show better approximation quality than low-rank matrices.
- Circulant and Toeplitz matrices show comparable approximation qualities to that of more general **W**(**G**,**H**) which is why they are the major focus of our work.

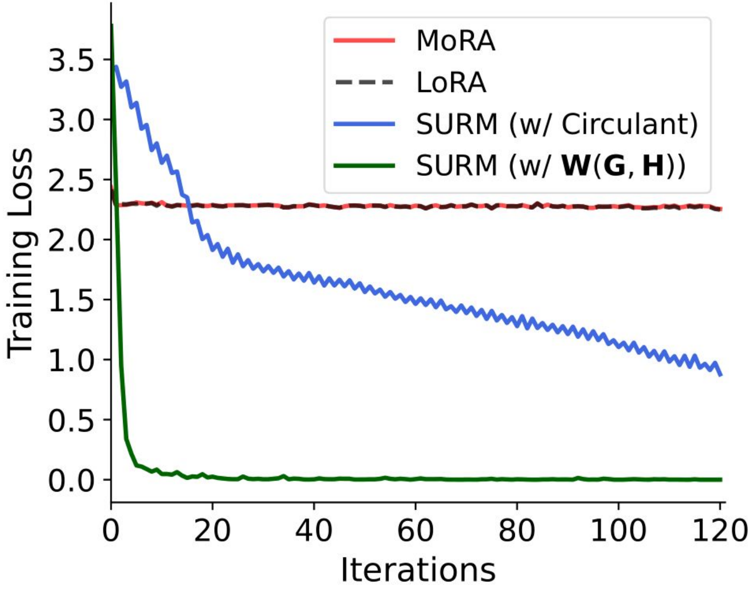
Low-Rank Matrices can struggle to fit the data

- Create a pinwheel dataset and we fit the data with a simple neural network with one hidden layer.
- Vary the type of the hidden layer with low rank matrices and SURMs.



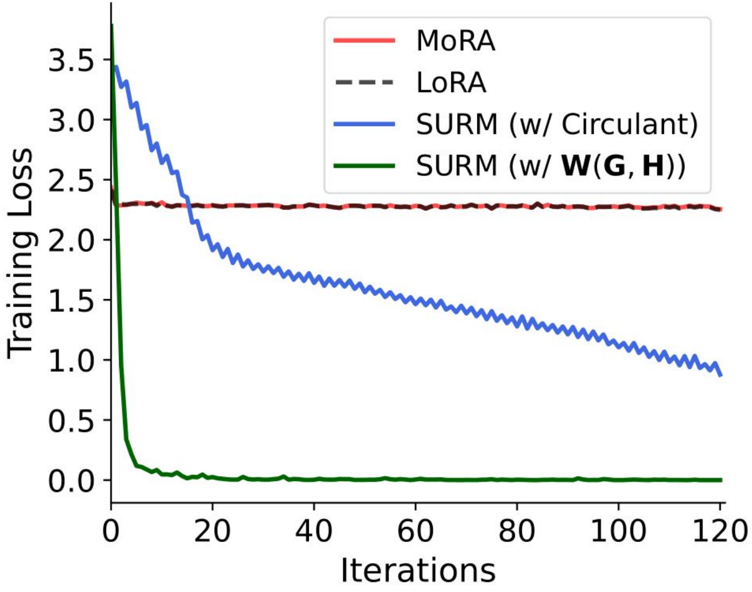
Fitting a synthetic dataset with Llama-2-7b

- Investigate if large ranks are needed for learning new tasks.



Fitting a synthetic dataset with Llama-2-7b

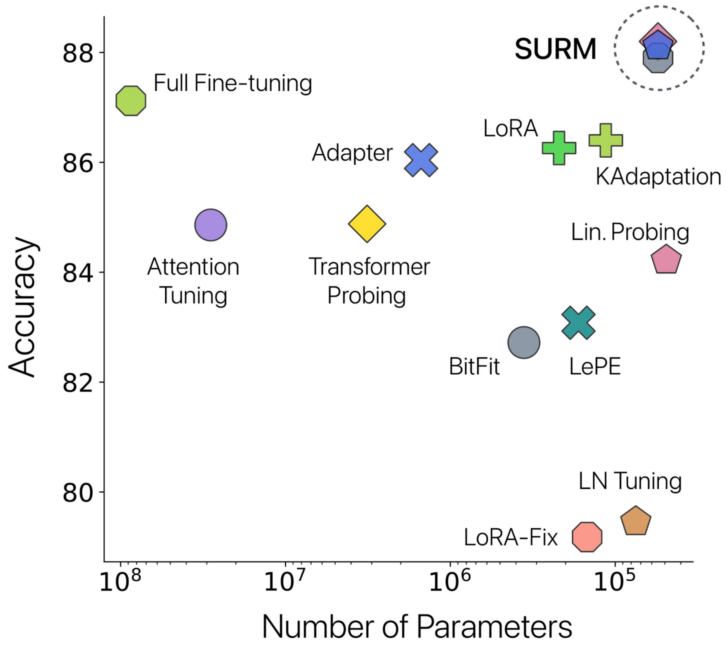
- Investigate if large ranks are needed for learning new tasks.



TLDR : Higher ranks and more expressive matrices may be required to learn tasks outside of its training distribution.

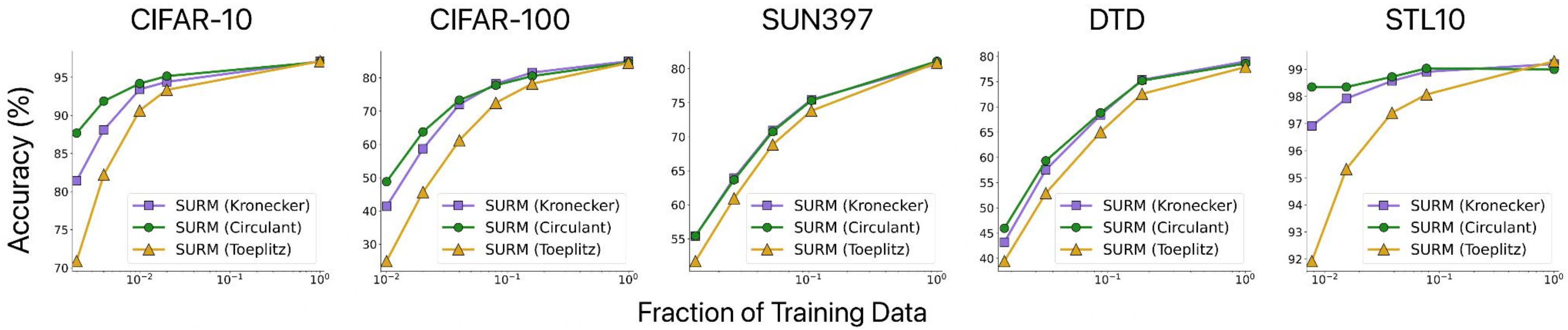
Vision Experiments

- Image Classification on CiFAR-10, CiFAR-100, DTD, SUN397 and STL10.



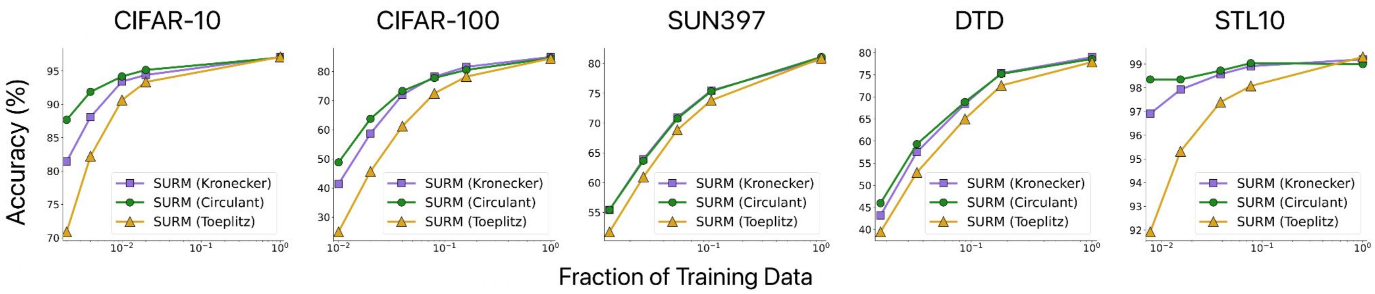
Vision Experiments

- **Low Resource setting :**



Vision Experiments

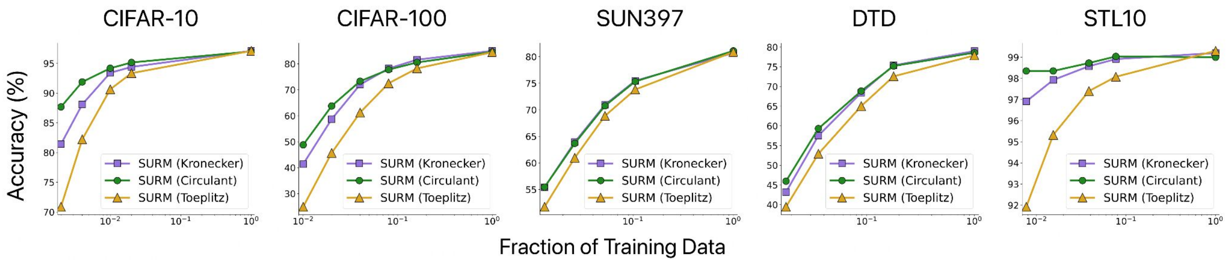
- **Low Resource setting :**



- **Large Data regimes :** SURM match performance to full fine tuning on ImageNet and INaturalist 2021 datasets while using only **.06%** of training parameters.

Vision Experiments

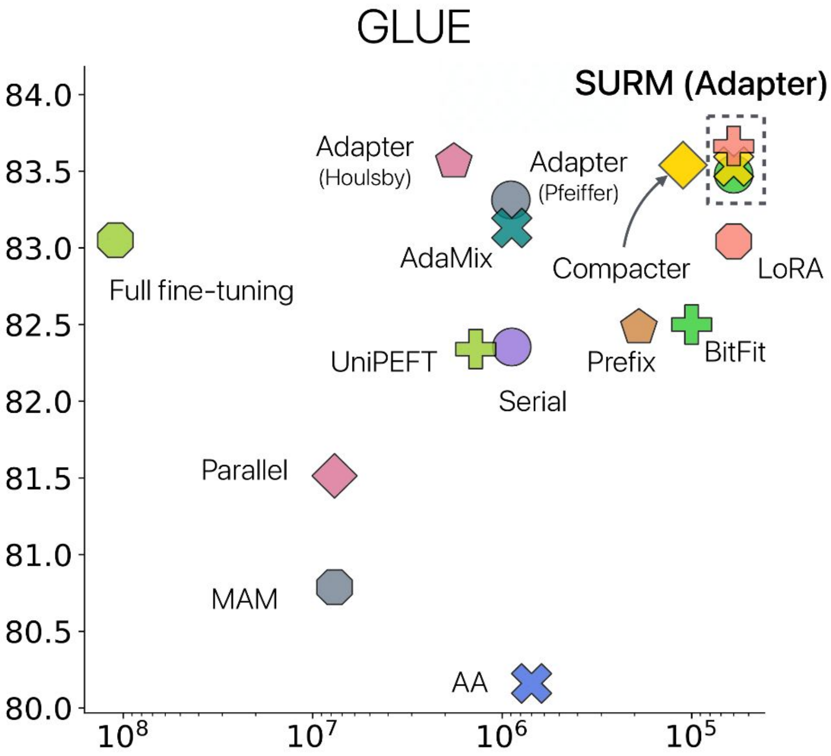
- **Low Resource setting :**



- **Large Data regimes :** SURM match performance to full fine tuning on ImageNet and INaturalist 2021 datasets while using only **.06%** of training parameters.

- **Medical Image Segmentation :** SURMs match performance of specialized architectures like U-Nets and outperforms LoRA on the Synapse multi-organ segmentation dataset.

NLP Experiments



- Results on the GLUE benchmark with SURM-Adapters.
- SURM (integrated into LoRA) outperforms the baseline LoRA, under the same parameter budget.

For more results see our paper and come to our poster session

Thank you!