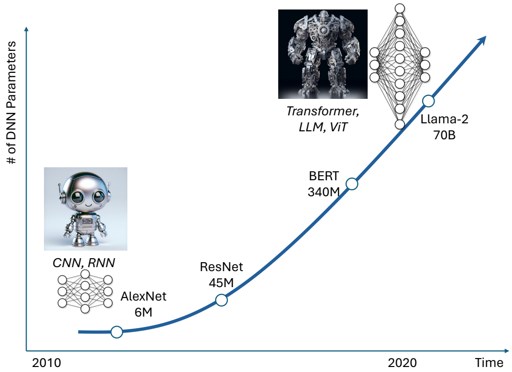
Differentiable Learning of Generalized Structured Matrices for Efficient Deep Neural Networks

Changwoo Lee, Hun-Seok Kim

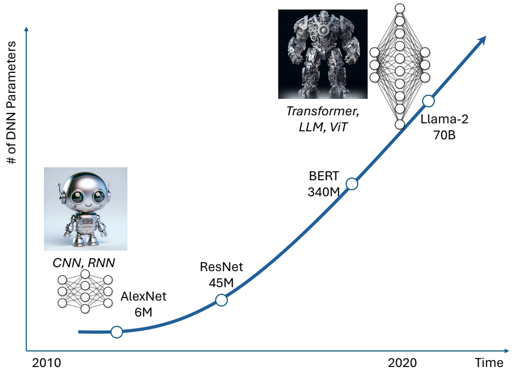
University of Michigan

ICLR 2024

UNIVERSITY OF MICHIGAN



Larger Deep Neural Networks (DNNs) { Better Accuracy 💪



Larger Deep Neural Networks (DNNs) { Better Accuracy 💪 / Slower Inference, More Resources 💸

Problem

Goal: find a Deep Neural Network (DNN) with good performance and low cost of inference in floating-point operations (FLOPs)

$$\min_{f} \sum_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{D}} \mathrm{error}(f(\boldsymbol{x}), \boldsymbol{y}) \quad \text{s.t.} \quad \mathrm{cost}(f) \leq B,$$

where f is a Deep Neural Network with L layers.

Problem

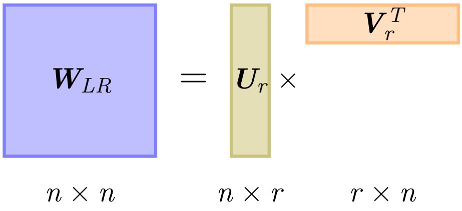
Goal: find a Deep Neural Network (DNN) with good performance and low cost of inference in floating-point operations (FLOPs)

$$\min_{f} \sum_{\boldsymbol{x},\boldsymbol{y}\sim\mathcal{D}} \operatorname{error}(f(\boldsymbol{x}), \boldsymbol{y}) \quad \text{s.t.} \quad \operatorname{cost}(f) \leq B,$$

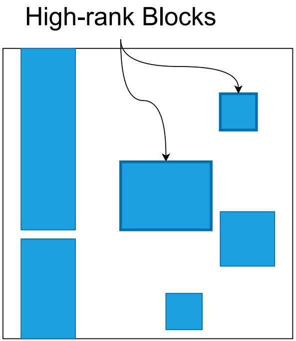
where f is a Deep Neural Network with L layers.

Use *structured matrices* for linear layers $\boldsymbol{W}\boldsymbol{x}$ to satisfy $\operatorname{cost}(f) \leq B$.

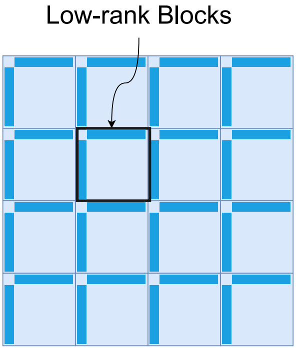
- $\operatorname{cost}(\boldsymbol{W}_{\text{structured}}\boldsymbol{x}) \ll \operatorname{cost}(\boldsymbol{W}_{\text{dense}}\boldsymbol{x})$



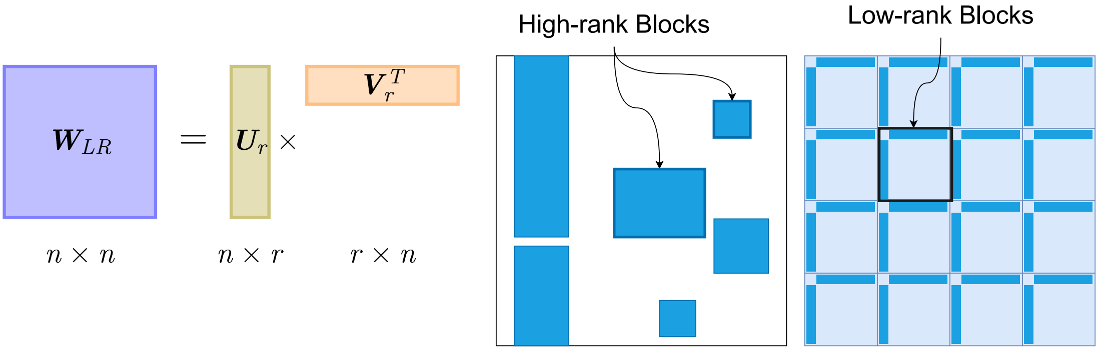
Low-Rank (LR)



Block Sparse (BSP)



Block Low-Rank (BLR)



Low-Rank (LR)

Block Sparse (BSP)

Block Low-Rank (BLR)

Optimal *Layer-wise* Structure Format for Better Accuracy-Efficiency Trade-off?

Challenges

Challenge 1. Discrete Exponential Search Space

- Search space size exponential to the number of weights.

Challenges

Challenge 1. Discrete Exponential Search Space

- Search space size exponential to the number of weights.

Challenge 2. Lack of Structured Matrix Format

- Structured matrices are hand-designed from human insight.

Challenges

Challenge 1. Discrete Exponential Search Space

- Search space size exponential to the number of weights.

Challenge 2. Lack of Structured Matrix Format

- Structured matrices are hand-designed from human insight.

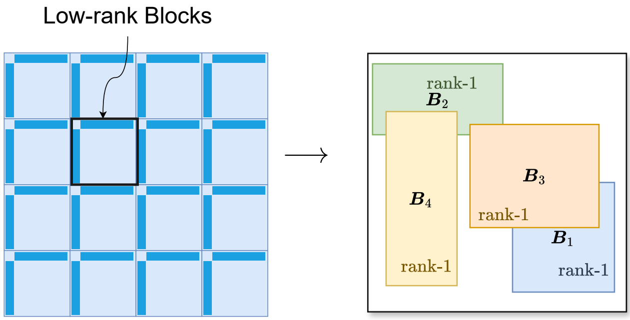
Solution: Differentiable Learning Approach

Generalize, **Parameterize**, then **Optimize** Structure by Gradient Descent!

$\downarrow$

Gaudi-GBLR Matrix

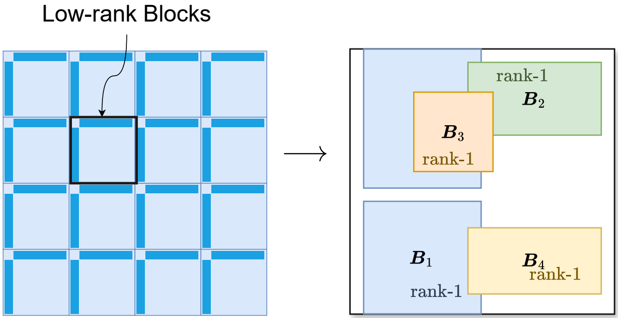
Generalized Block Low-Rank (GBLR) Matrix



Block Low-Rank

Generalized Block Low-Rank

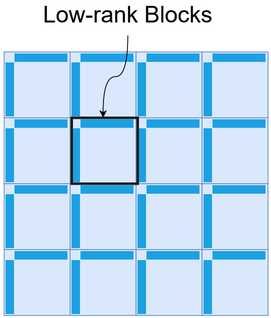
Generalized Block Low-Rank (GBLR) Matrix

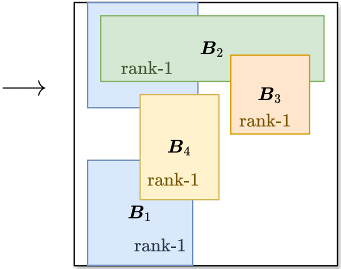


Block Low-Rank

Generalized Block Low-Rank

Generalized Block Low-Rank (GBLR) Matrix

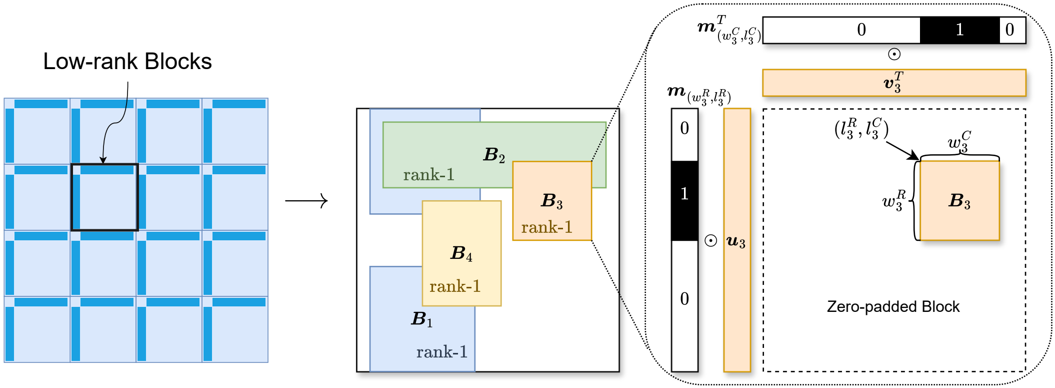




Block Low-Rank

Generalized Block Low-Rank

Block by Mask



Block Low-Rank

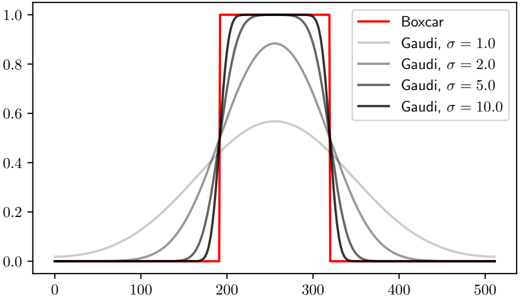
Generalized Block Low-Rank

Faster Mult when
Low Rank Blocks

Faster Mult when
Small Mask Widths

Differentiable Mask

Position of a Block = Two Gaussian-Dirichlet (Gaudi) Masks



- A Gaudi mask is fully differentiable with respect to the width and the starting point of the non-zero elements.

Gaudi-GBLR Matrix

A GBLR matrix with Gaudi masks $\tilde{\boldsymbol{m}}^{\sigma}_{(w_k^R, l_k^R)}, \tilde{\boldsymbol{m}}^{\sigma}_{(w_k^C, l_k^C)}$:

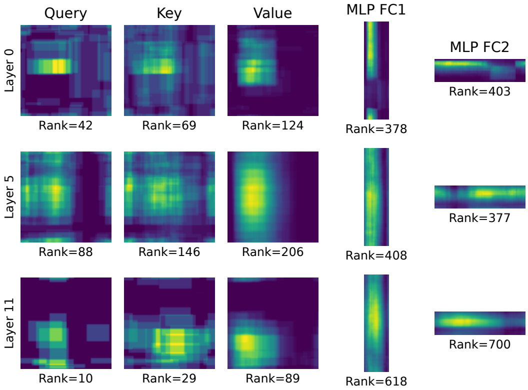
$$\boldsymbol{W}^{\boldsymbol{\theta}} = \sum_{k=1}^{K} \left(\tilde{\boldsymbol{m}}^{\sigma}_{(w_k^R, l_k^R)} \odot \boldsymbol{u}_k \right) \left(\tilde{\boldsymbol{m}}^{\sigma}_{(w_k^C, l_k^C)} \odot \boldsymbol{v}_k \right)^T .$$

$\mathrm{cost}(\boldsymbol{W}^{\boldsymbol{\theta}} \boldsymbol{x}) \propto$ sum of the widths of the masks.

Training Efficient DNN by Gaudi-GBLR Weights

1. Replace weights of a DNN f to Gaudi-GBLR matrices.
2. Penalize the sum of the width of the Gaudi masks during the training at each gradient descent step.

Learned Block Layout

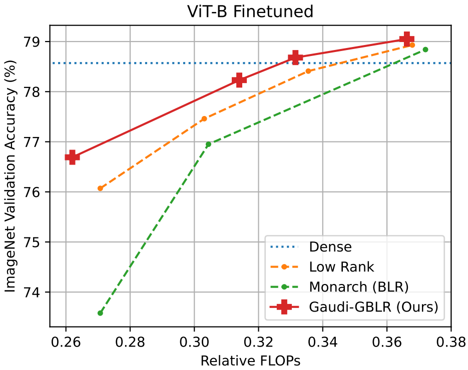


Learned Block Layout of ViT-Base on ImageNet

The brighter, the more overlaps among blocks.

Vision Tasks

- ImageNet Classification (1,000 classes)
- Replace weights of ViT-Base to Gaudi-GBLR matrices and fine-tune.
- Better accuracy-efficiency trade-off than fixed structured matrices.

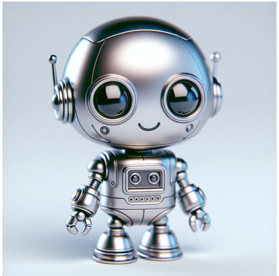


Language Task

- Language Generation-WikiText103
- Better performance while using fewer computations.

Perplexity by weight type of GPT-2 after fine-tuning on WikiText103.

Weight Type	Perplexity (↓)	Relative FLOPs
Dense	19.36	100%
Low Rank	19.48	43.75%
Monarch	20.56	43.75%
Gaudi-GBLR	**19.24**	**43.7**%



Thank you!

Visit us at our poster on
Wed 8 May 10:45 a.m.-12:45 p.m. CEST

Paper

Code