Forget the Data and Fine-Tuning! Just Fold the Network to Compress


Dong Wang
TU Graz


Haris Šikić
TU Graz

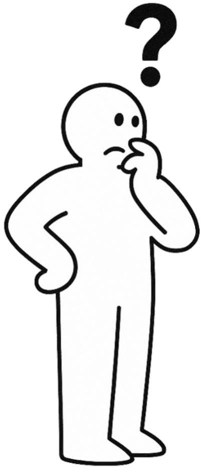

Lothar Thiele
ETH Zürich

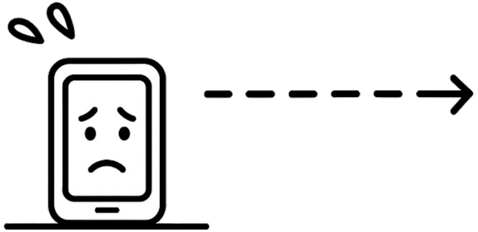

Olga Saukh
TU Graz
Complexity Science Hub Vienna

Paper website

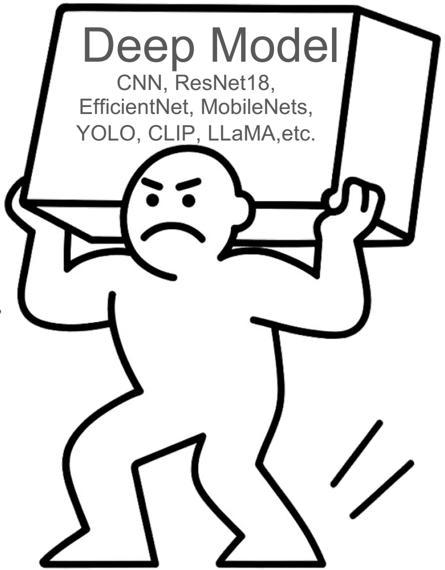
Motivation: model compression

So powerful…
But so over-sized.
How to compress a
deep learning model?





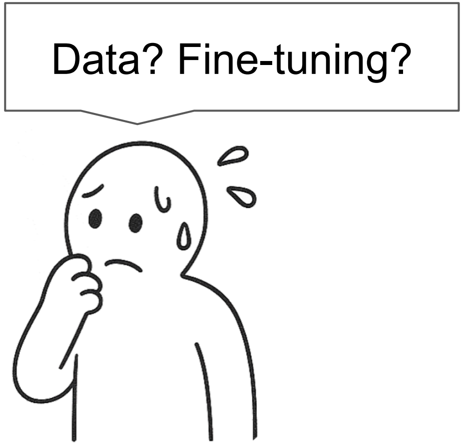
Edge Devices



Classical solutions: need data or fine-tuning

Full model

-0.6	3.1	7.2
-5.2	4.3	2.8
3.3	9.1	-0.9



0.0	3.1	7.2
-5.2	4.3	0.0
3.3	9.1	0.0

Model Pruning

➡️ Zero out less-important neurons
❌ Needs data and fine-tuning

117	183	255
35	205	178
187	255	112

Model Quantization

➡️ Use low precision
❌ Needs data or quantization-aware training

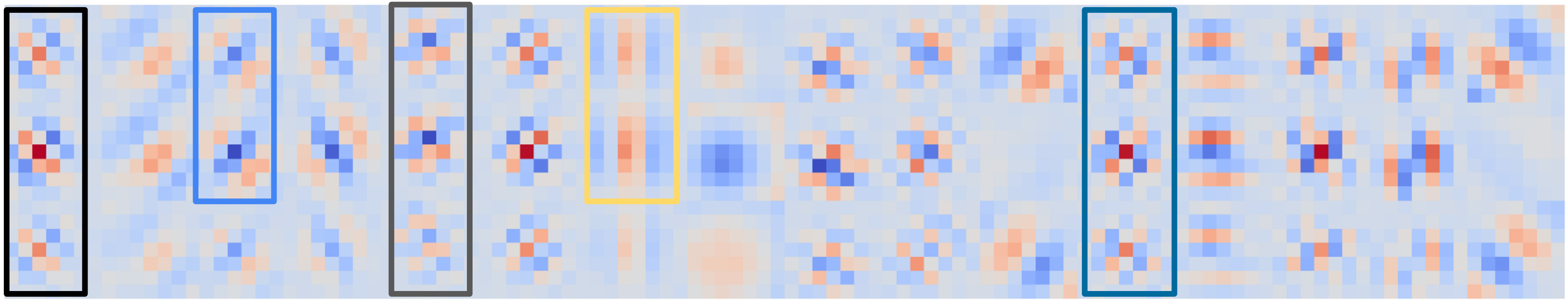
-0.5	4.1
-6.1	2.9
2.3	8.8

Knowledge Distillation

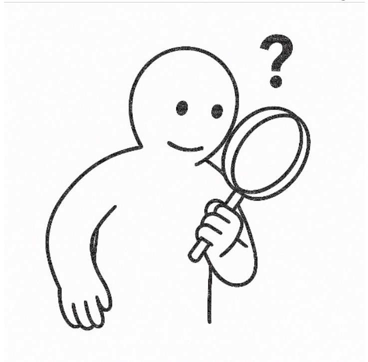
➡️ Train a small model imitating the big model
❌ Needs data and training

Similar patterns in well-trained weight

The 1st Conv layer in a ResNet18 trained on ImageNet

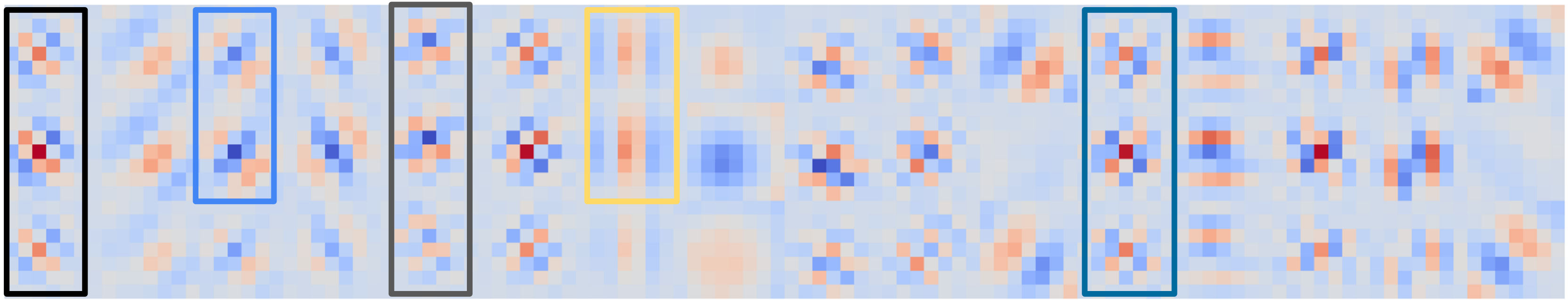


Question: “Why not fold these similar structures instead of zeroing them out?”



Similar patterns in well-trained weight

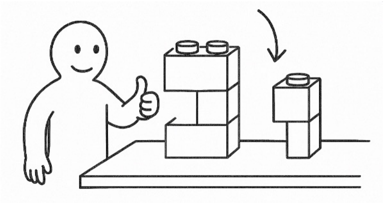
The 1st Conv layer in a ResNet18 trained on ImageNet



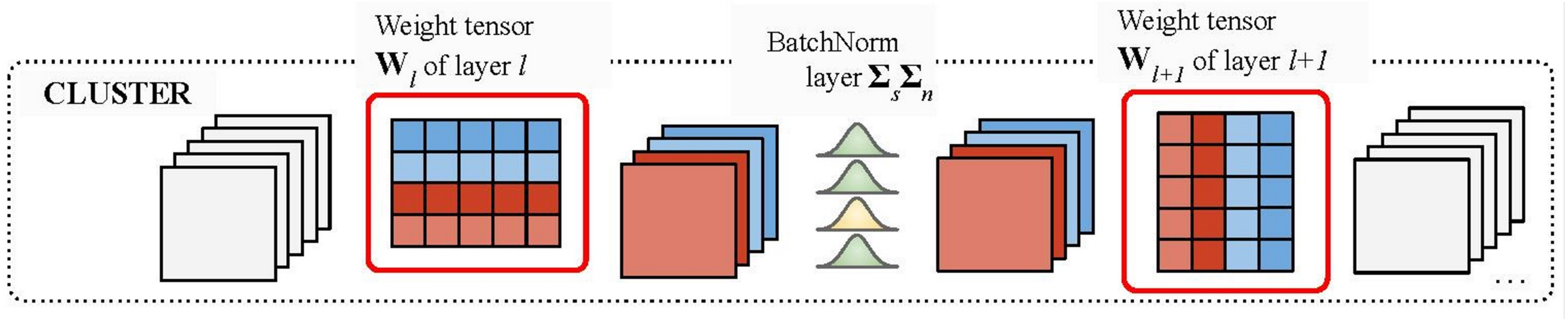
Question: “Why not fold these similar structures instead of zeroing them out?”

Our contribution: “A data-free and fine-tuning-free model compression method.”

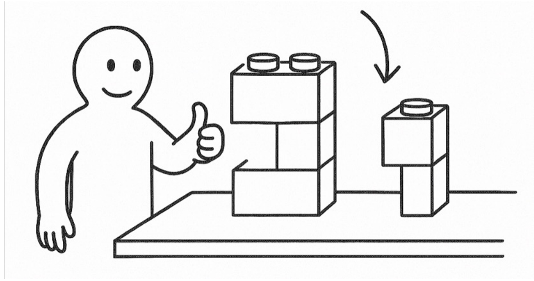
Model Folding!



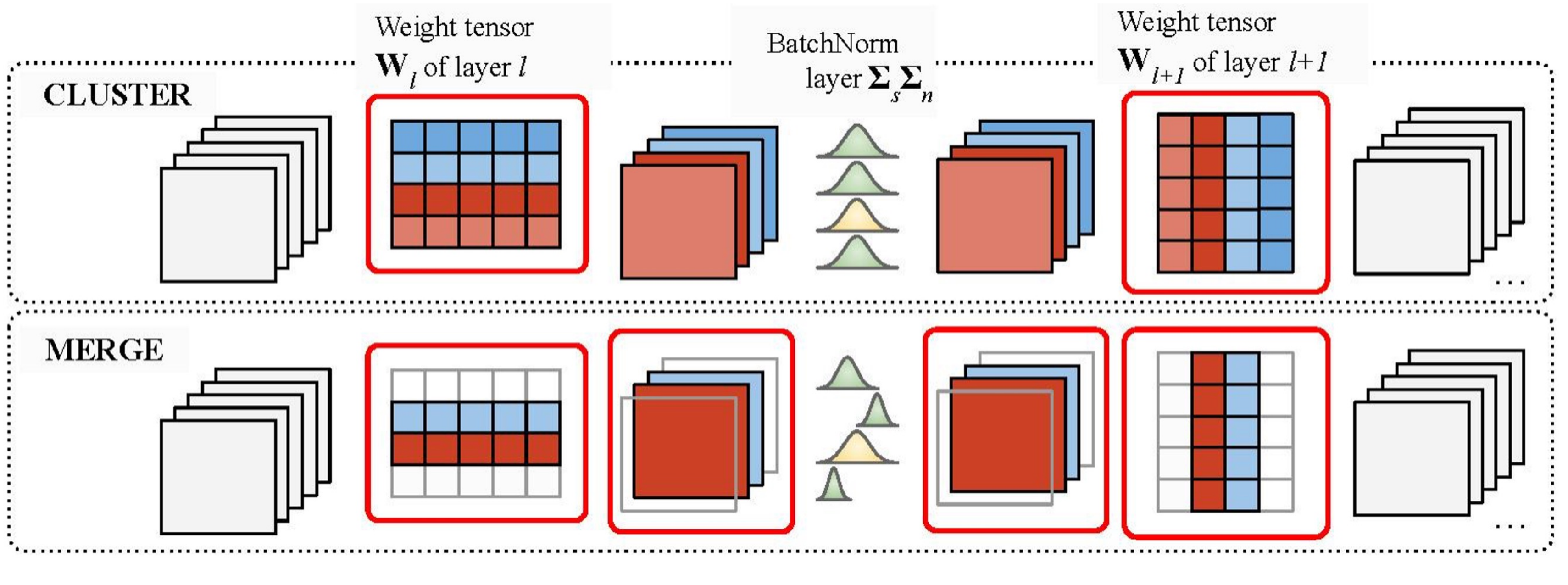
Model Folding

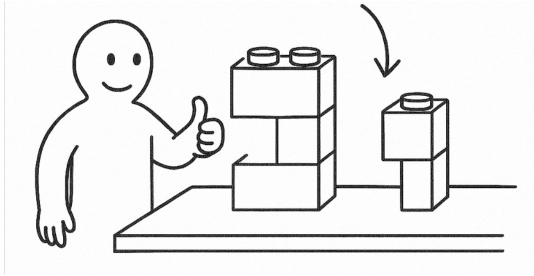


Cluster weights by k-means



Model Folding

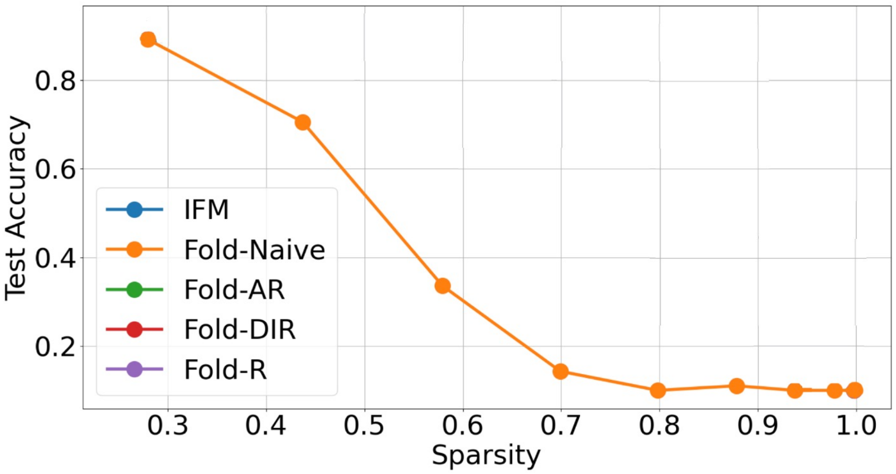




Merge the clustered groups

Model Folding

Fold ResNet18-CIFAR10

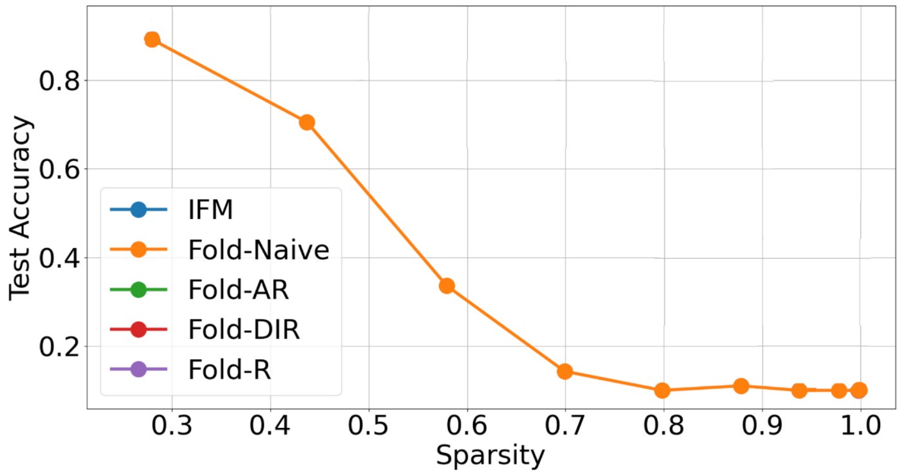


Naive model folding does not work well at high sparsity.

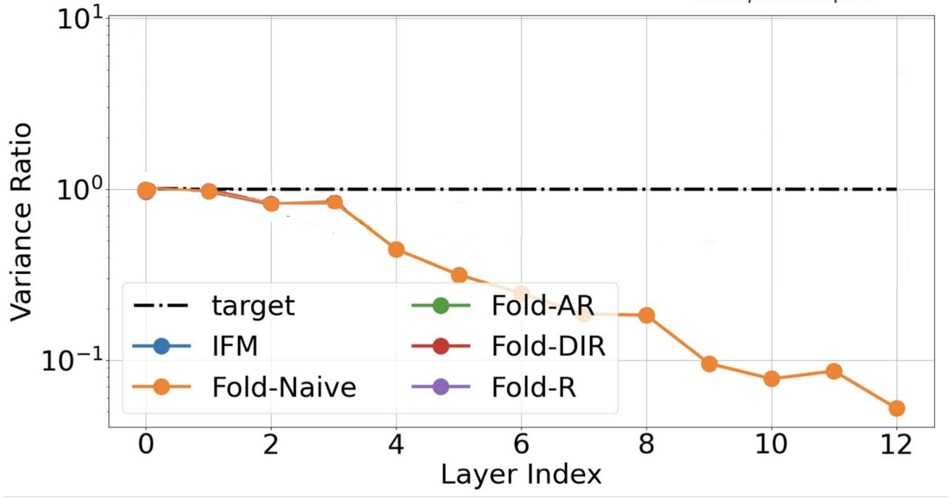
Why?

Model Folding

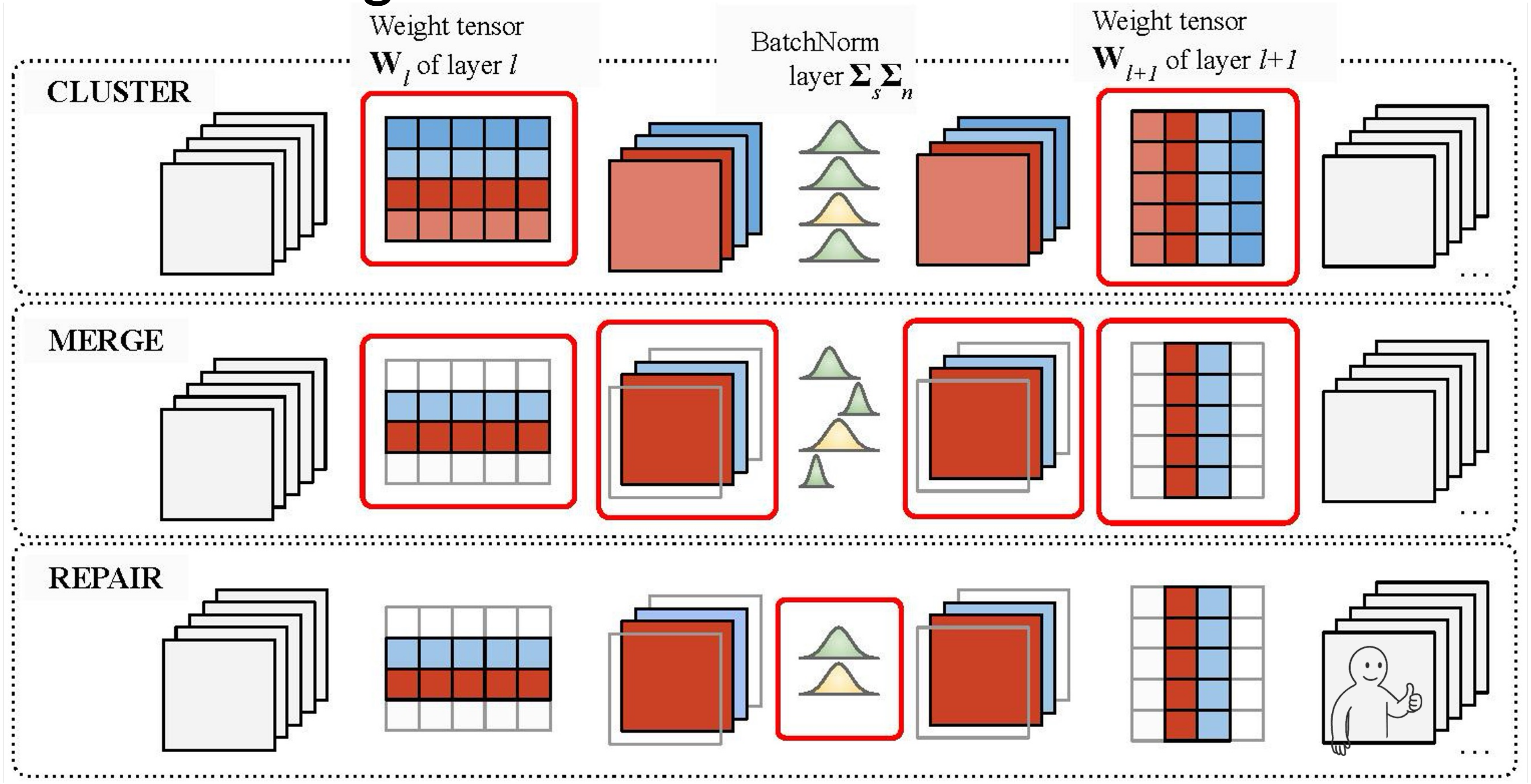
Fold ResNet18-CIFAR10



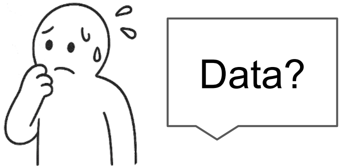
Variance collapse!



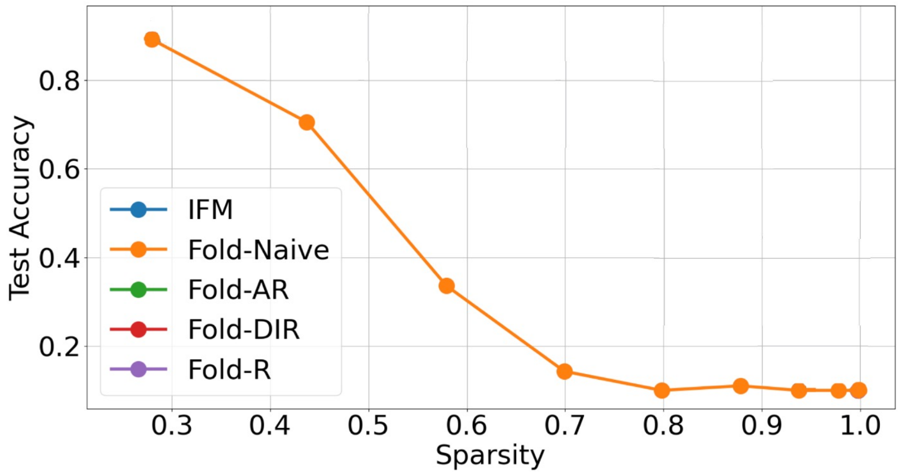
Model Folding with REPAIR



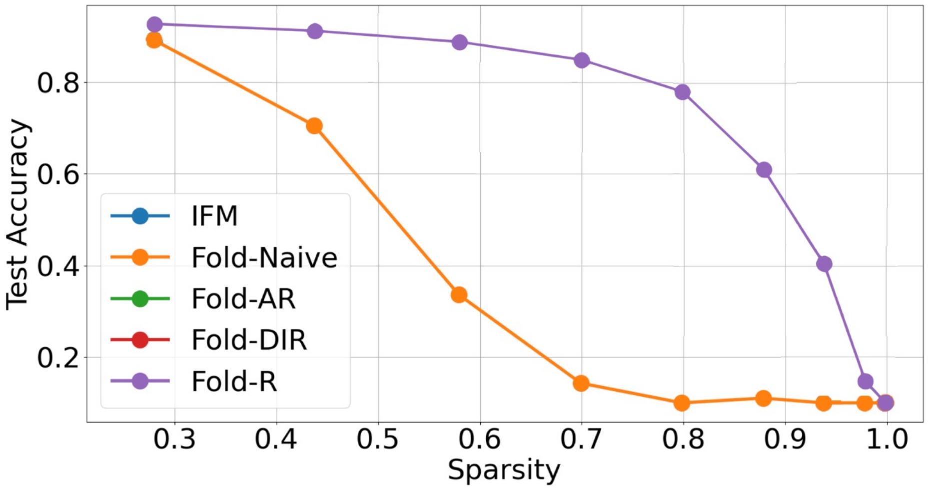
Model Folding with REPAIR



Fold ResNet18-CIFAR10



Fold ResNet18-CIFAR10 with REPAIR



Jordan, Keller, et al. "Repair: Renormalizing permuted activations for interpolation repair." 2022.

➔ Rescale each channel in the folded network such that its mean/std matches the uncompressed network

Model Folding with Data-Free REPAIR

Fold-AR = Approximate REPAIR

Algorithm 2 Fold-AR for a Single Layer

Require: $\mathbf{W}_{\sigma,l}$, $\mathbf{W}_{b,l}$, $\mathbf{W}_l$, $\mathbf{W}_{l+1}$ ▷ Input components of the layer

Compute the normalized weight matrix: $\bar{\mathbf{W}}_l \leftarrow \mathbf{W}_{\sigma,l}\mathbf{W}_l$

Construct the combined weight matrix: $\hat{\mathbf{W}}_l \leftarrow [\mathbf{W}_{l+1}^T \quad \bar{\mathbf{W}}_l \quad \mathrm{diag}(\mathbf{W}_{b,l})]$

Solve the clustering problem:

$$\begin{aligned} \min_{\mathbf{U}_l} \quad & \|\hat{\mathbf{W}}_l - \mathbf{C}_l\hat{\mathbf{W}}_l\|_F^2 \\ \text{s.t.} \quad & \mathbf{C}_l = \mathbf{U}_l(\mathbf{U}_l^T\mathbf{U}_l)^{-1}\mathbf{U}_l^T \\ & (\mathbf{U}_l)_{i,j} \in \{0,1\} \quad \sum_j (\mathbf{U}_l)_{i,j} = \mathbf{1} \end{aligned}$$

Update the scaling matrix: $\mathbf{W}_{b,l} \leftarrow (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{W}_{b,l}\mathbf{U}$

Update the second-layer weights: $\mathbf{W}_{l+1}^T \leftarrow \mathbf{U}^T\mathbf{W}_{l+1}^T$

Update the current-layer weights: $\bar{\mathbf{W}}_l \leftarrow (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\bar{\mathbf{W}}_l$

for $c = 1, \ldots, m$ **do** ▷ Adjust scaling factors for each cluster

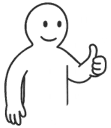
  Compute cluster size: $N_c \leftarrow \sum_i I(\mathbf{U}_{i,c} = 1)$ ▷ $I(\cdot)$ is the indicator function

  Compute intra-cluster correlation:

$$E[c] \leftarrow \frac{1}{N_c^2 - N_c}\sum_{i,j}\frac{\bar{\mathbf{w}}_{l,i}\cdot\bar{\mathbf{w}}_{l,j}^T}{\sqrt{\|\bar{\mathbf{w}}_{l,i}\|^2\|\bar{\mathbf{w}}_{l,j}\|^2}} I(\mathbf{U}_{i,c} = \mathbf{U}_{j,c} = 1) I(i \neq j)$$

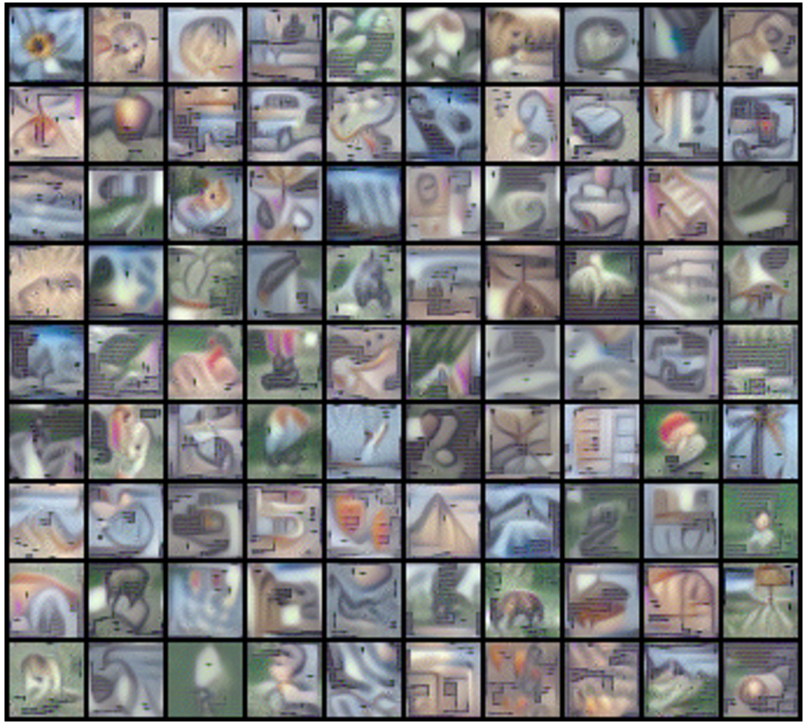
  Update the scaling factor for cluster c:

$$\mathbf{W}_{\sigma,l,c,c} \leftarrow \mathbf{W}_{\sigma,l,c,c}\frac{N_c}{\sqrt{N_c + (N_c^2 - N_c)E[c]}}$$



end for

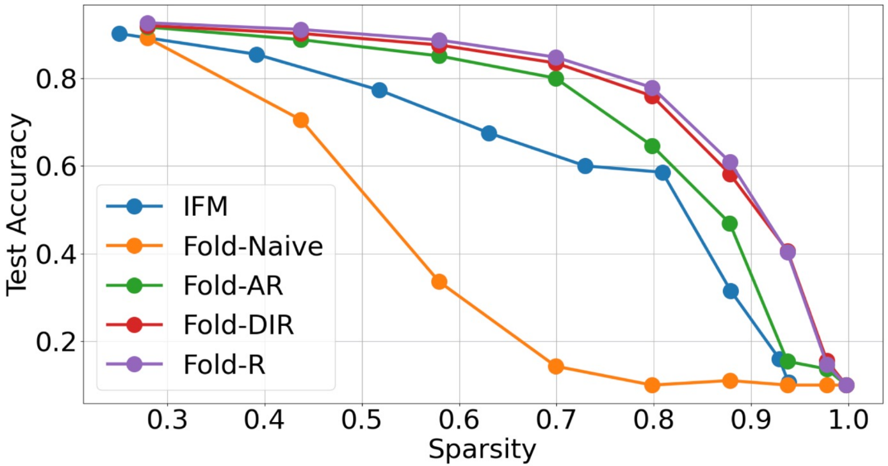
Fold-DIR = Deep Inversion REPAIR



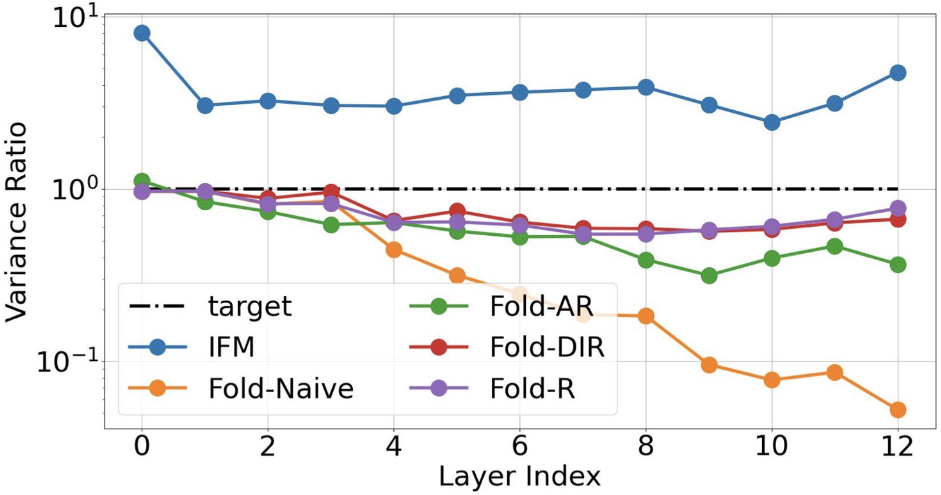
H. Yin et al, Dreaming to distill: Data-free knowledge transfer via deepinversion, 2020

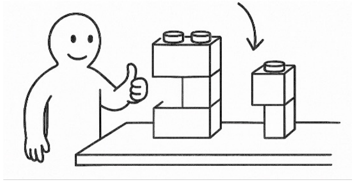
Model Folding with Data-free REPAIR

Fold ResNet18-CIFAR10 with Data-free REPAIR

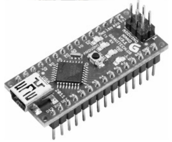


Repaired Variance





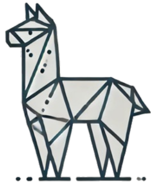
Model Folding is data-free and Fine-tuning-free



Model folding in combination with standard pipelines for edge deployment → 2.5x faster, 1.8x less RAM

Sparsity	10%			25%			50%			70%		
	Runtime	RAM	Flash	Runtime	RAM	Flash	Runtime	RAM	Flash	Runtime	RAM	Flash
NVIDIA Jetson Nano [NVI24]	2ms	59.5K	3.4M	2ms	55.7K	2.8M	1ms	48.0K	1.9M	1ms	36.5K	1.2M
ESP-EYE [Esp24]	2591ms	59.5K	3.4M	1868ms	55.7K	2.8M	1532ms	48.0K	1.9M	1186ms	36.5K	1.2M
Arduino Nano 33 BLE Sense [Ard24]	6831ms	59.5K	3.4M	3726ms	55.7K	2.8M	4218ms	48.0K	1.9M	2969ms	36.5K	1.2M

Model Folding on LLaMA without any data usage or fine-tuning



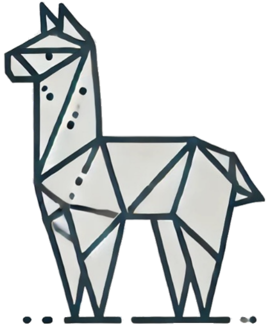
Prune ratio	Method	Data usage	WikiText2↓	BoolQ	WinoGrande	ARC-e	ARC-c	Average↑
0%	LLaMA-7B (Touvron et al., 2023a)	/	5.68	75.05	69.93	75.34	41.89	65.55
20%	Magnitude Prune	/	36136	43.21	49.40	27.23	21.59	35.36
20%	LLM-Pruner (Ma et al., 2023)	Gradients	10.53	59.39	61.33	59.18	37.18	54.27
20%	FLAP (An et al., 2023)	Calibration	6.87	69.63	68.35	69.91	39.25	61.79
20%	Wanda_sp (Sun et al., 2023)	Calibration	8.22	71.25	67.09	71.09	42.58	63.00
20%	SliceGPT (Ashkboos et al., 2024)	Calibration	7.00	57.80	67.96	62.67	36.01	56.11
20%	ShortGPT (Men et al., 2024)	Calibration	15.48	62.17	67.40	58.88	31.91	55.09
20%	Model Folding	/	13.33	62.29	62.19	49.83	26.37	50.17

Summary

1. **Model Folding** is a data-free and fine-tuning-free model compression method.
2. **Fold-AR, Fold-DIR** are data-free REPAIR approximation methods.
3. Model folding surpasses the performance of SOTA data-free model compression.

Thank you!

Dong Wang
TU Graz
https://wangdongdong.wang



Paper website

ICLR
International Conference On Learning Representations