WACV WAIKOLOA HAWAII JAN 3-7 • 2023

Project website

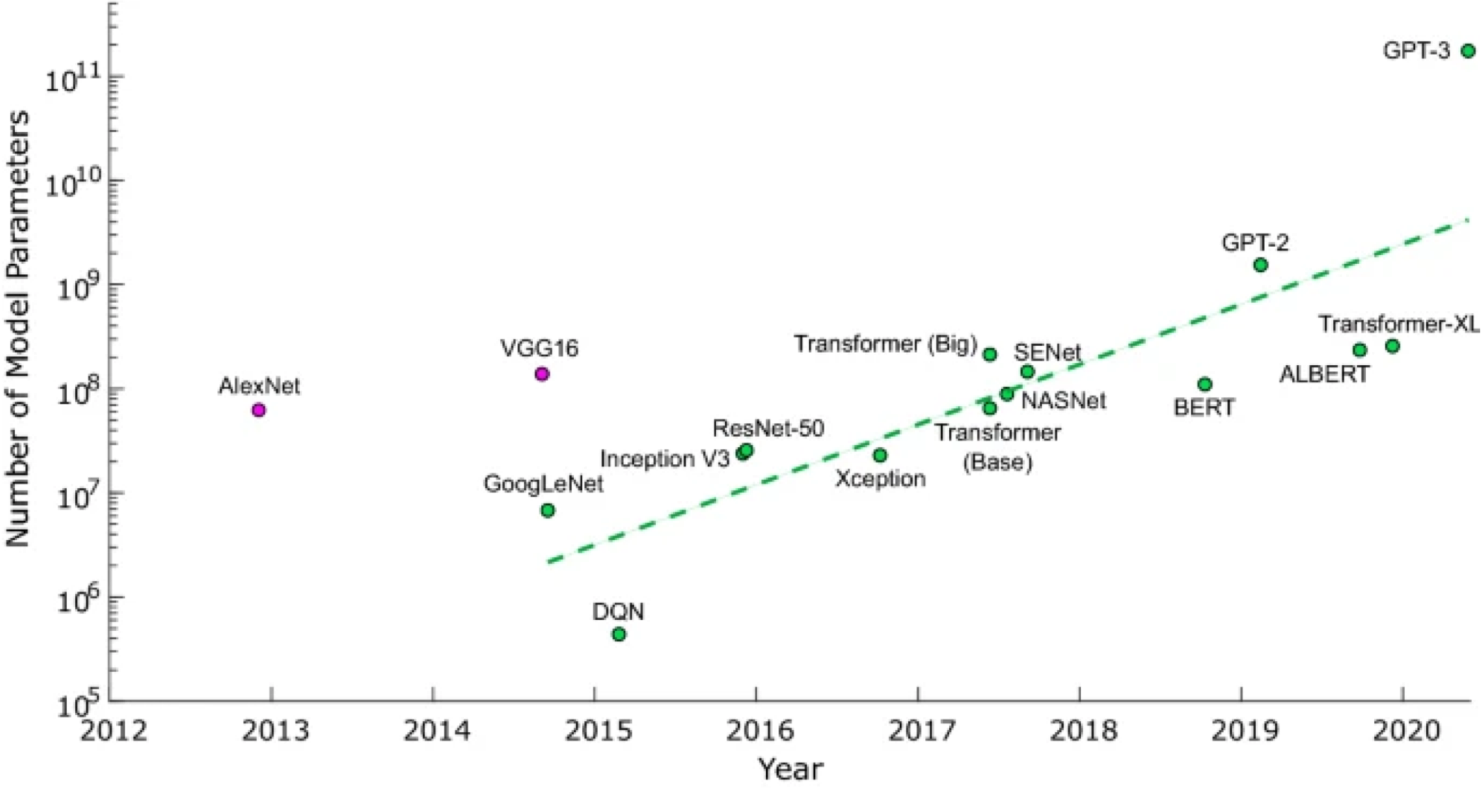
Compact and Optimal Deep Learning with Recurrent Parameter Generators

Jiayun Wang[*1] Yubei Chen[*3,4] Stella X. Yu[1,2] Brian Cheung[5] Yann LeCun[3,4]

[1]UC Berkeley / ICSI [2]University of Michigan [3]Meta [4]New York University [5]MIT

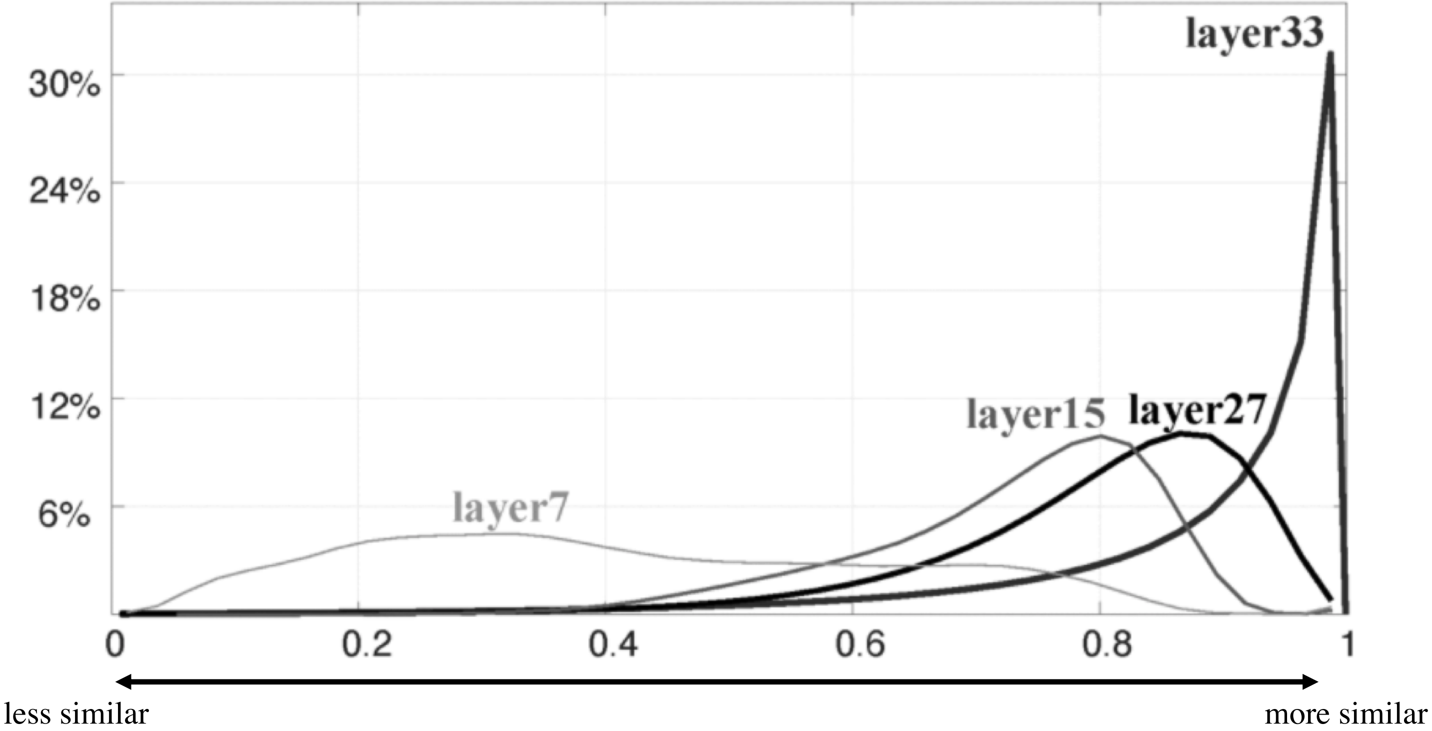
* indicates equal contribution

Exponential Growth in Neural Network Size



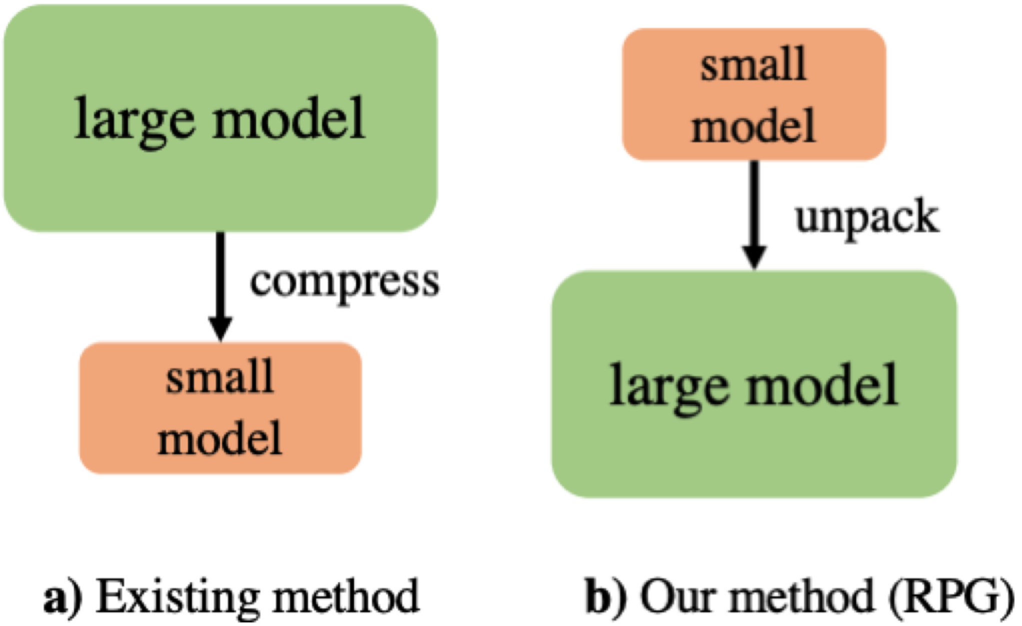
Freely scalable and reconfigurable optical hardware for deep learning. Bernstein et al. (2021)

Neural Nets are Redundant: Filter Similarity Increases with Depth

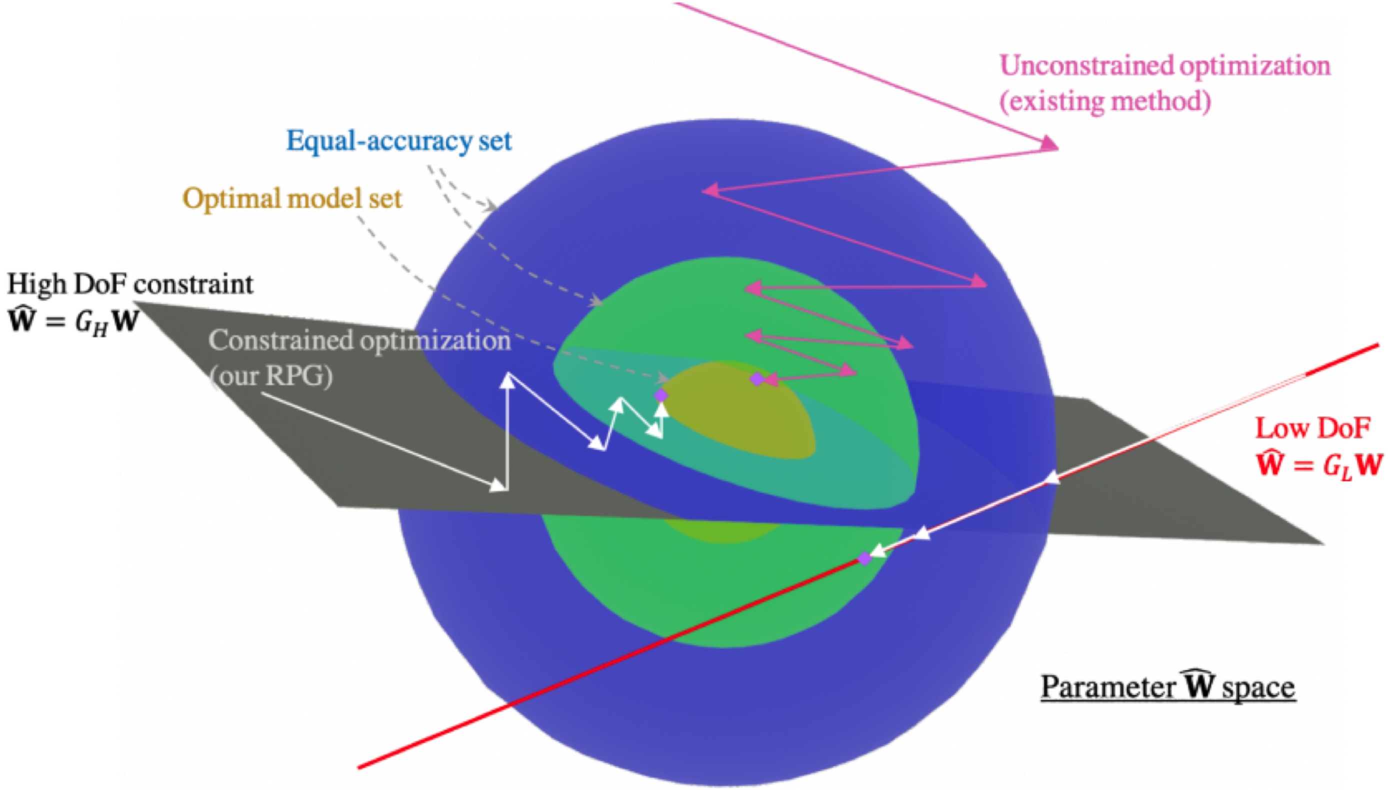


Histogram of Filter Similarity (ResNet34)

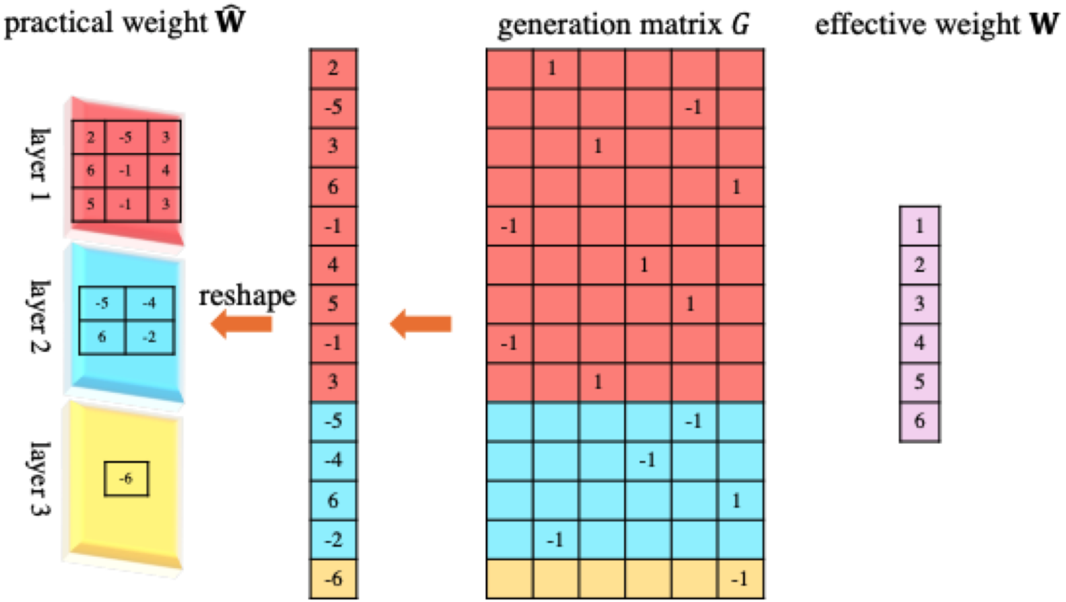
New Paradigm for Compact and Optimal Deep Learning



Linearly Constrained Neural Optimization

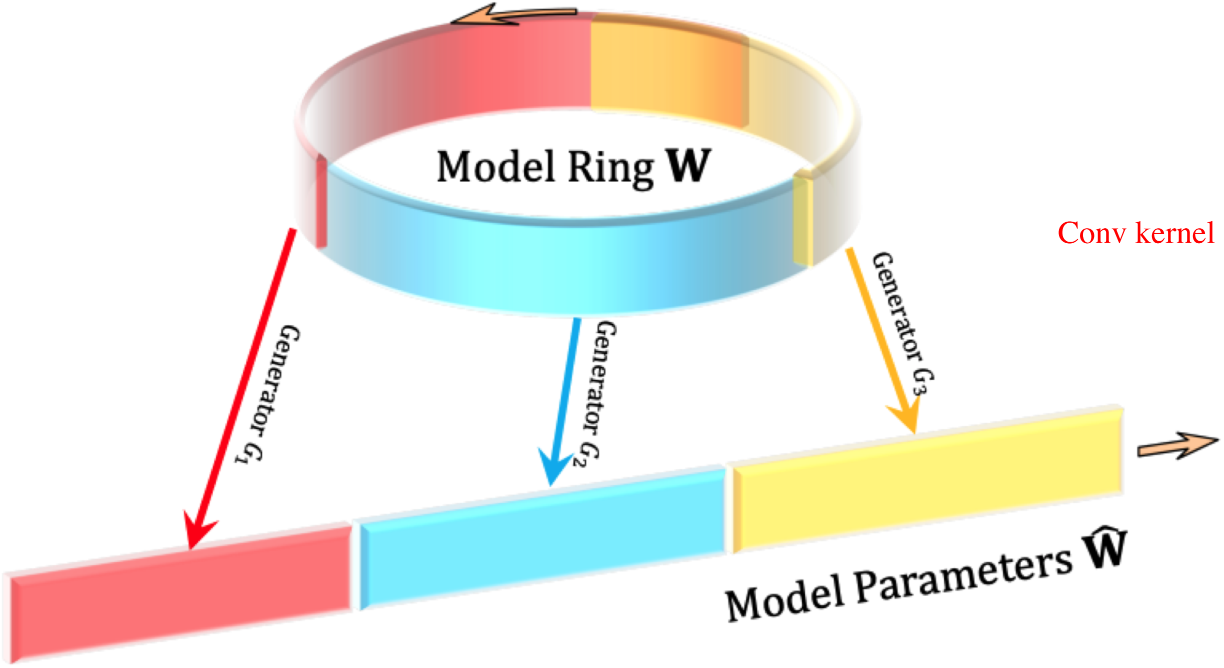


Linearly Constrained Neural Optimization



Networks are optimized with a linear constraint $\hat{\mathbf{W}} = G\mathbf{W}$
The constrained (practical) parameter $\hat{\mathbf{W}}$ of each network layer was generated by the generating matrix $\boldsymbol{G}$ from the free (effective) parameter $\mathbf{W}$, which is directly optimized.
$\hat{\mathbf{W}}$ is unpacked large model parameter while the size of $\mathbf{W}$ is the model DoF (degree of freedom).

Recurrent Parameter Generator (Special Case)



Model parameters is a combination of conv kernels:

$$\widehat{W} = [\mathbf{K}_1, \boldsymbol{K}_2, \dots, \boldsymbol{K}_L]$$

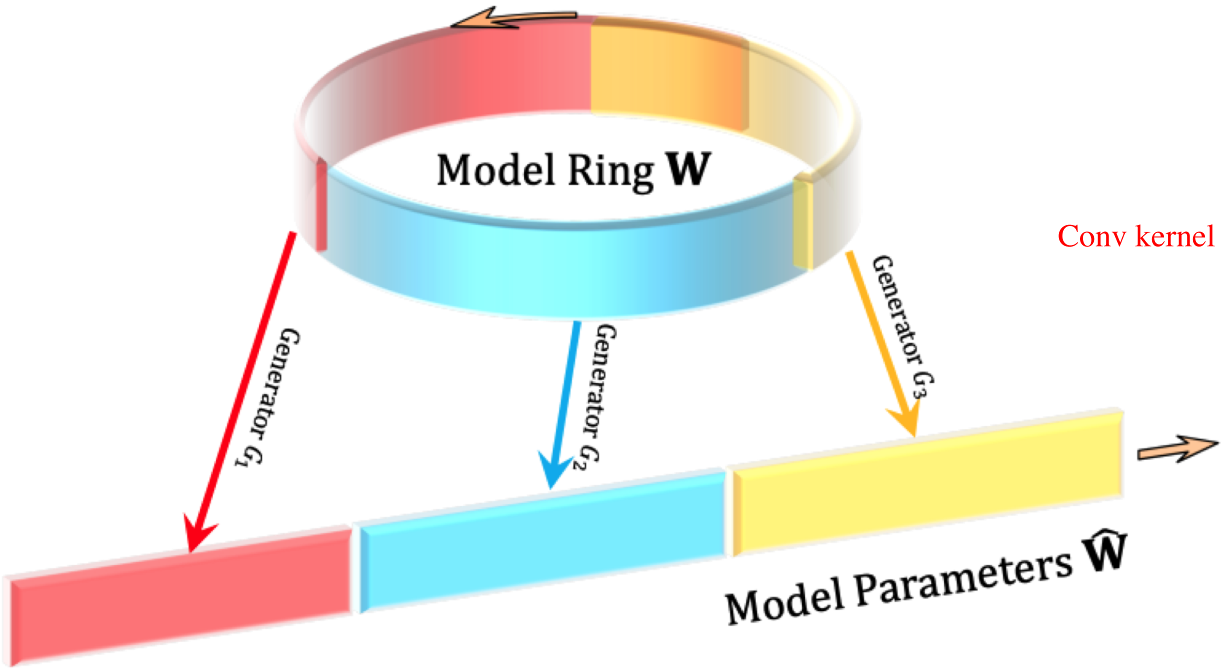
Generating matrix

$$\mathbf{K}_i = G_i \cdot \mathbf{W}, i \in \{1, \dots, L\}$$

Conv kernel at layer *i*

(shared) Model ring

Recurrent Parameter Generator (Special Case)



Model parameters is a combination of conv kernels:

$$\widehat{W} = [\mathbf{K}_1, \boldsymbol{K}_2, \dots, \boldsymbol{K}_L]$$

Generating matrix

$$\mathbf{K}_i = G_i \cdot \mathbf{W}, i \in \{1, \dots, L\}$$

Conv kernel at layer *i* (shared) Model ring

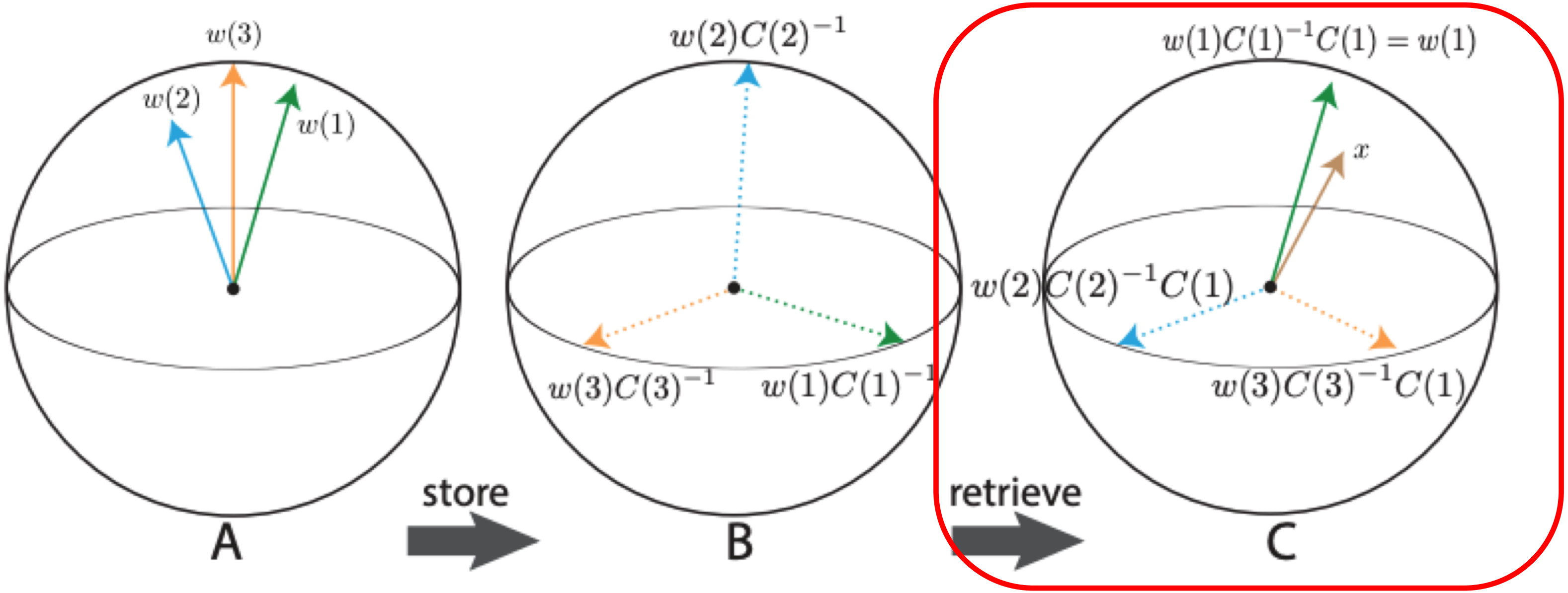
Destructive weight sharing:

$$G_i \in \{b \circ p | b \in B(N_i), p \in P(N_i)\}$$

Random sign flip permutation

We prove the orthogonality of two generated kernels (see Appendix)

Why Orthogonal? Retrieval from the Associative Memory



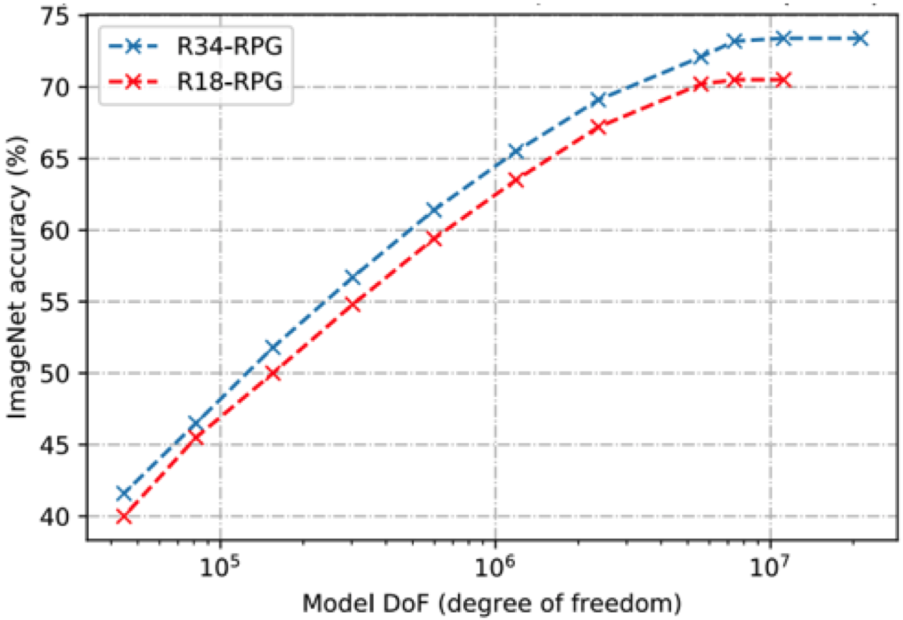
Superposition of many models into one. [Cheung et al. (2020)]

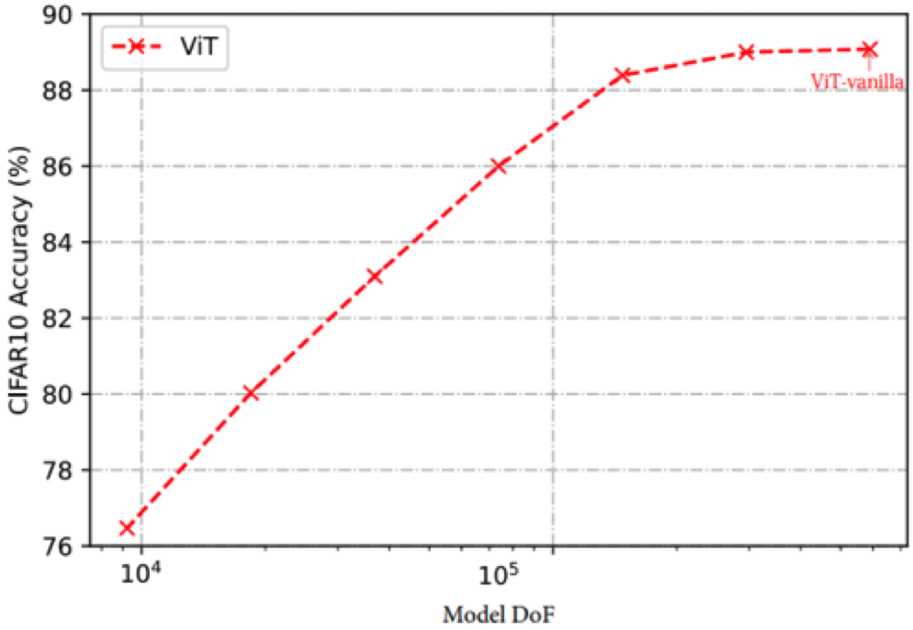
Extreme Model DoF Compression

Acc. (%)	R18-RPG			R18-vanilla	R34-RPG			R34-vanilla
ImageNet	40.0	67.2	70.5	70.5	41.6	69.1	**73.4**	**73.4**
CIFAR100	60.2	75.6	77.6	77.6	61.7	76.5	**78.9**	**79.1**
Model DoF	45K	2M	5.5M	11M	45K	2M	11M	21M

For ImageNet classification, RPG achieves 96% of ResNet18's performance with only 18% DoF (the equivalent of one convolutional layer)
RPG achieves 52% of ResNet34's performance with only 0.25% DoF!

Log-Linear DoF-Accuracy Relationship





- Accuracy and model DoF follow a *power law*.
- The exponents of the power laws are the same for ResNet18-RPG and ResNet34-RPG on ImageNet.
- RPG enables *under-parameterized* models for l*arge-scale* datasets such as ImageNet, which may unleash new findings.

RPG Performs Better at the Same Model DoF

Image Classification

	DoF	Acc. (%)
R18-vanilla	11M	77.5
R34-RPG.blk	11M	78.5
R34-RPG	11M	**78.9**
R34-random weight share	11M	74.9
R34-DeepCompression [23]	11M	72.2
R34-Hash [12]	11M	75.6
R34-Lego [67]	11M	78.4
R34-vanilla	21M	**79.1**

Pose Estimation

Acc. (DoF)	CPM [62]	RPG	No shared w.
1x sub-net	84.7 (3.3M)		
2x sub-nets	86.1 (3.3M)	86.5 (3.3M)	87.1 (6.7M)
4x sub-nets	86.5 (3.3M)	87.3 (3.3M)	88.0 (13.3M)

Multitask Regression

RMSE (%)	Depth	Normal
Vanilla model	25.5	41.0
RPG with shared BN	24.7	40.3
Reuse & new BN	24.0	39.4
Reuse & new BN & perm. and reflect.	**22.8**	**39.1**

RPG can be Pruned and Quantized for Faster Runtime

Pruning RPG

fine-grained pruning

	acc before	acc after ↓ DoF	acc drop	model DoF
R18-IMP [18]	92.3	90.5	1.8	274k
R18-RPG	95.0	93.0	2.0	274k

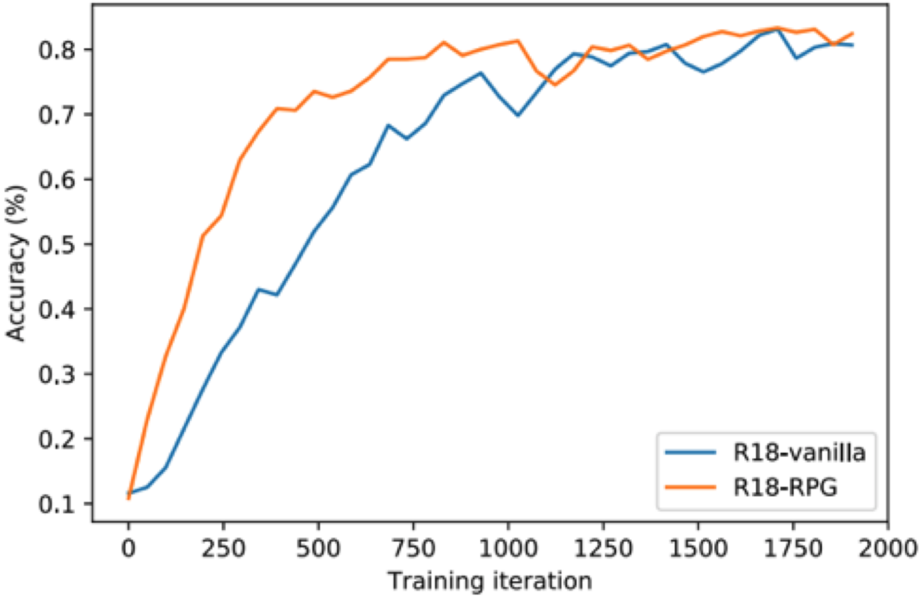
coarse-grained pruning

	DoF before pruning	Pruned acc.	FLOPs
R18-Knapsack	11.2M	69.35%	1.09e9
Pruned R18-RPG	5.6M	69.10%	1.09e9

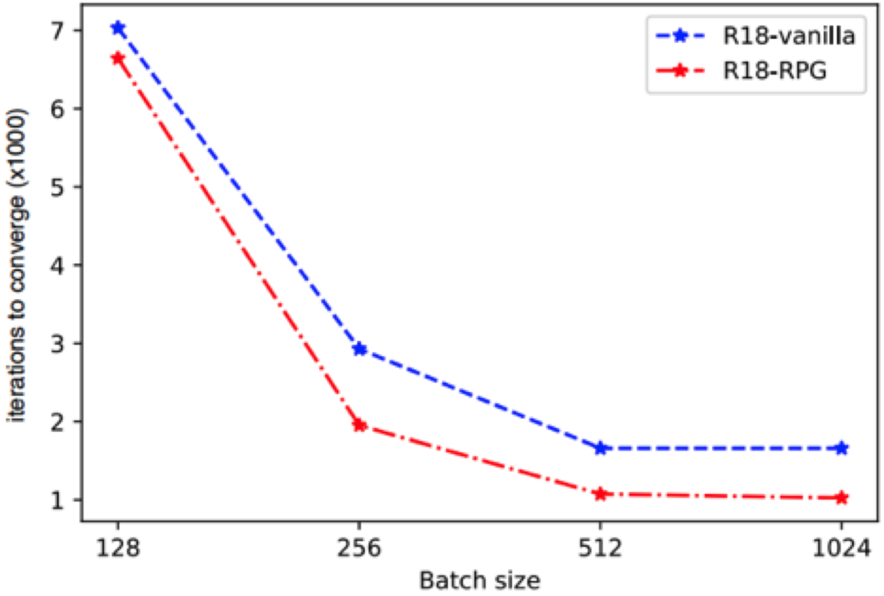
Quantize RPG

	# Params	Acc before	Acc after ↓ quantization	Acc drop
R18-vanilla	11M	69.8	69.5	0.3
R18-RPG	5.6M	70.2	70.1	0.1

RPG Converges Faster

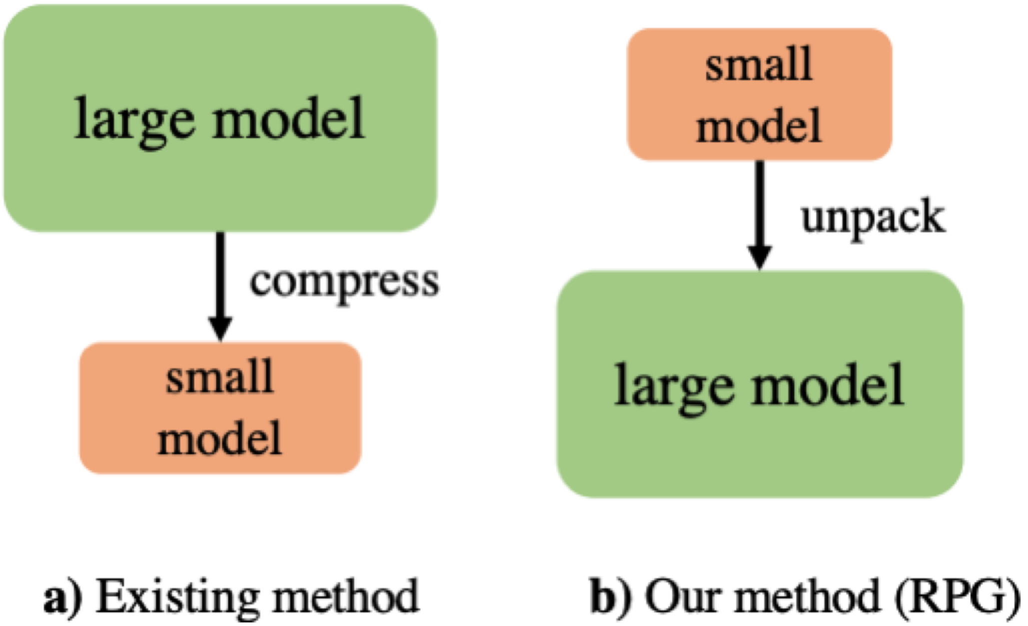


RPG converges faster than the vanilla model



RPG converges faster for different batch sizes

Summary: New Paradigm for Compact and Optimal Deep Learning



RPG directly optimizes a lean model with a small DoF.
Faster and higher accuracy.