Samsung Research

Data-Free Model Pruning at Initialization via Expanders

James Stewart, Umberto Michieli and Mete Ozay

Samsung Research UK

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Abstract

Model Pruning at Initialization (PaI) trains sparse networks to comparable accuracy with respect to their dense counterparts. We investigate data-free PaI based on the expansion properties of network graphs. In particular,

- We propose a stronger model (**RReg**) for generating expanders, which we then use to sparsify a variety of mainstream CNN architectures;
- We demonstrate that accuracy is an increasing function of expansion in a sparse model;
- We analyse the superior performance of **RReg** over the strong naïve random baseline and alternative models.

Pruning at Initialization (PaI)

Many pruning paradigms

Pruning paradigm	Weight source	Mask source	Train sparse network?
Pruning after training	Converged net	Converged net	Yes (fine-tune)
PaI – Sparse Selection	Initialized net	Initialized net	No
PaI – Sparse Training #1	Initialized net	Converged net	Yes
PaI – Sparse Training #2	**Initialized net**	**Initialized net**	**Yes**

Our Focus: Data-Free Pruning at Initialization of randomly selected weights.

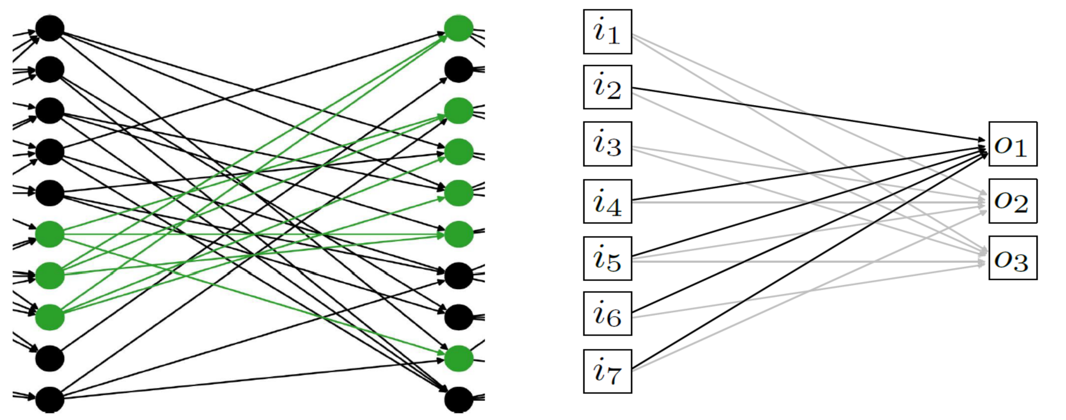
Our Three Steps:

1) Initialize random weights
2) Compute pruning mask
3) Train the sparse network to convergence

Computing the pruning mask via expander graphs

To capture the graph structure of a neural network, we model individual layers as bipartite graphs, as in [1].

Definition (α-expander). Let $d \in \mathbb{N}_{\geq 3}$ (degree) and $\alpha \in \mathbb{R}_{\geq 0}$. We say that a d-regular bipartite graph $G = (V_G^0, V_G^1, E_G)$ is an α-expander if, $\forall i \in \{0,1\}$ and $\forall S \subseteq V_G^i$ with $|S| \leq |V_G^i|/2$, we have that $|\partial S| \geq \alpha |S|$, where ∂S denotes the set of vertices connected to S.



(a) **Linear** layer: graph expansion allows subsets of input neurons to interact with larger subsets of output ones (highlighted in green).

(b) **Conv** layer: a filter is applied to input channels $\{i_j\}_{j=2,4,5,6,7}$, to obtain output channel o_1, as determined by the graph structure.

Main benefits:

→ Expanders graphs are simultaneously **sparse** yet **highly connected**.

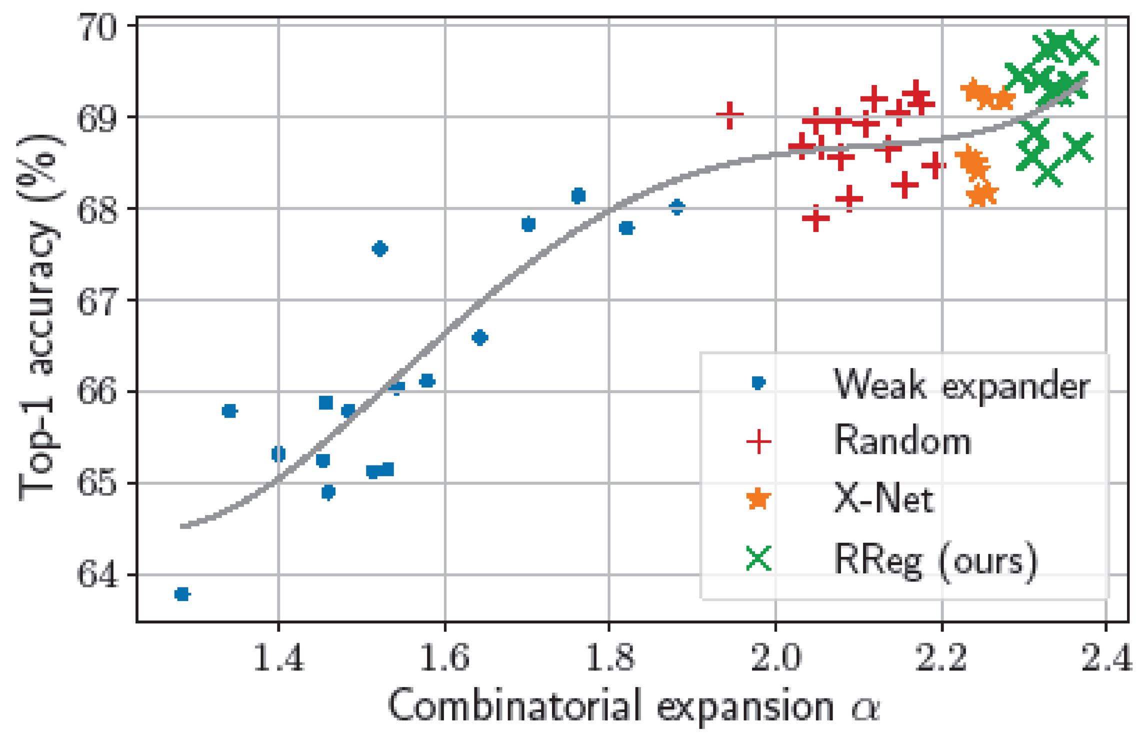
- subsets of neurons to interact with a larger subset of other neurons,
- higher **feature shareability** and **flow of information** through the network

Methods to achieve expansion property:

For fixed *n* and *d* :

- **Random model [2]** connects every pair of vertices with an edge independently with probability *2d/n.*
- **X-Net [1]** chooses a random *n*-vertex *d*-left-regular (*i.e.*, every left vertex has degree *d*) graph uniformly at random from the set of all such graphs.
- **Rreg (ours)** chooses a random *n*-vertex *d*-regular graph uniformly at random from the set of all such graphs.

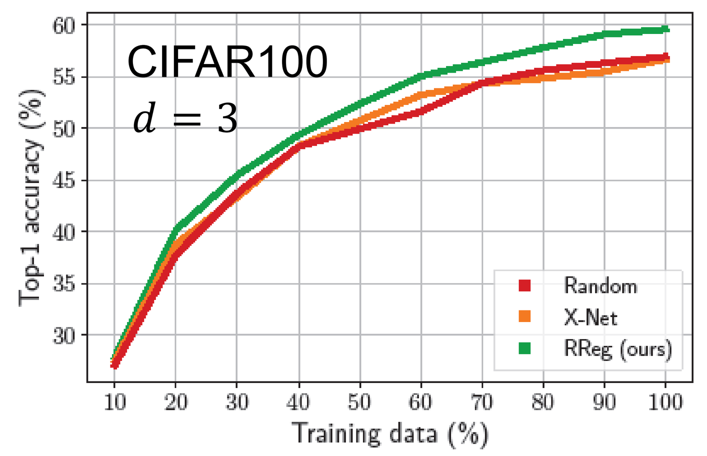
Model	Regularity	α	Edges
Random	random	low	$d \cdot n/2$
X-Net	d-left-regular	medium	$d \cdot n/2$
RReg (ours)	d-regular	high	$d \cdot n/2$



Results

PP: remaining parameters [%]

Δ: relative gain vs. random



Method	PP	d	CIFAR10 Acc	CIFAR10 Δ_R	CIFAR10 Δ_{RAR}	CIFAR100 Acc	CIFAR100 Δ_R	CIFAR100 Δ_{RAR}
Original	100	-	94.24	-	-	74.16	-	-
Random [23]	13.79 ±0.11	60	**93.50**	-	-	**72.97**	-	-
X-Net [25]			93.46	-0.04	-0.62	72.73	-0.33	-0.89
RReg (ours)			**93.50**	+0.00	+0.61	72.72	-0.34	-0.04
Random [23]	7.07 ±0.20	30	92.51	-	-	69.81	-	-
X-Net [25]			92.65	+0.15	+1.87	69.80	-0.01	-0.03
RReg (ours)			**93.05**	+0.58	+5.44	**70.15**	+0.49	+1.16
Random [23]	3.62 ±0.15	15	91.30	-	-	66.58	-	-
X-Net [25]			91.38	+0.09	+0.92	66.81	+0.35	+0.69
RReg (ours)			**91.50**	+0.22	+1.39	**67.72**	+1.71	+2.74
Random [23]	0.79 ±0.03	3	85.81	-	-	56.98	-	-
X-Net [25]			86.06	+0.29	+1.76	56.69	-0.51	-0.67
RReg (ours)			**87.02**	+1.41	+6.89	**59.61**	+4.62	+6.74

Tiny-ImageNet:

Method	VGG16 PP [%]	VGG16 Acc	MN PP [%]	MN Acc	RN18 PP [%]	RN18 Acc	RN34 PP [%]	RN34 Acc	RN50 PP [%]	RN50 Acc	RN101 PP [%]	RN101 Acc	RN152 PP [%]	RN152 Acc	WRN28-10 PP [%]	WRN28-10 Acc	WRN40-14 PP [%]	WRN40-14 Acc	Avg Acc	Avg Δ_R
Original	100	40.03	100	54.75	100	53.88	100	56.95	100	57.08	100	60.13	100	61.29	100	46.27	100	49.04	53.27	-
Random [23]	0.79 (±0.03)	22.84	19.33 (±0.10)	21.81	3.45 (±0.05)	44.02	2.33 (±0.02)	**47.34**	14.24 (±0.06)	46.77	8.52 (±0.08)	49.56	6.62 (±0.08)	49.05	1.04 (±0.02)	35.5	0.65 (±0.01)	39.51	39.63	-
X-Net [25]		23.20		19.49		42.69		46.97		45.36		49.45		49.47		35.57		40.21	39.16	-1.20
RReg (ours)		**25.31**		**34.26**		**44.30**		46.30		**48.27**		**51.04**		**51.21**		**36.50**		**40.54**	**41.94**	+5.11

RReg sparse network can achieve higher accuracy at a same number of parameters than their shallower and narrower fully-connected counterparts.

Type	Model	P	Acc	Δ_R	Δ_{RAR}	Model	P	Acc	Δ_R	Δ_{RAR}
Orig	RN50	23.9	59.36	-	-	WRN28-4	5.9	43.16	-	-
RReg	RN152	21.9	62.01	+4.46	+6.52	WRN28-10	5.9	43.76	+1.39	+1.06

Conclusion

✓ We proposed RReg to generate sparse layers with optimal expansion properties.

✓ We showed that classification accuracy is an increasing function of graph expansion

✓ RReg shows consistent improvement over strong baselines [1-2].

[1] Prabhu et al. *Deep expander networks: Efficient deep networks from graph theory,* ECCV, 2018.

[2] Liu et al. *The unreasonable effectiveness of random pruning: Return of the most naive baseline for sparse training,* ICLR, 2022.