N2NSkip: Learning Highly Sparse Networks using Neuron-to-Neuron Skip connections

Arvind Subramaniam **Avinash Sharma**

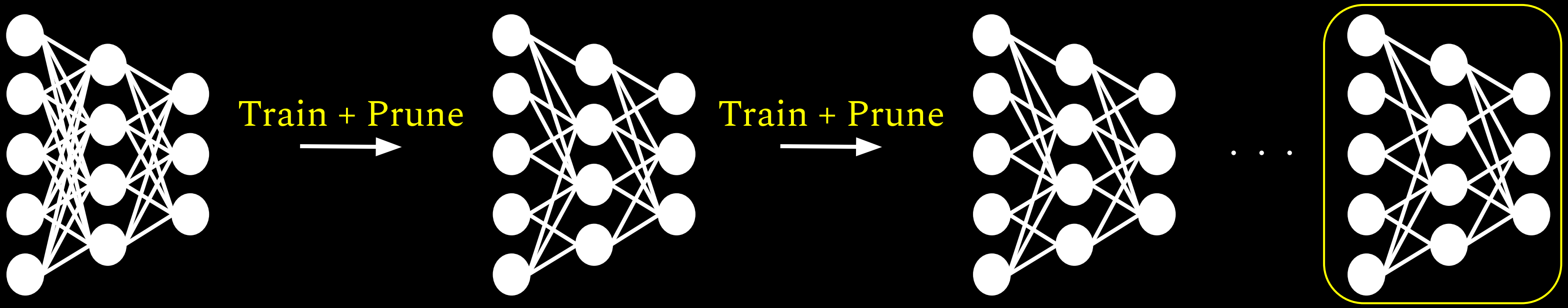
Centre for Visual Information Technology (CVIT)

IIIT Hyderabad

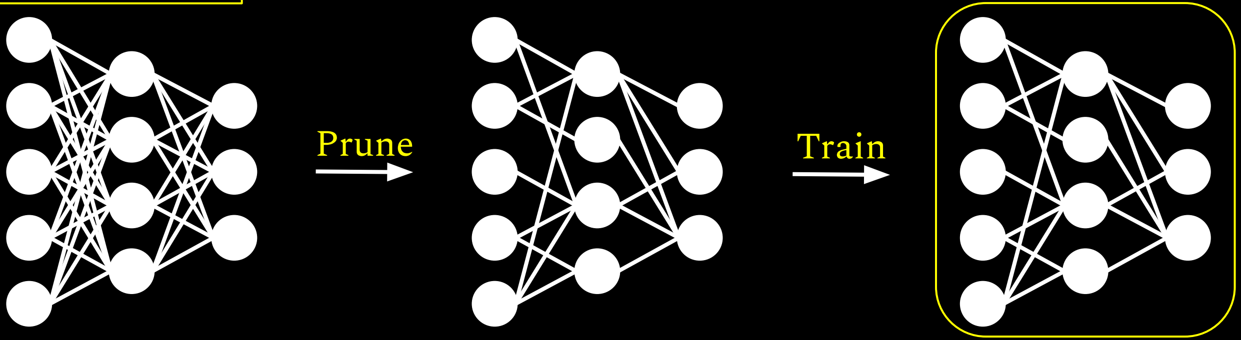
India

Network Pruning and Motivation

Iterative Pruning



One-shot Pruning



Why Single shot prune-train?

Advantages

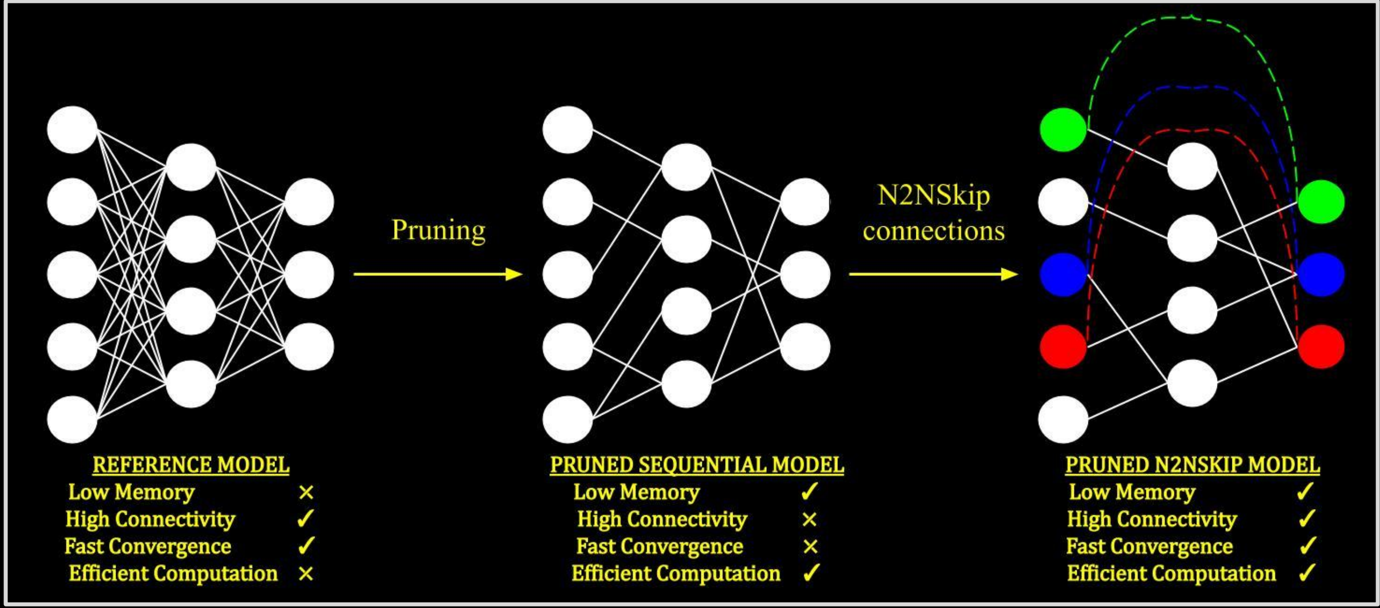
1. Relatively lower computation
2. Faster training time
3. Generic network structure

Disadvantages

1. Inferior overall connectivity of the pruned network
2. Slower convergence

Q. Is it possible to prune a network at initialization (prior to training) while maintaining rich connectivity, and also ensure faster convergence?

Single shot Prune-train + Rich connectivity



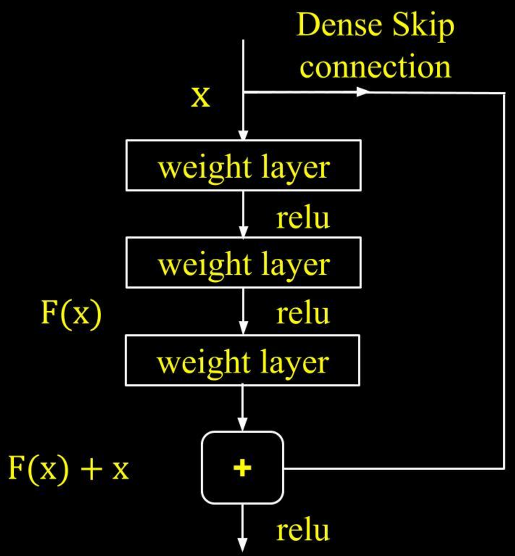
Neuron-to-Neuron Skip (N2NSkip) connections

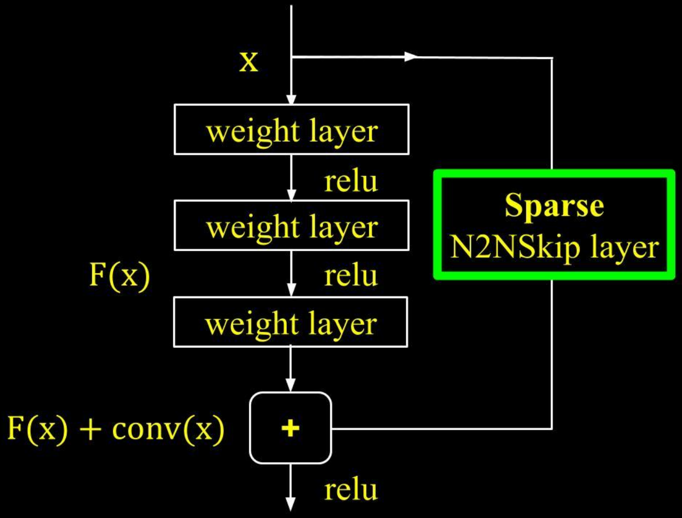
(Overall sparsity is maintained after adding N2NSkip connections)

Neuron-to-Neuron Skip (N2NSkip) connections

Skip connections in ResNet, where the dense output activation of a layer l is merely added to the output of the layer l + k.

For a given sparsity, neurons in layer l are randomly connected to neurons in layer l + k, ***while maintaining overall sparsity of the network***.





Contributions

1. **Superior Accuracy**

2. **Faster Convergence**

3. **Enhanced Connectivity**

Experimental Results

Preliminary Pruning Strategy

1. Randomized Pruning (RP)
2. Connection Sensitivity Pruning (CSP)

Datasets

1. CIFAR-10
2. CIFAR-100
3. ImageNet

1. CIFAR-10 and CIFAR-100

Model	Method	CIFAR-10			CIFAR-100		
		10%	5%	2%	10%	5%	2%
VGG19 (143M)	Baseline	93.16 ± 0.12	-	-	74.09 ± 0.15	-	-
	RP	92.08 ± 0.36	89.43 ± 0.75	86.52 ± 1.75	71.23 ± 0.26	69.82 ± 0.65	55.43 ± 1.94
	N2NSkip-RP	**92.92 ± 0.19**	**92.65 ± 0.25**	**91.12 ± 0.36**	**72.67 ± 0.23**	**72.13 ± 0.31**	**61.21 ± 0.42**
	CSP	92.79 ± 0.23	92.14 ± 0.47	90.35 ± 0.98	72.83 ± 0.27	71.92 ± 0.68	59.92 ± 1.21
	N2NSkip-CSP	**93.02 ± 0.13**	**92.86 ± 0.19**	**92.12 ± 0.29**	**73.72 ± 0.16**	**73.05 ± 0.25**	**65.45 ± 0.41**
ResNet50 (23M)	Baseline	95.33 ± 0.11	-	-	74.94 ± 0.13	-	-
	RP	88.53 ± 0.21	86.17 ± 0.39	83.33 ± 0.93	67.72 ± 0.25	62.28 ± 0.42	51.11 ± 1.01
	N2NSkip-RP	**91.59 ± 0.16**	**89.14 ± 0.24**	**87.67 ± 0.51**	**70.45 ± 0.14**	**67.56 ± 0.28**	**60.19 ± 0.51**
	CSP	93.15 ± 0.15	92.25 ± 0.27	89.12 ± 0.36	69.29 ± 0.22	65.73 ± 0.34	55.02 ± 0.79
	N2NSkip-CSP	**94.37 ± 0.12**	**93.59 ± 0.21**	**92.26 ± 0.31**	**72.37 ± 0.15**	**70.43 ± 0.27**	**63.16 ± 0.48**

Table 1: Test Accuracy of pruned ResNet50 and VGG19 on CIFAR-10 and CIFAR-100 with either RP or CSP as the preliminary pruning step.

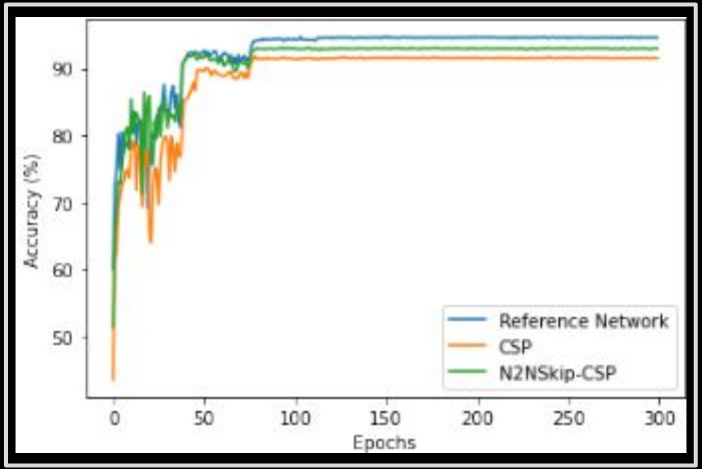
ImageNet

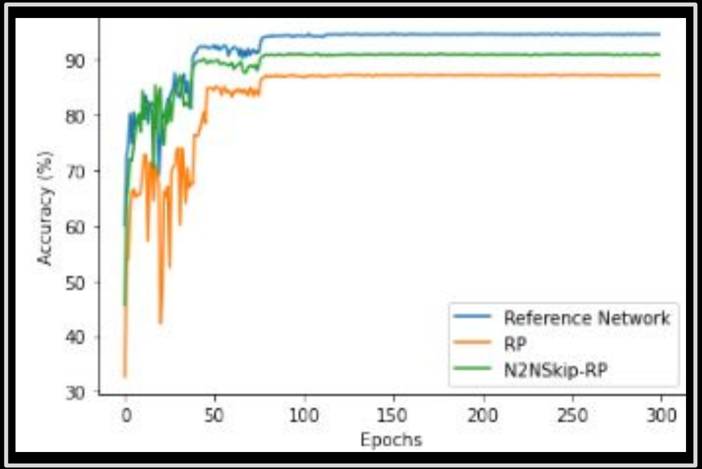
Model	Method	Density			Model	Method	Density		
		50%	30%	20%			50%	30%	20%
	Baseline	74.70 ± 0.26	-	-		Baseline	74.70 ± 0.28	-	-
ResNet50	CSP	73.42 ± 0.29	70.42 ± 0.37	68.67 ± 0.65	ResNet50	RP	72.46 ± 0.32	68.65 ± 0.45	65.32 ± 0.97
(23M)	**N2NSkip-CSP**	$\mathbf{74.59 \pm 0.22}$	$\mathbf{72.89 \pm 0.33}$	$\mathbf{72.09 \pm 0.45}$	(23M)	**N2NSkip-RP**	$\mathbf{74.12 \pm 0.29}$	$\mathbf{71.19 \pm 0.39}$	$\mathbf{70.03 \pm 0.51}$

Table 2: Test Accuracy of pruned ResNet50 on ImageNet with either CSP (left) or RP (right) as the preliminary pruning step.

Larger increase in accuracy at network densities of 5% and 2%, as compared to 10%, regardless of the preliminary one-shot pruning paradigm used.

2. Faster Convergence





(a) N2NSkip-RP vs RP on ResNet50 (b) N2NSkip-CSP vs CSP on ResNet50

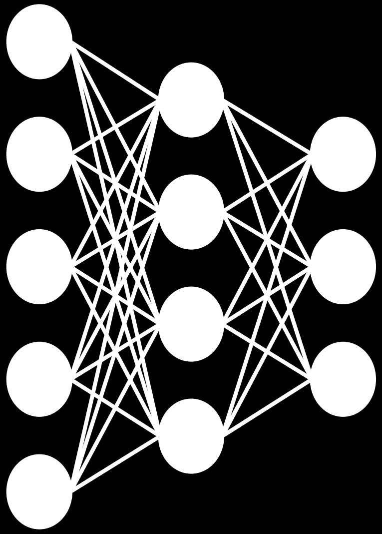
Accuracy of N2NSkip networks during the first fifty epochs is nearly equal to the baseline accuracy.

3. Connectivity Analysis

- Providing a novel framework that compares relative connectivity of each pruned network (wrt to the reference network).

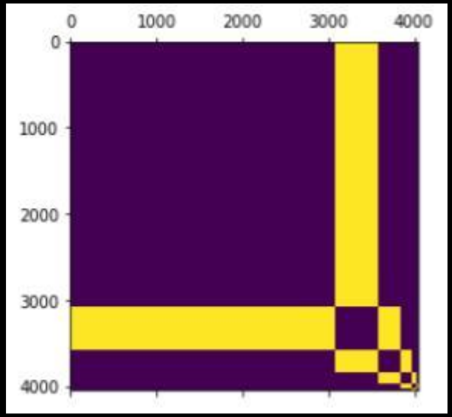
- Concept of heat diffusion to gauge network connectivity of classical DNNs.

Connectivity Analysis Pipeline



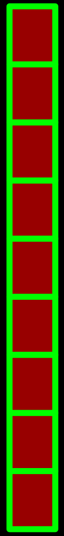
Pruned network
(after training)

Step 1



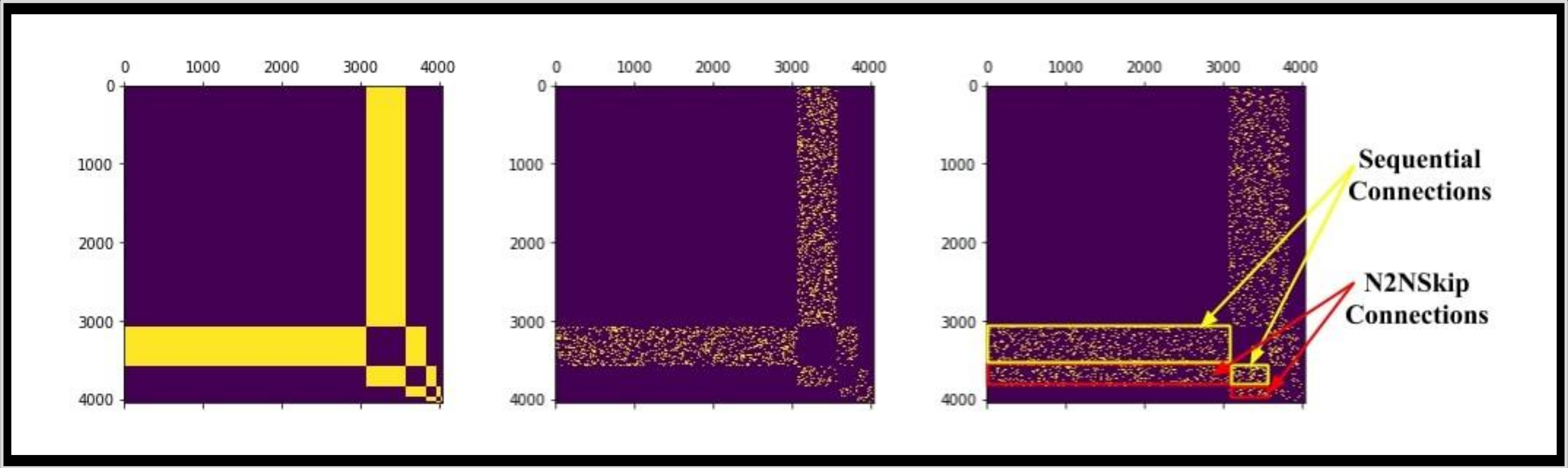
Adjacency matrix

Step 2



Heat Signature

1. Obtaining Adjacency Matrices



Reference Network **Pruned Network** **Pruned Network + N2NSkip**

Each adjacency matrix is an nxn dimensional matrix, where n in the total number of channels/neurons in the network.

Obtaining Heat Signatures

Constructing the graph Laplacian matrix and using its spectral embedding:

$$L = D - W$$

$$H(t) = Ue^{-\Lambda t}U^T$$

$$\boxed{S = H(t)A} \longrightarrow \boxed{\textbf{Heat Signature}}$$

Λ - diagonal matrix of corresponding eigenvalues.

A - n x 1 Binary matrix that assigns each node as source (1) or sink (0).

S - n x 1 matrix which gives an estimate of the heat signature of the network.

Comparing Heat Signatures

$$S = H(t)A$$

$$F = \|S_{\text{reference}} - S_{\text{prune}}\|_2$$

The relative connectivity of the pruned network with respect to the reference network is determined by the Frobenius norm of their respective heat signatures.

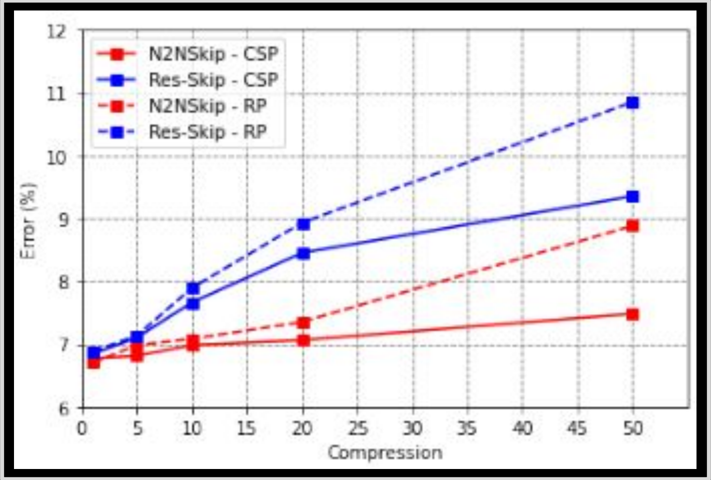
A lower value of F indicates superior overall connectivity

Comparing Heat Signatures

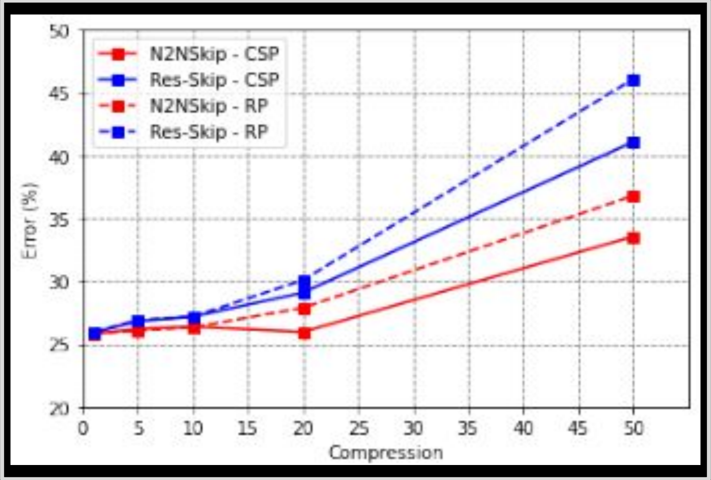
Model	Method	Density			
		50%	10%	5%	2%
VGG19	RP	3.6×10^{-3}	4.2×10^{-1}	2.3×10^{-1}	6.5×10^{0}
	N2NSkip-RP	$\mathbf{2.8\times10^{-6}}$	$\mathbf{4.9\times10^{-5}}$	$\mathbf{9.9\times10^{-4}}$	$\mathbf{1.3\times10^{-3}}$
	CSP	7.9×10^{-3}	7.1×10^{-5}	9.1×10^{-2}	2.3×10^{0}
	N2NSkip-CSP	$\mathbf{1.4\times10^{-6}}$	$\mathbf{2.5\times10^{-5}}$	$\mathbf{6.2\times10^{-5}}$	$\mathbf{3.3\times10^{-4}}$
ResNet50	RP	4.4×10^{-3}	3.9×10^{-2}	4.5×10^{-1}	1.2×10^{1}
	N2NSkip-RP	$\mathbf{8.1\times10^{-6}}$	$\mathbf{5.5\times10^{-5}}$	$\mathbf{3.8\times10^{-4}}$	$\mathbf{5.6\times10^{-3}}$
	CSP	7.9×10^{-3}	6.7×10^{-5}	6.1×10^{-2}	9.2×10^{0}
	N2NSkip-CSP	$\mathbf{5.3\times10^{-6}}$	$\mathbf{1.7\times10^{-5}}$	$\mathbf{4.2\times10^{-5}}$	$\mathbf{8.9\times10^{-4}}$

Table 3: Difference in connectivity of pruned models with respect to the reference network at saturated heat distribution. The difference is minimum for N2NSkip networks, thereby indicating superior overall connectivity in the model.

Comparison with conventional skip connections



(a) N2NSkip vs Res-Skip on VGG19 (CIFAR-10)



(b) N2NSkip vs Res-Skip on VGG19 (CIFAR-100)

At higher sparsity levels, N2NSkip connections result in lower performance degradation as compared to ResSkip connections.

Summary

N2NSkip connections act as sparse weighted skip connections between sequential layers of the network.

Adding N2NSkip connections to pruned networks provides:

1. Superior Test Performance
2. Faster Convergence
3. Enhanced Overall Connectivity

Deep Learning can greatly benefit from similar explorations in graph theory.

Thank you!!!