Exploring synapse pruning for energy-efficiency in Neural Networks (NN)

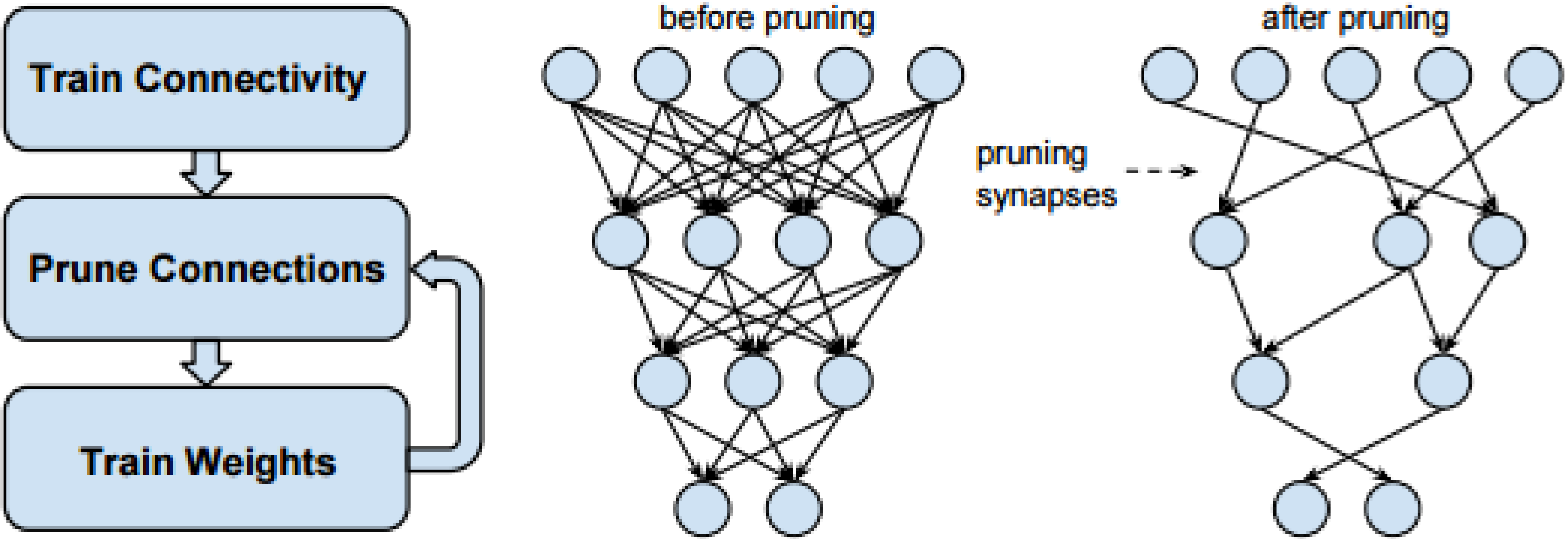
Aayush Ankit, Chankyu Lee, Priyadarshini Panda
ECE, Purdue University

Motivation

- General purpose computing systems involve frequent and high b andwidth data transfers between memory and the computation c ore.
- Typical data flow :- DRAM – SRAM – Computation Core.
- **DRAM fetches extremely energy hungry**. Hence need efficient ways to store data in the memory.

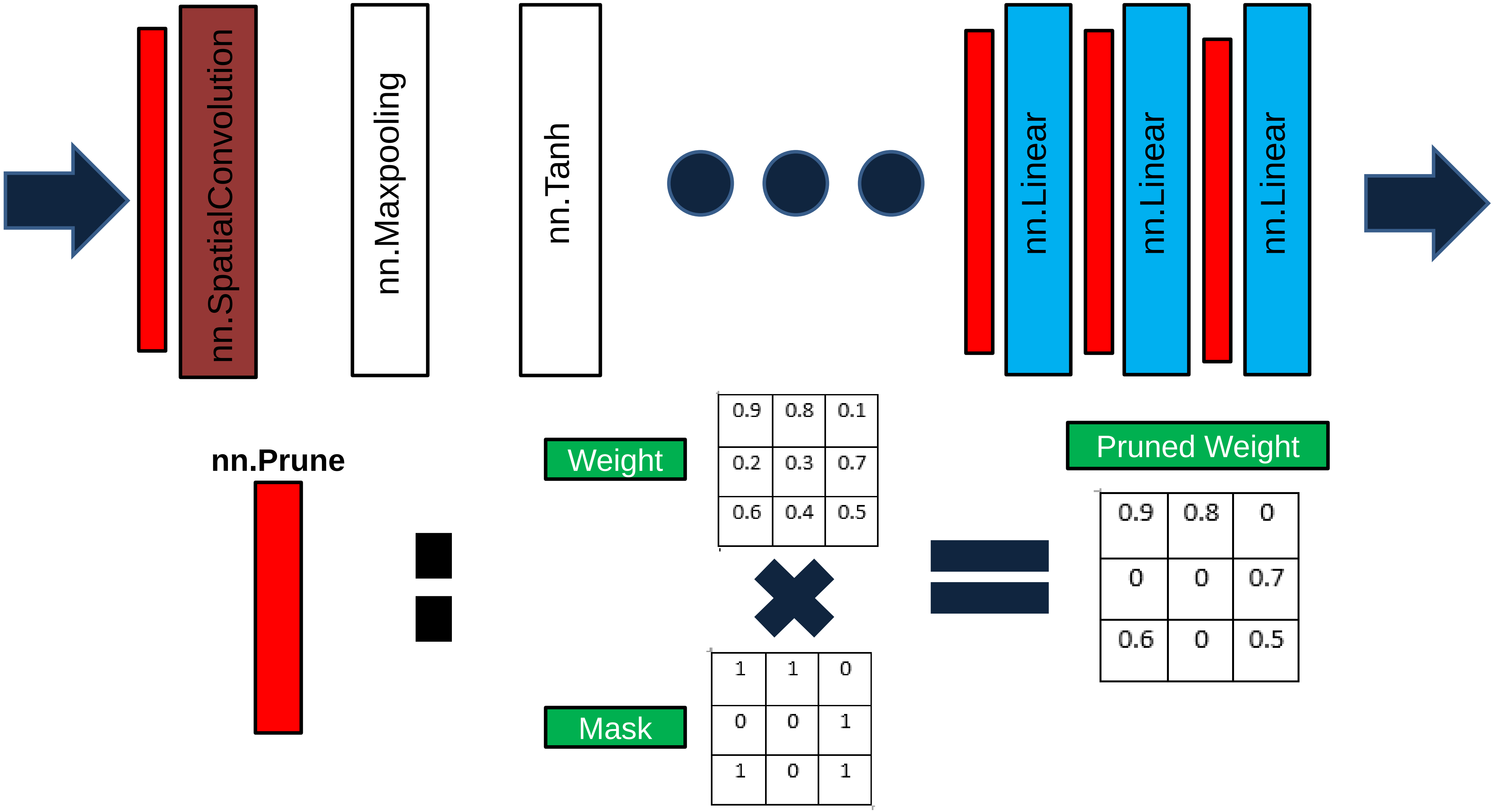
- **Pruning exposes the sparsity in the NN** thereby, providing **ave nues for data compression**.
- Data Compression can be significant driven (low overhead) for e xample – Frequent Pattern/Value compression, other encoding s chemes etc.

What is pruning a network ?

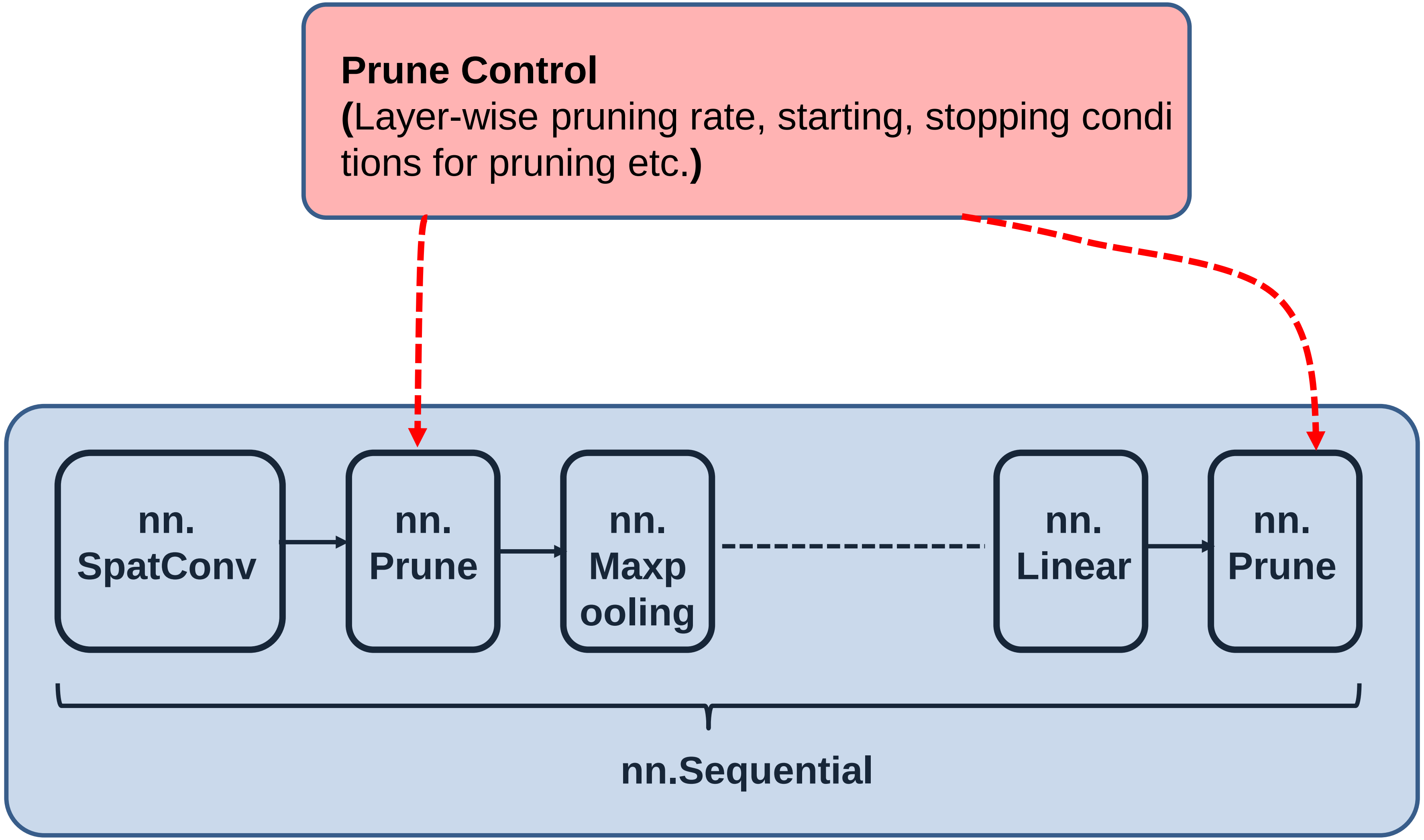


S. Han et. al. “*Learning both Weights and Connections for Efficient Neural Networks*” NIPS 2015

Our Implementation



Our Implementation (Continued)



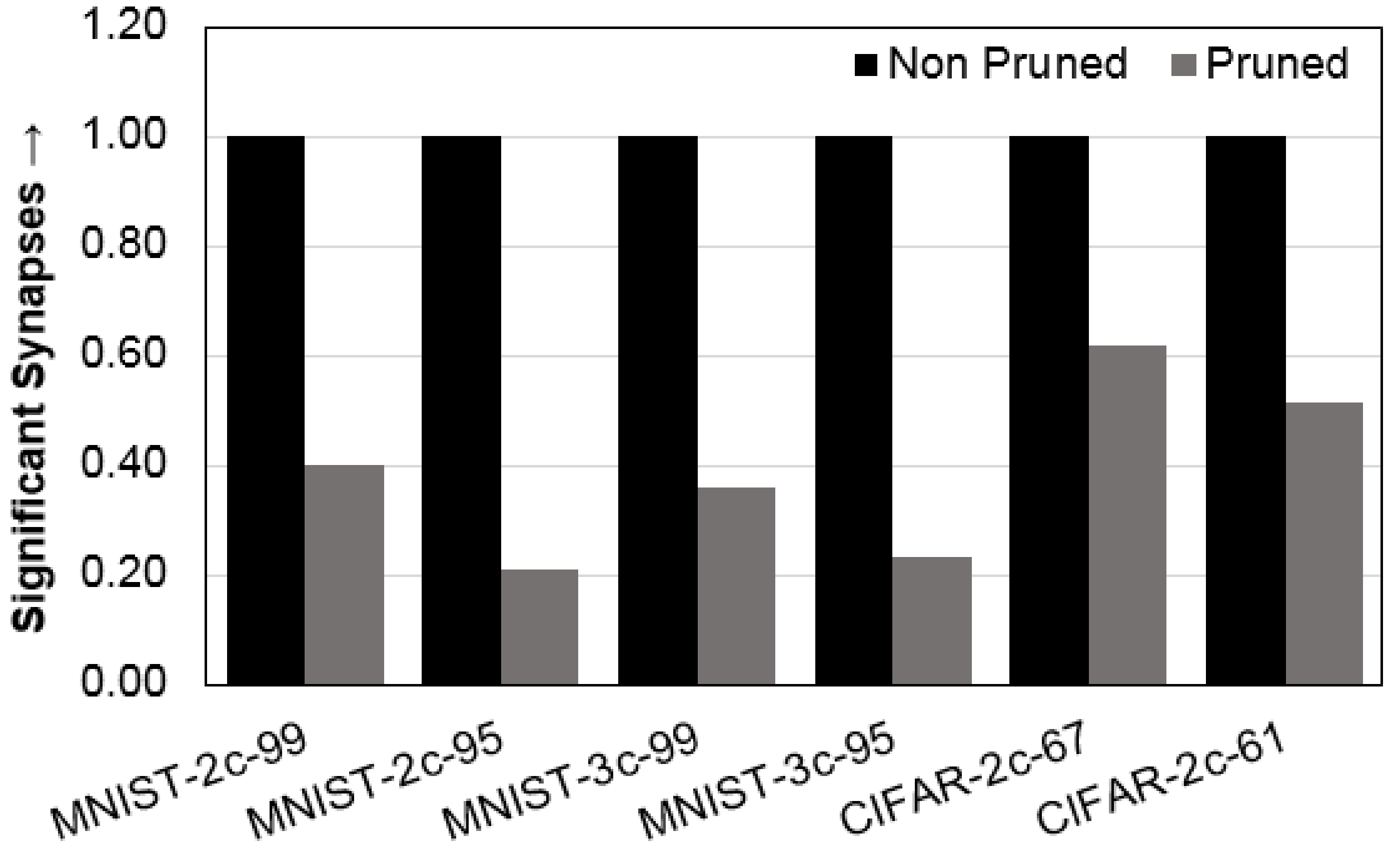
Experimental Methodology

- The “**nn.Prune**” functionality was implemented using masks to modify the “**forward**” and “**backward**” function in **Torch**.

- Simulation benchmarks :-

Dataset	Neural Network (NN) structure
MNIST	28x28 – 6c5 – 16c5 – 120 – 84 – 10
MNIST	28x28 – 6c3 – 16c3 – 36c4 – 240-120 – 84 – 10
CIFAR-10	3x32x32 – 6c5 – 16c5 – 120 – 84 – 10
CIFAR-10	3x32x32 – 6c3 – 16c3 – 36c4 – 240-120 – 84 – 10

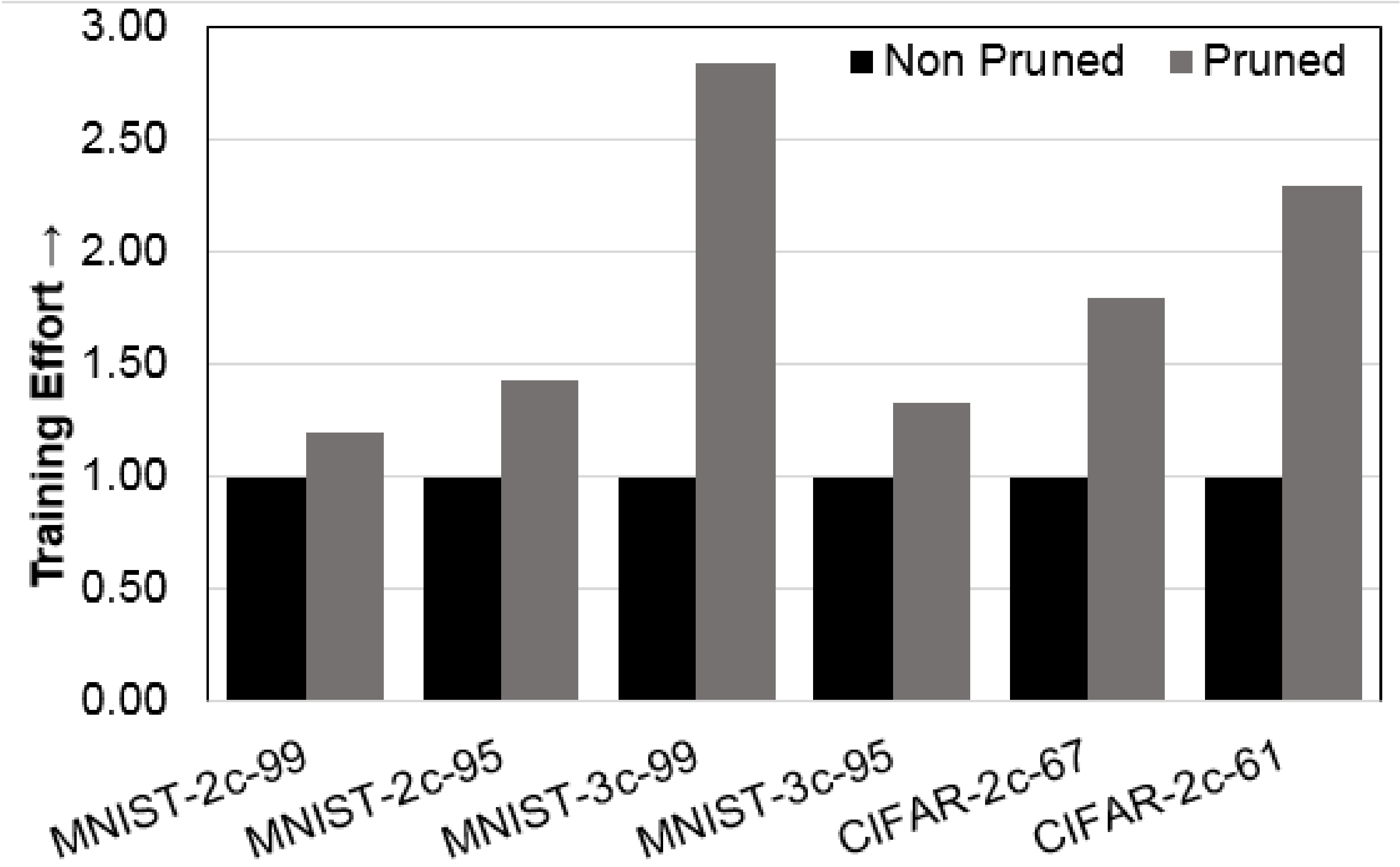
Result #1. Benefits from Pruning



Inferences:-

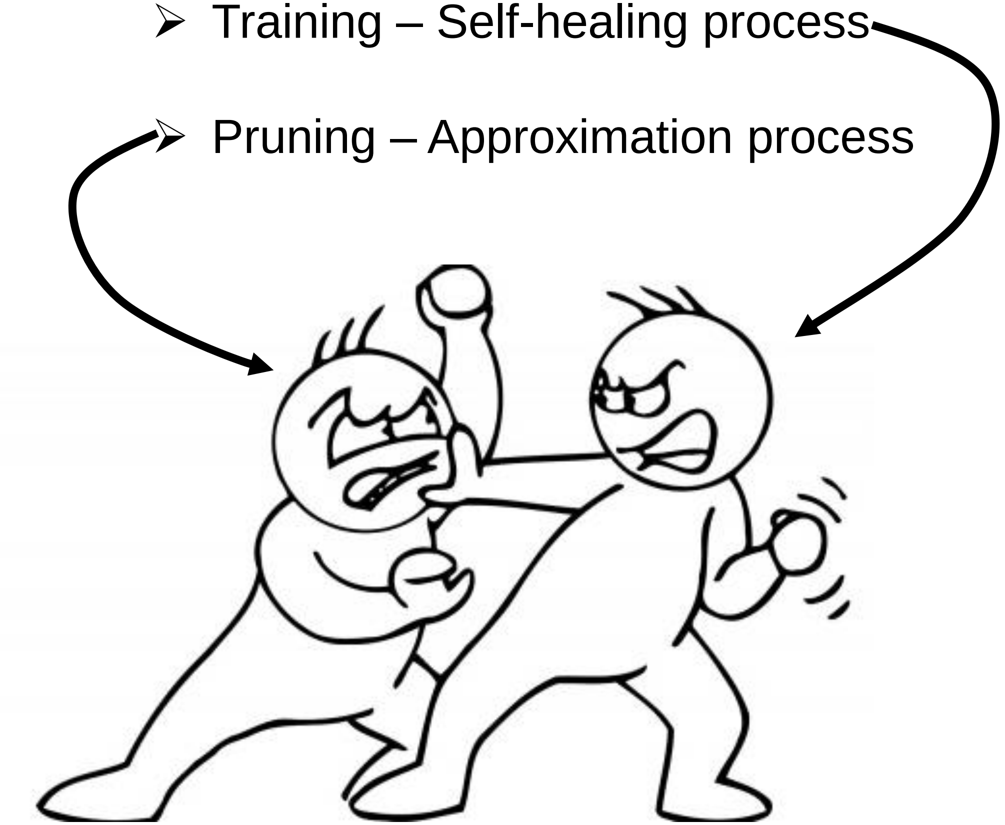
- Pruning can potentially le ad to memory savings fro m 38% to 79% (61% on a verage) across all the be nchmarks.
- Lesser the accuracy req uirement, more is the sc ope of pruning.
- More number of layers (r edundancy), more is the scope of pruning.

Result #2. Training Effort

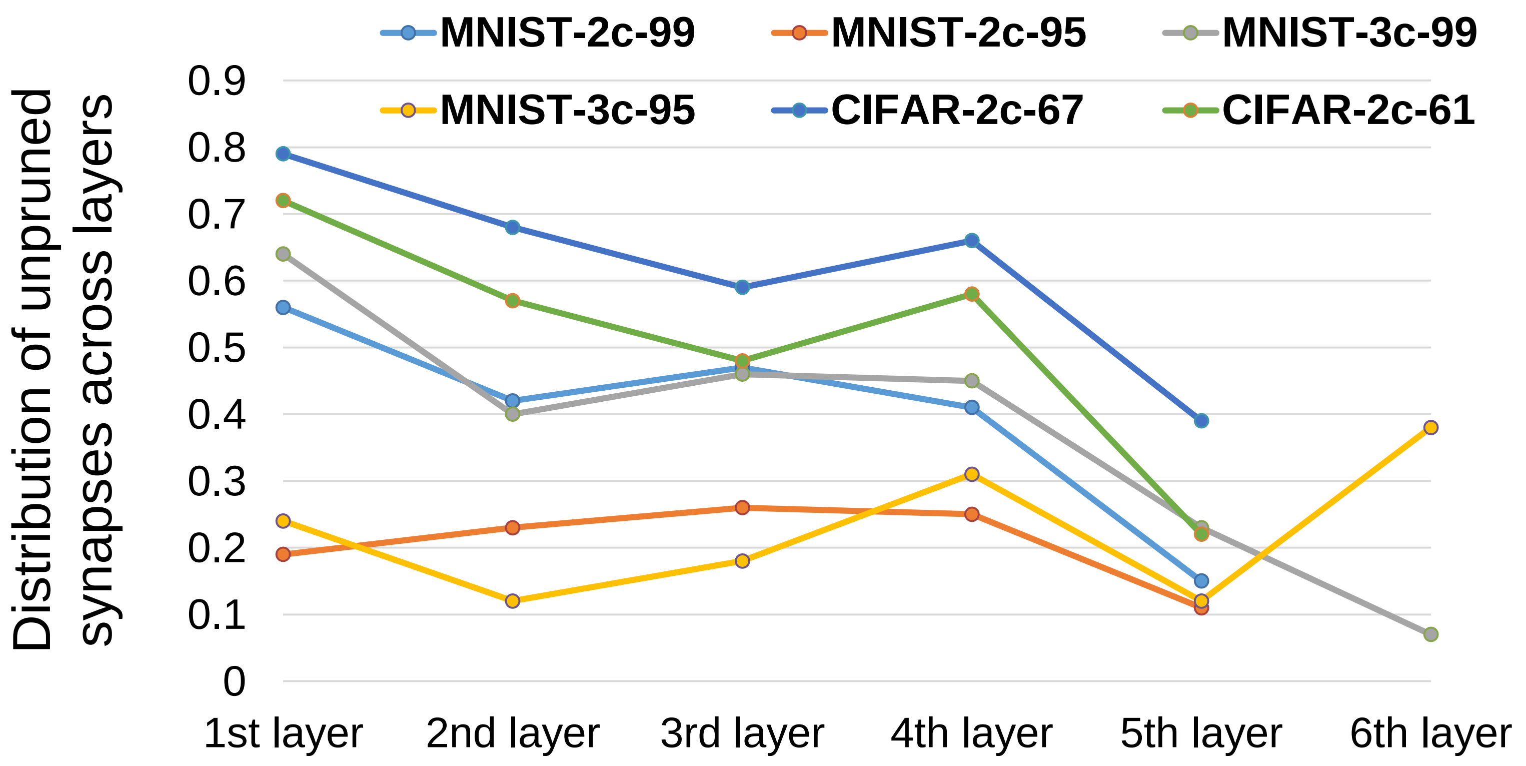


Pruning the neural network demands more training effort(time) to reach iso-accuracy.

- Training – Self-healing process
- Pruning – Approximation process



Result #3. Synapse distribution across layers after training



Key points:-

- This is one of the possible distributions which leads to our result and may not be the golden result.
- We do not infer the signific ance of one layer w.r.t. other layers in the NN fro m this graph as that requir es further rigorous simulati ons.

Conclusions

- Pruning is a simple yet powerful technique to realize application aware flexible NN architectures (connectivity).

- Pruning and Training complement each other to obtain a optimally trained neural network.

- Amount of parameter reduction is a strong function of network str ucture, accuracy and training effort.

Future work

- Develop efficient pruning control techniques based on the trainin g dynamics.

- Analyze pruning on bigger networks.

Questions ?