LG전자 AI Scientist 고급과정

Pruning (Part II)

서울대학교 컴퓨터공학부 이 영 기

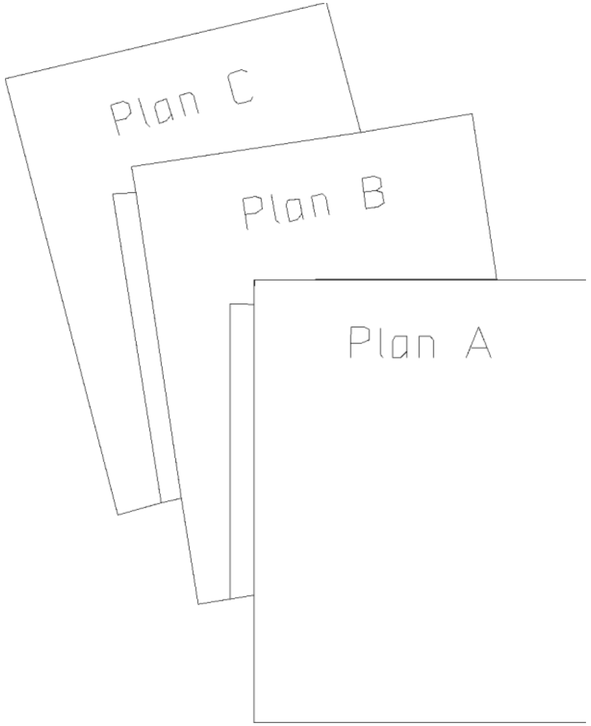
Human-Centered Computer Systems Lab

SEOUL NATIONAL UNIVERSITY

Overview

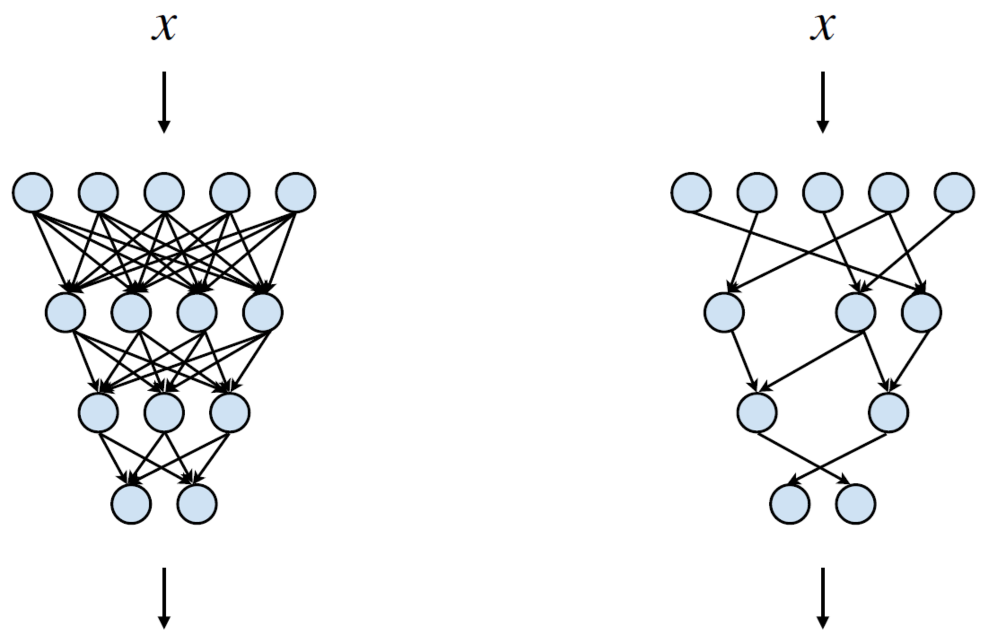
- Objective
  - To discuss how to select pruning ratio and how to fine-tune in NN pruning
  - To understand Lottery ticket Hypothesis

- Content
  - Determine the Pruning Ratio
  - Fine-tune/train Pruned Neural Network



Recap: Pruning Problem Formulation

- In general, we could formulate the pruning as follows:
  $\underset{\mathbf{W}_P}{\arg\min}\, L(\mathbf{x}; \mathbf{W}_P)$ subject to $\|\mathbf{W}_p\|_0 < N$
- L represents the objective function for neural network training.
- X is input, W is original weights, W_p is pruned weights.
- $\|W_p\|_0$ calculates the #nonzeros in W_p, and N is the target #nonzeros .

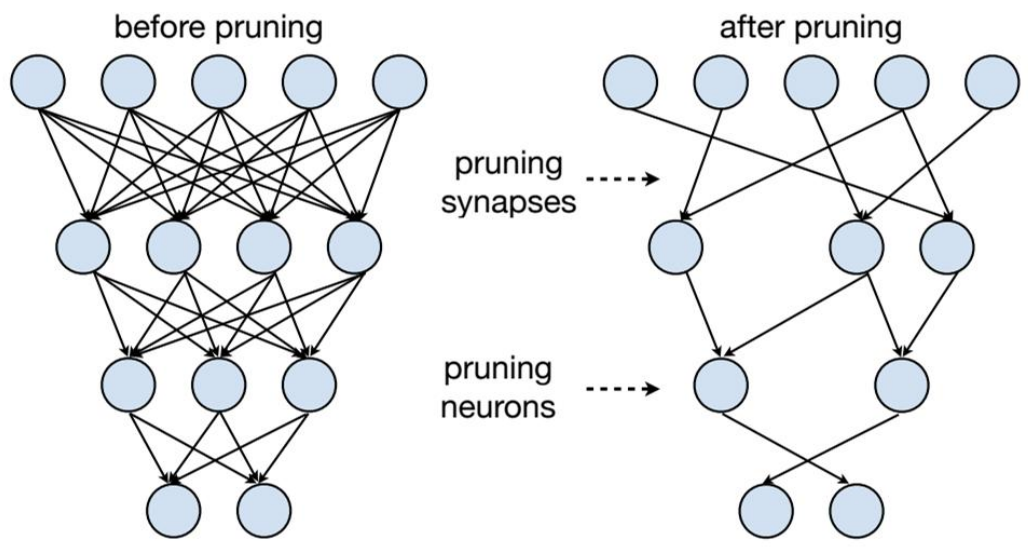


$\underset{\mathbf{W}}{\arg\min}\, L(\mathbf{x}; \mathbf{W})$

$\underset{\mathbf{W}_P}{\arg\min}\, L(\mathbf{x}; \mathbf{W}_P)$
$s.t. \|\mathbf{W}_P\|_0 \leq N$

Neural Network Pruning

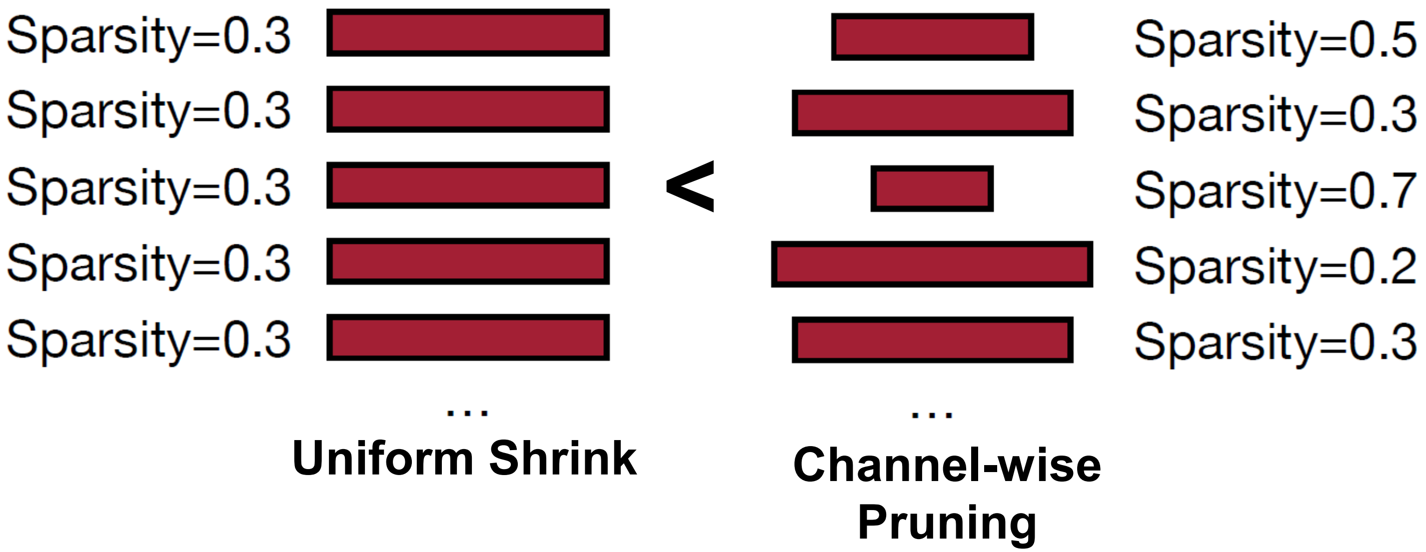
- ~~Introduction to Pruning~~
  - ~~What is pruning?~~
  - ~~How should we formulate pruning?~~
- ~~Determine the Pruning Granularity~~
  - ~~In what pattern should we prune the neural network?~~
- ~~Determine the Pruning Criterion~~
  - ~~What synapses/neurons should we prune?~~
- Determine the Pruning Ratio
  - What should the target sparsity be for each layer?
- Fine-tune/Train Pruned Neural Network
  - How should we improve the performance of pruned models?

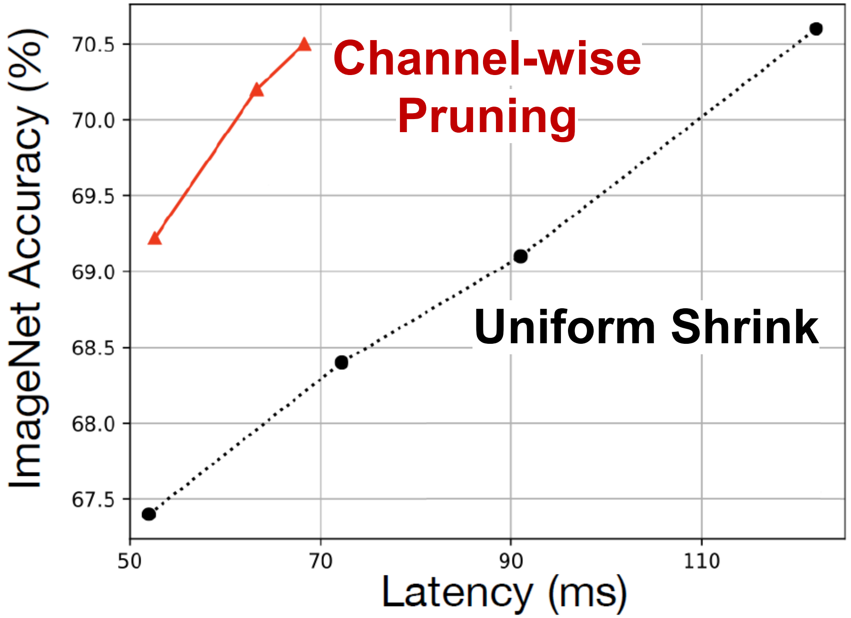


Prune 30%?
Prune 50%?
Prune 70%?

How Should We Select Ratios For Each Layer?

- Non-uniform pruning is better than uniform shrinking





Analyze the sensitivity of each layer

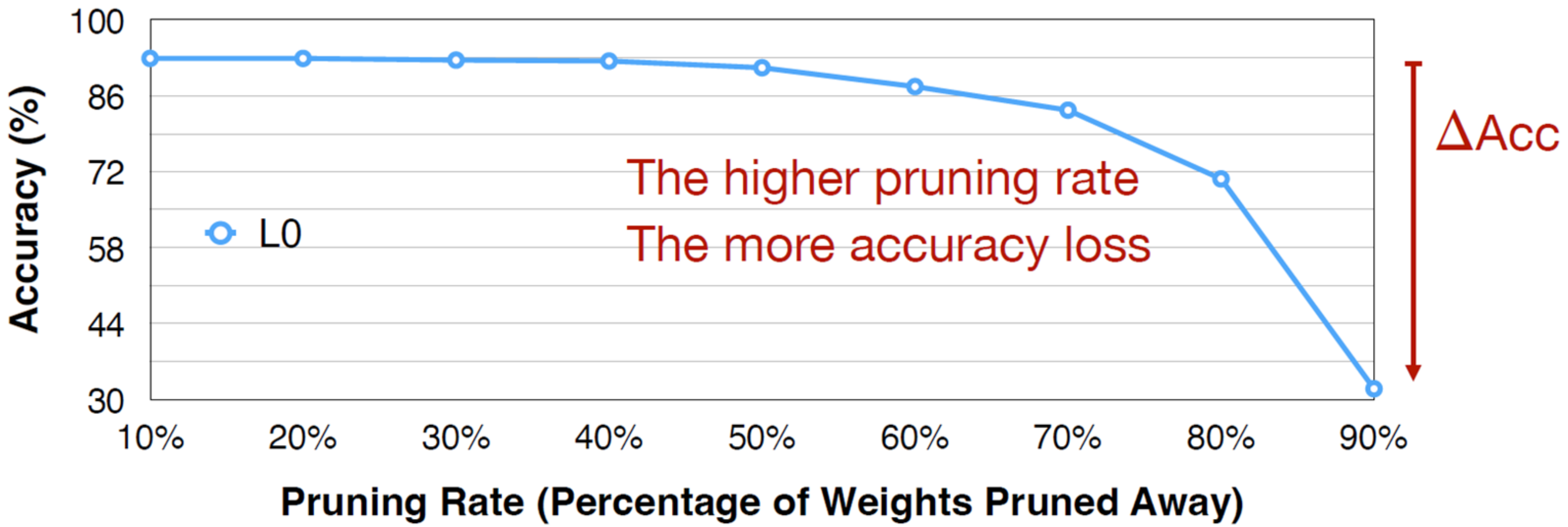
Analyze the Sensitivity of Each Layer

- We need different pruning ratios for each layer since each layer has different sensitivity to pruning
  - Some layers are more sensitive (e.g., the first layer)

- We can perform **sensitivity analysis** to determine the per-layer pruning ratio

Analyze the Sensitivity of Each Layer

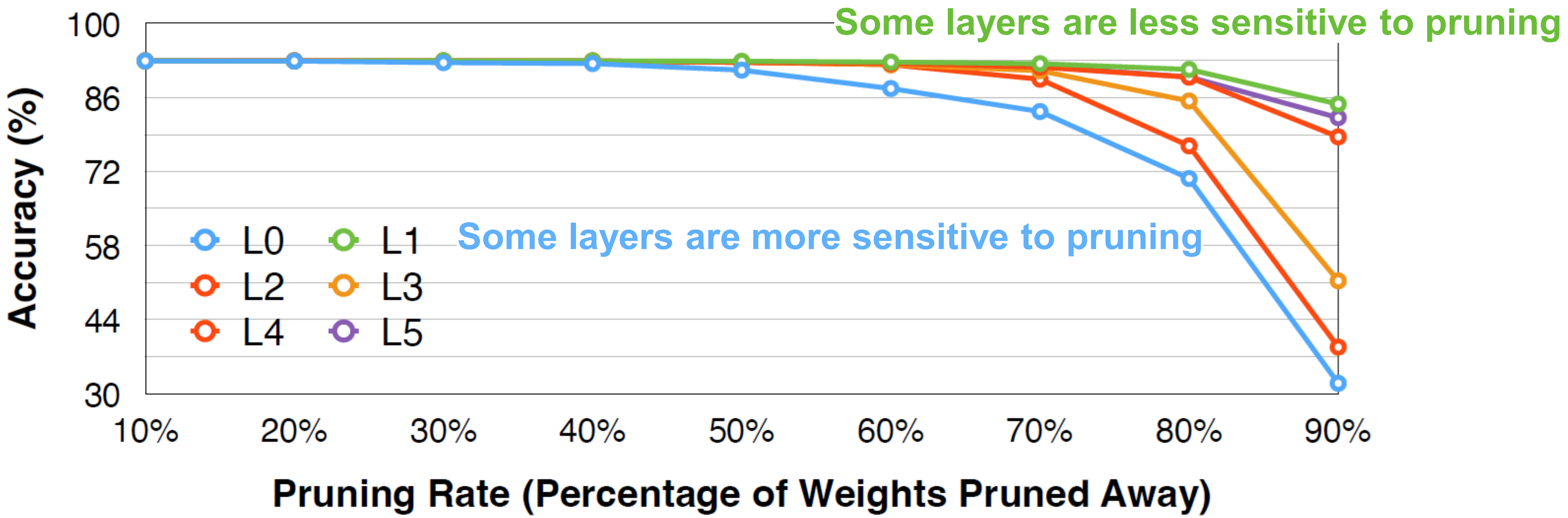
- We can perform **sensitivity analysis** to determine the per-layer pruning ratio
  - Pick a layer L_i in the model
    - Prune the layer L_i with pruning ratio $r \in \{0, 0.1, 0.2, ..., 0.9\}$ (or other strides)
    - Observe the accuracy degradation, ΔAcc_r^i for each pruning ratio



* Analysis on VGG-11 on CIFAR-10 dataset

Analyze the Sensitivity of Each Layer

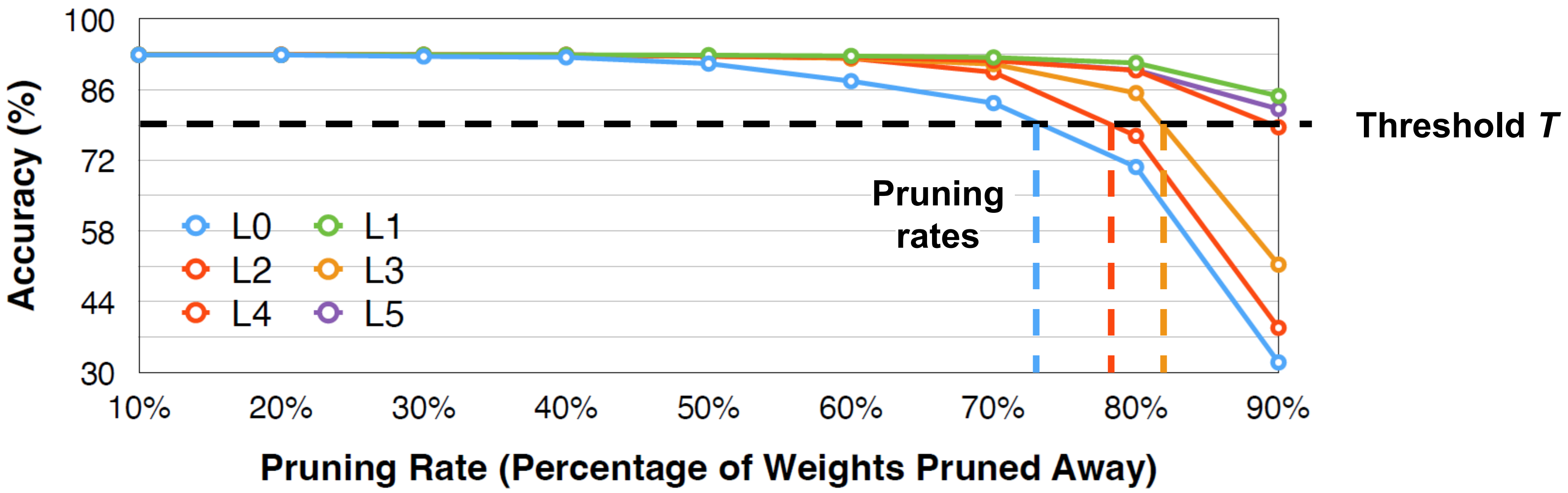
- We can perform **sensitivity analysis** to determine the per-layer pruning ratio
  - Pick a layer L_i in the model
    - Prune the layer L_i with pruning ratio $r \in \{0, 0.1, 0.2, \ldots, 0.9\}$ (or other strides)
    - Observe the accuracy degradation, ΔAcc_r^i for each pruning ratio
  - Repeat the process for all layers



* Analysis on VGG-11 on CIFAR-10 dataset

Analyze the Sensitivity of Each Layer

- We can perform **sensitivity analysis** to determine the per-layer pruning ratio
  - Pick a layer L_i in the model
    - Prune the layer L_i with pruning ratio $r \in \{0, 0.1, 0.2, \ldots, 0.9\}$ (or other strides)
    - Observe the accuracy degradation, ΔAcc_r^i for each pruning ratio
  - Repeat the process for all layers
  - Pick a degradation threshold T such that the overall pruning rate is desired



* Analysis on VGG-11 on CIFAR-10 dataset

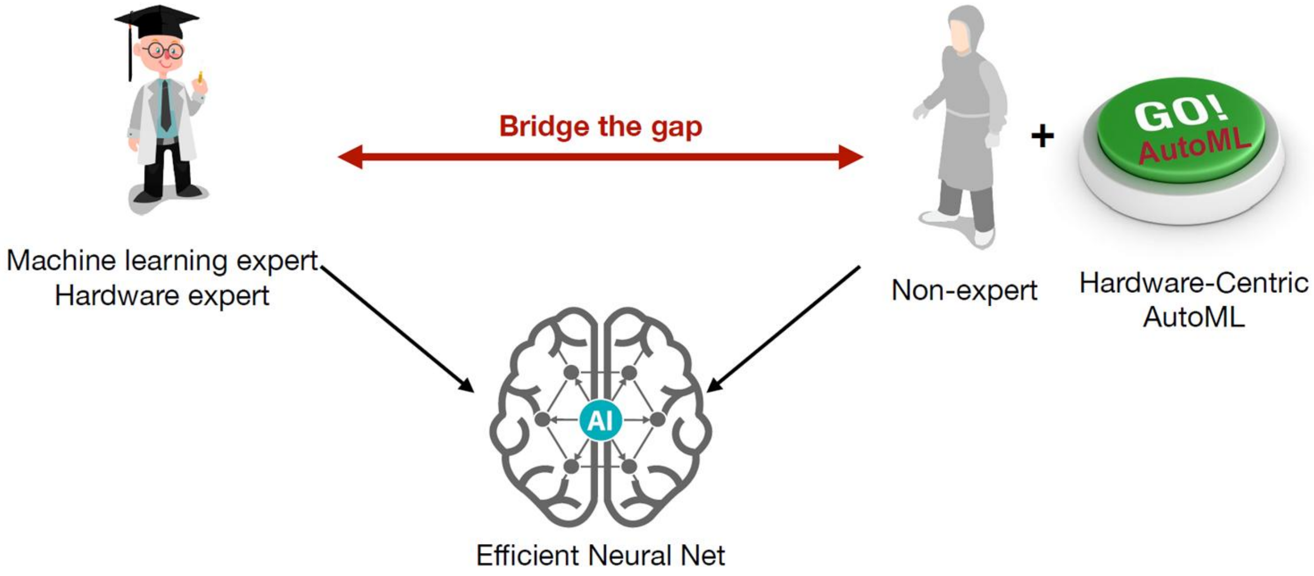
Is this Optimal?

- Maybe not. We do not consider the interaction between layers.
- Can we go beyond the heuristics? Yes, automatic pruning!

Automatic Pruning

- Given an overall compression ratio, how do we select per-layer pruning ratios?
  - Sensitivity analysis ignores the interaction between layers sub-optimal solution

- Conventionally, such process relies on human expertise with trials and error
  - Can we d



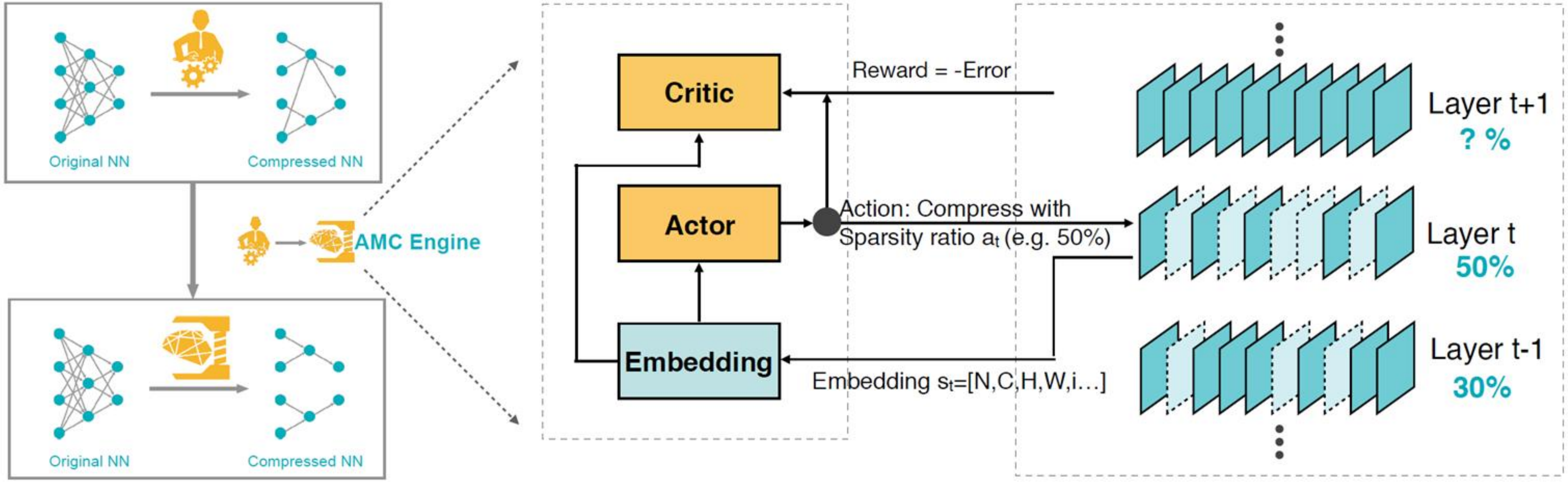
He *et al.*, "AMC: AutoML for Model Compression and Acceleration on Mobile Devices", ECCV 2018

AMC: AutoML for Model Compression

- Pruning as a reinforcement learning problem

Model Compression by Human:
Labor Consuming, Sub-optimal

e.g.) channel pruning

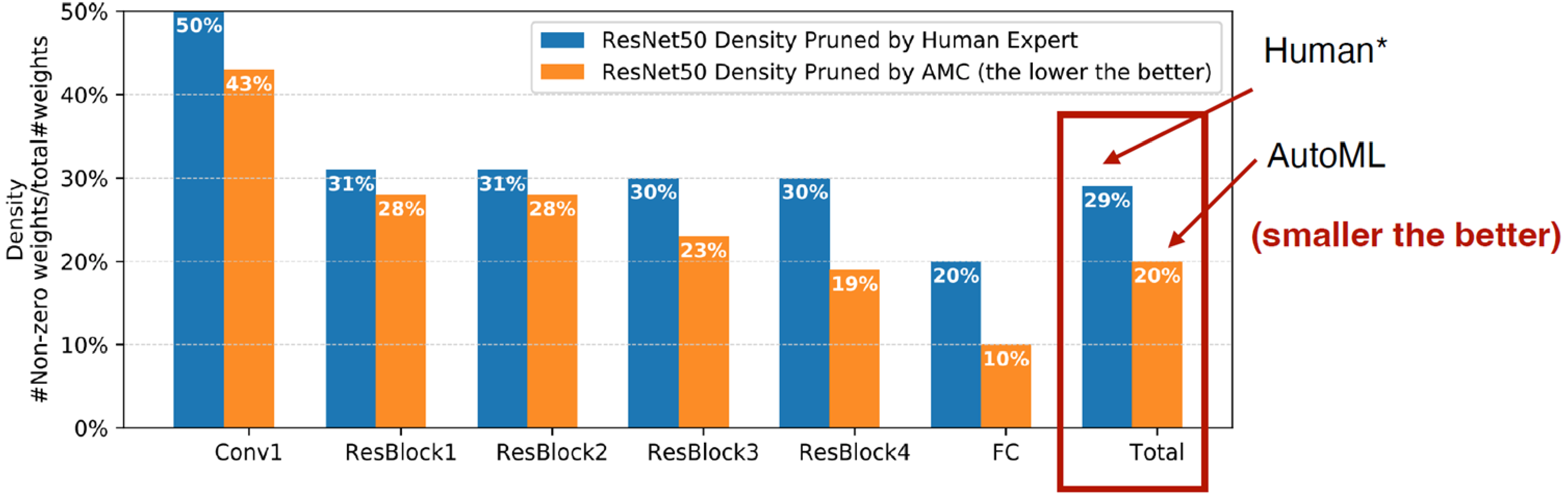


Model Compression by AI:
Automated, High Compression Rate, Faster

He *et al.*, "AMC: AutoML for Model Compression and Acceleration on Mobile Devices", ECCV 2018

AMC: AutoML for Model Compression

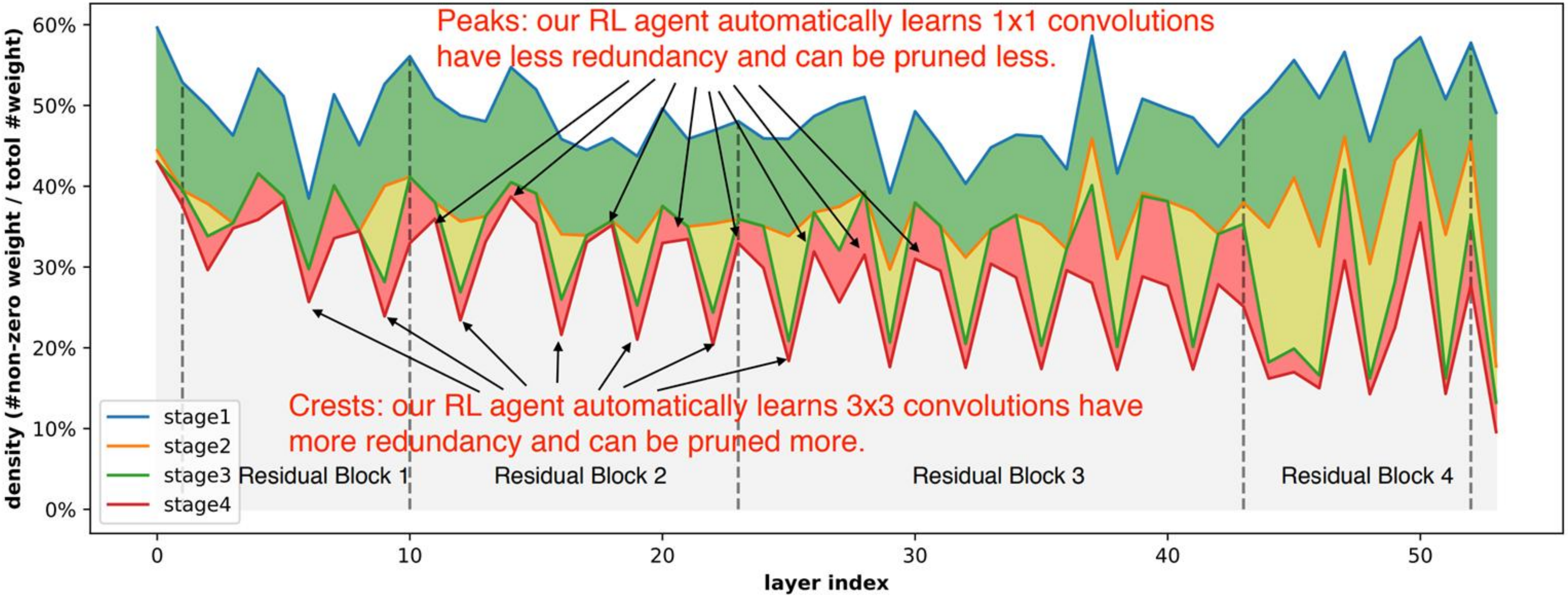
- AMC shows higher compression rate than human expert (or heuristic method)



He *et al.*, "AMC: AutoML for Model Compression and Acceleration on Mobile Devices", ECCV 2018

AMC: AutoML for Model Compression

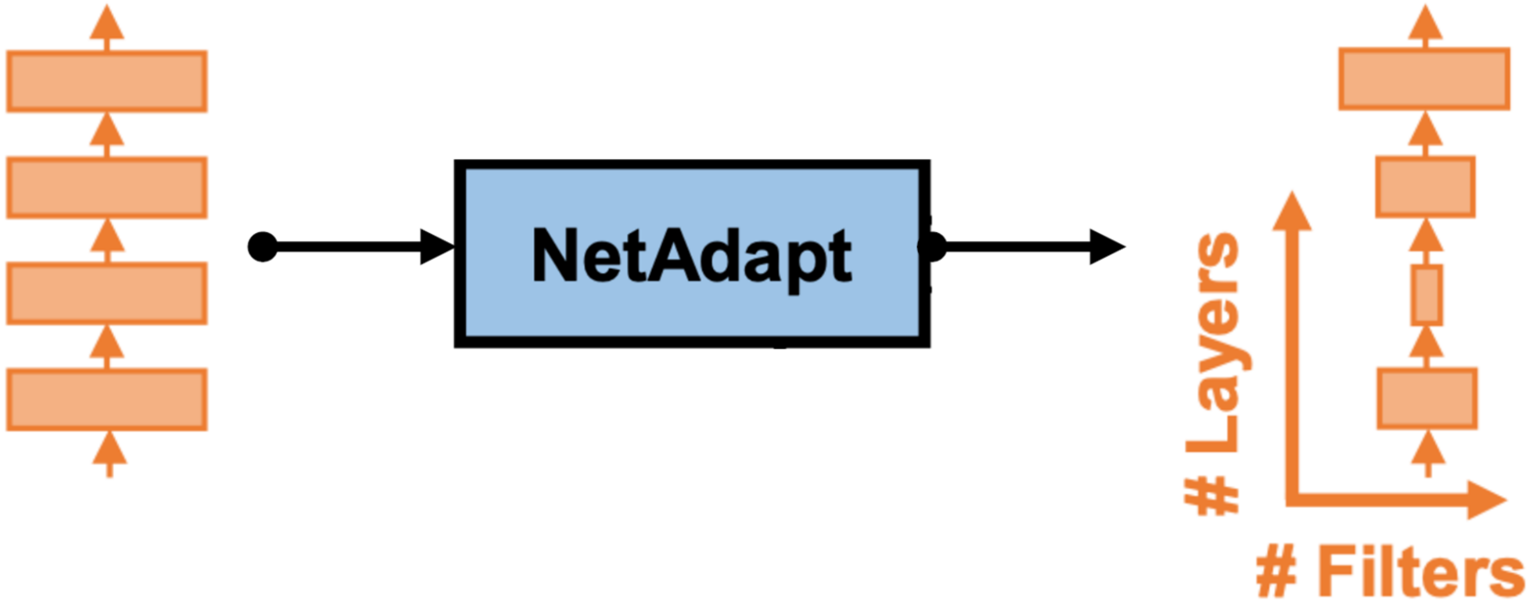
- The reinforcement learning agent automatically learns that 3x3 convolution has more redundancy than 1x1 convolution and can be pruned more



He *et al.*, "AMC: AutoML for Model Compression and Acceleration on Mobile Devices", ECCV 2018

NetAdapt

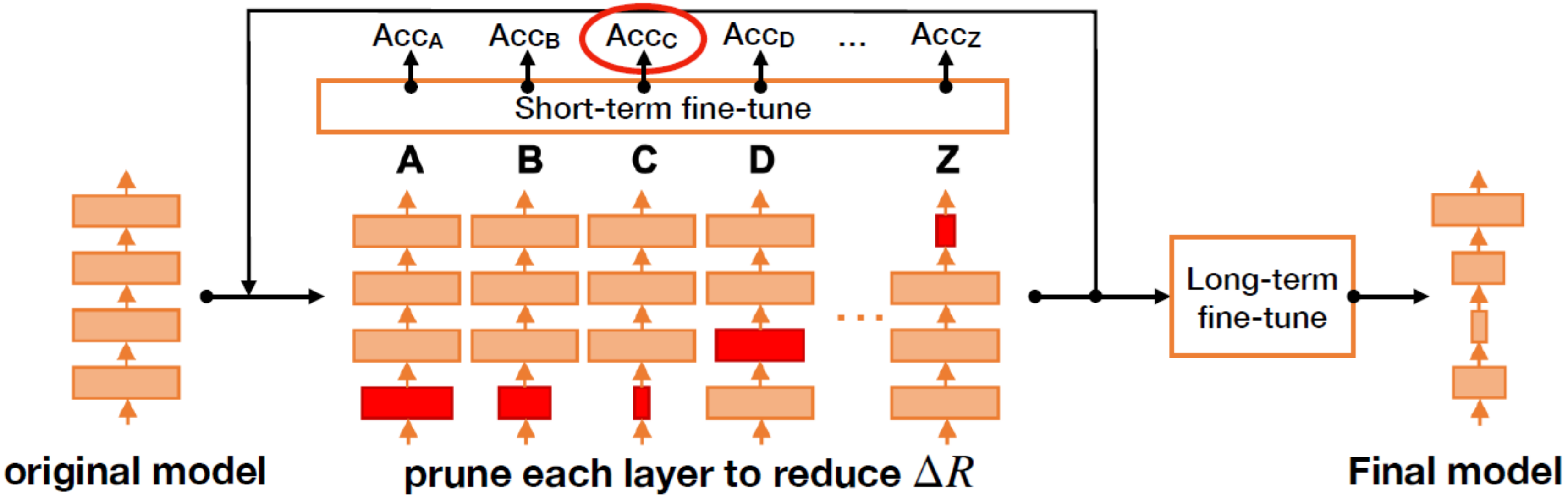
- The goal of NetAdapt is to find a per-layer pruning ratio to meet a global resource constraint (e.g., latency, energy, …)
- The process is done iteratively
- We take **latency** constraint as an example



Yang et al., "NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications", ECCV 2018

NetAdapt

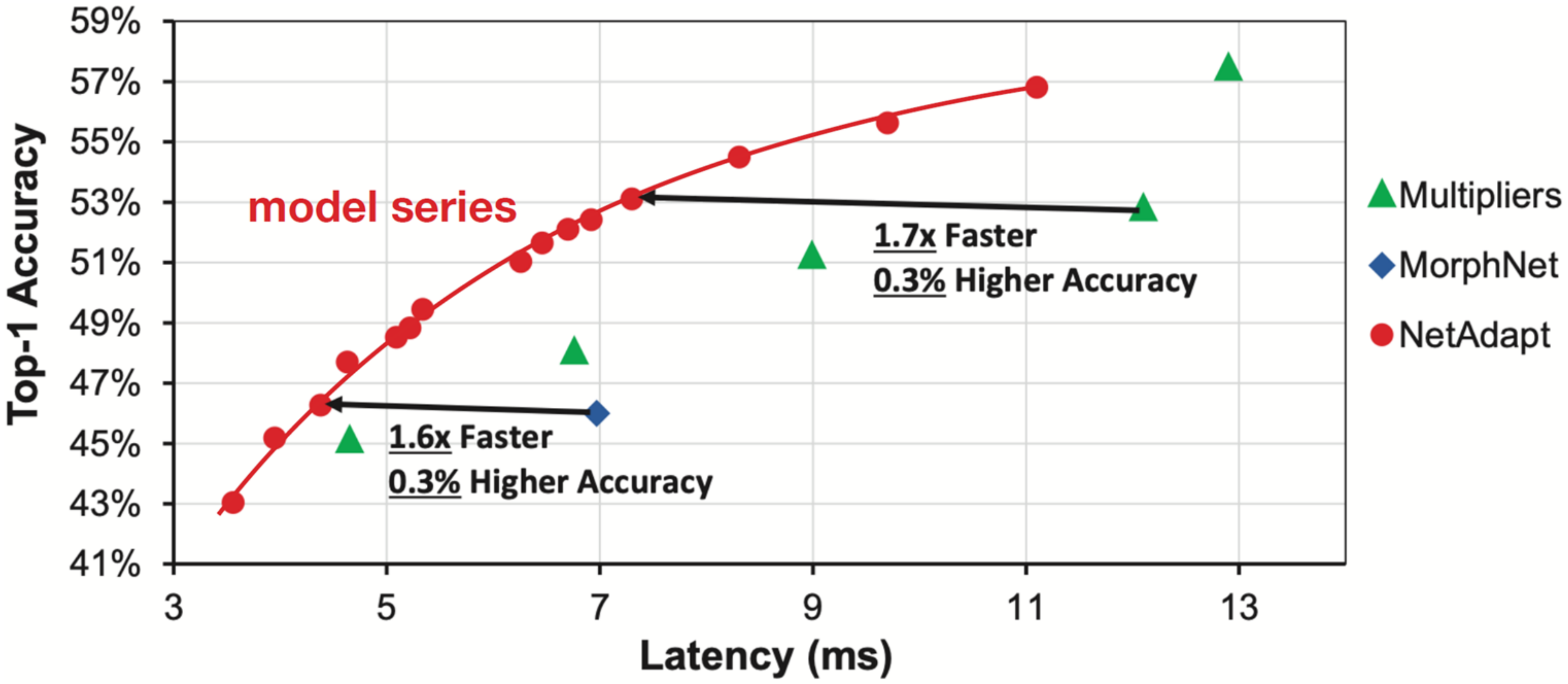
- For each iteration, we aim to reduce the latency by a certain amount ΔR (manually defined)
  - For each layer L_k (k in A-Z in the figure)
    - Prune the layer s.t. the latency reduction meets ΔR (based on a pre-built lookup table)
    - Short-term fine-tune model (10k iterations); measure accuracy after fine-tuning
  - Choose and prune the layer with the highest accuracy
- Repeat until the total latency reduction satisfies the constraint
- Long-term fine-tune to recover accuracy



Yang et al., "NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications", ECCV 2018

NetAdapt

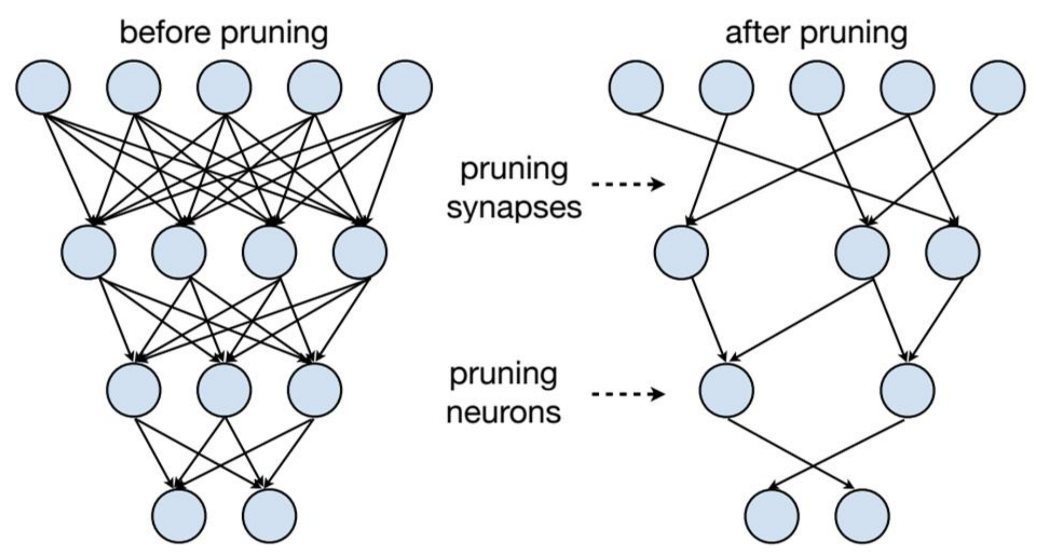
- The iterative nature allows us to obtain a serial of models with different costs
  - #models = #iterations



Yang et al., "NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications", ECCV 2018

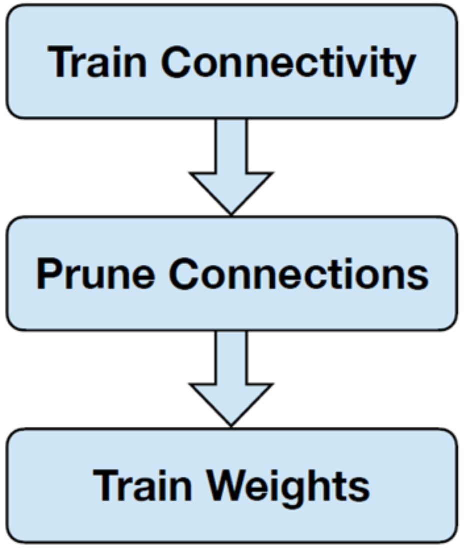
Neural Network Pruning

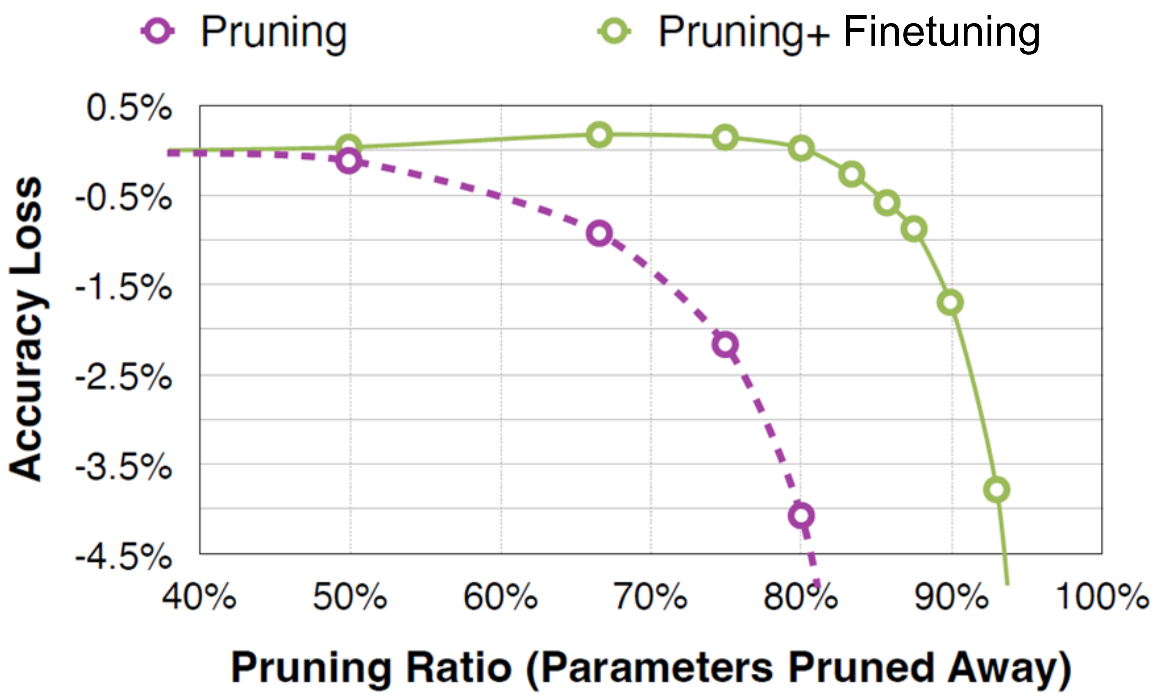
- ~~Introduction to Pruning~~
  - ~~What is pruning?~~
  - ~~How should we formulate pruning?~~
- ~~Determine the Pruning Granularity~~
  - ~~In what pattern should we prune the neural network?~~
- ~~Determine the Pruning Criterion~~
  - ~~What synapses/neurons should we prune?~~
- ~~Determine the Pruning Ratio~~
  - ~~What should the target sparsity be for each layer?~~
- Fine-tune/Train Pruned Neural Network
  - How should we improve the performance of pruned models?



Fine-tuning Pruned Neural Networks

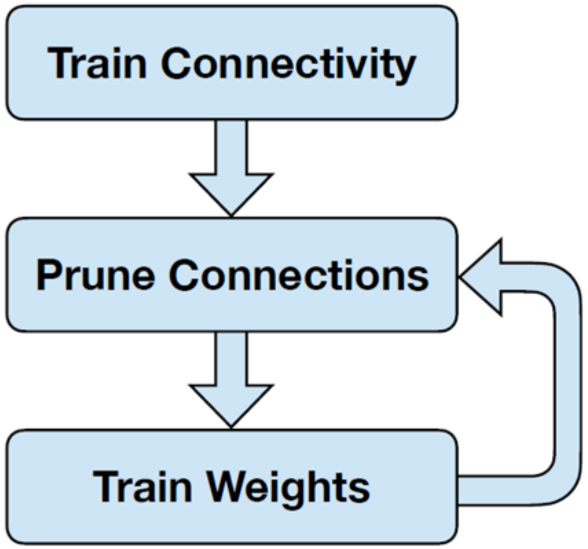
- The model performance may decrease after pruning.
- Fine-tuning the pruned neural network will help recover the accuracy and enables a higher pruning ratio.

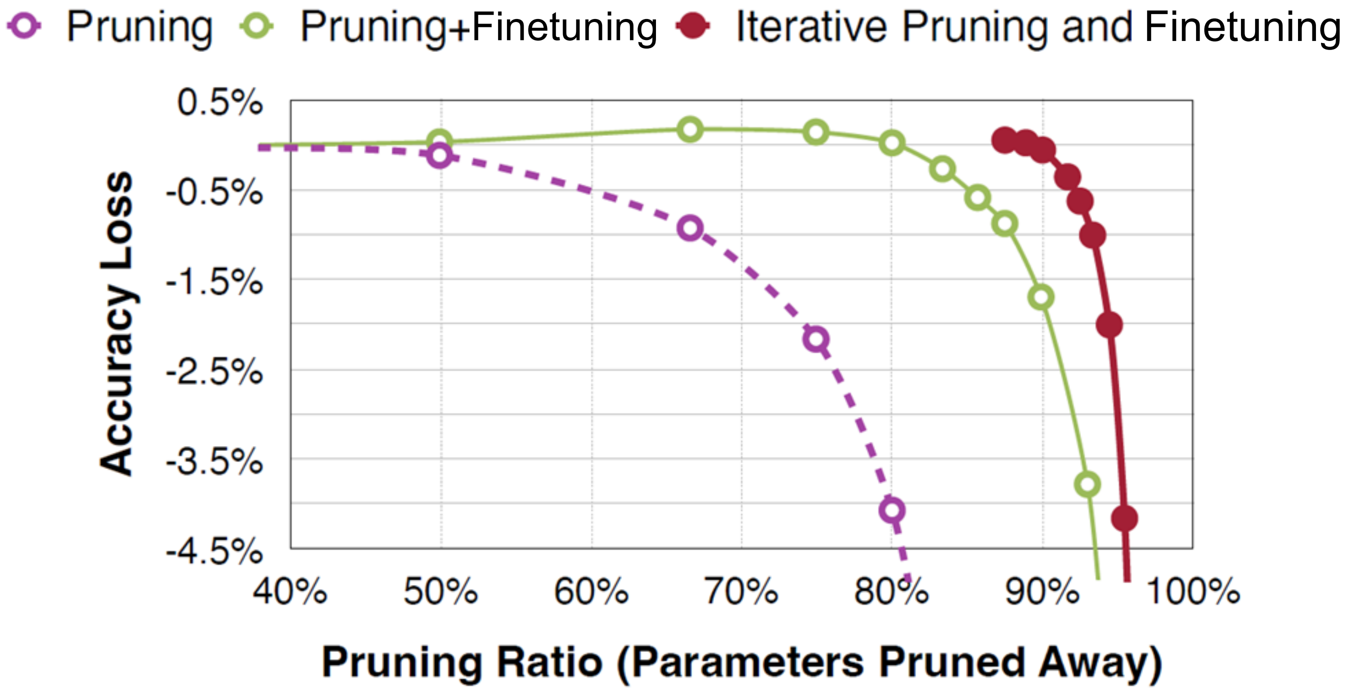




Iterative Pruning

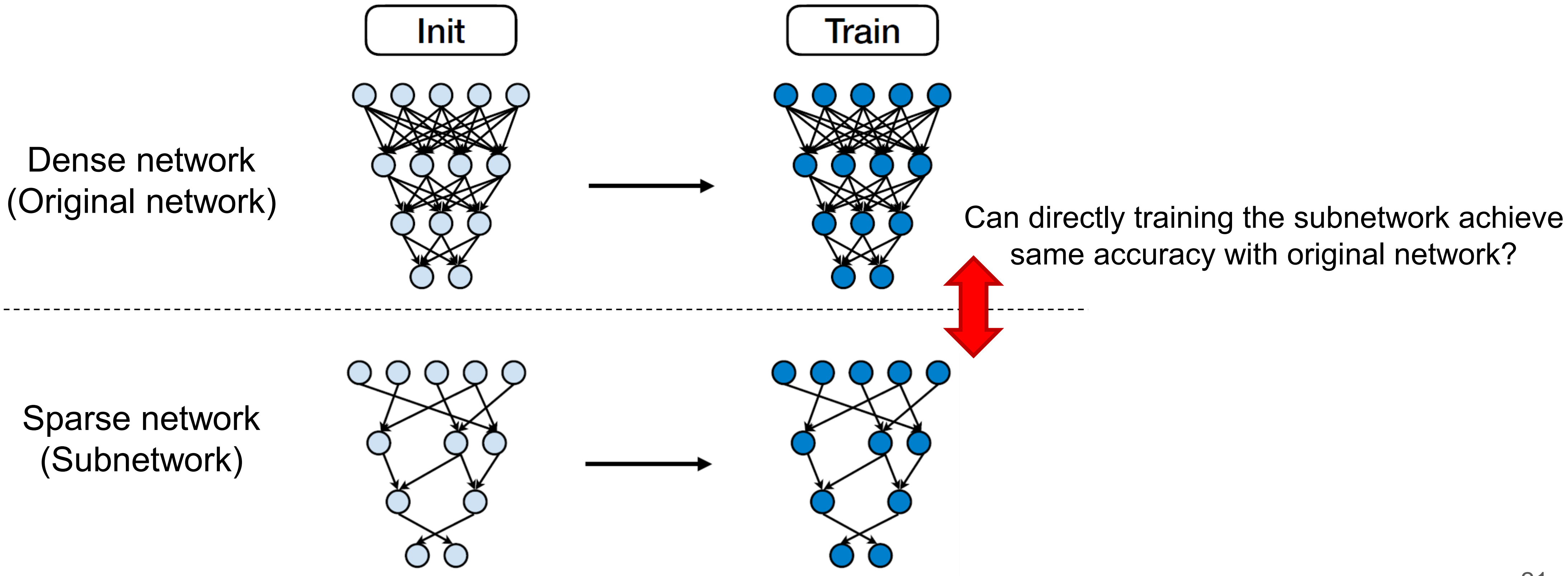
- Consider pruning followed by fine-tuning is one iteration.
- Iterative pruning gradually increases the target sparsity in each iteration.
  - boost pruning ratio from 5x to 9x on AlexNet compared to single-step aggressive pruning.





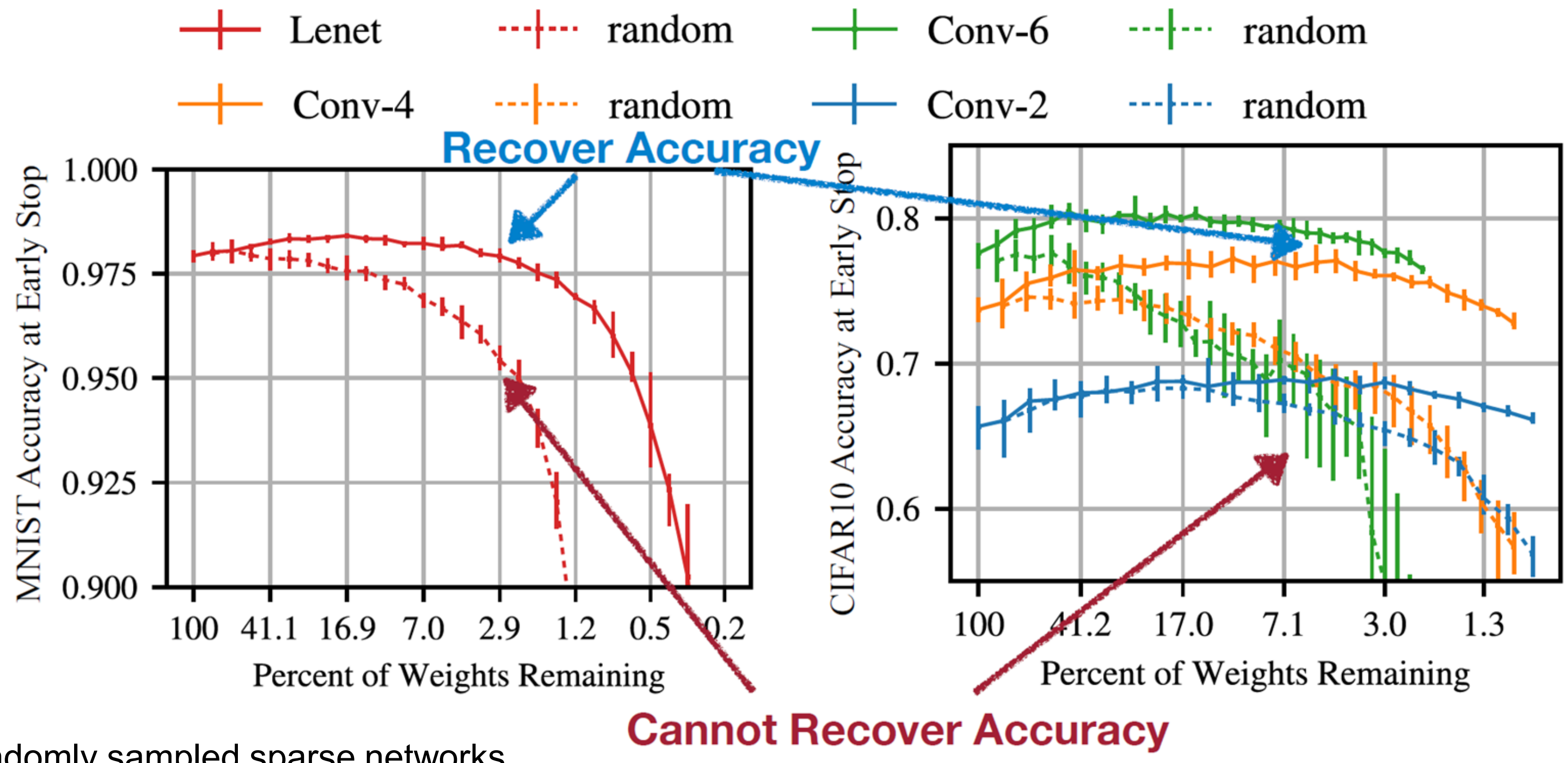
Can we Train a Sparse Neural Network from Scratch?

- Neural network pruning shows that a neural network can be reduced in size.
- Question: Can we directly train this sparse network from scratch?



Can we Train a Sparse Neural Network from Scratch?

- Experience tells us that the architectures uncovered by pruning are harder to train from the start reaching lower accuracy than the original networks.

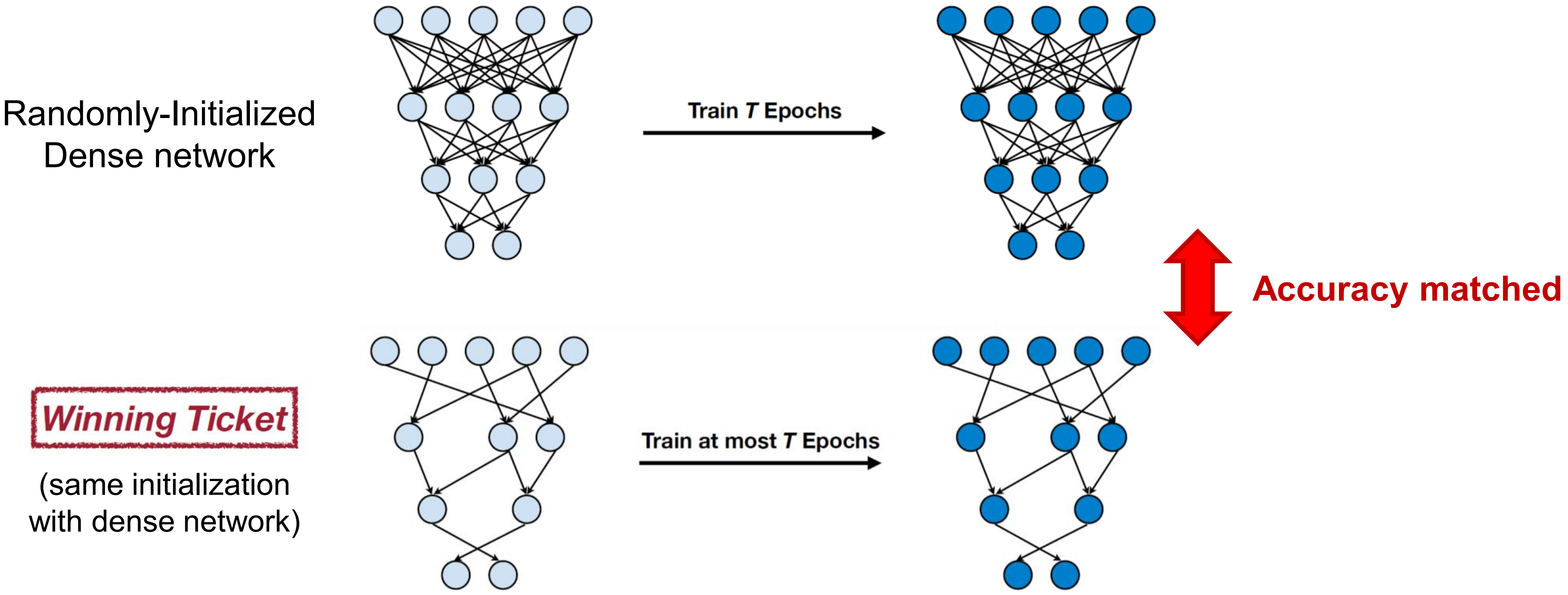


Dashed Lines : randomly sampled sparse networks
Solid Lines : correct sparsity mask found by training

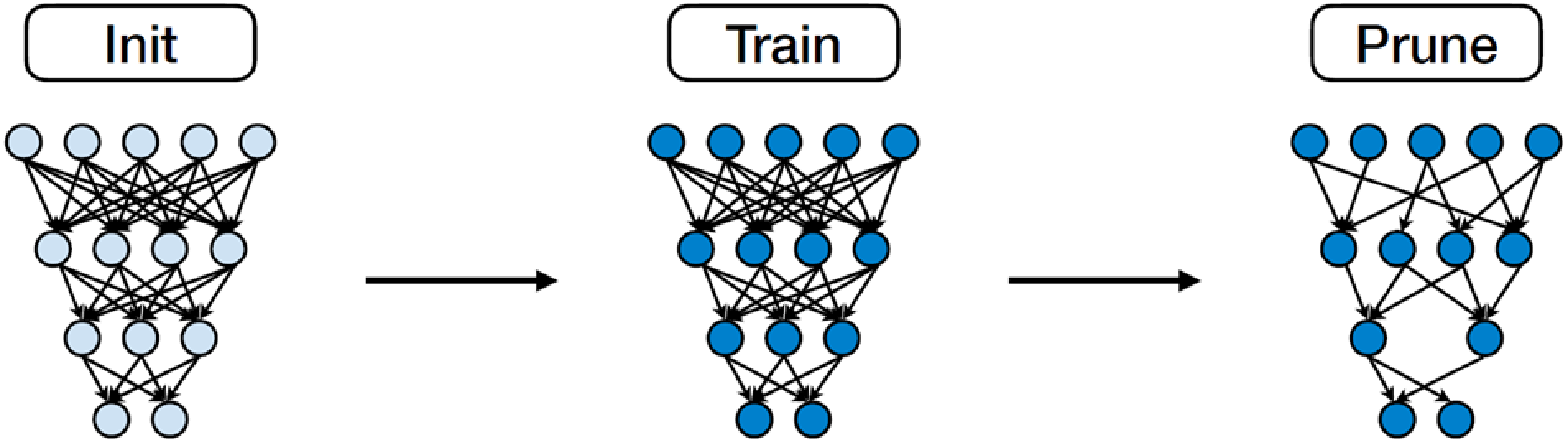
Frankle *et al.*, “The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks”, ICLR 2019

Lottery Ticket Hypothesis

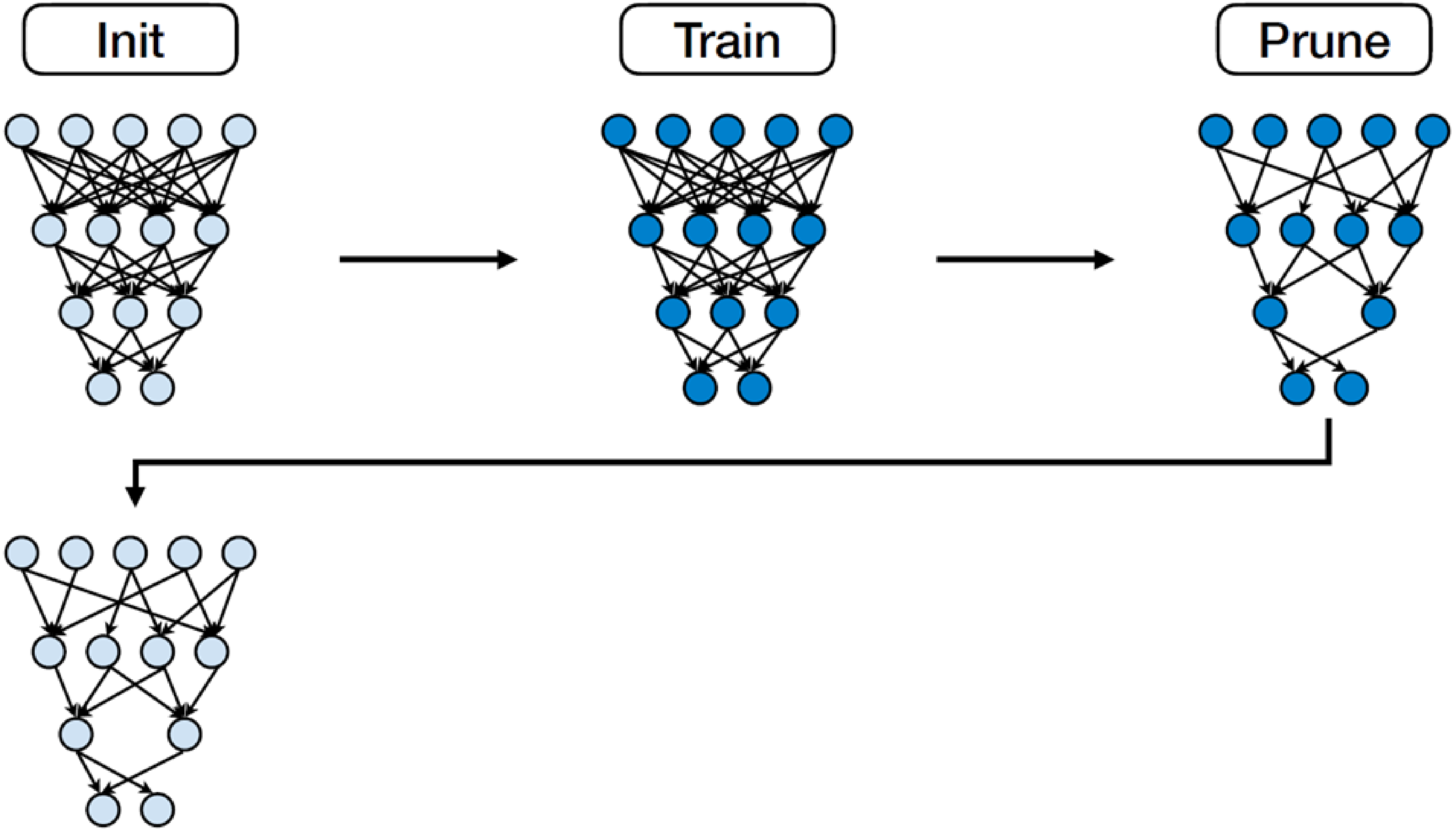
"A randomly initialized, dense neural network contains **a subnetwork** that is initialized such that—when **trained in isolation**—it can **match the test accuracy** of the original network after training for **at most the same number of iterations**."



Iterative Magnitude Pruning

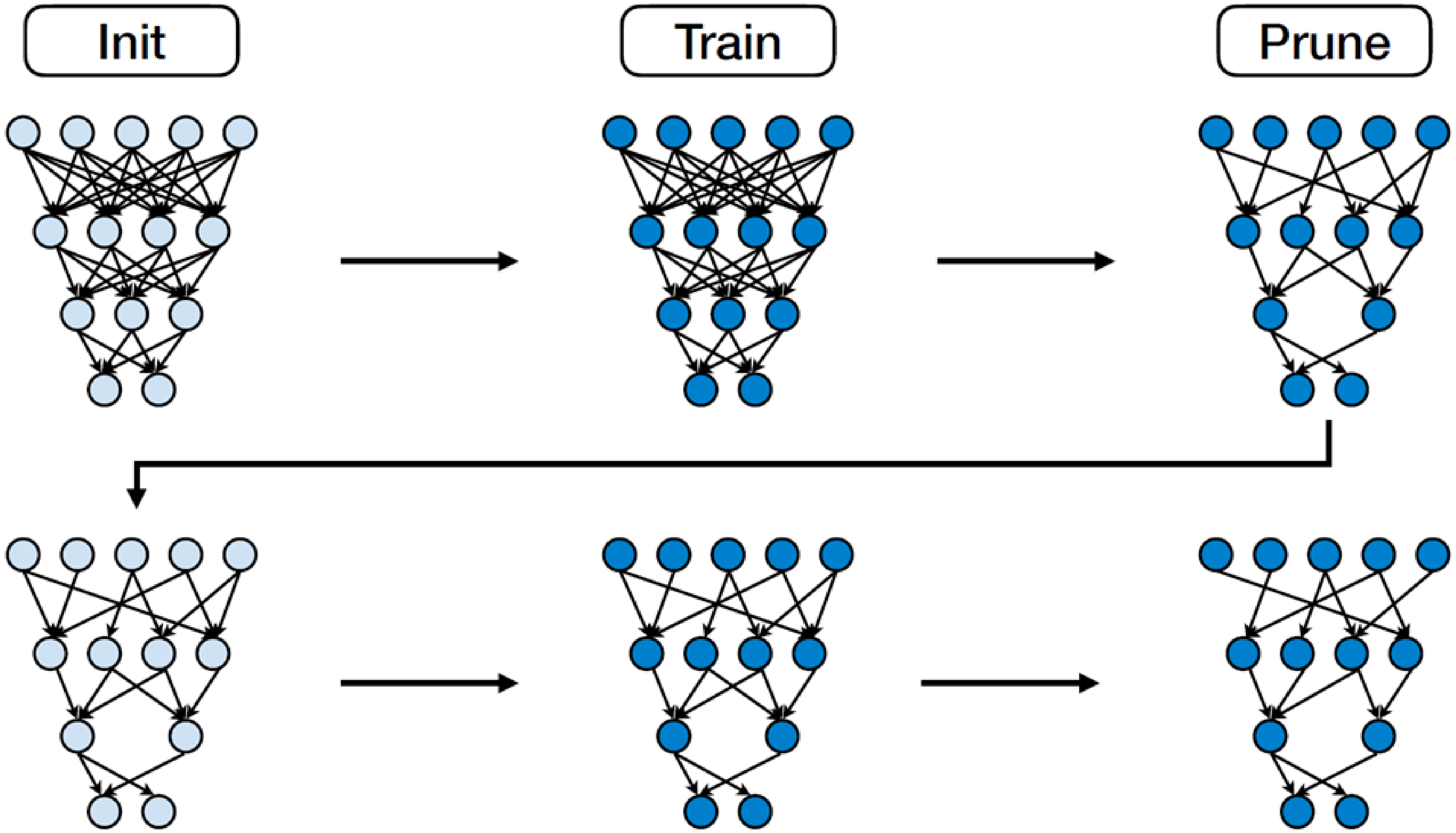


Iterative Magnitude Pruning

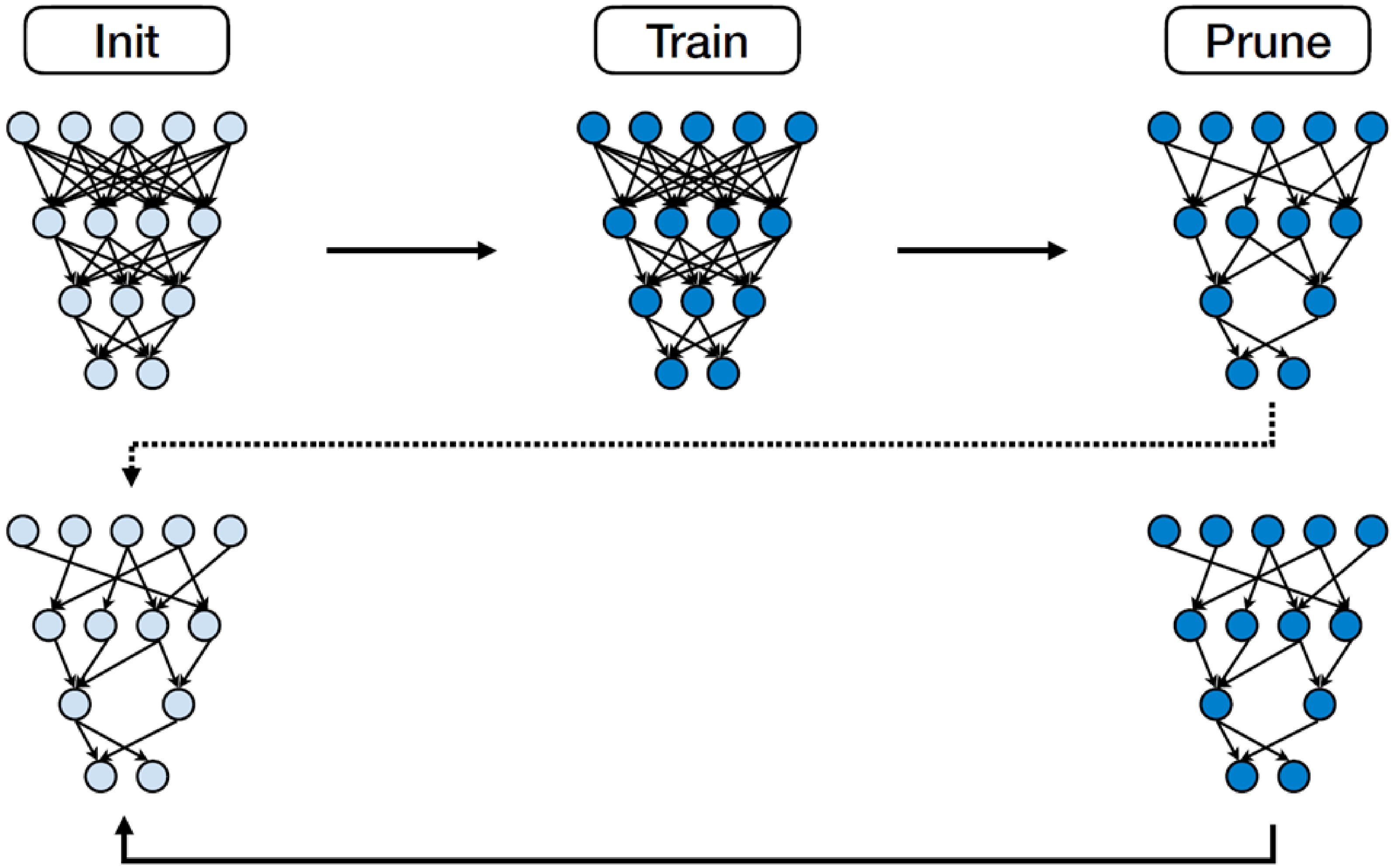


Iterative Magnitude Pruning

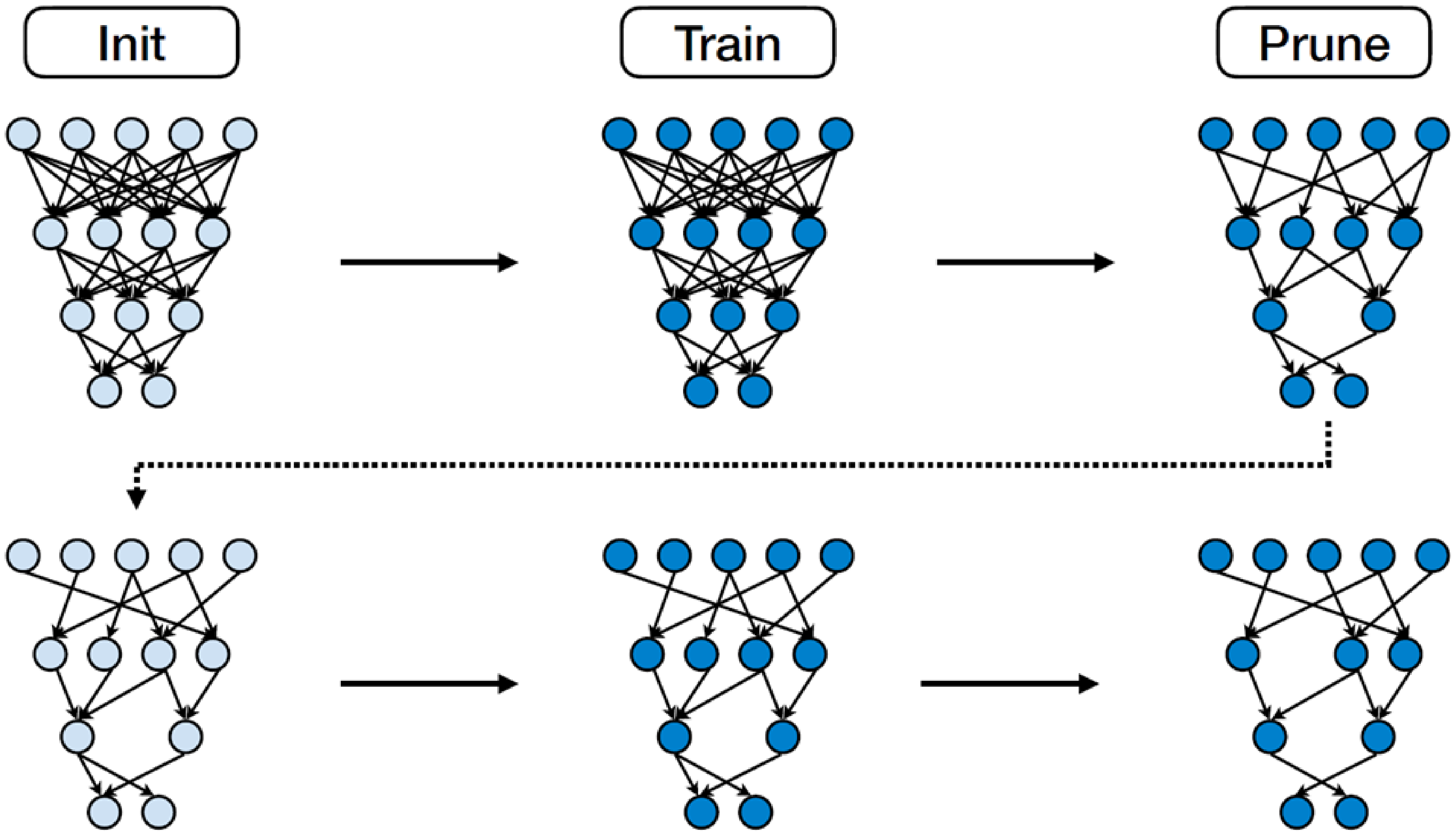
- Iterative Magnitude Pruning



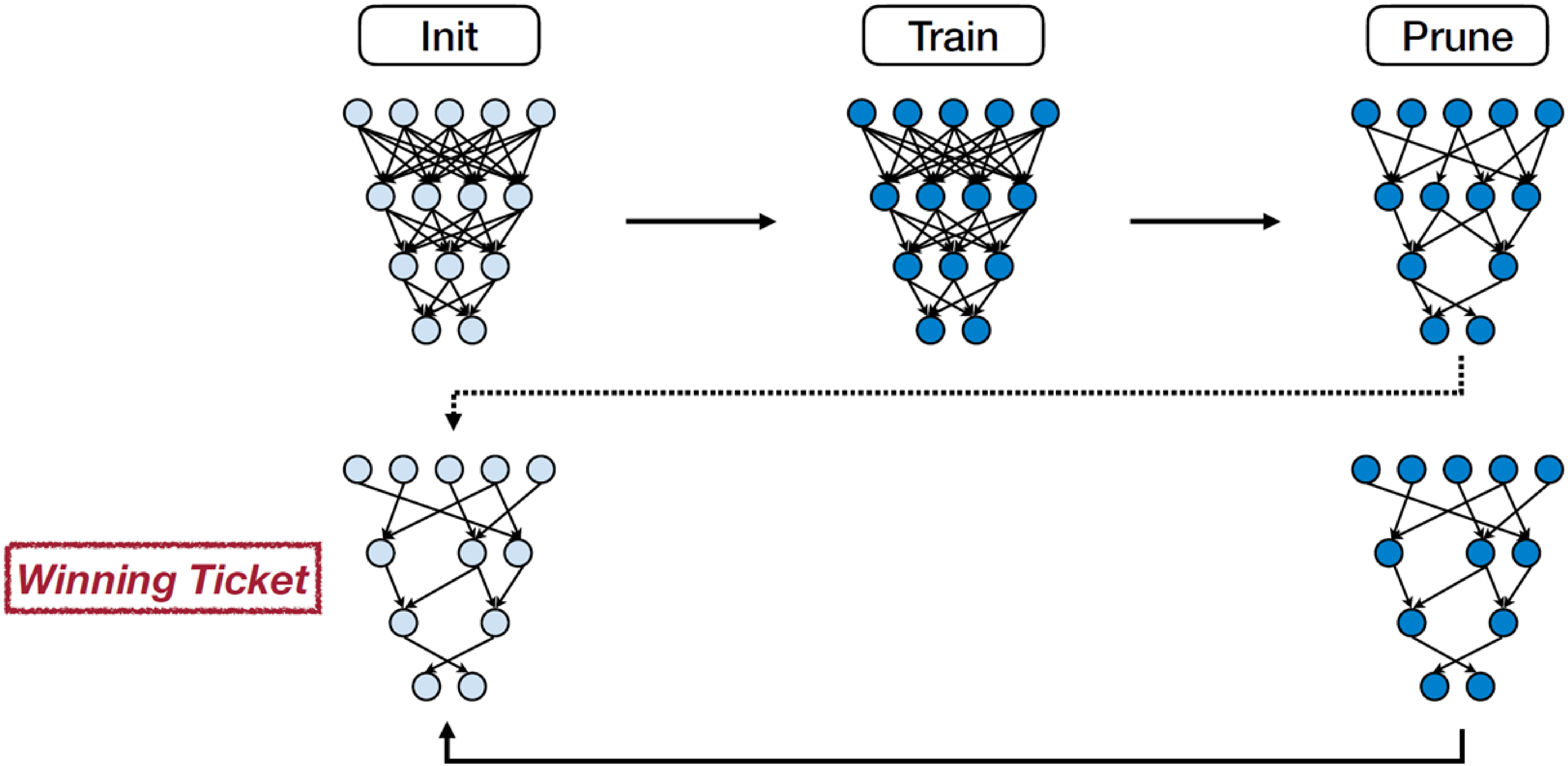
Iterative Magnitude Pruning



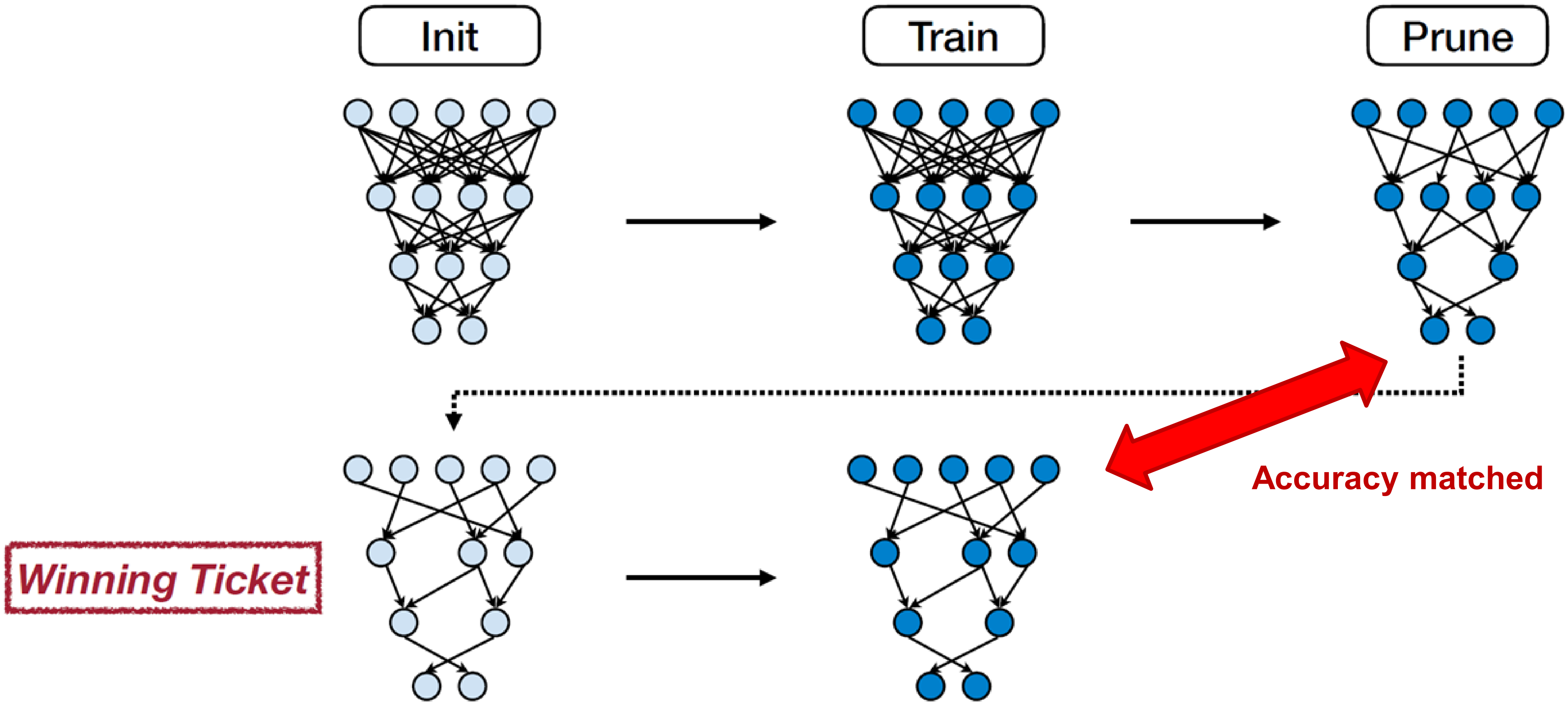
Iterative Magnitude Pruning



Iterative Magnitude Pruning

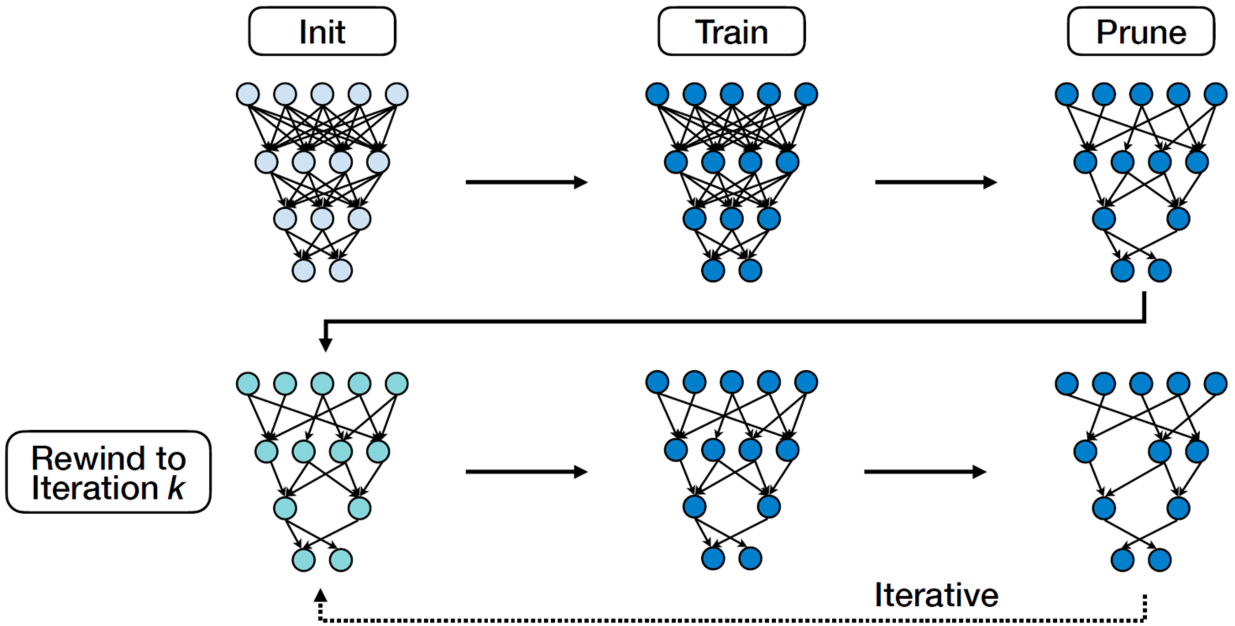


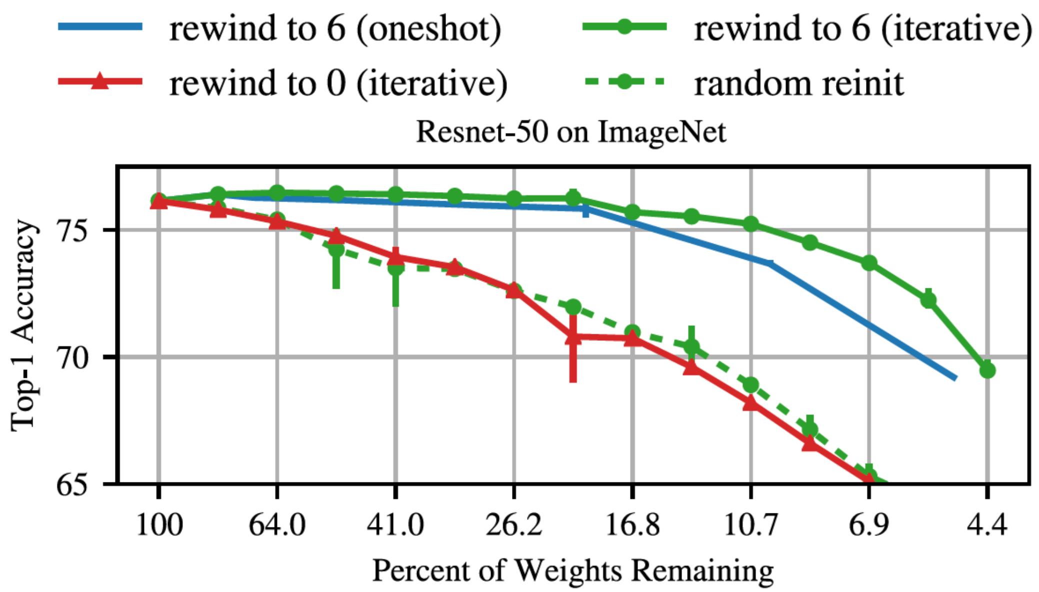
Iterative Magnitude Pruning



Iterative Magnitude Pruning

- Resetting the weights to the very initial value works for small-scale tasks (e.g., MNIST), but fails on deep networks.
- Instead, it is possible to robustly obtain pruned subnetworks by resetting the weights to the **values after a small number of k training iterations.**





Summary

- In this lecture, we learned:
  - Each layer in neural network has different sensitivity to pruning
  - Automated ways to find pruning ratios
  - Performance improvement after pruning
  - Lottery ticket hypothesis