MACQUARIE University
SYDNEY·AUSTRALIA

Topic: The Lottery Ticket Hypothesis

FINDING SPARSE, TRAINABLE NEURAL NETWORKS

DATE: 1st, March, 2024

Model Pruning

A MATURE METHOD

- A method to reduce parameters and increase speed (Mostly inference speed)

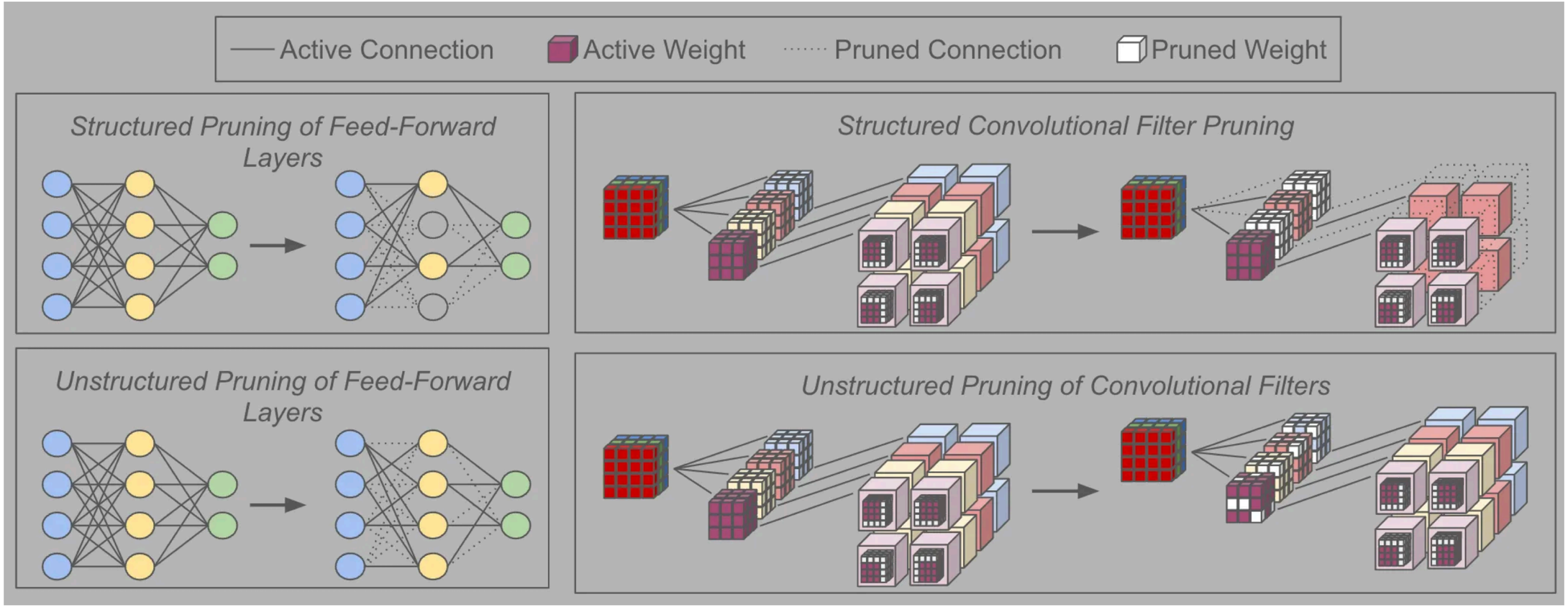
- Most famous one might be the paper *Deep Compression* by Song Han et. al,
  - *Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding*

- One-shot VS Iterative Magnitude Pruning (IMP)
  - 1.Begin with a fully-trained, dense model
  - 2.Select and prune the lowest magnitude weights in the network
  - 3. Fine-tune/train the resulting subnetwork to convergence
  - 4. Repeat steps (2)-(3) until the desired pruning ratio is achieved

- Structured VS Unstructured Pruning

Model Pruning

A MATURE METHOD



The Lottery Ticket Hypothesis (LTH)

MACQUARIE University
SYDNEY·AUSTRALIA

A HYPOTHESIS

The 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.*

More formally, consider a dense feed-forward neural network $f(x;\theta)$ with initial parameters $\theta = \theta_0 \sim \mathcal{D}_\theta$. When optimizing with stochastic gradient descent (SGD) on a training set, f reaches minimum validation loss l at iteration j with test accuracy a. In addition, consider training $f(x; m\odot\theta)$ with a mask $m \in \{0,1\}^{|\theta|}$ on its parameters such that its initialization is $m \odot \theta_0$. When optimizing with SGD on the same training set (with m fixed), f reaches minimum validation loss l' at iteration j' with test accuracy a'. The lottery ticket hypothesis predicts that $\exists\, m$ for which $j' \leq j$ (*commensurate training time*), $a' \geq a$ (*commensurate accuracy*), and $\|m\|_0 \ll |\theta|$ (*fewer parameters*).

The Lottery Ticket Hypothesis (LTH)

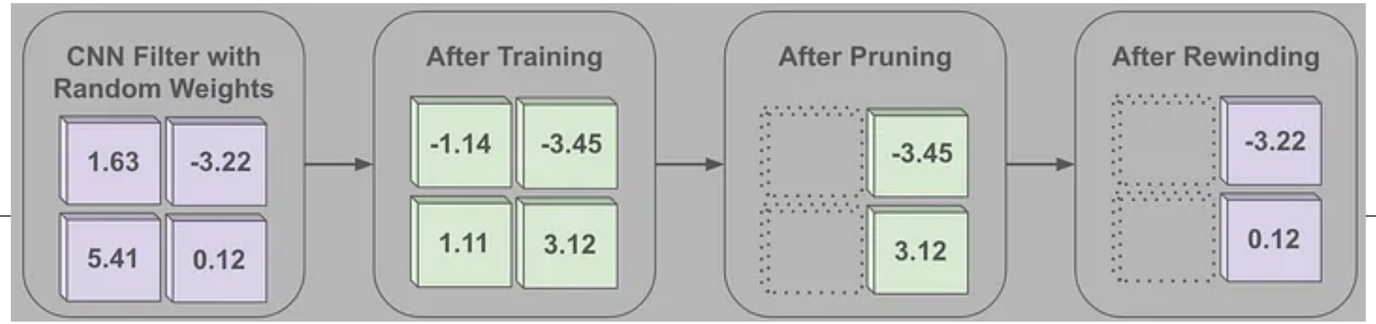
MACQUARIE University
SYDNEY·AUSTRALIA

A HYPOTHESIS

Identifying winning tickets. We identify a winning ticket by training a network and pruning its smallest-magnitude weights. The remaining, unpruned connections constitute the architecture of the winning ticket. Unique to our work, each unpruned connection's value is then reset to its initialization from original network *before* it was trained. This forms our central experiment:

1. Randomly initialize a neural network $f(x; \theta_0)$ (where $\theta_0 \sim \mathcal{D}_\theta$).
2. Train the network for j iterations, arriving at parameters θ_j.
3. Prune $p\%$ of the parameters in θ_j, creating a mask m.
4. Reset the remaining parameters to their values in θ_0, creating the winning ticket $f(x; m \odot \theta_0)$.

As described, this pruning approach is *one-shot*: the network is trained once, $p\%$ of weights are pruned, and the surviving weights are reset. However, in this paper, we focus on *iterative pruning*, which repeatedly trains, prunes, and resets the network over n rounds; each round prunes $p^{\frac{1}{n}}\%$ of the weights that survive the previous round. Our results show that iterative pruning finds winning tickets that match the accuracy of the original network at smaller sizes than does one-shot pruning.



The Lottery Ticket Hypothesis (LTH)

MACQUARIE University
SYDNEY·AUSTRALIA

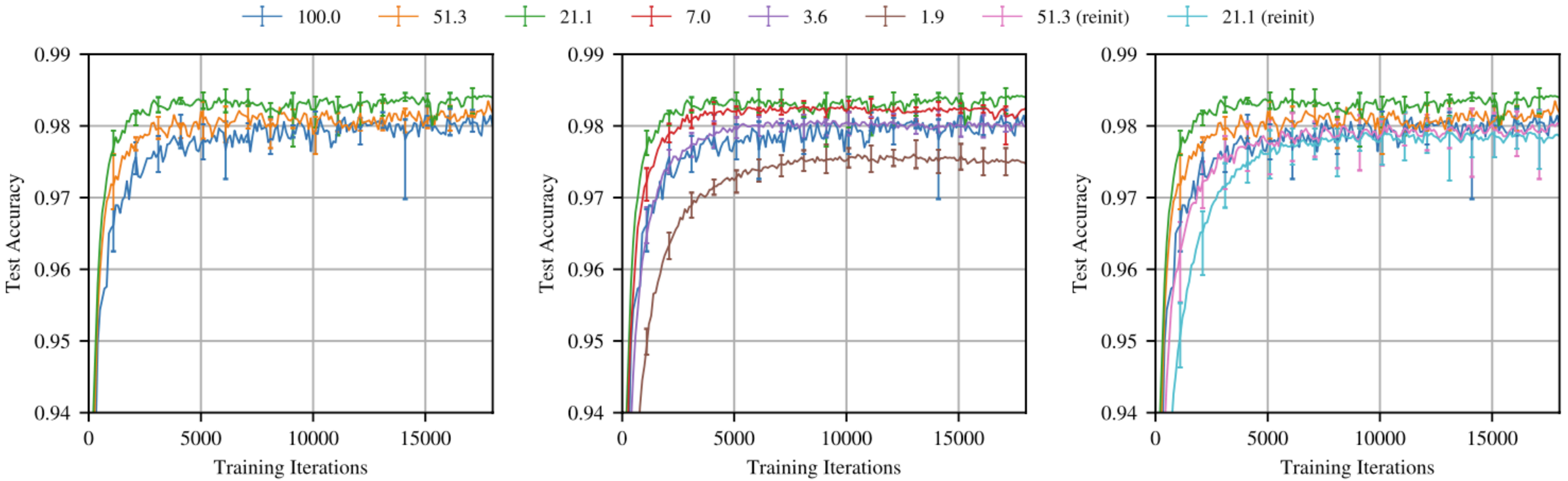
A HYPOTHESIS

Network	Lenet	Conv-2	Conv-4	Conv-6	Resnet-18	VGG-19
Convolutions		64, 64, pool	64, 64, pool 128, 128, pool	64, 64, pool 128, 128, pool 256, 256, pool	16, 3x[16, 16] 3x[32, 32] 3x[64, 64]	2x64 pool 2x128 pool, 4x256, pool 4x512, pool, 4x512
FC Layers	300, 100, 10	256, 256, 10	256, 256, 10	256, 256, 10	avg-pool, 10	avg-pool, 10
All/Conv Weights	266K	4.3M / 38K	2.4M / 260K	1.7M / 1.1M	274K / 270K	20.0M
Iterations/Batch	50K / 60	20K / 60	25K / 60	30K / 60	30K / 128	112K / 64
Optimizer	Adam 1.2e-3	Adam 2e-4	Adam 3e-4	Adam 3e-4	← SGD 0.1-0.01-0.001 Momentum 0.9 →	
Pruning Rate	fc20%	conv10% fc20%	conv10% fc20%	conv15% fc20%	conv20% fc0%	conv20% fc0%

Figure 2: Architectures tested in this paper. Convolutions are 3x3. Lenet is from LeCun et al. (1998). Conv-2/4/6 are variants of VGG (Simonyan & Zisserman, 2014). Resnet-18 is from He et al. (2016). VGG-19 for CIFAR10 is adapted from Liu et al. (2019). Initializations are Gaussian Glorot (Glorot & Bengio, 2010). Brackets denote residual connections around layers.

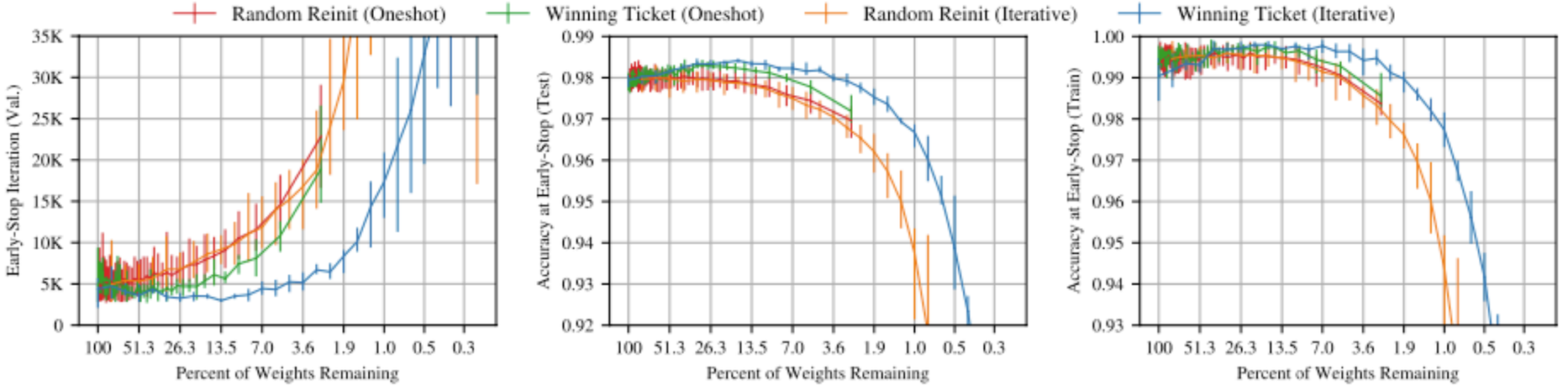
The Lottery Ticket Hypothesis (LTH)

A HYPOTHESIS

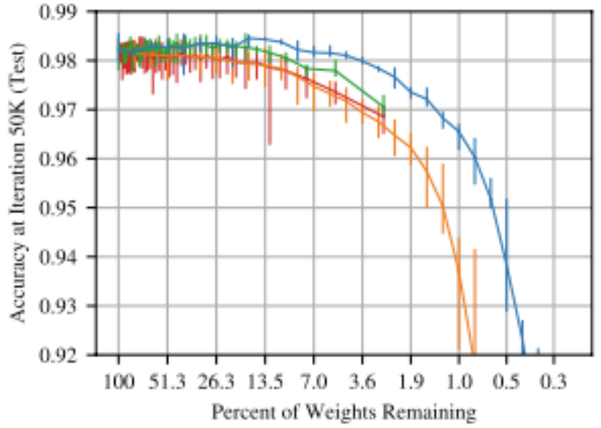


The Lottery Ticket Hypothesis (LTH)

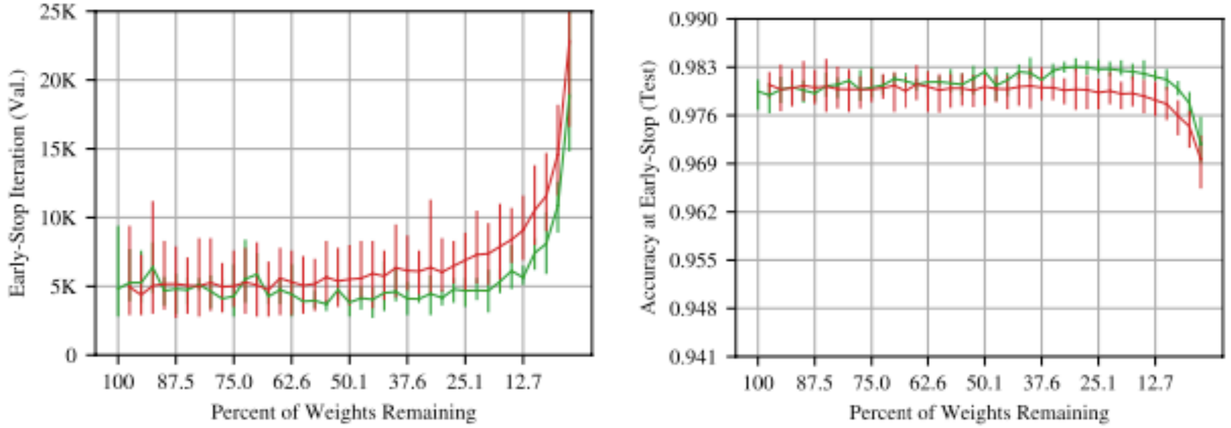
A HYPOTHESIS



(a) Early-stopping iteration and accuracy for all pruning methods.



(b) Accuracy at end of training.



(c) Early-stopping iteration and accuracy for one-shot pruning.

The Lottery Ticket Hypothesis (LTH)

A HYPOTHESIS

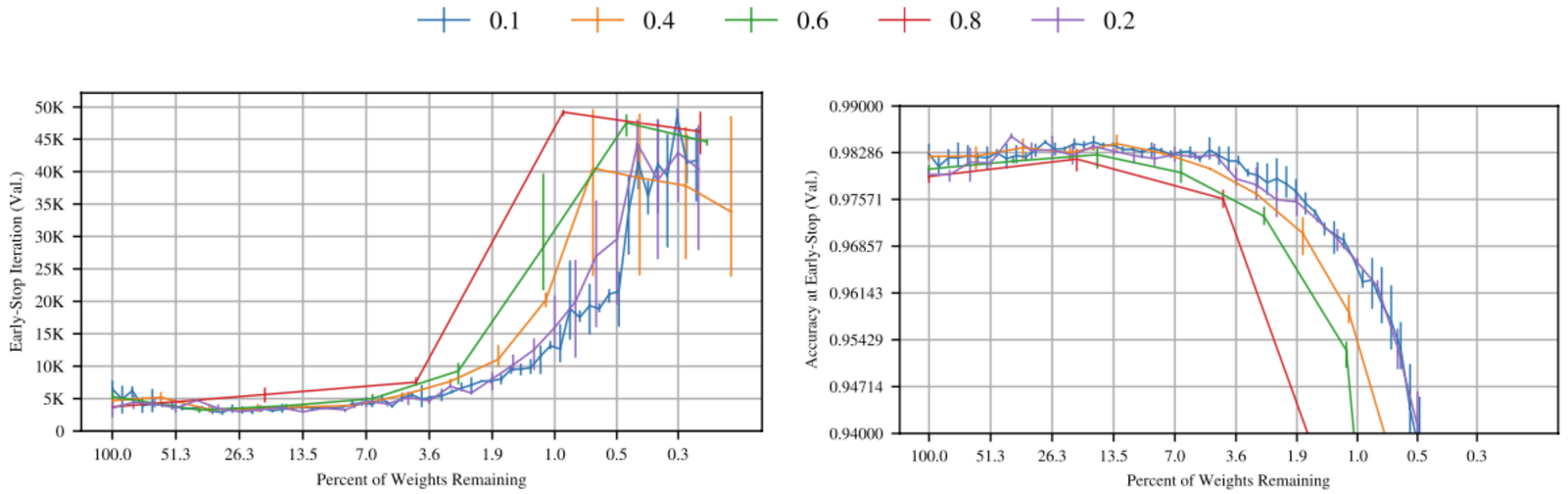


Figure 29: The early-stopping iteration and validation accuracy at that iteration of the iterative lottery ticket experiment when pruned at different rates. Each line represents a different *pruning rate*—the percentage of lowest-magnitude weights that are pruned from each layer after each training iteration.

The Lottery Ticket Hypothesis (LTH)

A HYPOTHESIS

Global pruning. On Lenet and Conv-2/4/6, we prune each layer separately at the same rate. For Resnet-18 and VGG-19, we modify this strategy slightly: we prune these deeper networks *globally*, removing the lowest-magnitude weights collectively across all convolutional layers. In Appendix I.1, we find that global pruning identifies smaller winning tickets for Resnet-18 and VGG-19.

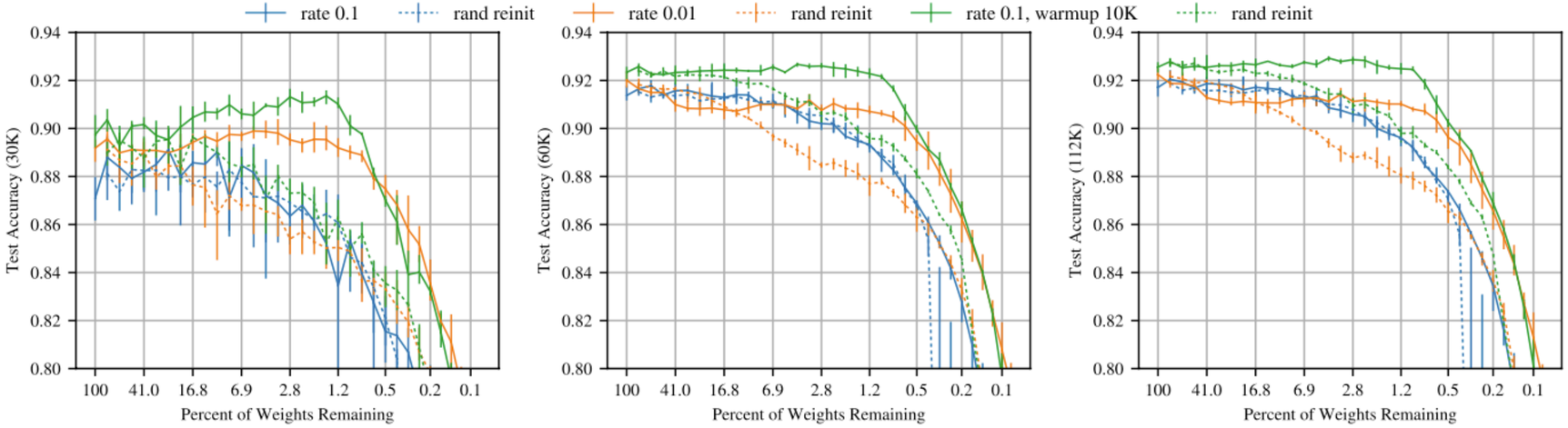


Figure 7: Test accuracy (at 30K, 60K, and 112K iterations) of VGG-19 when iteratively pruned.

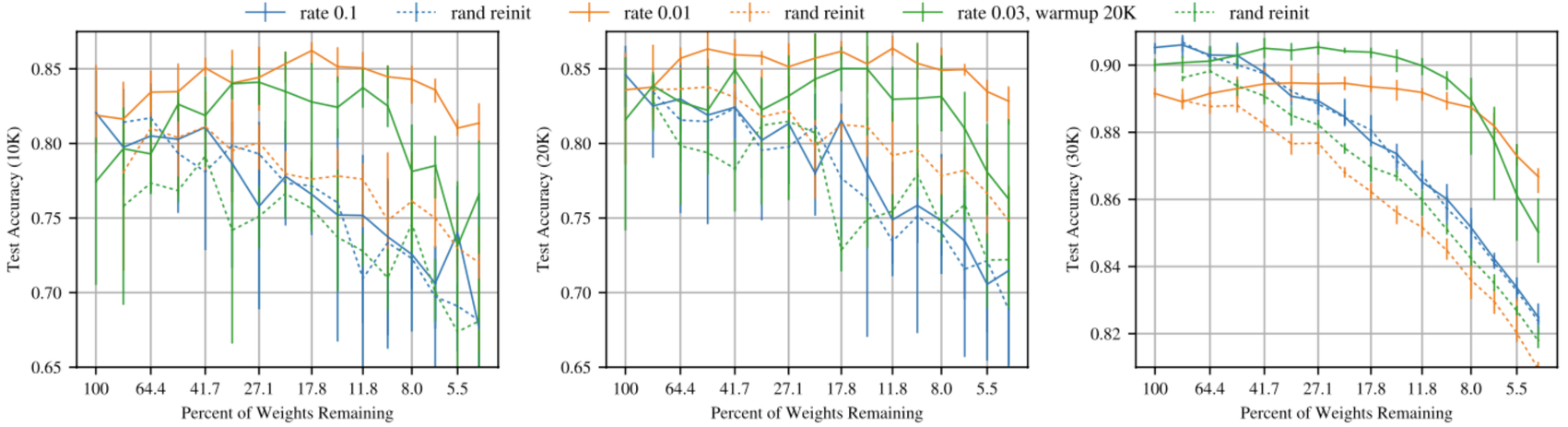


Figure 8: Test accuracy (at 10K, 20K, and 30K iterations) of Resnet-18 when iteratively pruned.

Discussion

A HYPOTHESIS

- The importance of winning ticket initialization
- The importance of winning ticket structure
- The improved generalization of winning tickets

Discussion

MACQUARIE University
SYDNEY·AUSTRALIA

SOME PERSONAL THOUGHTS ABOUT THE PAPER AND THE IDEA

- Is this paper cherry picking when doing the experiment?
  - Is the network too large? (ResNet18: 11.5M, ResNet20: 0.27M)
  - No imagenet dataset (Conducted in the following work, *Stabilizing the Lottery Ticket Hypothesis*)

- There's also different opinion about reuse original init value
  - same year, same conference, (*Rethinking the Value of Network Pruning*)

- The idea itself is useful, and there are many derivative research outputs
  - *The Lottery Ticket Hypothesis for Object Recognition*
  - *The Lottery Ticket Hypothesis for Pre-trained BERT Networks*
  - etc.

MACQUARIE University
SYDNEY·AUSTRALIA

Thank you for listening