The Lottery Ticket Hypothesis

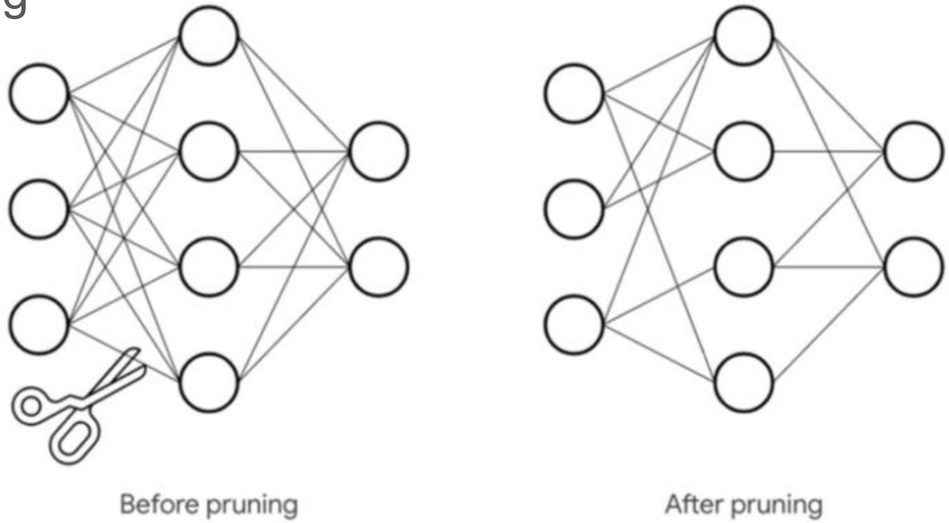
Lebedyuk Eva

2023

Network pruning

Standard pruning methods:

- One-shot pruning
- Iterative pruning



The Lottery Ticket Hypothesis

- ➔ Consider a dense neural network ***f(x; θ)*** with initial parameters $\boldsymbol{\theta = \theta_0 \sim D_\theta}$
- ➔ When optimizing with 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*θ)*** with a mask m ∈ $\{0, 1\}^{|\theta|}$ on its parameters such that its initialization is $\boldsymbol{m*\theta_0}$
- ➔ 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’***

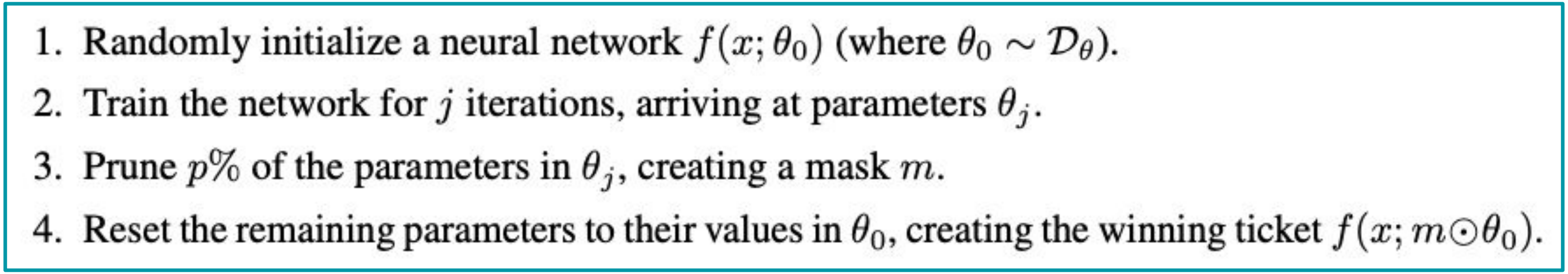
The lottery ticket hypothesis predicts that ∃

- m for which j’ ≤ j (commensurate training time)
- a’≥ a (commensurate accuracy)
- ||m||0 << |θ| (fewer parameters)

Winning tickets

- We find that a standard pruning technique automatically uncovers such trainable subnetworks from fully-connected and convolutional feed-forward networks.
- We designate these trainable subnetworks ***$f(x; m*\theta_0)$*****,** ***winning tickets***

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 $f(x; m \odot \theta_0)$.

Results

We identify winning tickets in:

- fully-connected architecture for MNIST
- convolutional architectures for CIFAR10

The winning tickets we find are ***10-20%*** (or less) of the size of the original

Winning tickets in fully-connected networks

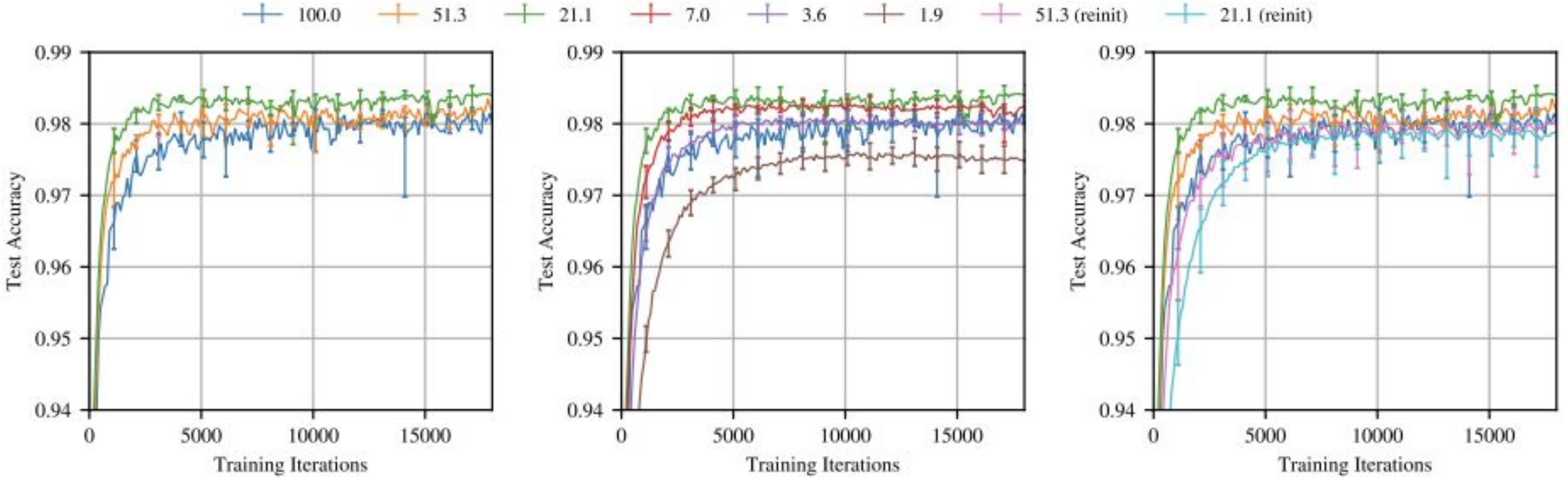
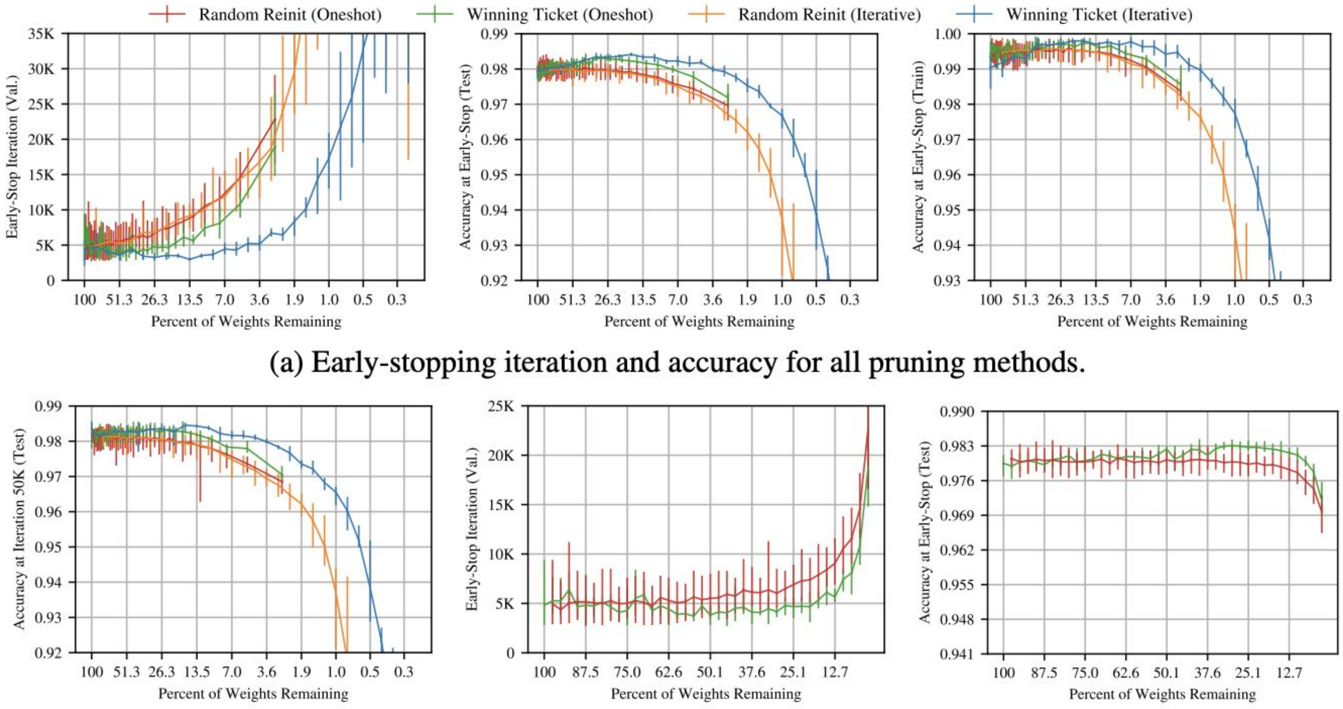


Figure 3: Test accuracy on Lenet (iterative pruning) as training proceeds. Each curve is the average of five trials. Labels are P_m—the fraction of weights remaining in the network after pruning. Error bars are the minimum and maximum of any trial.

Winning tickets in fully-connected networks



(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.

Figure 4: Early-stopping iteration and accuracy of Lenet under one-shot and iterative pruning. Average of five trials; error bars for the minimum and maximum values. At iteration 50,000, training accuracy ≈ 100% for $P_m \geq 2\%$ for iterative winning tickets (see Appendix D, Figure 12).

Winning tickets in convolutional networks

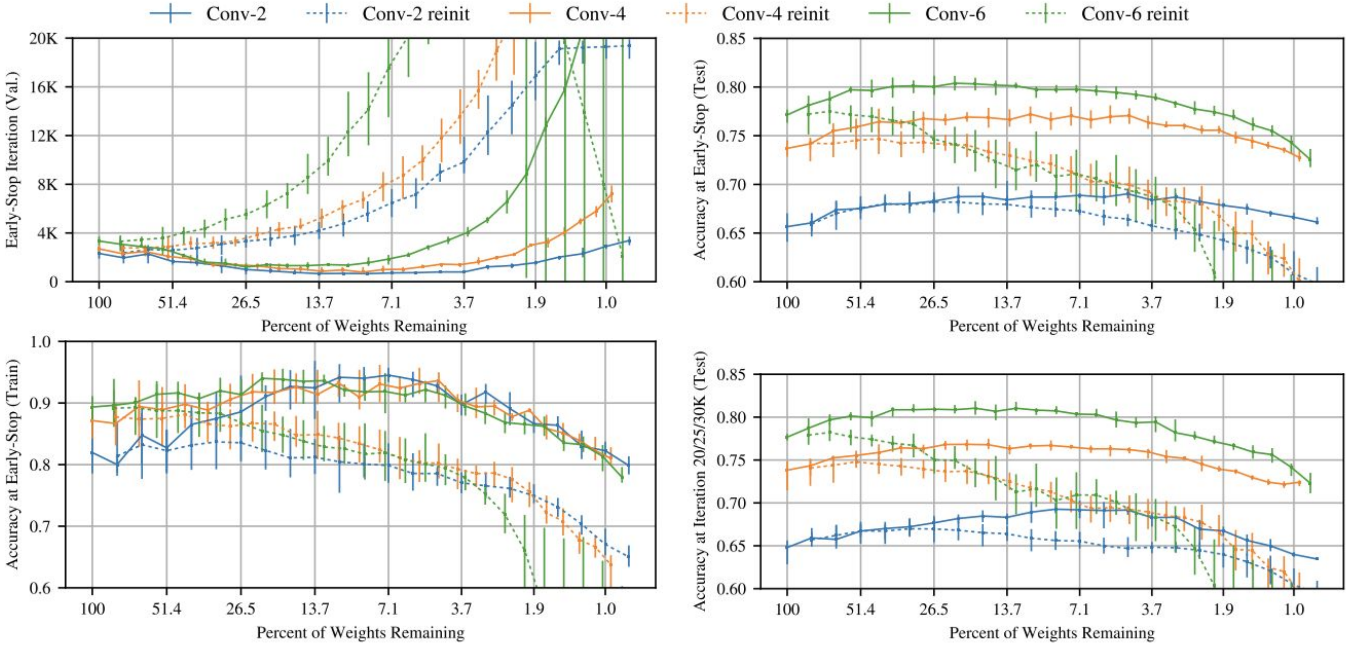


Figure 5: Early-stopping iteration and test and training accuracy of the Conv-2/4/6 architectures when iteratively pruned and when randomly reinitialized. Each solid line is the average of five trials; each dashed line is the average of fifteen reinitializations (three per trial). The bottom right graph plots test accuracy of winning tickets at iterations corresponding to the last iteration of training for the original network (20,000 for Conv-2, 25,000 for Conv-4, and 30,000 for Conv-6); at this iteration, training accuracy $\approx 100\%$ for $P_m \geq 2\%$ for winning tickets (see Appendix D).

Winning tickets in convolutional networks

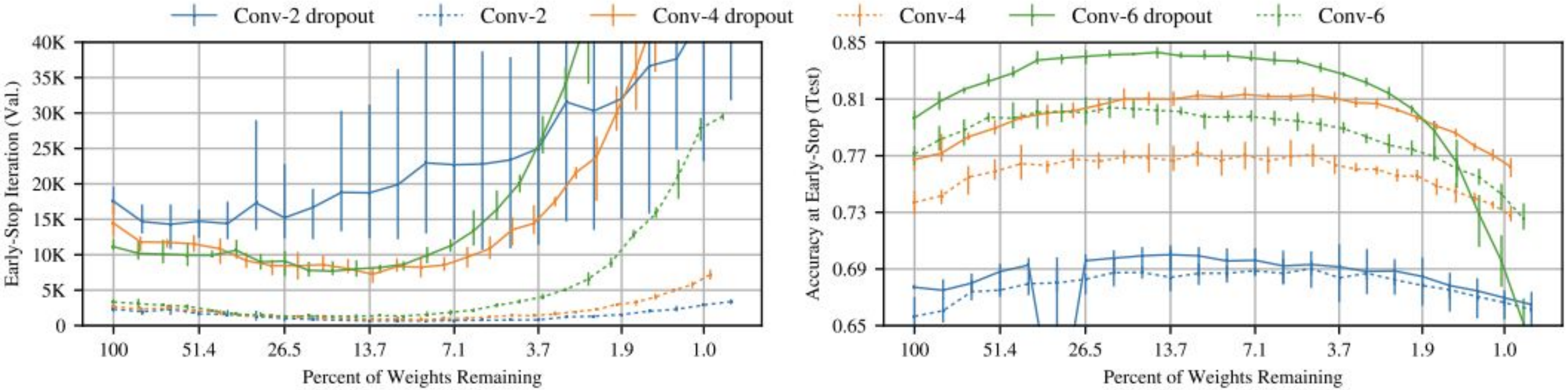


Figure 6: Early-stopping iteration and test accuracy at early-stopping of Conv-2/4/6 when iteratively pruned and trained with dropout. The dashed lines are the same networks trained without dropout (the solid lines in Figure 5). Learning rates are 0.0003 for Conv-2 and 0.0002 for Conv-4 and Conv-6.

Networks, used in practice

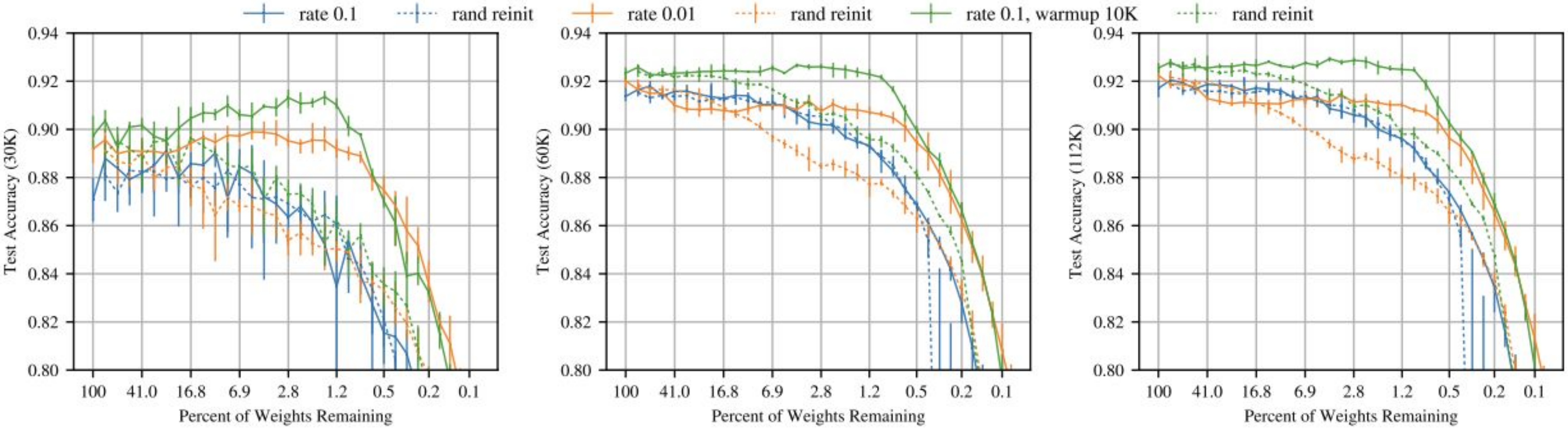


Figure 7: Test accuracy (at 30K, 60K, and 112K iterations) of VGG-19 when iteratively pruned.

Discussion

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

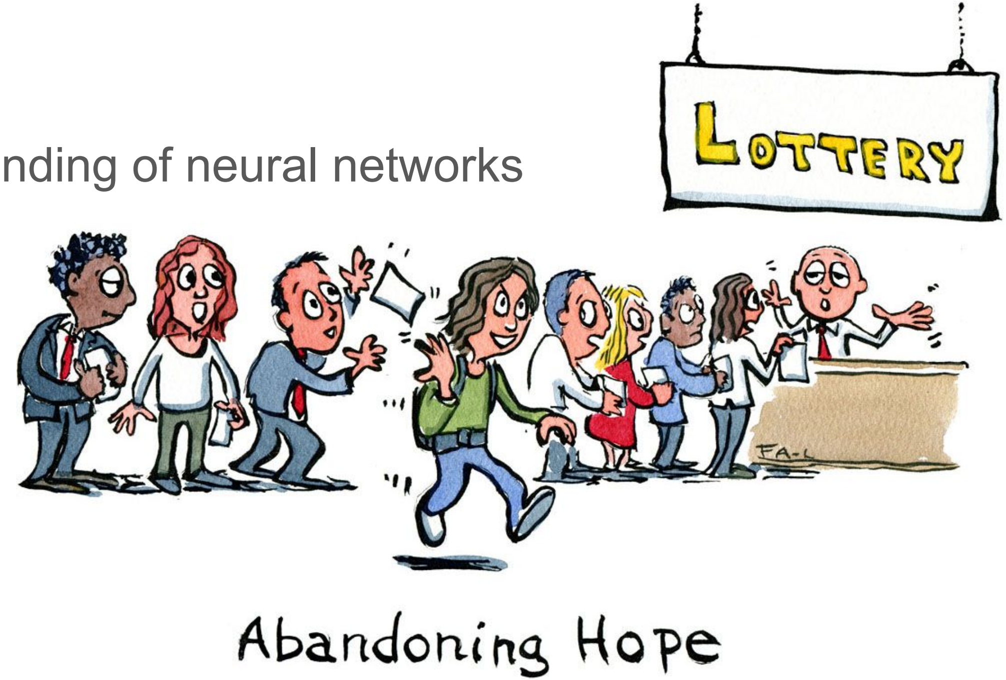
Limitations and future work

- We intend to explore more efficient methods for finding winning tickets for larger datasets
- We intend to study other pruning methods from the extensive contemporary literature
- We plan to explore why warmup is necessary and other improvements to our scheme for identifying winning tickets in deeper networks

Implications

With winning tickets we can:

- Improve training performance
- Design better networks
- Improve our theoretical understanding of neural networks



Source

- https://arxiv.org/abs/1803.03635