Overparameterization is **not required** for **robustness**. **Winning lottery tickets** account for the network's accuracy, **train faster**, and may achieve **better robustness** with adversarial training.

Download the paper:

The Search for Sparse, Robust Neural Networks

Justin Cosentino*, Federico Zaiter*, Dan Pei†, Jun Zhu†

Department of Computer Science, Tsinghua University

Introduction

- Recent work on deep neural network pruning has shown there exist sparse subnetworks that achieve equal or improved accuracy, training time, and loss using fewer network parameters when compared to their dense counterparts.
- Orthogonal to pruning literature, deep neural networks are known to be susceptible to adversarial examples, which may pose risks in security- or safety-critical applications.
- We perform an extensive empirical evaluation and analysis testing the Lottery Ticket Hypothesis [1] with adversarial training and show this approach enables us to find sparse, robust neural networks.

Methodology

The Lottery Ticket Hypothesis (LTH)

- LTH proposes an unstructured pruning strategy for making neural networks sparser.
- It states that a randomly-initialized, dense neural network contains subnetworks that–when trained in isolation–achieve better accuracy and generalization because of the fortuitous initialization of their weights.

Adversarial Robustness

- We use the **Fast Gradient Sign Method (FGSM)** and **Projected Gradient Descent (PGD)** white-box attacks for generating adversarial examples.

Experimental Design

- We evaluate the adversarial robustness of networks produced from three iterative pruning strategies: **resetting**, **random initialization**, and **continued training**.
- We perform experiments on LeNet 300-100 with MNIST Digits and Fashion. We train the model with and without adversarial training for 20 pruning iterations.
- A pruning iteration consists of initializing the current iteration's parameters, training for 50,000 iterations, and pruning some percent of the model to get an updated mask.

Conclusions

- Unlike prior empirical studies [2] suggesting a trade-off between network sparsity and adversarial robustness, we find that overparameterization is not required for robustness.
- Winning lottery tickets not only account for the overall network's accuracy, but can also train faster and achieve similar–if not better–robustness with adversarial training.
- Future work includes finding novel pruning techniques specifically for robustness.

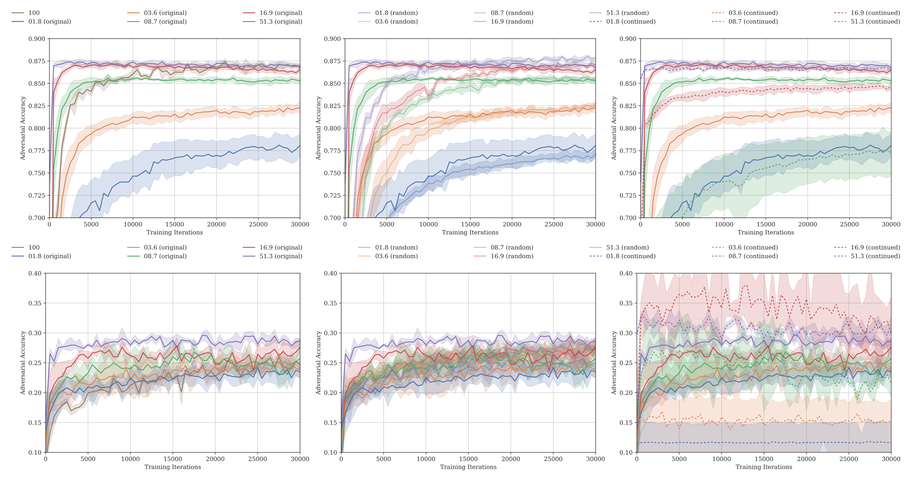

Figure 1. LeNet 300-100 adversarial test accuracy on MNIST Fashion with the FGSM (top) and PGD (bottom) attacks as training proceeds (left) and comparisons with the random (middle) and continued (right) pruning strategies. Labels are P_m: the fraction of weights remaining in the network after pruning.

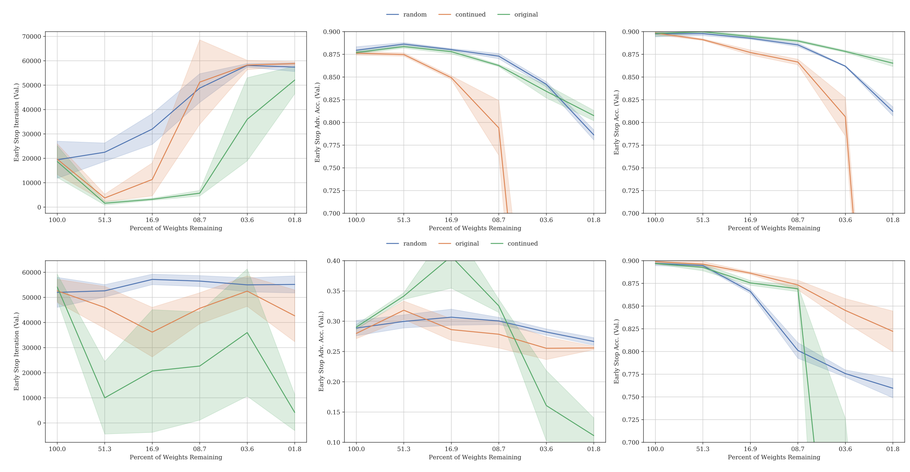

Figure 2. Early-stopping iteration (left), adversarial accuracy (middle), and natural accuracy (right) for each pruning strategy using LeNet 300-100 on MNIST Fashion with the FGSM attack (top) and PGD (bottom) attacks. Accuracy measures are taken at the early stopping iteration.

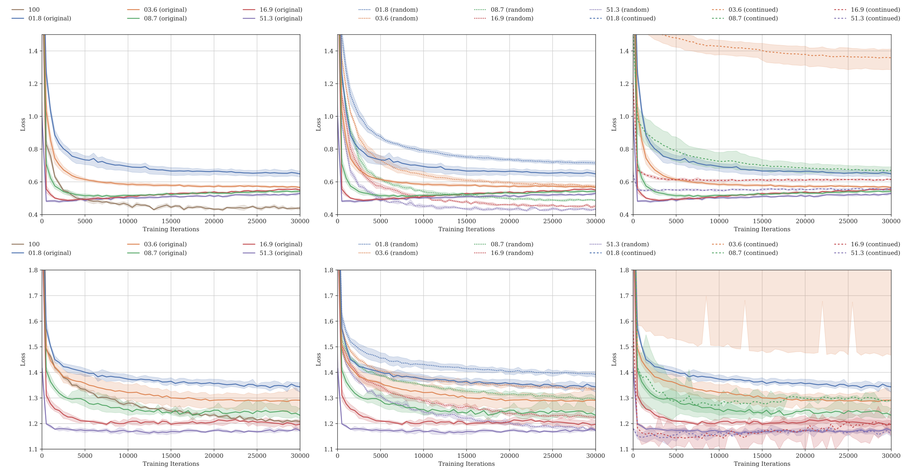

Figure 3. The adversarial validation loss data corresponding to Figure 2, i.e., the adversarial validation loss for LeNet 300-100 on MNIST Fashion with the FGSM (top) and PGD (bottom) attack as training proceeds (left) and comparisons with the random (middle) and continued (right) pruning strategies. Labels are P_m: the fraction of weights remaining in the network after pruning.

[1] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[2] Luyu Wang, Gavin Weiguang Ding, Ruitong Huang, Yanshuai Cao, and Yik Chau Lui. Adversarial robustness of pruned neural networks. 2018.