Deep Learning Reproduction Report

Margarita Grishechkina

The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks

(Frankle & Carbin, 2019)

In this paper, the authors introduce the term "winning ticket" as a metaphor, which describes a small subnetwork hidden inside a large, randomly initialised neural network. Just like buying a lottery ticket and hoping to get lucky, the idea is that some parts of the large network are already well-suited – or "lucky" – from the very beginning. These subnetworks, when reset to their original starting weights and trained again, can reach similar accuracy to the full model. The term highlights how these small networks, despite being simpler, may be all that is needed for effective learning if their initial weights are favourable.

Although the concept of a "winning ticket" involves removing parts of a neural network, it is different from commonly used methods such as dropout, which randomly disables some neurons during training to prevent overfitting, but all weights are restored during testing. In contrast, the winning ticket hypothesis involves permanently removing certain weights (after training), leaving a smaller network that can still learn effectively from scratch. Both techniques aim to reduce model complexity and improve generalisation, but the winning ticket approach finds a fixed, trainable subnetwork – not one that changes randomly during training.

In simple terms, the method involves iterative pruning, where the lowest-magnitude weights are removed after training. Then, the remaining weights are reinitialised to their original random values, and the smaller network is trained again from scratch. The authors released their implementation publicly, which supports reproducibility and makes it easier for others to try the pruning method. That said, reproducing the full results remains challenging in practice. The method is highly sensitive to the initial random seed, and it performs best on smaller datasets such as MNIST and CIFAR-10. On larger datasets like ImageNet or deeper architectures, it tends to perform poorly unless adjusted with techniques like learning rate rewinding. As well as that, it requires multiple full training runs, which makes the process computationally expensive and potentially impractical for large-scale applications. As for evaluation, the authors mainly use test accuracy and model sparsity (the percentage of weights pruned). This leads to some criticism – for example, the paper does not discuss robustness, calibration, or inference speed. It also could have included other metrics such as the generalisation gap or training time. Furthermore, there is little exploration of how stable or consistent the method is across different settings.

The authors tested their hypothesis on small-scale image classification tasks using LeNet (on MNIST) and Conv2/4/6, ResNet-18, and VGG-19 (on CIFAR-10). Convolutional layers were 3x3 and networks were pruned to 10-20% of their original size. These are small-scale classification tasks. No NLP models, transfer learning, or real-world deployment scenarios were tested. Also, the paper does not explore whether the idea generalises to other tasks such as segmentation or regression.

This leaves several opportunities for improvement. For example, future work could include benchmarks on more diverse datasets and tasks, study the long-term stability of winning tickets, try different pruning methods (not just based on weight magnitude), and provide reproducibility scripts for all figures and experiments.

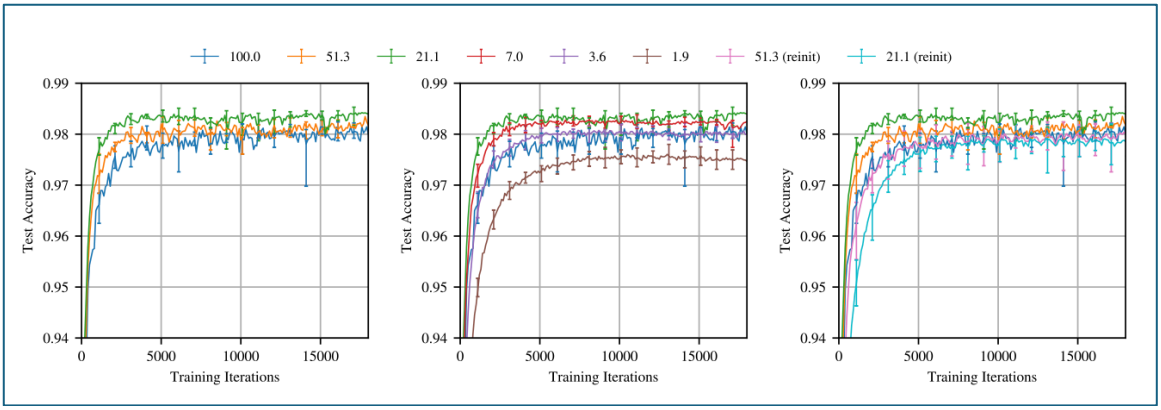

Figure 1, Test accuracy on Lenet (iterative pruning) as training proceeds. Each curve is the average of five trials. Labels are Pm—the fraction of weights remaining in the network after pruning. Error bars are the minimum and maximum of any trial. Figure adapted from Frankle & Carbin (2019).

A key result is shown in Figure 1, which plots the test accuracy of pruned subnetworks over training iterations. Even after removing a large percentage of weights (leaving only 21.1% of the original connections), these smaller networks still achieve nearly the same accuracy as the full model. When the same subnetworks are randomly reinitialised, their performance drops noticeably. This highlights that it is not just sparsity that matters, but also the specific initialisation – the right weights in the right places from the beginning.

While these findings are promising on small datasets, Gale et al. (2019) were unable to reproduce similar results on larger datasets like ImageNet without modifying the method, raising concerns about how well the hypothesis scales to more complex models. On the other hand, Brix et al. (2020) proposed a stabilised version of the hypothesis for Transformer architectures and reported encouraging results on machine translation tasks. This suggests that the core idea may still be valid if appropriately adapted to different model types.

Further perspective is offered by Hoefler et al. (2022), who describe magnitude-based pruning – like that used in the Lottery Ticket Hypothesis – as a "strong baseline." However, they also report that more advanced methods often outperform it on larger models and that such pruning approaches face clear scalability limitations. The authors call for the development of standardised benchmarks to enable more meaningful comparisons between pruning strategies. This reinforces earlier concerns and underlines the importance of evaluating pruning methods not only in terms of accuracy, but also in terms of reproducibility and practical deployment at scale.

References:

Brix, C., Bahar, P., & Ney, H. (2020). Successfully applying the stabilized lottery ticket hypothesis to the transformer architecture. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 3909–3915). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.360/

Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. In International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1803.03635

Gale, T., Elsen, E., & Hooker, S. (2019). The state of sparsity in deep neural networks. arXiv preprint arXiv:1902.09574. https://arxiv.org/abs/1902.09574

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peste, A. (2022). Sparsity in deep learning: Pruning and growth for efficient inference. Journal of Machine Learning Research, 22(241), 1–124. https://www.jmlr.org/papers/volume22/21-0366/21-0366.pdf