Winning Weights: A Review of The Lottery Ticket Hypothesis

Søren Winkel Holm

June 20, 2022

Introduction

While Deep Learning (DL) is celebrated as a universal technological step forward, making complex modelling available to many industries requiring less use of domain-specialised engineers, the computational cost of the training procedure of Deep Artificial Neural Network (DNN)'s makes the technology unavailable for low-resource actors. The price of large language modelling projects such as Google's T5 is estimated in the region of \$ 10 M and such tech companies have a leading role in research. While the processing of big data is naturally computationally intensive, much of training cost is caused by a large number of epochs being required before convergence. The Lottery Ticket Hypothesis (LTH) points to DNN learning dynamics that are keeping this number high.

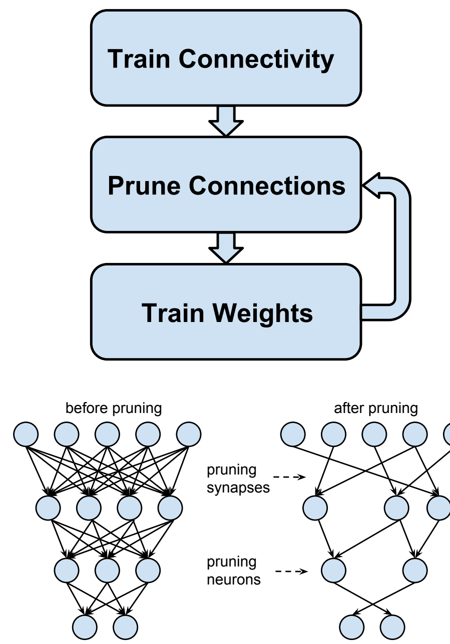

Figure 1: The original iterative pruning method where connectivity training corresponds to standard full training of a dense DNN [Han+15, Fig. 2 and 3].

To help availability, rich actors often share the trained model parametrizations but even in this case, application might be widely inaccessible because of computational costs of inference using these large parametrizations. This problem has been attacked using model pruning, reducing trained model size while retaining performance, but the expensive training of the full DNN has generally been required. LTH explains why an initial full training is generally required and opens up for researching how to train efficient parametrizations from scratch. These ideas and methods will here be reviewed.

Fundamental Concepts

DNN pruning refers to disabling particular model connections $w_i \leftarrow 0$ possibly to improve generalization, reducing memory constraints in inference and lowering inference computation [LDS89]. Pruning during training is related to regularization e.g. using dropout, while pruning after fully training a dense network parametrization $q(w) = \tilde{w}$ is often motivated by computational cost, and might require some fine-tuning to limit the decrease in accuracy [Lan20]. Disabling unnecessary weights is a way to learn the connectivity of a DNN and can be performed iteratively based on magnitude such as

$$\forall i \text{ s. t. } |w_i^{(t)}| < k \text{ let } w_i^{(t+1)} \leftarrow 0, \quad (1)$$

where each iteration is followed by fine-tuning, k is a threshold set to e.g. $k = s\sqrt{\mathrm{Var}[w]}, s = \frac{1}{2}$ [Han+15; Zmo+19]. The procedure is stopped when a specified compression level of performance drop is reached [Han+15] as shown on Figure 1. The pruned parametrizations $w^{(p)}$ can be represented using a mask m, $w^{(p)} = m \odot w$ and (1) is thus dubbed the masking criterion. Pruning might be structured locally by assigning layer-specific thresholds and target compression levels or by fixing parts of the DNN [Han+15]. The resulting sparse DNN $f(x; m \odot w)$ is called a subnetwork of the full, trained $f(x; w)$

Simple pruning approaches have empirically been shown to work well across network types and learning tasks with compression rates of $\sim \times 10$ resulting in accuracy drops ~ 1 % [Bla+20, Fig. 7]. These results do not arise when training from the start with such pruned networks [Li+16, Chap. 4], [Han+15, Chap. 3.3]. LTH gives an explanation for this effect by postulating the existence of a *winning ticket* for a randomly initialized, dense DNN $f(x; w^{(0)}), w^{(0)} \sim \mathcal{D}_w$. A winning ticket is a subnetwork $f(x; m \odot w^{(0)})$ that can be trained by itself and reach same generalization error as the full network in the same number of epochs or less. The name thus implies the existence of an initialization lottery where specific combinations of connection masks and weight prior realisations allow learning. In this context, standard pruning techniques find winning tickets by first learning the entire, dense w and then after training finding m.

LTH implies that m can be computed from a full training after which the weights can be *rewinded* to $w^{(0)}$ at which point the training $m \odot w^{(0)}$ should result in a performant network.

State of the Art

Evidence for ticket existence

LTH was presented by Frankle and Carbin [FC19] in 2019 where empirical evidence for the hypothesis was presented on MNIST and CIFAR-10. An effect was seen when comparing training of rewinded weights of a winning ticket to random reinitialization. Using iterative pruning, winning tickets were found for all tried DNN's and these were found to learn faster than full networks, but for deep networks such as VGG-19, finding the tickets was sensitive to learning rate setup and required warmup steps [FC19, Chap. 4].

In follow-up work, the robustness of this iterative search for winning tickets was improved by introducing a procedure called *instability analysis* where the impact of Stochastic Gradient Descent (SGD) noise such as minibatch order and augmentations was investigated [Fra+20]. This analysis showed that for many deeper networks stability against SGD noise occurs after a number of training steps k. For LTH to hold robustly on these DNN's, rewinding of weights was changed from $m \odot w^{(0)}$ being the winning ticket to $m \odot w^{(k)}$ being winning. Thus, the winning ticket was not shown to exist at initialization but slightly-trained winning subnetworks were empirically found across challenging datasets and network sizes. These winners were dubbed winning matching tickets instead of winning lottery tickets and this weaker hypothesis has been called The Lottery Ticket Hypothesis with Rewinding (LTH-R) [Lan20] .

Concurrently, analysis quantifying the performance of winning tickets compared to previous pruning methods was performed by Renda, Frankle, and Carbin [RFC20]. The rewinding to winning matching tickets and retraining of LTH-R ended up outperforming standard pruning that fine-tunes final weights. Furthermore, the rewinding approach was superior in a limited-budget setting across Natural Language Processing (NLP) and Computer Vision (CV) tasks, and it was concluded that LTH-R was State of The Art (SOTA) for pruning in terms of accuracy, compression and computational cost [RFC20, Chap. 6] [Lan20].

Early ticket identification

The simple train-rewind-retrain approach used to empirically demonstrate the existance of winning matching tickets was revealed to give strong pruning performance across tasks without need for hyperperameter tuning [RFC20, Chap. 6]. However, this method requires full convergence of the network before identifying the optimal ticket. Training of sparse DNN's would improve dramatically if the winning ticket mask m could be found before training.

In 2020, identification of winning tickets was performed early in training by You, Li, et al. [You+20]. These Early-Bird (EB) tickets were found using a mask distance measure between epochs. A mask m_t was at each epoch computed and the Hamming distance between the binary matrices m_{t-1} and m_t was used as a ticket search criterion [You+20, Chap. 3.3]. Search was stopped when the criterion was under $\epsilon = 0.1$ for five consecutive epochs, resulting in an algorithm that succesfully found winning tickets at much less computational cost. Across CV tasks, EB tickets performed at the same accuracy level compared to standard winning tickets and other pruning techniques while using less than half the number of computations.

Also in 2020, two approaches attempted to fully exploit LTH by finding m at initialization were presented. One by Wang, Wang, Zhang, and Grosse [Wan+20] called Gradient Signal Preservation (GraSP) which required computation of a Hessian-gradient product $\mathbf{Hg}$ using a batch of training data after which m is constructed by tresholding network scores $w \odot \mathbf{Hg}$. The score computation was theoretically motivated through linearised training dynamics [Wan+20, Chap. 4.1] which have been described for wide networks using a kernel over training data [Lee+19]. Another approach also analysed gradient flow at initialization, but this method named Iterative Synaptic Flow Pruning (SynFlow) produced by Tanaka, Kunin, Yamins, and Ganguli [Tan+20], did not use any training data. The researchers identified a key problem with aggressive pruning that especially must be addressed when designing a priori pruning mechanisms: *layer-collapse*, wherein an entire layer is pruned. Using avoidance of this problem a guiding principle, synaptic saliency, a score metric which provenly did not induce layer-collapse. The ticket was then constructed by optimising the weights against a synaptic saliency loss function based on layer-wise products of absolute weight values.

Both methods were tested on CV tasks and achieved comparative performance to standard LTH with SynFlow outperforming all other methods at extreme compression ratios where GraSP and standard magnitude-based LTH suffer from layer-collapse [Tan+20, Chap. 7]. For multiple architectures, especially of the ResNet type, GraSP is, however, slightly SOTA at more reasonable compression ratios of $\times 100$ to $\times 10$[Wan+20, Tab. 4][Tan+20, Fig. 6]. Also, the pruning algorithm Single-shot Network Pruning (SNIP) contains ideas used in both these methods and performs similarly to GraSP also using training data, but is not formulated in the LTH context [LAT19].

Open Problems

- Sparsity not computationally optimal
- LTH Not proven [Tan+20] proves other things. Look at https://arxiv.org/pdf/1810.02340.pdf

References

[Bla+20] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. "What is the State of Neural Network Pruning?" In: *Proceedings of Machine Learning and Systems*. Ed. by I. Dhillon, D. Papailiopoulos, and V. Sze. Vol. 2. 2020, pp. 129–146. URL: `https://proceedings.mlsys.org/paper/2020/file/d2ddea18f00665ce8623e36bd4e3c7c5-Paper.pdf`.

[FC19] Jonathan Frankle and Michael Carbin. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks". In: *International Conference on Learning Representations*. 2019. URL: `https://openreview.net/forum?id=rJl-b3RcF7`.

[Fra+20] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. "Linear Mode Connectivity and the Lottery Ticket Hypothesis". In: *ArXiv* abs/1912.05671 (2020).

[Han+15] Song Han, Jeff Pool, John Tran, and William J. Dally. "Learning Both Weights and Connections for Efficient Neural Networks". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 1135–1143.

[Lan20] Robert Tjarko Lange. "The Lottery Ticket Hypothesis: A Survey". In: *Rob's Homepage* (June 27, 2020). URL: `https://roberttlange.github.io/posts/2020/06/lottery-ticket-hypothesis` (visited on 06/10/2022).

[LAT19] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. "SNIP: Single-shot Network Pruning based on Connection Sensitivity". In: *ArXiv* abs/1810.02340 (2019).

[LDS89] Yann LeCun, John S. Denker, and Sara A. Solla. "Optimal Brain Damage". In: *NIPS*. 1989.

[Lee+19] Jaehoon Lee et al. "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent". In: *NeurIPS*. 2019.

[Li+16] Hao Li et al. "Pruning Filters for Efficient ConvNets". In: *CoRR* abs/1608.08710 (2016). arXiv: `1608.08710`. URL: `http://arxiv.org/abs/1608.08710`.

[RFC20] Alex Renda, Jonathan Frankle, and Michael Carbin. "Comparing Rewinding and Fine-tuning in Neural Network Pruning". In: *ArXiv* abs/2003.02389 (2020).

[Tan+20] Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, and Surya Ganguli. "Pruning neural networks without any data by iteratively conserving synaptic flow". In: *ArXiv* abs/2006.05467 (2020).

[Wan+20] Chaoqi Wang, ChaoQi Wang, Guodong Zhang, and Roger B. Grosse. "Picking Winning Tickets Before Training by Preserving Gradient Flow". In: *ArXiv* abs/2002.07376 (2020).

[You+20] Haoran You et al. "Drawing early-bird tickets: Towards more efficient training of deep networks". In: *ArXiv* abs/1909.11957 (2020).

[Zmo+19] Neta Zmora et al. "Neural Network Distiller: A Python Package For DNN Compression Research". In: (Oct. 2019). URL: `https://arxiv.org/abs/1910.12232`.