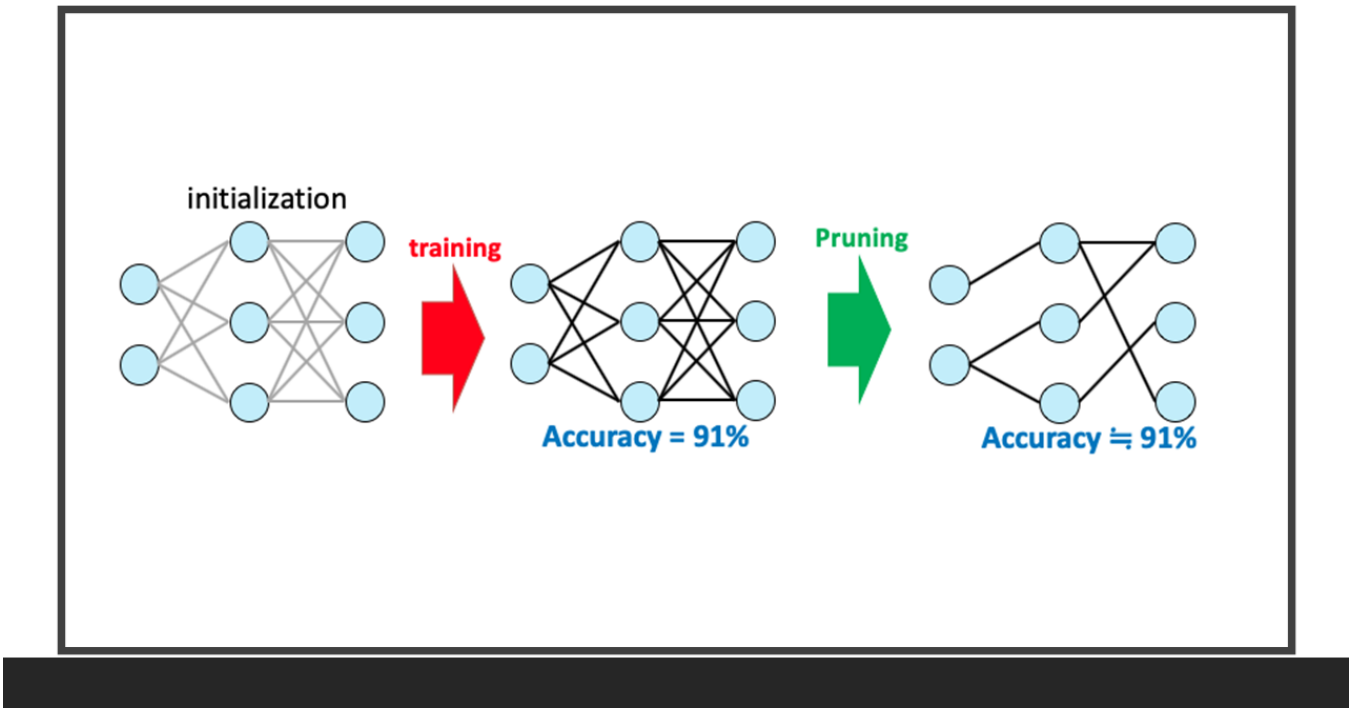
One of the Most Important Papers in Modern Machine Learning was Published Almost a Year Ago

The Lottery Ticket Hypothesis changed the way we think about training machine learning models.

Jesus Rodriguez
May 11 · 6 min read ★



Source: https://medium.com/analytics-vidhya/rigging-the-lottery-training-method-of-high-speed-and-high-precision-sparse-networks-without-2340b0461d55

Every once in a while we encountered some new research ideas that challenge the foundation of the core principles of machine learning. That was a case almost a year ago when researchers from the Massachusetts Institute of Technology(MIT) published the Lottery Ticket Hypothesis which became one of the most important papers of 2019. Today, I would like an analysis I published during that time.

Training machine learning models is one of the most challenging and computationally

training of machine learning models under the single axiomatic assumption that training should cover the entire model. About a year ago, AI researchers from the Massachusetts Institute of Technology(MIT) published a paper that challenges that assumption and proposes a smarter and simpler way to train neural networks by focusing on subsets of the model. Within the AI community, the MIT thesis has come to be known by the catchy name of the Lottery Ticket Hypothesis.

The process of training machine learning models is one of the areas in which data scientists often face the compromise between theory and the constraints of real world solutions. More often than not, a neural network architecture that seems ideal for a specific problem can't be fully implemented because the cost of training would be prohibited. Typically, the initial training of a neural networks requires large datasets and days of expensive computation usage. The result are very large neural network structures with connections between neuros and hidden layers. This structure often needs to be subjected to optimization techniques to remove some of the connections and adjust the size of the model.

One question that bothered AI researchers for decades is whether we actually need those large neural network structures to begin with. Obviously, if we connect almost every neuron in an architecture, we are likely to achieve a model that can perform the initial task but the cost might be prohibited. Couldn't we start with smaller, leaner neural network architectures to begin with? This is the essence of the Lottery Ticket Hypothesis.

The Lottery Ticket Hypothesis

Using an analogy from the gambling world, the training of machine learning models is often compared to winning the lottery by buying every possible ticket. But if we know how winning the lottery looks like, couldn't we be smarter about selecting the tickets?

In machine learning models, training processes produced large neural network structures that are the equivalent to a big bag of lottery tickets. After the initial training, models need to undergo optimization techniques such as pruning that remove unnecessary weights within the network in order reduce the size of the model without sacrificing performance. This is the equivalent of searching for the winning tickets in the bag and getting rid of the rest. Very often, pruning techniques end up producing neural network structures that are 90% smaller than the original. The obvious question

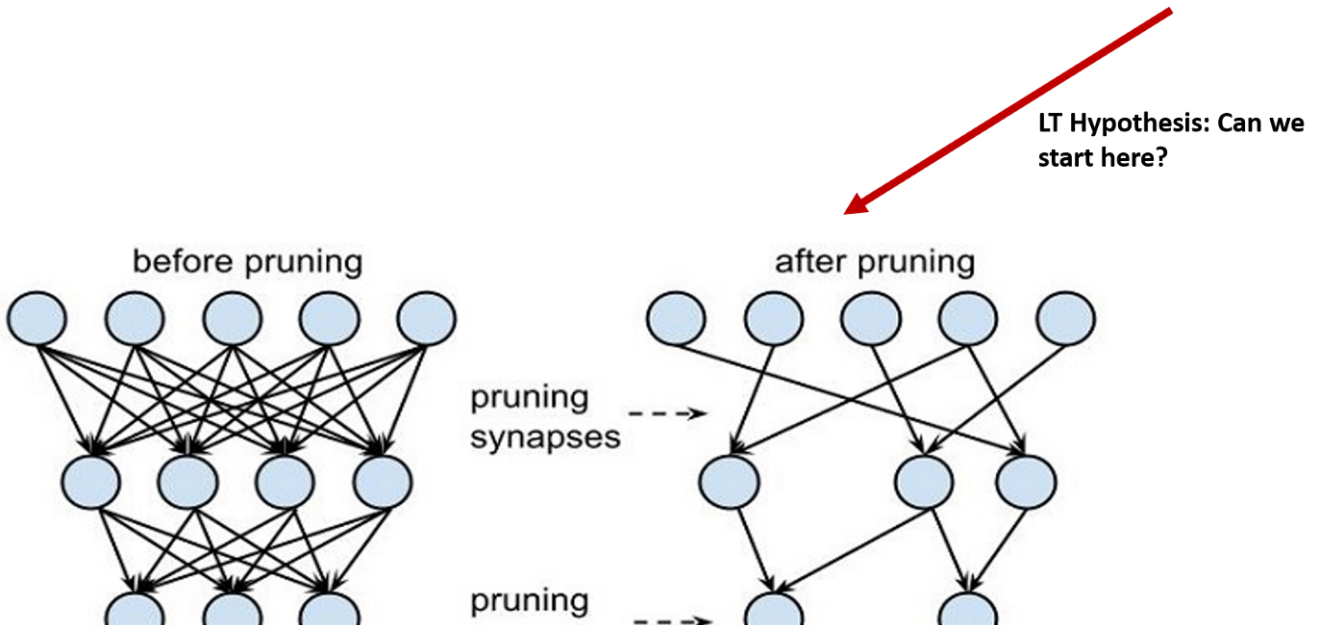
practical experiences in machine learning solutions show that the architectures uncovered by pruning are harder to train from the start, reaching lower accuracy than the original networks. So you can buy a big bag of tickets and work your way to the winning numbers but the opposite process is too hard. Or so we thought 😉

The main idea behind MIT's Lottery Ticket Hypothesis is that, consistently, a large neural network will contain a smaller subnetwork that, if trained from the start, will achieve a similar accuracy than the larger structure. Specifically, the research paper outlines the hypothesis as following:

· ***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.*

In the context of the paper, the small subnetwork is often referred to as the *winning ticket*.

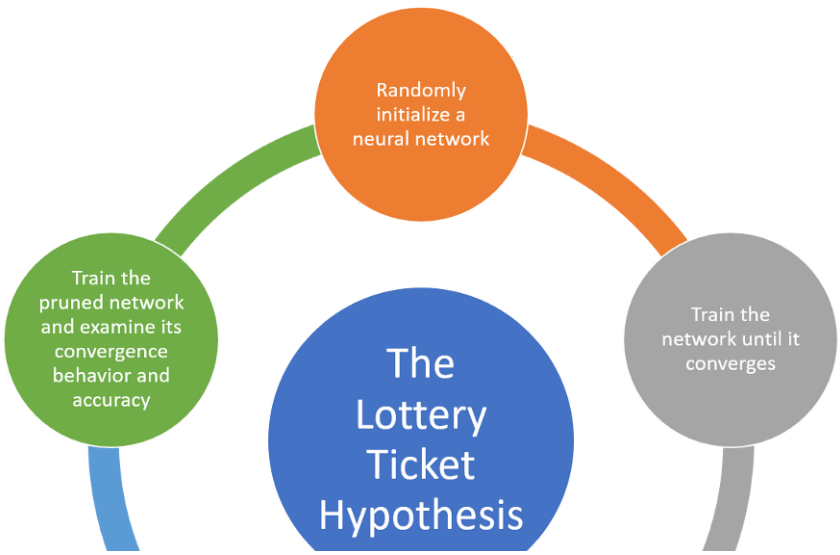
Consider a neural network in the form of *f(t, a, p)* in which *t= training time, a= accuracy and p= parameters*. Now consider *s* being the subset of all trainable neural networks from the original structure generated by pruning process. The Lottery Ticket Hypothesis tells us that there is *a f'(t', a', p') € s* in a way that *t' <= t, a'>= a and p' <= p*. In simple terms, conventional pruning techniques unveiled neural network structures that are smaller and simpler to train than the original.



If the Lottery Ticket Hypothesis is true, then the next obvious question is to find the strategy to identify the winning ticket. The process for this involves an iterative process of smart training and pruning which can be summarized in the following five steps:

1. Randomly initialize a neural network.
2. Train the network until it converges.
3. Prune a fraction of the network.
4. To extract the winning ticket, reset the weights of the remaining portion of the network to their values from (1) — the initializations they received before training began.
5. To evaluate whether the resulting network at step (4) is indeed a winning ticket, train the pruned, untrained network and examine its convergence behavior and accuracy.

This process can be run one time or multiple. In the one-shot pruning approach, the network is trained once, p% of weights are pruned, and the surviving weights are reset. Although one-shot pruning is certainly effective, the Lottery Ticket Hypothesis paper showed the best results when the process was applied iteratively over n rounds; each round prunes p1/n% of the weights that survive the previous round. However, one-shot pruning often produced very solid results without the need of computationally expensive training.



The MIT team tested the Lottery Ticket Hypothesis across a group of neural network architectures and the results showed that the pruning method was not only able to find to optimize the architecture but to find the winning ticket. Look at the following chart as

Observe two main things in these results. The winning tickets, without the remaining redundancy of the wide network, train faster than the wide network. In fact, the skinnier they are, the faster they train (within reason). However, if you reinitialize the networks' weights randomly (control), the resulting nets now train slower than full network. Therefore, pruning is not just about finding the right architecture, it's also about identifying the 'winning ticket', which is a particularly luckily initialized subcomponent of the network.

Based on the experimental results, the MIT team expanded their initial hypothesis with what they called the Lottery System Conjecture which express the following:

· **The Lottery Ticket Conjecture:** Returning to our motivating question, we extend our hypothesis into an untested conjecture that SGD seeks out and trains a subset of well-initialized weights. Dense, randomly-initialized networks are easier to train than the sparse networks that result from pruning because there are more possible subnetworks from which training might recover a winning ticket.

The conjecture seems to make sense conceptually. The larger the pool of pruned subnetworks, the better chances to find a winning ticket.

The Lottery Ticket Hypothesis could become one of the most important machine learning research papers of recent years as it challenges the conventional wisdom in neural network training. While pruning typically proceeds by training the original network, removing connections, and further fine-tuning, the Lottery Ticket Hypothesis tells us that optimal neural network structures can be learned from the start.