



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Achieving results with untrained neural networks using Supermasks

Independent Study Proposal

im Arbeitsbereich Knowledge Technology, WTM

Prof. Dr. S. Wermter

Department Informatik

MIN-Fakultät

Universität Hamburg

vorgelegt von

Vincent Rolfs

am

09.11.2020

Gutachter: Prof. Dr. S. Wermter

Dr. M. Kerzel

Vincent Rolfs

Matrikelnummer: 6789106

Contents

1	Introduction	1
2	Approach	2
2.1	Step one: Finding the pruning treshold	2
3	Evaluation	3
3.1	Dataset	4
3.2	Baseline neural network	4
3.3	Training and early stopping	4
3.4	Training and early stopping	5
3.5	Lottery tickets	5
3.6	Hyperoptimization	5
4	Conclusion	6
	Bibliography	6

1 Introduction

Interest in decreasing the size of neural networks that are used to solve certain problems has existed since at least the year 1990 [1]. Different methods have been proposed that can be used to achieve a decrease in network size of up to 90% while retaining performance [1] [4] [5] [7].

This has multiple advantages, which can be classified according to the type of pruning that is applied. If the network is pruned after training while otherwise keeping the trained weights constant, this may result in reduced storage size, less energy consumption and faster computation during the application phase [3]. If it is possible to retrain the smaller network from scratch, this may result in less overfitting because the number of parameters has decreased [4]. And if a smaller but still effective network topology can be chosen *before training*, this can reduce the training time, because less parameters have to be optimized.

However, conventional pruning techniques have the problem that, when re-training from scratch on the smaller topology, the performance gets considerably worse [3]. These pruning techniques are only amenable for pruning the weights computed by an initial training run, and do not work well if the aim is to retrain once again on the smaller topology. Therefore, they cannot be used to choose a good topology before the training starts, because the correct parameters for the smaller network can only be discovered by training the larger network.

In contrast to this negative state of affairs, in their paper "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks" [3], the authors present a hypothesis that essentially claims that pre-training pruning nonetheless has a lot of potential:

The Lottery Ticket Hypothesis [3]. *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.*

The authors further show that the subnetwork mentioned in the hypothesis can be uncovered by training the dense network and then setting a percentage $p\%$ of the parameters with smallest magnitude to zero (and freezing them, so that they are not trained anymore). If the remaining parameters are then set to their initial values (the random values they had before the training started), and one retrains the smaller network starting with these initial values, then the training time will usually be less and the test performance better compared to the original, dense, trained network [3].

It is imperative here that the non-frozen parameters are reset to their *initial*, random values, not newly chosen random values. If the latter is chosen, the performance is much worse and the "Lottery Ticket" has not been uncovered [3]. In other words, we should not only search for efficient network topologies, but also for good initial values.

This curious fact has been further investigated by Zhou et al. [9]. They realized that the subnetworks obtained through pruning show a performance that is

already significantly higher than chance *before they are trained*. More specifically, a randomly-initialized, dense network is first trained. The trained parameters are then ranked by a value given by

$$\text{sign}(w_{\text{initial}}) \cdot w_{\text{trained}}$$

so that parameters with large magnitude that retained their sign are ranked high. A percentage $p\%$ of the lowest-ranking parameters is set to zero and frozen, the rest is reset to a constant with the same sign as their initial value. The resulting network is **not** trained again, rather it is being evaluated directly. Note that the values of the parameters at that point come directly from the initial random initialization, the only training information comes from deciding which of these random values to set to zero. Nonetheless, this technique achieves a remarkable 86% accuracy on MNIST, and 41% on CIFAR-10. The authors call the corresponding 0-1-masks "Supermasks".

In this paper, we provide the following contributions to the new theory of Lottery Ticket pruning:

- We describe a complete methodology for applying Lottery Ticket pruning to a trained neural network. The methodology crucially uses the Supermask idea for finding the pruning threshold.
- We show that the methodology applied on MNIST outperforms hyperoptimization of the network layer sizes: Our method can yield a significantly smaller size while retaining the same accuracy and training speed.
- As part of this, we reproduce the main results regarding Lottery Tickets [3] and Supermasks [9] on MNIST: Supermasks do indeed yield untrained networks with higher-than-random accuracy, which can be trained to achieve faster training and higher accuracy with less parameters, compared to the unpruned networks.

2 Approach

In this section we describe a complete methodology for applying Lottery Ticket pruning to a trained neural network. We assume that we have a fully connected multilayer perceptron which has already been trained to satisfaction and should be pruned now. Importantly, we also assume that the weights with which the network had been initialized before training have been saved.

2.1 Step one: Finding the pruning threshold

As the first step of the methodology, we want to find an optimal pruning threshold. For this, we use the theory of Supermasks [9], which essentially says that during the procedure for finding Lottery tickets outlined in [3], the candidates that are found achieve a performance that is already significantly better than random *before they are trained*. It is therefore natural to simply use the candidate that achieves the highest *untrained* performance as the Lottery ticket, which will be trained to get a

highly performant pruned network. This is useful because the computation of the different candidates as well as their evaluation is not computationally expensive, since no training has to take place.

We now give more details about this step. Following the Lottery ticket approach, we want to take our initial network parameters, set some of them to zero and freeze them there, and train the resulting network, which therefore effectively has a smaller size. In the original Lottery ticket paper [3], this is done by setting all parameters to zero which had a small magnitude after training. How many parameters are set to zero depends on the specific threshold – a higher threshold corresponds to more aggressive pruning. For example, suppose the initial parameters are 0.1, 0.2, 0.3 and the parameters after training are $-0.2, 0.05, 0.3$. If we prune with a threshold of $t = 0.1$, we would get the parameters 0.1, 0, 0.3. Starting with these values, we would then train the network, but the second value would be frozen at 0.

The Supermask approach [9] provides two further insights. First, the authors show that instead of setting those parameters to zero whose magnitude after training $|w_{\text{trained}}|$ is minimal, one gets better results by setting those parameters to zero for which the value

$$\text{sign}(w_{\text{initial}}) \cdot w_{\text{trained}}$$

is minimal. In other words, we keep those values which both have a large magnitude after training and retain their sign, and set the other parameters to zero. Following the example above, if the initial parameters are 0.1, 0.2, 0.3 and the parameters after training are $-0.2, 0.05, 0.3$, then pruning with a threshold $t = 0.1$ yields 0, 0, 0.3 – in this case, the first value is pruned as well, because it changed signs.

The second, and most crucial, insight in [9] is that even though the parameters we get through this method are just a version of the initial, random parameters with some set to zero, they nonetheless show a surprising performance when evaluated. This is where the term Supermask comes from: We take the initial, random parameters and multiply them with a mask of zeros and ones, and suddenly we get surprisingly great performance: In [9], the authors report an accuracy of 86% on MNIST.

This leads us to a straightforward way of choosing a threshold: We simply compute the masked networks for different thresholds, and then (without training!) check their performance on the validation set. We then choose the threshold the yielded the highest performance. More to the point, we choose the masked network corresponding to that threshold, and train it to yield our pruned network.

3 Evaluation

Our basic setup follows closely the one in [9].

3.1 Dataset

All experiments are done with the well established MNIST dataset [2]. It contains a collection of grayscale images of size 28×28 , which each correspond to exactly one of ten possible digits 0 to 9. The dataset has already been split into a training and a testing part; it contains 60.000 training examples and 10.000 testing examples. We are using the training examples for training the network, and the testing examples for validation of the network performance during training.

3.2 Baseline neural network

In order to apply the pruning techniques and supermasks, we need a baseline neural network. As in [9], we chose a fully connected neural network with three layers. The input size is $28 \times 28 = 784$, the first layer has size 200, the second has size 30 and the output layer has size 10, since there are ten classes in the dataset. Compared to [9] we chose a smaller value for the first and second layer size: In that paper, the sizes are 300 and 100 respectively. This was done because early experimentation with simple hyperoptimization showed that we could achieve the same performance with our smaller architecture during training. Since we are testing pruning capabilities, it is good to start with a network size that cannot be reduced trivially.

3.3 Training and early stopping

Since we need a meaningful trained baseline that does not overfit, we use the same early stopping criterion as in [9]. After saving the initial, random parameters, the neural network is first trained for 20.000 iterations, using the Adam optimizer [6] with learning rate 0.0012 (as in [9]) and the PyTorch cross entropy loss [8]. Each iteration corresponds to the processing of one batch of training data; one batch contains 60 training examples. This initial training certainly leads to overfitting, which can be seen by plotting the validation loss, which is evaluated every 100 iterations.

Because of this, we then record at which iteration the minimal validation loss was achieved. For all random seeds, this always happens around iteration 5000. We then retrain the network from scratch – starting from the saved, initial parameters – for that many iterations. While the training will never go exactly the same the second time around due to randomness, this seems to work well in practice both in our experiments and in [9], yielding a trained, non-overfit baseline network.

The whole training procedure is performed ten times using different random seeds. Before training, the validation accuracy is 0.101 ± 0.016 . After training, the validation accuracy is 0.977 ± 0.001 .

3.4 Training and early stopping

The next step of our methodology is the application of the supermasks. To that end, we apply the following procedure taken from [9]. First, we reset the baseline network to its initial parameters, which were randomly initialized, but saved before the training started. Next, we choose a threshold t and iterate through the parameters w_1, \dots, w_n of the network. We set a new value for each parameter according to the formula

$$w_i^{\text{new}} := \begin{cases} w_i^{\text{initial}} & \text{if } \text{sign}(w_i^{\text{initial}}) \cdot w_i^{\text{trained}} \geq t \\ 0 & \text{otherwise.} \end{cases}$$

In effect, we keep parameters the same which had a large magnitude after training, and still had the same sign as the initial value. The other parameters are set to zero. The resulting neural network is identical to the initial, randomly initialized network, with some parameters set to zero. Effectively, a zero-one mask was applied to the initial, untrained network – the so-called Supermask.

We then evaluate the accuracy of this masked network on the validation set. Of course, the performance depends on the threshold: For very large thresholds, all parameters are set to zero, so the accuracy will be close to 0.1. For very small (negative) thresholds, all parameters are kept at their initial value, which was randomly chosen at the beginning, so we still expect an accuracy close to 0.1. For intermediate thresholds, something remarkable happens: The masked network achieves a validation accuracy that is significantly better than 0.1, even though all the parameter values come from a random distribution, and the only training information lies in the decision which parameters to set to zero. The performance for different thresholds of the masked networks across all ten random seeds are shown in XXX.

3.5 Lottery tickets

The next step is to apply the knowledge we have about the performance of the masked networks. We wish to find a "Lottery Ticket" – this is essentially a masked network which has been trained, where the parameters that are zero are frozen during training. When trying to prune a network using lottery tickets, one is faced with the one question that always arises during pruning: How small should the resulting network be? Here, we use the validation results of the supermasks as a guideline: Out of all the masked networks resulting from different thresholds, we choose the masked network that has the highest validation accuracy before training. We then train this masked network for 5000 iterations and observe the validation loss.

3.6 Hyperoptimization

In order to have a comparization for the quality of the pruning produced by the lottery ticket approach, we employ a simple hyperoptimization scheme to optimize

the network size. Whereas the initial sizes for layers 1 and 2 were 200 and 30 respectively, for the hyperoptimization we allow all combinations of 50, 100, 150 and 5, 15, 25 respectively. For each of these size combinations, we take a randomly initialized network of that size, and train it for 5000 iterations, observing the validation loss.

4 Conclusion

References

- [1] Yann Le Cun, John S. Denker, and Sara A. Solla. “Optimal Brain Damage”. In: *Advances in Neural Information Processing Systems 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, pp. 598–605. ISBN: 1558601007.
- [2] Li Deng. “The mnist database of handwritten digit images for machine learning research”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142.
- [3] Jonathan Frankle and Michael Carbin. “The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks”. In: *International Conference on Learning Representations*. 2019.
- [4] Song Han et al. “Learning both Weights and Connections for Efficient Neural Network”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015. URL: <https://proceedings.neurips.cc/paper/2015/file/ae0eb3eed39d2bcef4622b2499a05fe6-Paper.pdf>.
- [5] Babak Hassibi and David Stork. “Second order derivatives for network pruning: Optimal Brain Surgeon”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Hanson, J. Cowan, and C. Giles. Vol. 5. Morgan-Kaufmann, 1993. URL: <https://proceedings.neurips.cc/paper/1992/file/303ed4c69846ab36c2904d3ba8573050-Paper.pdf>.
- [6] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: <http://arxiv.org/abs/1412.6980>.
- [7] Hao Li et al. “Pruning Filters for Efficient ConvNets”. In: *International Conference on Learning Representations*. 2017.
- [8] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.

- [9] Hattie Zhou et al. “Deconstructing Lottery Tickets: Zeros, Signs, and the Supermask”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019, pp. 3597–3607.