ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

Pruning Neural Networks with Supermasks

Vincent Rolfs, Matthias Kerzel and Stefan Wermter *

University of Hamburg - Department of Informatics
Vogt-Koelln-Str.30, 22527 Hamburg - Germany

Abstract.

The Lottery Ticket hypothesis by Frankle and Carbin states that a randomly initialized dense network contains a smaller subnetwork that, when trained in isolation, will match the performance of the original network. However, identifying this pruned subnetwork usually requires repeated training to determine optimal pruning thresholds. We present a novel approach to accelerate the pruning: By methodically evaluating different Supermasks, the threshold for selecting neurons as part of a pruned Lottery Ticket network can be determined without additional training. We evaluate the method on the MNIST dataset and achieve a size reduction of over 60% without a drop in performance.

1 Introduction

Interest in decreasing the size of neural networks has existed since at least the year 1988 [1, 2]. Different methods have been proposed that can be used to achieve a decrease in network size of up to 90% while retaining performance [3, 4, 5]. Size reduction has multiple advantages: 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 [6]. If it is possible to retrain the smaller network from scratch, this may result in less overfitting because the number of parameters has decreased [7]. Moreover, if a smaller but still effective network topology can be chosen *before training*, this can reduce the training time because fewer parameters have to be optimized.

One state-of-the-art pruning approach, by Frankle and Carbin [6], is based on their *Lottery Ticket hypothesis*, which states that a randomly initialized dense network contains a subnetwork that can be trained in isolation and will match the performance of the original network. The subnetwork can be uncovered by training the dense network and setting a percentage of the parameters with the smallest magnitude to zero, and freezing them. If the remaining parameters are then set to their initial values, and one retrains the smaller network, then the training time and test performance will usually improve. However, selecting the correct neurons in this step requires repeated training and evaluation of subnetwork candidates.

To avoid this repeated training, we adopt a finding from Zhou et al. [8], who show that these subnetworks perform significantly better than chance *before they*

*The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169).

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

are trained. Zhou et al. take a randomly-initialized and trained dense network; the trained parameters are ranked by a value given by $\text{sign}(w_{\text{initial}}) \cdot w_{\text{trained}}$ so that parameters with a 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 can be evaluated directly. Zhou et al. call the corresponding 0-1-masks "Supermasks".

We propose a novel pruning method of using Supermasks to determine pruning thresholds for Lottery Tickets without additional training. The values of the parameters come directly from the initial random initialization; the only training information comes from deciding which of these random values to set to zero. We further 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.

2 Methodology

We describe a methodology for neural network pruning based on the Lottery Ticket hypothesis and Supermasks. We assume that we have a fully connected multilayer perceptron that has been trained and should be pruned. Importantly, we also assume that the weights with which the network had been initialized before training have been saved.

Following the Lottery Ticket hypothesis, we take our initial network parameters, set some of them to zero and freeze them, and then train the resulting network. The result will effectively have a (much) smaller size. There are different methods for choosing which of the initial parameters to prune: In [6], this is done by selecting all parameters which had a small magnitude after training. How many parameters are set to zero depends on the specific threshold. In [8], the authors show that instead of setting those parameters to zero whose magnitude after training is minimal (i.e., $|w_{trained}|$ is minimal), one gets better results by setting those parameters to zero for which the value $\text{sign}(w_{initial}) \cdot w_{trained}$ is minimal. In other words, we keep those values that have both a large magnitude after training and retain their sign and set the other parameters to zero. In any case, it is necessary to choose a good threshold to apply the method effectively.

2.1 Selecting the threshold

We describe a novel method for selecting a good pruning threshold based on insights from [8]: When choosing the parameters as outlined above, the result is just a version of the initial, random parameters with some set to zero. However, without further training, the resulting networks nonetheless already perform surprisingly well. 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 achieve surprisingly good performance.

This leads us to our novel way of threshold selection: We compute the masked networks for different thresholds and then (without training!) check their per-

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

formance on the validation set. We choose the threshold that 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. Both the computation of the masked networks and the evaluation of them are computationally cheap since no training is required.

2.2 Summary of steps

The algorithm for applying Lottery Ticket pruning is as follows:

1. Take a fully connected multilayer perceptron, initialize it with random values and save these initial values.
2. Train the network to satisfaction.
3. For many different thresholds t, compute the parameters for a masked network M_t using the formula
$$w_i^{\text{new}} := \begin{cases} w_i^{\text{initial}} & \text{if } \operatorname{sign}\left(w_i^{\text{initial}}\right) \cdot w_i^{\text{trained}} \geq t, \\ 0 & \text{otherwise.} \end{cases}$$
4. Evaluate all the networks M_t on a validation set and choose the best-performing network M_*.
5. Freeze all the parameters of M_* which are exactly zero, so that they don't change during training.
6. Train M_* to satisfaction.

3 Experiment and Evaluation

To evaluate our method, we apply it to the well-established MNIST dataset [9] ten times using different random seeds. The dataset contains a collection of 28×28 grayscale images of single digits. It contains 60.000 training examples and 10.000 testing examples, which are used for validation. For each of these runs, we also perform a hyperoptimization grid search with the goal of reducing the layer size as a comparison.

3.1 Baseline neural network, training and early stopping

We chose a fully connected neural network with three layers as a baseline architecture. 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 the dataset has ten classes.

Because we need a meaningful trained baseline that does not overfit, we use early stopping. After saving the initial, random parameters, the neural network is first trained for 20.000 iterations, using the Adam optimizer [10] with learning rate 0.0012, a batch size of 60 (as in [8]) and cross-entropy loss. As in [8], we determine the number of training epochs according to the validation loss and then retrain the network from scratch starting from the saved, initial parameters for as many epochs as determined. For all random seeds, the minimal validation loss is achieved around iteration 5000. Across random seeds, the validation accuracy is 0.101 ± 0.016 before training and 0.977 ± 0.001 after training, where the $\pm$ sign denotes the standard deviation.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

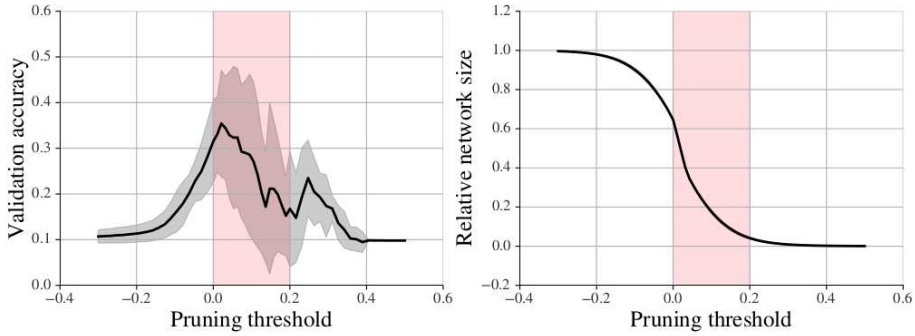

Fig. 1: Left: The validation accuracies of the untrained, masked networks for different thresholds aggregated across all ten seeds. Even though the masked networks were never trained explicitly, they reach an accuracy significantly better than chance level (0.1). Right: Relative sizes of the networks after pruning with different thresholds. The variance across seeds is too small to be visible. The colored area indicates the thresholds that are part of the evaluation.

3.2 Supermasks

For the computation of the masked networks, we consider thresholds t from the set $\{0, 0.01, 0.02, \ldots, 0.2\}$. As described above, we take the initial parameters of the network and set some of them to zero, using the formula described in Sect. 2.2. The exact size reduction achieved by a particular threshold depends on the initialization, but the variance is low across seeds: $t = 0$ results in a relative size of $64.6\% \pm 0.4$, while $t = 0.2$ results in $4.1\% \pm 0.6$. The size for different thresholds is visualized in Figure 1 (right). For each of the ten networks, we determine the threshold which resulted in the highest validation accuracy. Here, the variance is high across the different seeds. The mean threshold is found to be at 0.049 ± 0.049, resulting in a relative network size of $38.7\% \pm 17.9$. The resulting accuracy of these masked and otherwise untrained networks is $42.6\% \pm 15.2$, where the smallest value observed is 19.1% and the largest is 70.0%. Fig. 1 (left) shows the accuracies obtained by pruning with different thresholds. This result is in line with [8]: Masking the initial, random values does indeed lead to high accuracy, even before training.

3.3 Resulting pruned networks

After selecting a threshold using the Supermask evaluation, the corresponding masked network is trained (steps 5 and 6 of our methodology). For comparison, we also use a hyperoptimization grid search which evaluates smaller values for the baseline network's layer sizes. While the baseline uses sizes of 200 and 30 neurons for the first and second layers, respectively, the hyperoptimization additionally considers all possible combinations of $(50, 100, 150)$ for the first layer and $(5, 15, 25)$ for the second layer. We are thus comparing each Lottery Ticket

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

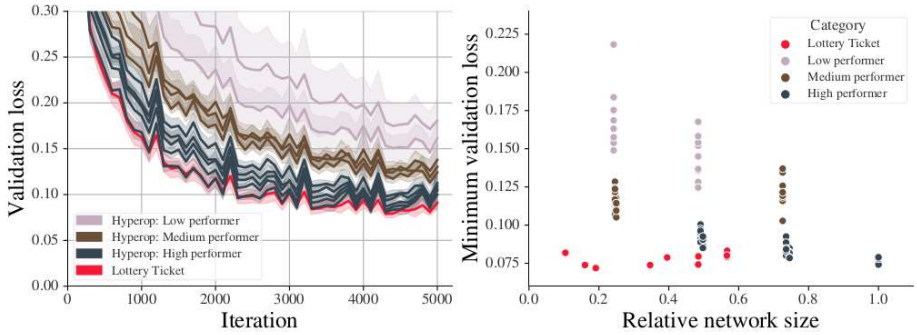

Fig. 2: Left: The validation loss during training of pruned networks and hyperoptimized networks, evaluated every 100 iterations across all ten random seeds. The hyperoptimized networks are grouped into three categories based on their training performance. The performance of the pruned networks is at least as good as all other hyperoptimizations. Right: The relative network size and minimum validation loss of the pruned networks and the hyperoptimizations. All 10 seeds are visualized. It is apparent that the Lottery Tickets tend to outperform the hyperoptimizations both regarding size and validation performance.

network to $1 + 3 \cdot 3 = 10$ other architectures.

To train the pruned network, we take the best-performing masked network from the previous step and train it for 5000 iterations. For comparison, we also train the ten networks with different layer sizes for 5000 iterations, starting from random initializations. The validation loss for all runs is visualized in Figure 2 (left). We have grouped the different hyperoptimizations into three groups according to their performance. The graph shows that the pruned network trains at least as quickly and effectively as all the hyperoptimization networks. After 5000 iterations, the pruned network achieves a validation loss of 0.091 ± 0.007 while the original baseline network architecture (without any layer size reduction) achieves 0.109 ± 0.019. The "High performer" group of hyperoptimizations as a whole achieves a final validation loss of 0.106 ± 0.016. At the same time, the pruned networks that were found by our method are generally smaller (pruned more aggressively) than the results from the hyperoptimization, as shown in Fig. 2 (right). The optimal Supermasks, and therefore also the pruned networks, have a relative network size of $38.7\% \pm 17.9$.

Our method produces networks that are comparable in size to the low-performing hyperoptimizations, while they tend to outperform even the high-performing hyperoptimizations. See Fig. 2 (right), where the minimum validation loss is compared to the sizes of the tested network. The graph shows that the pruned networks based on Lottery Tickets tend to achieve a better performance while simultaneously being smaller in size compared to the hyperoptimizations.

ESANN 2021 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Online event, 6-8 October 2021, i6doc.com publ., ISBN 978287587082-7.
Available from http://www.i6doc.com/en/.

4 Conclusion

We present a novel methodology for pruning neural networks based on the Lottery Ticket hypothesis and Supermasks. Our method determined the threshold for selecting neurons to be part of a pruned Lottery Ticket network by evaluating different Supermasks. By using Supermasks, the network alleviates the need for additional training time for each evaluated subnetwork. Our result outperforms hyperoptimization grid search on a dense neural network trained on the MNIST dataset by yielding on average 60% smaller networks without a reduction in classification accuracy. In future work, we will conduct experiments on larger datasets and comparisons with other pruning methods to further improve the understanding of the merits of Lottery Ticket-based pruning.

References

[1] Michael C Mozer and Paul Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In *Advances in neural information processing systems*, pages 107–115, 1989.

[2] Yann Le Cun, John S. Denker, and Sara A. Solla. *Optimal Brain Damage*, pages 598–605. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[3] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *arXiv preprint arXiv:2003.03033*, 2020.

[4] Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. In S. Hanson, J. Cowan, and C. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5. Morgan-Kaufmann, 1993.

[5] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and H.P. Graf. Pruning filters for efficient convnets. In *International Conference on Learning Representations*, 2017.

[6] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

[7] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[8] Hattie Zhou, Janice Lan, Rosanne Liu, and Jason Yosinski. Deconstructing lottery tickets: Zeros, signs, and the supermask. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 3597–3607. Curran Associates, Inc., 2019.

[9] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.