COMP 576 Final Project Proposal

[redacted]

November 2020

1 Abstract

There has been great new found interest in neural network pruning techniques sparked by Lottery Ticket hypothesis [1], which shows that there exists a sub-network within any neural network that when trained to full accuracy, achieves comparable and even better performance. This line of methods can reduce the parameter counts of trained networks by over 90%, decreasing memory requirements and improving inference speed without compromising accuracy. There has been numerous pruning techniques proposed on this topic that aims at more efficient pruning and better final accuracy.

We present an in depth analysis on the different design choices between these proposed pruning techniques and present two new algorithms, Double-pruning and Scored-pruning, that aims to solve the problem of premature pruning and allows reintroduction of lost weights into the neural network.

2 Background/ Motivation

In modern Deep Learning research, model size is strongly correlated with the predictive power of the model. The Lottery Ticket Hypothesis challenged this idea by proposing that there exists a smaller network within the larger network in which the same accuracy can be achieved by simply training the smaller network, the "winning ticket". If this hypothesis is true and researchers can develop efficient algorithms to find such a smaller network, then we can save a huge amount of computation power during training. Furthermore, reducing the model size would enable training to be performed in resource-constrained scenarios, such as on mobile devices, which can further unleash the power of Deep Learning. Therefore, this work has sparked many new innovations in the field. More hypotheses have been made in light of this, including but not limited to the Strong Lottery Ticket Hypothesis, which claims that if we create a large enough network we can prune to find a network with the same accuracy as the full network even without training. Theorists are attempting to bound the size of the original network where we can find such an efficient sub-network while practitioners are experimenting various pruning techniques in different settings such as CNN for applications in Computer Vision etc.

3 State of the art methods/ Tension/ Detailed related work

Currently there are numerous pruning technologies. The most basic being pruning by the norm of the weight. We can prune by different criteria (l_1, l_2 or even more complicated weights), and do one-shot or multi-shot pruning. We also have the choice to keep the pruned weights, rewind them to initialization, or randomly re-initialize them. Also there is the choice between structured pruning and unstructured pruning. For example, Li [2] proposed pruning by filters in CNN for optimal performance.

Current state-of-the-art approach suggests that we can keep the trained weights in multi-shot pruning to achieve best performance. While unstructured pruning achieves better performance for Multi-Layer Perceptrons, the weights learned here are sparse in nature and hence less efficient in decreasing memory cost. On the other hand, structured pruning achieves the added benefit of much efficient memory use while taking minor performance penalty.

4 Broader Scope and Goal

Network pruning, a systematic way of removing parameters in Neural Networks, has drawn huge amount of attentions in the research communities. Lottery Ticket Hypothesis [1] has inspired many works in both theoretical side and empirical side. In [3], authors proved that training is not necessary if presented with a large enough network and the proof was improved in [4] shortly afterwards. People have tried to prune CNN [2], which has various applications in Computer Vision.

In our project, we want to analyze the existing pruning techniques, which were tested on different benchmarks and therefore lack direct comparisons, and examine two novel pruning strategies: 1. We want to train the pruned network and reintroduce the highest norm weights. 2. We want to reward the weights that produce values that align best with the gradient, and re-introduce the highest weights with the highest scores. These two strategies are built on existing techniques and would be interesting to the community if we can show good results empirically.

5 Proposed solution and major contributions

One of the major concern in pruning methods is that by removing weights too early, we might prematurely remove weights that are important for the network. So we are proposing a way for us to re-introduce some pruned weights later in the pruning process if they prove to be useful. So we are proposing two algorithms that reintroduces lost weights pruned early in the pruning process.

Double-pruning trains the pruned network as well as the kept network, and reintroduces weights with large norm back into the network. Scored-pruning rewards a weight if it produce result that aligns with the gradient, and reintroduces edges with the highest score after several iteration. They both aim to solve the premature pruning problem.

One of the other issue for research around LTH and pruning methods is there is not a unified benchmark for comparing different settings and pruning methods. Each work presents their unique set of experiments that shows improved performance. In our report, we propose our benchmark that surveys different pruning methods and compare them to show their comparative advantages.

6 Proposed experiments-datasets, tasks, architectures, expected results,

Due to the computation limitations, we will be experimenting using fully connected networks. Google Speech dataset is a good dataset for Multi-Layer Perceptron workloads. We are not using image datasets as that would be better suited for CNN. We might also include MNIST for completeness since there are a few previous works that base their experiments on MNIST dataset. We want to show that our pruned network is converging to comparative accuracy to the full model.

In our experiment setup, we will be examining different methods based on the number of epochs (number of passes through the dataset) required to achieve target accuracy. For the pruned network, we want to prune the same percentage. Admittedly, for different pruning points, pruning in the later stage will give better accuracy, but early pruning will give lower computation budget. We will keep that in mind when we are comparing the performances. We will be comparing the number or epochs and accuracy while varying pruning strategy. We want to show what methods give us the best final accuracy, and what methods can reduce the computation budget while having very little or no accuracy penalty.

7 Project execution plan

The main goal and targets for this projects are:

- Survey the state-of-the-art pruning algorithms and understand their design and implementation choices.
- Implement LTH pruning algorithm and benchmark tests that evaluated the performance of the pruned network.
- Implement Double-pruning and Scored-pruning algorithm on the LTH framework
- Conduct experiments that benchmark different pruning techniques and compare to show their comparative advantages.
- Formulate final report

8 Feasibility and limitations of approach (GPUs, CPUs needed, no of experiments)

Due to the limitation of computing power, we will conduct our experiments on fully connected networks (MLPs). In the future, we can apply these ideas to CNN as well.

9 Potential impact

This work will provide a better understanding of different pruning techniques based on Lottery Ticket Hypothesis. We hope to simplify training and achieve better final accuracy with our newly propose algorithm. This will allow better future analysis of novel LTH-based algorithms and improve the pruning efficiency and final accuracy of such models. This line of work is particularly impactful as it can harness the accuracy of large neural networks and concentrate them into small networks that are memory efficient and fit into our everyday mobile devices and better integrate into our daily life.

References

[1] Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. *arXiv:1803.03635 [cs]*, March 2019. arXiv: 1803.03635.

[2] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets, 2017.

[3] Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, and Ohad Shamir. Proving the Lottery Ticket Hypothesis: Pruning is All You Need. *arXiv:2002.00585 [cs, stat]*, February 2020.

[4] Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, and Dimitris Papailiopoulos. Optimal lottery tickets via subsetsum: Logarithmic over-parameterization is sufficient, 2020.