E0:270 - Machine Learning - Neural Network Pruning Techniques

Braunstein, Cameron [1] **Nair, Abhishek** [2] **Shaw, Vishal** [2]

Abstract

This project investigated various methods of pruning a neural network. We have found that The abstract should be ideally less than 10 short sentences. Among other things, provide information about what is your objective and how much progress has been made.

1. Introduction

The PDF report should contain at most two pages excluding references and appendix and at most four pages including everything. Write about a subset of the following things that applies to your project:

1. Problem Statement
2. Motivation
3. Literature Review
4. Dataset description
5. Preliminary Results
6. Future Work

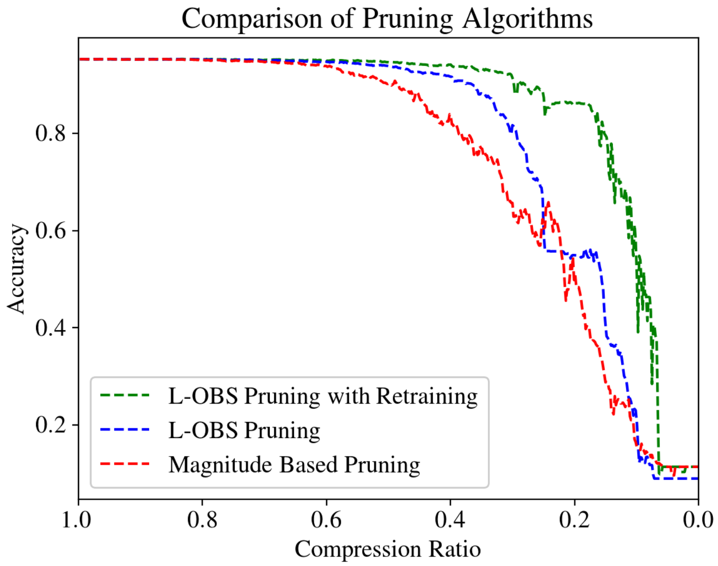
2. Literature Review

The Optimal Brain Surgeon algorithm (OBS) as described by Hassibi, Stork and Wolff (put citation), prunes weights based on their calculated effect on the error using a second order Taylor expansion of the error function, and adjusts the unpruned weights to compensate. Layerwise OBS (L-OBS) as described by Dong, Chen and Pan simplifies the OBS algorithm by only considering a weight's effect on local error. They demonstrate that any increase in the global error from pruning a weight is reasonably bounded above by the corresponding increase in local error by pruning that weight. Because of this, L-OBS approximates OBS and is more computationally feasible.

3. Model description

For our experiments, we trained a 748-300-100-10 feed foward neural network to classify the MNIST dataset. We then tested several pruning algorithms on the network. We implemented magnitude based weight pruning as a control. We ran a simplified version of the L-OBS algorithm, which recalculated the inverse Hessian only after every 2000 pruned weights. We also ran this simplified L-OBS algorithm, with a 3 titration retraining after every 2000 pruned weights and then a recalculation of the inverse Hessian. In the original L-OBS algorithm, the inverse Hessian is recalculated after every pruned weight. However this change was made out of computational necessity.

4. Preliminary Results



Our control, the Magnitude Based Pruning, dropped in accuracy the fastest, followed by our simplified L-OBS algorithm and then simplified L-OBS with retraining. The L-OBS with retraining graph is particularly jagged, as retraining every 2000 prunings resulted in spikes of accuracy, particularly as the network became more sparse. Interestingly, all algorithms stayed closed to their initial test accuracy until a compression ratio of approximately 0.6. This suggests that the network has redundancies which can be

[1] Mathematics Department, Indian Institute of Science, Bangalore [2] Department of Computer Science and Automation, Indian Institute of Science, Bangalore. Correspondence to: Braunstein, Cameron <comeronb@iisc.ac.in>, Nair, Abhishek <abhishek-nair@iisc.ac.in>, Shaw, Vishal <vishalshaw@iisc.ac.in>.

Midterm Project Report, E0:270, Machine Learning, 2018.

eliminated before training begins.

5. Future Work

We hope to combine several of our tested algorithms to see if we can achieve even better compression without loss in accuracy.

You can additionally provide anything else that is relevant to your project but is not present in the list given above.

You can create various sections/subsections etc. to organize your report as per the need. Use `biblio.bib` file to provide references. The references should be provided by using cite keyword (**?**).

You can contact your project mentor if you have any confusion.

A. Optional Appendix (Ungraded)

Remove this part if you are not using an Appendix. Appendix is ungraded. The reader may wish to ignore the appendix altogether. Write everything that you think is important in the main report text only.