SPANDEX Write Up

Spiking Audio Neural DEnoising eXperiment – Georgia Tech Research Institute

Author: Michael Jurado (Michael.Jurado@gtri.gatech.edu)

Overview

Our primary goal for the Intel DNS Challenge was to create a light-weight, low power, high-performing neural audio denoiser. We achieved this through global unstructured synaptic pruning. This approach has implications far beyond the challenge and shows that SNN performance can be maintained even after substantial weight pruning. Our submission results consist of two SDNNs with 50% and 75% weight sparsity. We achieve these results through the use of a cubic pruning scheduler, which allows for a fully trained model to be pruned in only a handful of epochs.

Previous Work and Research Background

This research stems from an experiment to map a spiking neural network (SNN) variant of vgg16 onto the Loihi1 chip using Nengo. The key challenge was the large synaptic fanout in vgg16's convolutional layers, which made it tough to fit within the Loihi core's 128-to-1 synapse-to-neuron ratio constraint

To reduce the fanout of the network, we decided to employ both structured and unstructured pruning techniques. Structured pruning involves removing groups of parameters, such as filters in a convolutional layer, based on specific heuristics like the L1 Norm, which provide insight into the importance of those groups. Unstructured pruning is the process of pruning individual low importance weights without regard to the overall structure of the network. This later approach allowed us to achieve a 95% synaptic sparsity for vgg16.

After modifying Nengo-Loihi's synaptic partitioning code to exploit this sparsity, we were able to successfully port vgg16 to Loihi1. The takeaway from this experimentation was twofold. First, large synaptic fanouts can hinder the efficient core utilization of large networks. Second, the Loihi chipset can take advantage of synaptic weight sparsity, provided there is an efficient neural core compiler to optimize the mapping

Technical Approach

After benchmarking a single epoch of the baseline SDNN solution, I realized that due to time constraints, I could only train my SNNs for exactly 5 epochs. Since I couldn't train my own architecture from scratch, I decided to prune the baseline solution. The baseline solution already represents one of the smallest SNN parameterizations in the competition. Any large amount of pruning applied to the model would result in a remarkably small and low-power audio denoiser.

Given this limited time budget, I decided that the iterative unstructured pruning approach I employed for cifar10 was infeasible since it requires alternating between pruning and training epochs. Therefore, I initially employed a linear pruning schedule. This meant that every training batch of the neural network would be accompanied by a small pruning step, in which the smallest absolute magnitude weights would be set to zero. After running a single epoch of this approach, I realized that the SI-SNR was steadily decreasing over the iterations. The most likely explanation for this is that the SNN was unable to

recover lost performance due to pruning since its parameters were constantly being changed. After some research, I found a technique called Gradual Magnitude Pruning (GMP) which aggressively prunes the NN early in the training and slowly transitions to the fine-tuning stage. Moreover, this technique calls for a pruning frequency parameter that allows the SNN to recover performance between every pruning step.

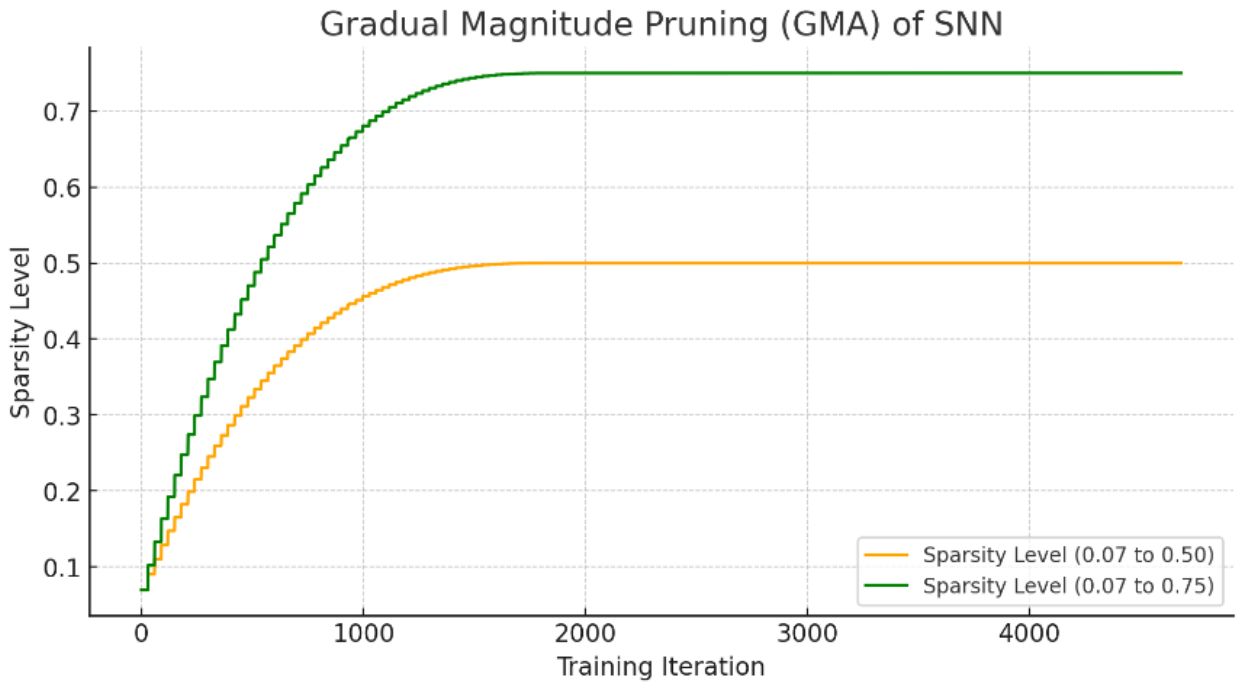

Figure 1: Baseline SDNN sparsity over the pruning and fine-tuning epochs. Gradual Magnitude Pruning employs a cubic scheduler.

Results

Our 50% and 75% sparsity SDNNs achieved a validation SI-SNR of ~12.3 and ~11.9 respectively. Given an effective mapping of the SNN to the Loihi2 chip, the size reduction and gained power efficiency should roughly mirror the resultant sparsity of the SNNs.

Future Work

Fully taking advantage of sparsity in SLAYER trained SNNs may require intensive infrastructure enhancements to netx, lava, and potentially lava-loihi. However, prior work has shown that it is indeed possible and highly advantageous to take advantage of this sparsity.

Our lab at Georgia Tech will continue to search for small, low-power, high performant architectures for Track II of the Intel DNS Challenge. We hope to couple this synaptic pruning with other SNN topologies besides just SDNNs.

This research is part of the SPANDEX-Hive IRAD at the Georgia Tech Research Institute (GTRI).