Efficient Deep Learning System Design

Linda Kebichi, YSP Student, *Lexington High School*
Kenneth Kalin, YSP Student, *Cambridge Rindge and Latin School*
Mengshu Sun, Ph.D. Candidate, *Northeastern University*
Jim Lin, Ph.D. Candidate, *Northeastern University*
Dr. Xue (Shelley) Lin, Electrical and Computer Engineering, *Northeastern University*

Northeastern University College of Engineering

Northeastern University Center for STEM Education

YOUNG SCHOLARS PROGRAM

NSF

ABSTRACT

The goal of this research is to solve one of the major issues that accompanies image recognition: how to develop an efficient model without losing too much accuracy. It often takes a large amount of computations for convolutional neural networks to discern the contents of an image, but through model compression and FPGA simulation this process can become much faster and more efficient. First, different pruning ratios were experimented with to develop a model with the right balance of accuracy and efficiency. Higher pruning ratios meant that a higher percentage of weights would be removed from the model, resulting in lower accuracy but greater efficiency. The convolutional neural networks were then run using FPGA development boards. The model ran much faster on the FPGA platform because it acts like a dedicated chip designed specifically to perform the tasks defined by the network model. The results show that having a pruning ratio between 0.7 and 0.8 on each layer and dedicating a chip entirely to image recognition on a certain dataset are the optimal conditions for achieving the highest accuracy and efficiency. The FPGA simulation indicates that more heavily pruned models have higher energy efficiency, perhaps making them a useful asset to mobile devices so as to not drain their battery.

INTRODUCTION

A convolutional neural network is a type of deep learning algorithm that allows computers to accurately solve complicated problems typically involving image recognition. Convolutional neural networks are trained using filters that detect certain features of an image so that the computer can recognize what is shown in pictures based on previous trends. However, these filters can be very large which makes it difficult for convolutional neural networks to run efficiently on devices without large processors. By pruning these models, unnecessary parameters can be removed from the overall structure of the network, allowing it to take up less space on a target device without compromising the overall accuracy of the network. Pruning checks to see which filter channels contain weights that are closest to zero and removes such channels since they are not particularly decisive in whether or not the model makes the correct distinction of an image.

A field-programmable gate array (FPGA) is an integrated circuit that can be designed specifically to perform the functions specified by the customer. In this case, the function of the chip is image recognition.

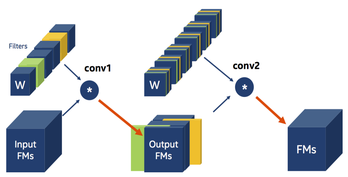

Figure 1: The two filters colored in green and yellow are removed, resulting in two fewer output feature maps in the sparse model.
Nervana Systems. 2018 https://nervanasystems.github.io/distiller/tutorial-struct_pruning.html

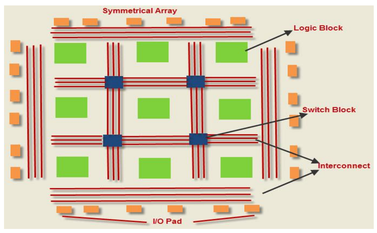

Figure 2: Basic FPGA architecture includes Configurable Logic Blocks, interconnection wires, and I/O pads.
Elprocus. 2020 https://www.elprocus.com/fpga-architecture-and-applications/

When simulated on an FPGA board, the image recognition software runs much more quickly than it does on a server, since the FPGA chip is dedicated to accomplishing a single task as opposed to the many tasks a server manages.

EXPERIMENTAL METHODS

To create a convolutional neural network with sufficient accuracy, various combinations of layers with different dimensions were tested. Several parameters were tested for each new layer added, and the results were graphed until a network with an overall accuracy of 80% on the CIFAR10 dataset was created. In addition, the neural network was trained with a varying number of epochs (a unit for measuring how much data the network is trained with) along with the batch sizes (amount of data per epoch) and learning rate (the interval by which the weights change in value). All of these variables determined how to train the network in a way that made it accurate enough without wasting too much time or CPU power (see figure 3). Decreasing the learning rate also resulted in a more accurate model because the weights did not change in value as drastically during training.

Next, this trained dense model was pruned to make the overall network smaller. For each layer of the network, the pruning ratio was increased until the model was at least 70% sparse. The model was then assessed for accuracy and repruned accordingly if it could handle more loss without becoming excessively inaccurate. By applying this method, a new network was created with about a 10% accuracy loss and a much faster runtime due to the size reduction (see figure 4).

The final phase was to modify an FPGA development board to accommodate the neural network. Both the dense and pruned models were run on this board, and the power consumption and number of computations required for each network were compared to evaluate how much more efficient the sparse model was compared to the dense model.

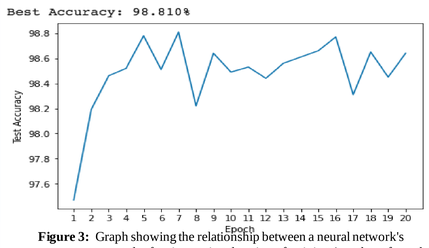

Figure 3: Graph showing the relationship between a neural network's accuracy as a result of an increasing duration of training (number of epochs). In this network, after the network received at least 10 epochs of training, no consistent improvement in image recognition was observed.

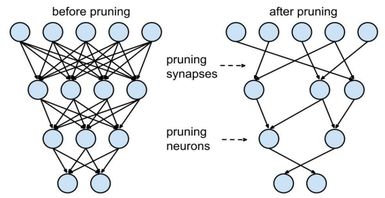

Figure 4: Illustration of a dense (unpruned) and a pruned neural network. The pruned network is much smaller and easier to traverse quickly, but loses accuracy as a result of the pruning process.
Singh. 2019 https://towardsdatascience.com/pruning-deep-neural-network-56cae1ec5505

RESULTS

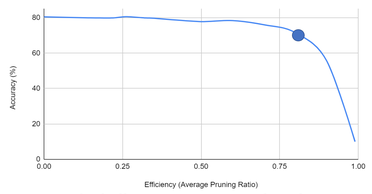

Figure 5: Graph of Efficiency vs. Accuracy for Cifar-10 on LeNet-5

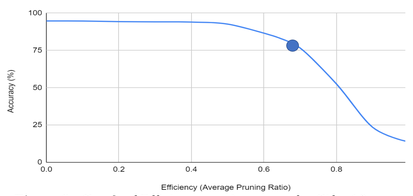

Figure 6: Graph of Efficiency vs. Accuracy for Cifar-10 on ResNet-18

- As pruning ratio increases accuracy decreases
- Settled on an average pruning ratio of 0.8 with 71.58% accuracy
- Accuracy drops sooner on ResNet-18 than it does on LeNet-5
- Settled on an average pruning ratio of 0.7 with 76.61% accuracy

RESULTS CONTINUED

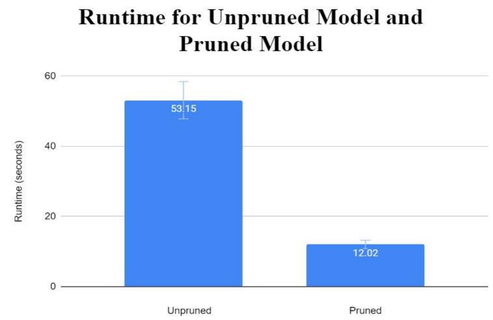

Figure 6: Comparison between runtimes for pruned and unpruned models. Pruned model is more than 4 times faster

Figure 7: Sample image run through convolutional networks

CONCLUSIONS

Multiple ways of making deep learning systems as efficient as possible while still maintaining substantial accuracy were developed through this research. It was found that the ideal pruning ratios on each layer depend not only on the type of dataset (Cifar-10 or MNIST), but also the type of neural network structure (LeNet-5 or ResNet-18). By increasing pruning ratios, deep learning algorithms can become much more efficient without suffering too much accuracy loss. Through pruning, a network that used similar power, ran more than 4.4 times as fast, and lost less than 10% accuracy was created. Another means by which to make such models more efficient is using FPGA chips designed specifically for the model. Dedicating an FPGA chip to image recognition is a much more efficient method than running deep learning software on a server. For both dense and sparse models, extensive testing to discover the optimal architecture and parameters (e.g. epochs, learning rate, batch size, etc.) is imperative to ensuring a high accuracy. In the future, column pruning could also be experimented with in comparison to filter pruning to see how they differ in terms of their effects on power consumption and accuracy.

REFERENCES

Sun, Mengshu, Zhao, Pu, Gungor, Mehmet, Pedram, Massoud, Leeser, Miriam, & Lin, Xue. (2020). 3D CNN Acceleration on FPGA using Hardware-Aware Pruning. *The 57th Annual Design Automation Conference 2020 (DAC 2020)*. Retrieved from http://par.nsf.gov/biblio/10146398

Wang, Yetang & Ye, Shaokai & He, Zhezhi & Ma, Xiaolong & Zhang, Linfeng & Lin, Sheng & Yuan, Geng & Tan, Sia & Li, Zhengang & Fan, Deliang & Qian, Xuehai & Lin, Xue & Ma, Kaisheng. (2019). Non-structured DNN Weight Pruning Considered Harmful.

ACKNOWLEDGEMENTS

Department of Engineering
Dr. Xue (Shelley) Lin – Assistant Professor
Mengshu Sun – Ph.D. Candidate
Jim Lin – Ph.D. Candidate

Department of STEM
Claire Duggan – Director of Programs and Operations
Salima Amiji, Natasha Zaarour, and Nicolas Fuchs – YSP Coordinators