Hardware Accelerator for Training Neural Networks

James Erik Groving Meade

Advisor: Jens Sparsø

This document outlines the current tentative plan for James Erik Groving Meade's Master's thesis. The objective of my Master's thesis is to design and implement a FPGA accelerator for improving energy and time requirements for training a neural network.

The past few years have been witness to a massive rise in interest regarding machine learning. In this new wave of AI, computer vision has been completely dominated by a new architecture known as convolutional neural network. These networks are often contain many hidden layers between input and output, hence the origin of the term "deep learning". These networks use convolutional filters to recognize features and classify images.

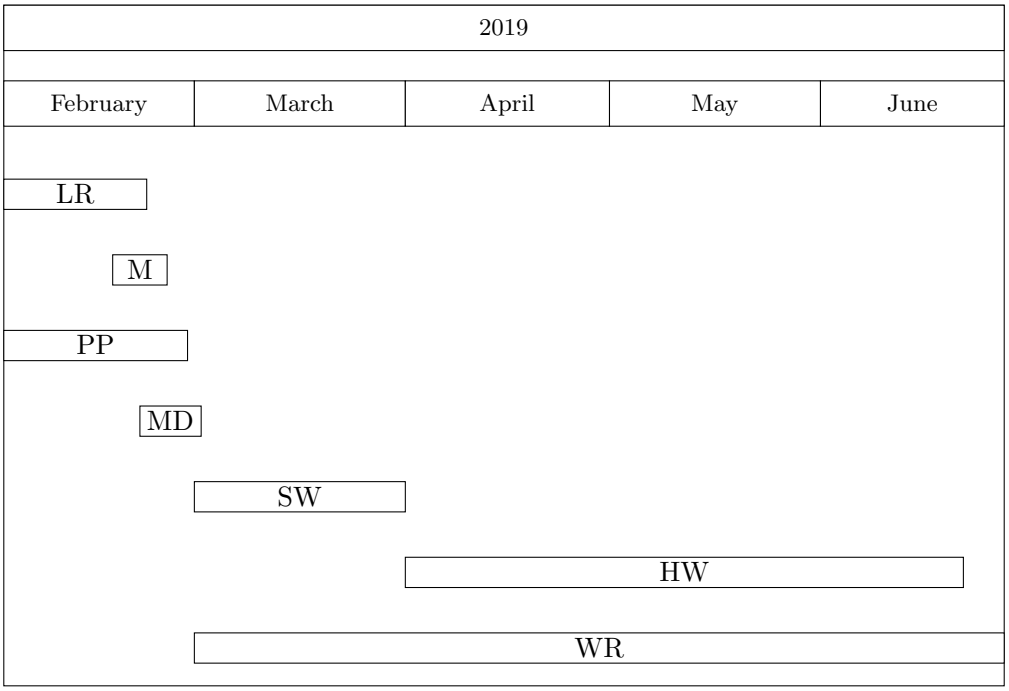
For these networks to operate, all the parameters and weights in the network must be trained using supervised learning and a large labeled training dataset. The training process in its current state can be rather slow and is often done using GPU's due to its massively parallel nature. We have only started to begin to see application-specific hardware being developed, and primarily in academia. Thus, the hardware side of neural networks is very much so in its infancy, as one must understand the intricacies of the high-level algorithms to be able to construct an efficient, targeted, low-level model on which to delegate the training workload.

While there are numerous chips that have been developed to perform inference such as the Eyeriss, Google TPU, and nn-X chips, not much focus has been given to performing the training of neural networks. There has been preliminary academic research such as the F-CNN FPGA model, but this work has not investigated reduced precision to improve speedup or energy efficiency. A paper by Courbariaux performed a software simulation of reduced precision and observed that low precision multiplications do not cause too much added error in many cases. Therefore, the proposed work shall focus on developing an FPGA-model that uses reduced precision and a modular, flexible architecture to achieve efficient and fast neural network training while minimizing the final error loss of the trained network.

Despite the dearth of research in application specific hardware for neural network training, there are grand use cases for this type of hardware. While the industry has defaulted to using GPU-based cloud computing centers, large energy savings could potentially be realized by switching to more specialized training hardware. This would result in greener and more cost-effective training and any changes in

speedup can be managed by adding more units due to the embarrassingly parallel nature of training neural networks.

Regarding the software engineering aspects of the project, weekly meetings have been arranged between the student and the advisor. Furthermore, all relevant papers, research, work, and code are being maintained in a git repository. Lastly, a tentative schedule for the timely completion of the thesis report has been created and is visible in the below figure.



LR Literature review
M Creating CPU/GPU models for benchmarking
PP Project plan
MD Hardware model design
SW Software implementation of the accelerator model
HW Hardware implementation on FPGA of the accelerator
WR Ongoing writing of report and reading literature

Figure 1: Current tentative schedule for completing the thesis