CS 205 High-Level Project Description - Group 15

We plan on parallelizing a program of a vision transformer. A vision transformer is a type of neural network architecture that is used for image classification and recognition tasks. This is a nuanced problem as there are abundant opportunities for both data and computational parallelization. There are various bottlenecks. There is a memory bottleneck in that the models can be pretty large. There is a bandwidth bottleneck in that we will use a PCI and a PCI bus inherently has limited bandwidth. And lastly there is a computational bottleneck in that we plan on using an FPGA which is resource limited. In terms of the parallelizable portions of our CNN itself, we can work to potentially parallelize certain sparse matrix multiplication, activation functions (like ReLu), and pooling techniques.

We will use the FPGA to exploit the parallelization. The custom hardware will be able to handle the computation. Furthermore, we plan to have a pipelined compute core that can handle multiple simultaneous computations. We will have a combination of both shared memory and distributed memory models. Since the FPGA has its own DRAM dimm, there will be distributed memory. We also plan to compare the performance results with a CPU implementation which has a shared memory model. Lastly, the custom FPGA hardware will be a MIMD machine if we decide to have multiple compute cores, otherwise we will have a SIMD machine.

We will try to use a small scale model as our priority to showcase our parallelization efforts and compare it between hardwares, instead of its accuracy. For vision transformers, the smallest implementation will be a ViT-Base 16x16 model, which contains around 86 million parameters, and can occupy around 330 MB of memory. In training, the majority of memory we require is for storing the input batches and gradients, however we plan to focus mainly on inference. Given that we are working with a relatively small model and only for inferencing, this should be reasonable given the resources we have on a CPU and FPGA, and should be more than enough on a GPU.

We plan on parallelizing a program of a vision transformer. A vision transformer is a type of neural network architecture that is used for image classification and recognition tasks. This is a nuanced problem as there are abundant opportunities for both data and computational parallelization. There are various bottlenecks. There is a memory bottleneck in that the models can be pretty large. There is a bandwidth bottleneck in that we will use a PCI and a PCI bus has limited bandwidth. And lastly there is a computational bottleneck in that we plan on using an FPGA which is resource limited. In terms of the parallelizable portions of our CNN itself, we can work to potentially parallelize certain sparse matrix multiplication, activation functions (like ReLu), and pooling techniques.