15-618 Fall 2022 Final Project Milestone Report

Junli Cao
junlicao@andrew.cmu.edu

Yuchen Wang
ywang7@andrew.cmu.edu

November 2022

1 Progress Summary

1. Implemented the baseline model, 2-layer MLP, supporting forward and backward passes on a single machine.
2. Implemented the data parallelism for the baseline 2-layer MLP model(without building the computational graph) on the MNIST dataset supporting an arbitrary number of workers in MPI in Python interface (mpi4py). Currently, we calculate the gradients of each layer by hand, which is hard to scale up with the number of layers.
3. Benchmarked the training, testing accuracies, and runtime on a local PC.

2 Preliminary Results

We can achieve data parallelism with a central parameter server. We benchmarked the running time on one node and four nodes with 95% accuracy on MNIST. We observe a 1.8x speedup when using four nodes. We do not achieve a higher speedup due to frequent communication (every step) in our current implementation and the bottleneck of the central node.

3 Updated Goals

1. Implement and benchmark the current design in C++ and compare the results with the Python implementation.
2. Modularize the calculation with backpropagation to support more layers without manually calculating the gradient of each layer.
3. Add a simple computational graph to support residual training layers.
4. Run experiments on the GHC cluster with more computational nodes and report the speedup concerning the number of workers.

5. Although we find it hard to implement model parallelism in C++, we may try to implement ring data parallelism and model parallelism in Python if we have time.