CMSC416: Introduction to Parallel Computing

Topic: Deep Learning in Parallel
Date: May 2, 2024

[redacted]

Deep Neural Networks

- Parameterized function approximator
  - Can work with very high dimensional data (image, text, etc)
- Computation organized in a sequence of layers with linear dependencies

Motivation

- Model sizes are increasing rapidly
- Many opportunities to exploit parallelism
  - Iterative process of training
  - Many iterations per epoch
  - Many layers in DNNs

Data Parallelism

- Each worker has an entire copy of the NN
- Divide data among GPUs and all reduce op to sync gradients
- Easy to use, but can't train models that don't fit on a single GPU

Inter-Layer Parallelism

- Distribute different layers to different GPUs and use P2P communication
- Break batch into multiple shards and process them in a pipelined fashion so multiple GPUs can be active at a time

Intra-Layer Parallelism

- Divide the work of each individual layer across multiple GPUs
- Essentially can be thought of as parallel matrix multiplication

Other

- Hybrid Parallelism: using two or more approaches in the same framework
- Designing User-Friendly and Communication-Efficient algorithms is of vital importance