15-418 Final Project Proposal

Manish Nagireddy (mnagired) and Ziad Khattab (zkhattab)

Title

An Exploration of Parallelism in Neural Networks

Summary

For our 15-418 final project, we are looking into potential axes of parallelism that exist within neural networks. We will be implementing neural networks in both C++ (via OpenMP) and Python (via PyTorch and mpi4py, an MPI package for Python) and measuring their performance on CPUs as well as GPUs.

Background

In recent years, we have seen the rapid increase of neural networks within the landscape of artificial intelligence, and more broadly within algorithmic decision making. However, because of the large size of these models (given by the number of parameters) as well as the large size of the data used to train them, performance of these so-called deep learning paradigms can be suboptimal without specific attention to parallelism.

Broadly, from the perspective of a neural network, there are two dimensions to parallelize: the data and the model.
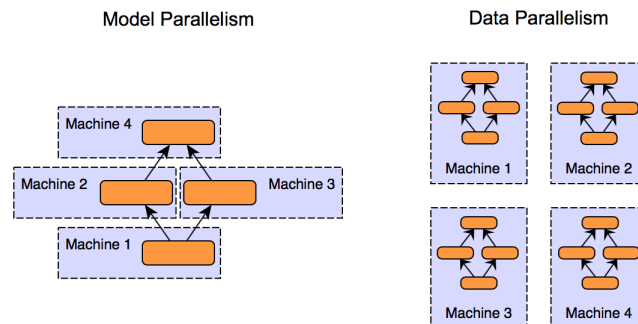
1. **Data Parallelism:**

   - given X machines/cores, split the data into X partitions and use the *same* model to train each partition of the data on a different device in parallel. Then, combine the resultant model weights from each partition
   - Note that this is model agnostic because it relies only on the data

2. **Model Parallelism:**

   - splitting the model into various partitions and assigning them to different machines.
   - NOTE: there are dependencies that are specific to specific model architectures, and so model parallelism is not really "parallel" because we are actually just assigning consecutive layers of a model to different devices. Some have referred to this as *model serialization*[1]
   - commonly used when the model is too large to fit on a single device

Refer to the figure[2] below for an illustration of data parallelism and model parallelism.

---

[1]Data Parallelism VS Model Parallelism in Distributed Deep Learning Training - Lei Mao's Log Book
[2]Link to Figure Citation

The Challenge

Resources

Goals and Deliverables

Platform Choice

Schedule