ABSTRACT

Deep neural networks (DNNs) are currently widely used for many artificial intelligence (AI) applications including computer vision, speech recognition, and robotics. While DNNs deliver state-of-the-art accuracy on many AI tasks, it comes at the cost of high computational complexity. Accordingly, techniques that enable efficient processing of DNNs to improve energy efficiency and throughput without sacrificing application accuracy or increasing hardware cost are critical to the wide deployment of DNNs in AI systems.

Training Deep Neural Network (DNN) is a computationally intensive task whose parallelization has become critical in order to complete the training in an acceptable time. However, there are two obstacles to developing a scalable parallel DNN in a distributed-memory computing environment. One is the high degree of data dependency exhibited in the model parameters across every two adjacent mini-batches and the other is the large amount of data to be transferred across the communication channel. This project uses parallelization strategy that maximizes the overlap of inter-process communication with the computation. The overlapping is achieved by using a thread per compute node to initiate communication after the gradients are available. The output data of backpropagation stage is generated at each model layer, and the communication for the data can run concurrently with the computation of other layers.This project implements a parallel deep learning algorithm for speech recognition.