Nimble: Lightweight and Parallel GPU Task Scheduling for DL

Paper Summary (ML Term Project)

Aniruddha Mahajan (2017A7PS0145P), Shreyas Srikrishna (2017A7PS0162P)

1 Introduction

Modern DL frameworks like Tensorflow support GPU-based neural network computation. They express the neural network as a computation graph, whose nodes represent DL operators like convolution and batch normalization. The framework goes through some preparation steps before submitting tasks to the GPU. This phase is referred to as **GPU Task Scheduling**.

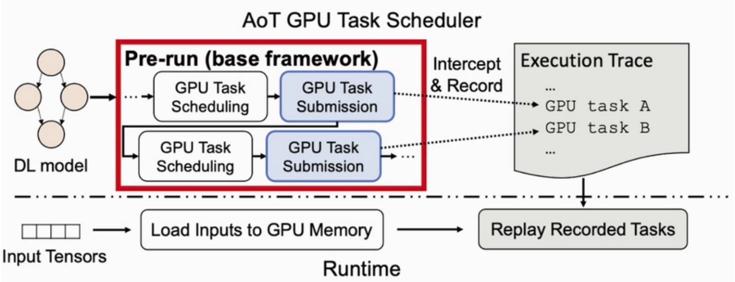
Current frameworks suffer from 2 main inefficiencies in task scheduling:

1. *Scheduling overhead* is found to be too high and leads to *idling* of GPUs.
2. GPU Tasks are executed *one at a time* and are not *parallelized*.

2 Solution - Nimble

Nimble is a deep-learning execution engine that schedules GPU tasks to run in parallel with minimal scheduling overhead for static neural networks. It uses the following two techniques:

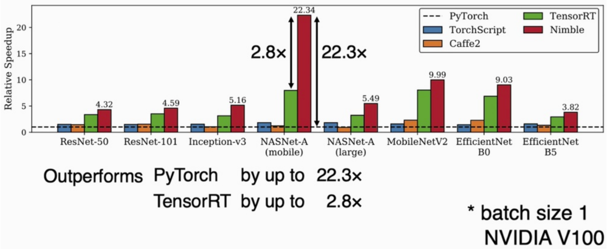
1. **Ahead-of-time scheduling**: Current DL frameworks run the same computation graph with same input/output tensor shapes repeatedly. Unlike this, Nimble runs the DL model *once* using the base framework, during which it intercepts and records the trace of GPU Tasks and constructs a task-schedule. This task-schedule is simply replayed for future iterations, thus avoiding the task-scheduling overhead.



2. **Automatic Multi-stream execution**: Using the computation graph (DAG of DL operators), Nimble runs an algorithm to find a stream assignment which maps each DL operator to a GPU stream while guaranteeing maximum parallelization of operators and minimum synchronizations across streams.

3 Results

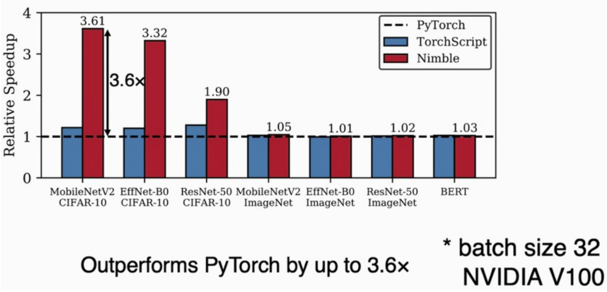
Nimble out-performs PyTorch by 22.3x times and TensorRT by 2.8x times during inference. It also speeds up training by 3.6x times compared to Pytorch.



Outperforms PyTorch by up to 22.3×
TensorRT by up to 2.8×

* batch size 1
NVIDIA V100

(a) Inference speedup



Outperforms PyTorch by up to 3.6×

* batch size 32
NVIDIA V100

(b) Training speedup