Power/Performance Analysis and Optimization for Deep Learning on a CPU-GPU Platform

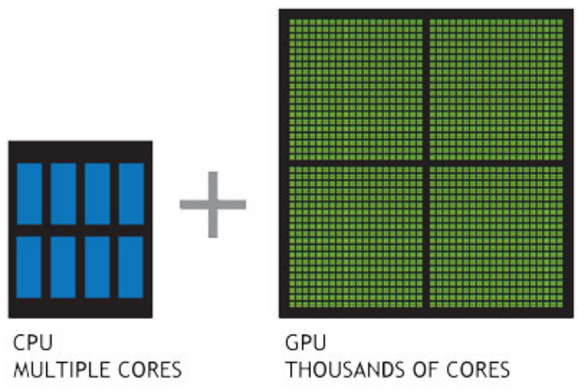
ECE 18-743 Poster Session
December 2017

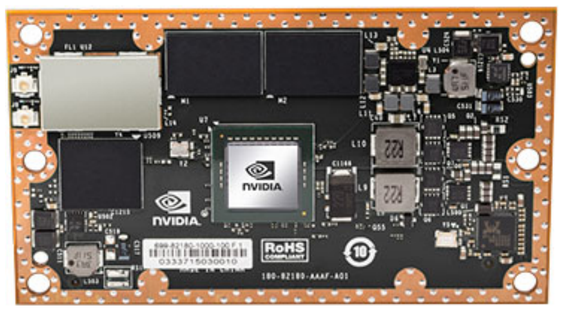
Ahmet Fatih Inci, Ting-Wu (Rudy) Chin

1. Goals

Objectives

- Understanding **what is left to be done on CPU** while performing inference on DNNs by using GPU





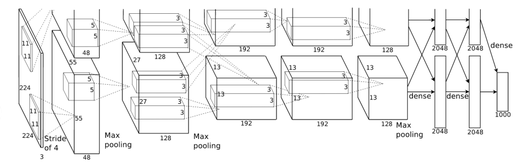
NVIDIA Jetson TX1

- Characterizing different DNN architectures on GPU and finding suitable benchmarks for CPU to utilize available resources under power/performance constraints

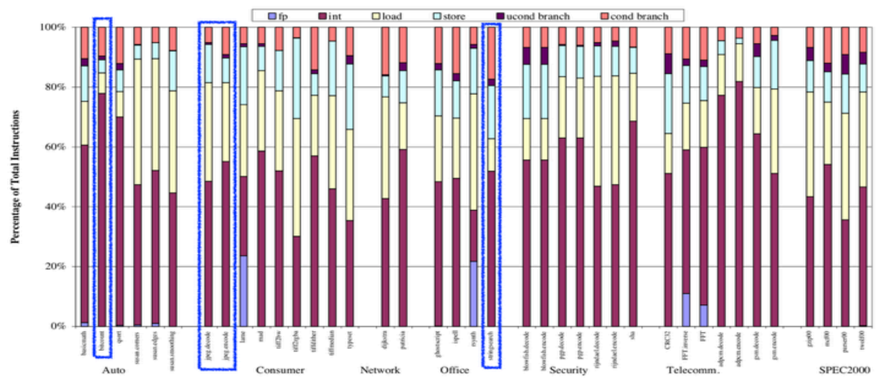
2. System Overview

Big Picture

- GPU
  - DNN Architectures in different size



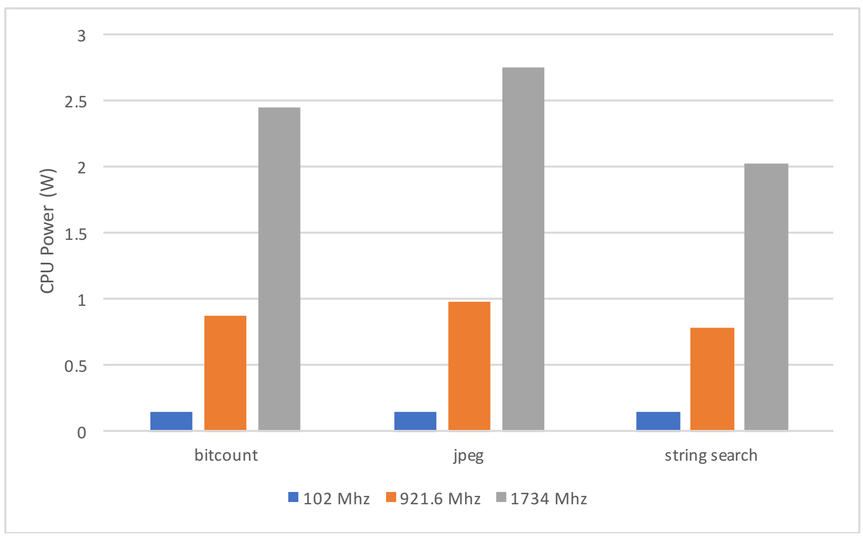
    - (AlexNet, GoogLeNet, ResNet-152)
  - Operating frequencies (76.8 MHz, 537.6 MHz, 998.4 MHz)
- CPU
  - PARSEC



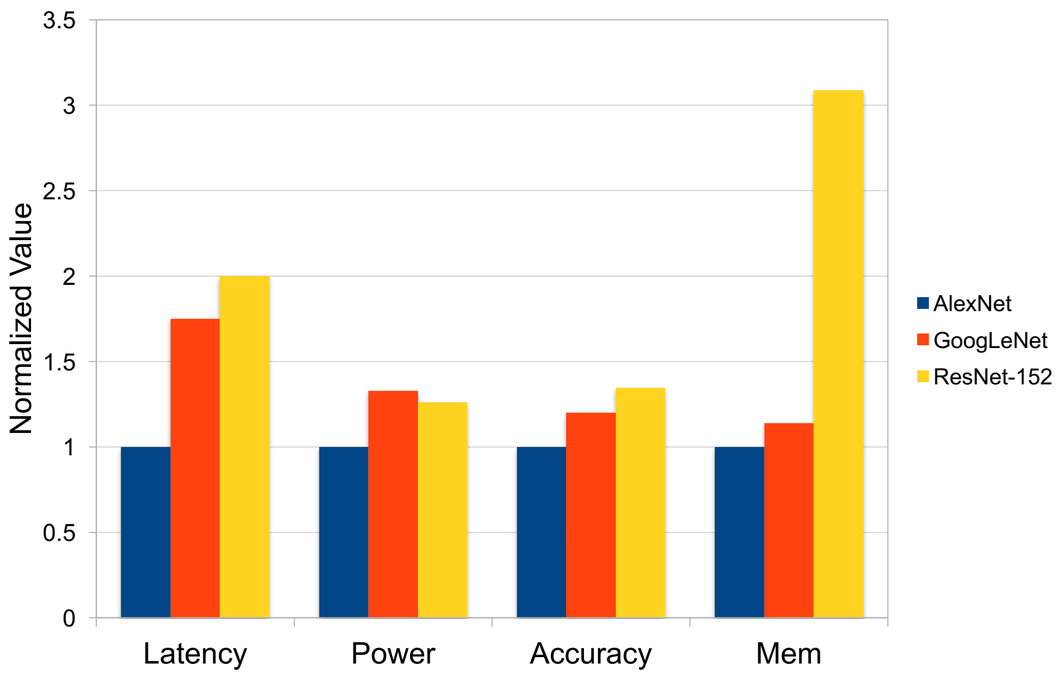
    - Compute-intensive (bitcount)
    - Memory-intensive (jpeg encode/decode)
    - Branch (stringsearch)
  - Operating frequencies (102 MHz, 921.6 MHz, 1.734 GHz)

3. Baseline Results

- CPU



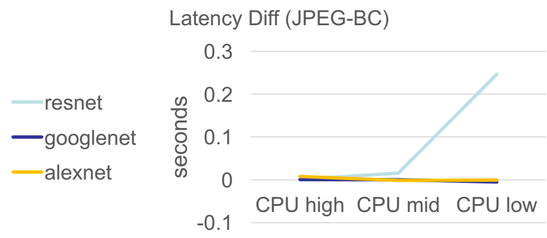
- GPU



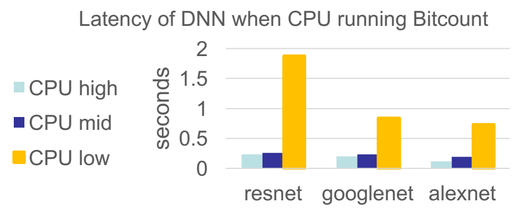
4. Results

Plots

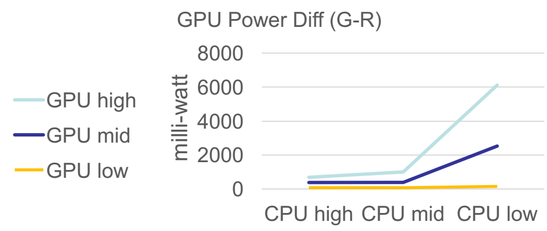
- When running memory-intensive DNNs, memory-intensive CPU benchmarks such as JPEG encode/decode adversely affects the DNN latency. We observe 13% performance difference.



- Though most of the tasks to be done in DNNs can be deployed to GPUs, CPU's frequency still play an important role due to data preparation. We observe up to 8.2x performance difference.



- GoogLeNet consumes more power than ResNet-152, which is deeper and larger, consistently. It exaggerates as the speed gap between CPU and GPU grows larger. Up to 6 W (2.8x).



Summary

- Our comprehensive analysis (81 results with different configurations) shows us the best CPU benchmark to run.
- Although people focus on GPU for DNN, CPU is not negligible.
- Memory consumption of both DNN and CPU benchmarks play an important role in overall power and performance.