Power/Performance analysis and optimization for deep learning on CPU-GPU platform

Ahmet Fatih Inci Ting-Wu (Rudy) Chin

18-743 Energy-Aware Computing

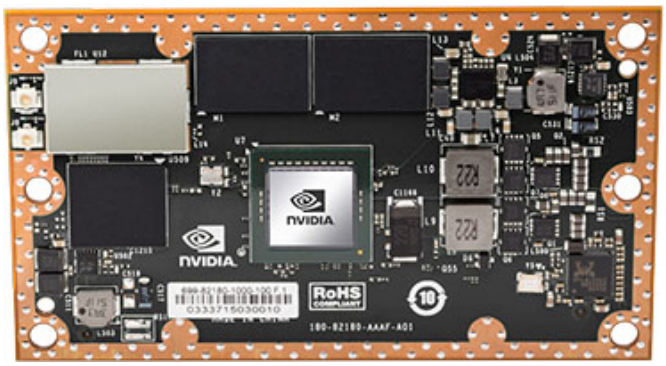
Project Website: https://github.com/afinci/18-743-Power-and-Performance-optimizations-for-DNNs-on-CPU-GPU

Outline

» Introduction
» CPU Benchmarks and Results
» GPU Benchmarks and Results
» Future Work

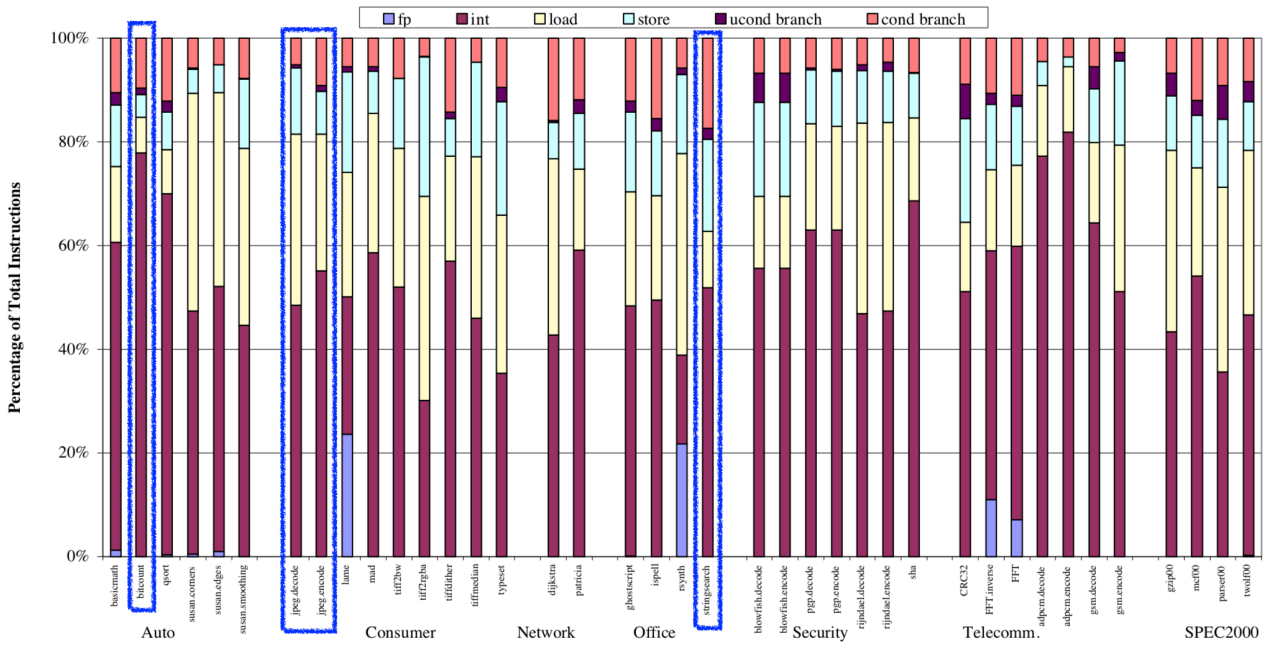
Introduction

» Profiling power/performance of embedded platform (TX1) while inferencing DNN on GPU and running benchmarks on CPU.



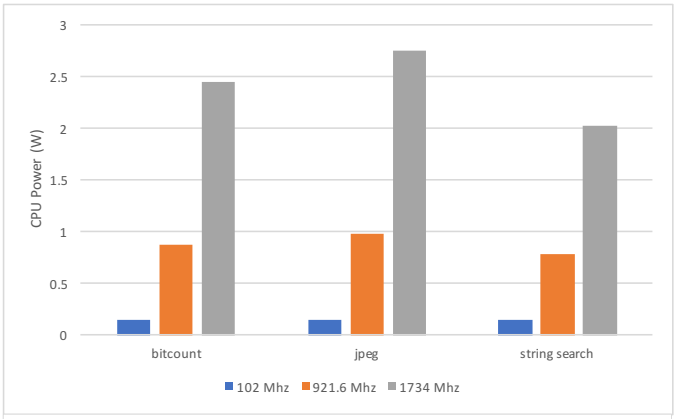
Nvidia Jetson TX1

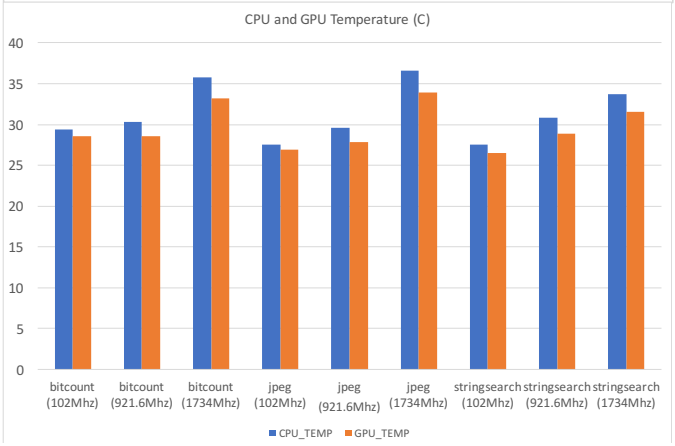
CPU Benchmarks



» bitcount (compute-intensive), jpeg (memory-intensive), string search (branch)

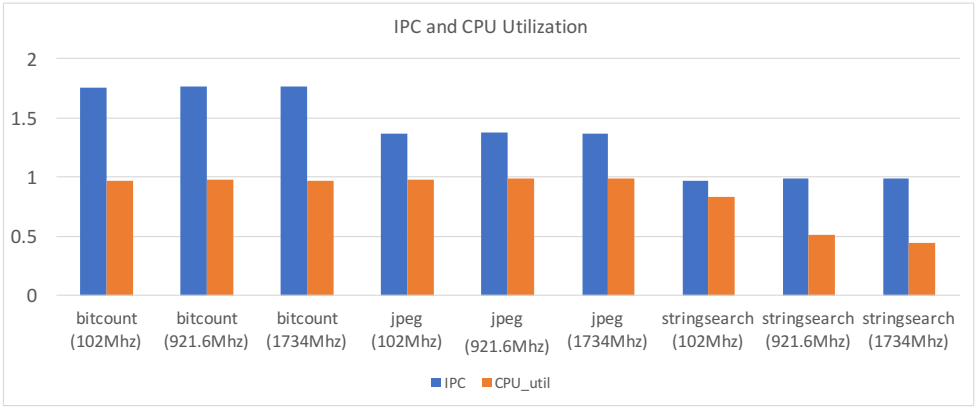
Results (CPU)





Carnegie Mellon

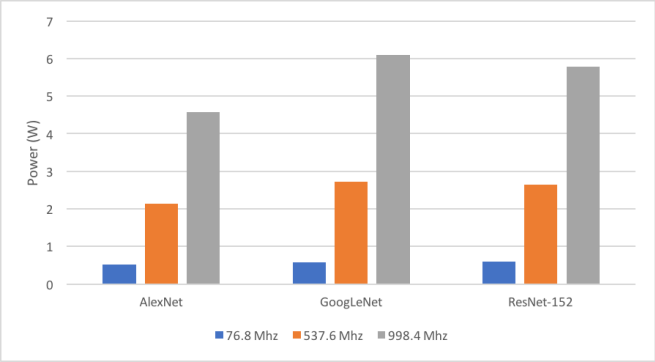
Results (CPU)

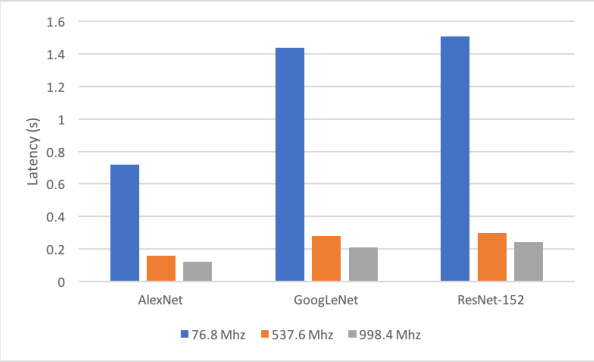


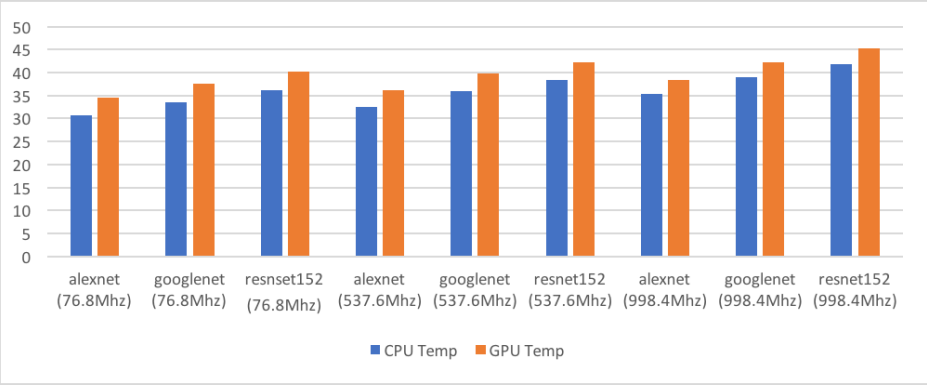
DNN Benchmarks

	Memory Overhead	Number of Layers
AlexNet	720 MB	7
GoogLeNet	820 MB	22
ResNet-152	2224 MB	152

Results (GPU)







Future Work

- M2
  - run CPU-GPU together by changing CPU-GPU frequency
  - analyze results

Q&A