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

Ahmet Fatih Inci and Ting-Wu Chin
Carnegie Mellon University
Department of Electrical and Computer Engineering
{ainci, tingwuc}@andrew.cmu.edu

Abstract

1. Introduction

2. Related Work

3. Methodology

In this project, Nvidia Jetson TX1 embedded platform will be used for power and performance analysis. We will run various deep neural network architectures for image classification task on GPU. We will use small, medium, and large deep neural networks to make a thorough power and performance analysis with different network characteristics. Caffe framework [1] will be used to implement deep neural networks on GPU. Similarly, we will use SPLASH2 benchmark suite [2] for CPU. Various SPLASH2 benchmarks will be used to cover different workload characteristics such as memory-intensive and compute-intensive workloads. Furthermore, we will run these CPU-GPU benchmarks with using three different frequency values for both CPU and GPU. First, we will run them individually to obtain power, performance, and temperature results for a baseline. Secondly, we will run them jointly to do analysis and optimization for deep learning on embedded CPU-GPU platforms.

Power values will be calculated by using current sensors in Nvidia Jetson TX1 platform. Temperature results also will be collected by using thermal sensors in TX1. We may also take off the heat sink to simulate embedded platforms which do not have heat sink. It will significantly increase temperature values. However, CPU and GPU will throttle themselves to not exceed thermal design power (TDP) constraint. Performance results for CPU will be calculated by using performance counters as an instruction per cycle (IPC) metric. Moreover, we will obtain CPU utilization results using system calls. Execution time of deep neural network will be calculated using Caffe framework as a performance metric.

4. Objectives and Deliverables

In this project, we try to understand how various CPU workloads and frequency affect GPU performance and system power under thermal design power (TDP) constraint. We fix GPU to run inference for a image classification task. Meanwhile, we try to run meaningful tasks on CPU to use remaining resources of the embedded platform. Moreover, we try to analyze how CPU-GPU workloads affect temperature values where CPU and GPU share the same power budget. Furthermore, we try to better understand what type of CPU workloads has more instruction per cycle (IPC).

5. Timeline

Timeline of the project tasks are listed below. Group members will work closely on each task to come up with a thorough analysis of power and performance results for deep learning on a CPU-GPU platform.

M1 - Choosing three SPLASH benchmarks (memory-intensive, compute intensive) for CPU

- Choosing three image classification deep neural networks with different scales for GPU

- Running CPU-GPU benchmarks individually to obtain the baseline for comparison purposes

M2 - Running CPU-GPU benchmarks jointly by changing frequency values for both CPU and GPU

- Analyzing power, performance, and temperature results

6. Conclusion

The authors will analyze power and performance results of a deep learning application on a CPU-GPU platform which is Nvidia's Jetson TX1 SoC in this case. Power and performance analysis will identify what is left to be done on CPU while GPU inferencing a deep neural network.

References

[1] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[2] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta. The splash-2 programs: Characterization and methodological considerations. In *Computer Architecture, 1995. Proceedings., 22nd Annual International Symposium on*, pages 24–36. IEEE, 1995.