Critique 1: In-Datacenter Performance Analysis of a Tensor Processing Unit

Summary

The paper is a study comparing the performance of a Tensor Processing Unit, a custom ASIC developed by Google, with a GPU (NVIDIA K80) and a CPU (Intel Haswell). The compute units are compared across workloads written in TensorFlow, which includes CNNs, MLPs and LSTMs. The paper is able to show that even though the GPU has much higher peak performance and memory bandwidth, the gain it has over the CPU is marginal, while the TPU is about 15X-30X faster, with 30X-80X better TOPS/Watt. This is due to the inference favoring response-times over throughput.

Strengths

- Lesser area and lower power usage as compared to a CPU or GPU.
- More multiplications per clock cycle leading to much faster inference.
- The TPU acts as a co-processor on the PCIe I/O bus, allowing it to be plugged into existing servers similar to a GPU, making it easy to deploy.

Weaknesses

- Low memory Bandwidth limits the utilisation of a TPU causing a bottleneck
- Low utilisation in case of sparse matrix multiplication
- Focuses on CNNs over MLPs and LSTMs, leading to under-utilisation in such cases

Improvements

- Speed up sparse matrix multiplication through special hardware and better representation
- Better Memory Architecture to overcome Memory Bottleneck-