Design Space Exploration of A Portable Tensor Processing Unit

Wesly Tonks
UC Davis
Davis, United States

done via an Avalon bus on an Altera DE-1 development kit. The TPU implements a fully pipelined 16 x 16 systolic array at a max clock speed of 115 MHz, yielding a noticeable speedup in fully connected neural network layers, when compared to the ARM A9 only.

This implementation allows useful neural Network applications such as image processing, object detection, speech-to-text, and more to be run in power and performance limited devices such as smart phones and embedded systems. The design is also scalable, allowing for a range of hardware configurations per application.

II. Background

A. Neural Networks

With the advent of big data applications becoming increasingly common, machine learning has proved to be the future of complex algorithms. Deep neural networks, a subspace of machine learning, are an attempt to reverse engineer the human brain and make computers learn and behave like humans. Deep neural networks need to first undergo training repeatedly with vast amounts of sample data, and self adjusted synapses (i.e. learning) based on

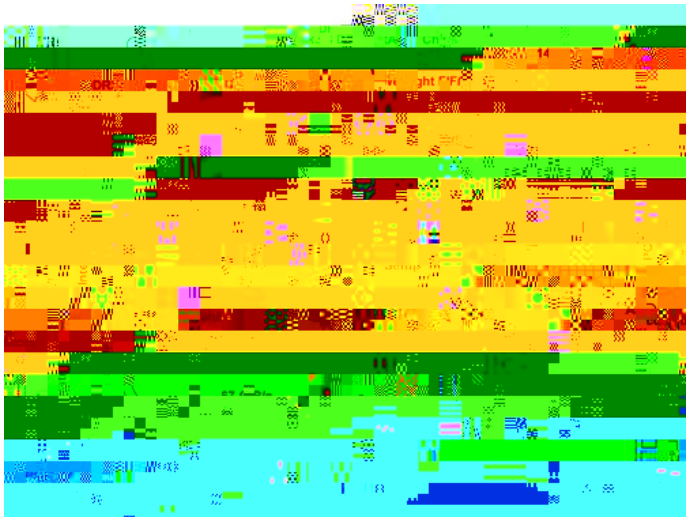

Fig. 1: High level view of Google's TPU architecture [1]

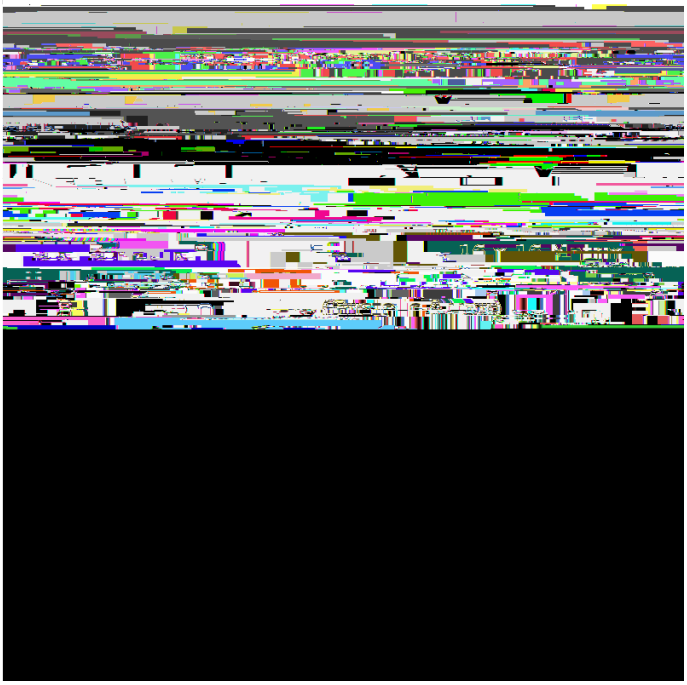

Fig. 2: High level view of system Architecture

B. Systolic Array

control modules receive their inputs from the bus control module, which can be considered the master controller in the TPU.

D. Software

To demonstrate the capabilities of our TPU, the plan was to train a convolutional neural network to benchmark on the TPU. Due to time constraints and technical difficulties, this did not happen, but throughout the process we developed a strong idea of how we can train a neural network and perform inference on the TPU. The convolutional neural network we developed is designed to perform classification of diagnoses on chest x-rays. The network takes a grayscale chest x-ray image as an input, and outputs any number of 15 classifications, including no diagnosis

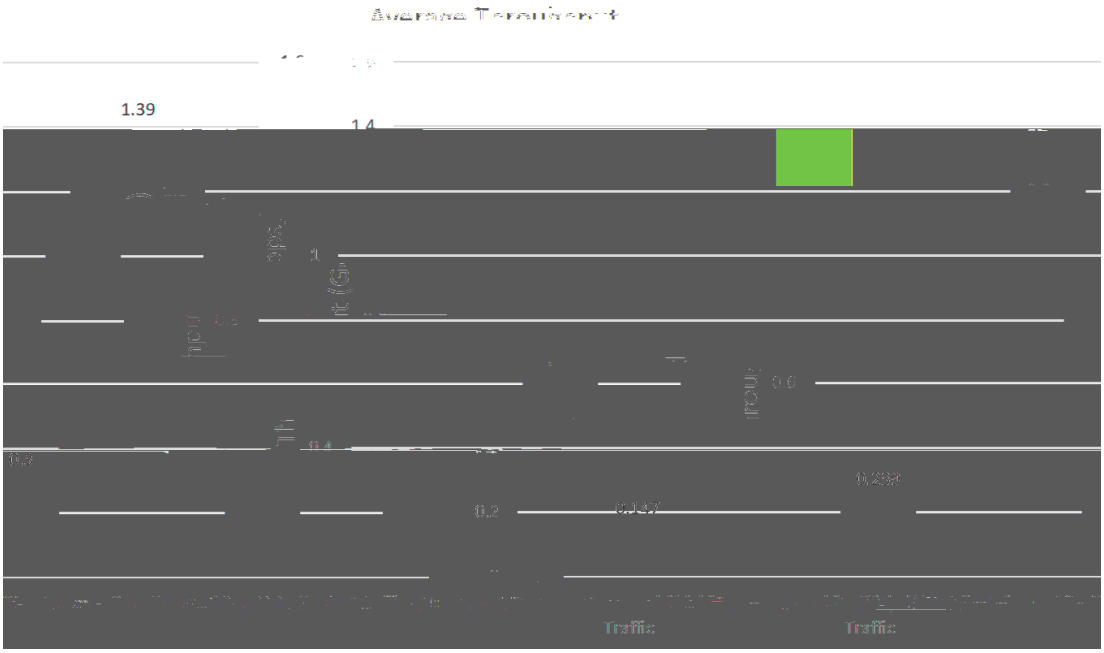

Fig. 6: Average throughput of matrix multiplications under several use cases

One last thing is that our group didnt have a change to fully implement the instruction set, but had nearly all the pieces there. Having a working instruction set would make the TPU a lot easier to control via software.