Convolutional Neural Net Implementation on FPGA's

Cory Nezin, Brenda So

September 23, 2017

Convolutional neural networks (CNN's) have recently gained momentum in the machine learning and artificial intelligence community. While CNN's provides state-of-the-art accuracy in many tasks, it comes with the cost of high computational and memory complexity which modern CPUs can barely handle. The two main contenders that achieve the computing ability required by neural nets are Field Programmable Gate Arrays (FPGA's) and Graphics Processing Units (GPUs). The comparison between FPGAs and GPUs is still an open research topic. Our project will explore ways to optimize neural network performance on FPGA's, and ultimately provide an application to demonstrate the power of neural networks on FPGA's.

Convolutional neural networks are a class of machine learning models which have shown great promise in image processing applications. CNN's take in a signal sampled in N-dimensional space and apply layers of transformations to produce a desired output. There are two main phases to CNN operation : training and inference. During training, CNN's learn another representation of the input signal – the representation they learn depends on the application of the CNN. During inference, CNN's will take the representation they learn along with the input signal and produce the desired output.

Due to their customizability and speed, FPGA's are very useful in prototyping DSP applications that require low latency. In our project, we will build a framework which can implement inference of CNN's on an FPGA. We will perform training on a GPU.

We are going to implement two applications of neural nets on FPGA – one for validation and benchmarking and the other for our actual application. For the former, we are going to use MNIST (Modified National Institute Standards and Technology database) handwritten digit database – a dataset commonly used for measuring CNN performance.

For the latter, we are going to connect a video camera to a neural net on the FPGA that implements style transfer in real time. Style transfer is a particular application of CNN's that can transfer an artistic style onto an image (see drawing). With this camera, we will be able to embed different styles of paintings onto a live feed from the camera, leading to interesting applications in the art community.

Convolutional Neural Networks on FPGA's

Brenda So, Cory Nezin