JLESC Collaboration proposal

Title: Investigation of heterogeneous platform for deep learning

Collaborators:
- Volodymyr Kindratenko, NCSA, senior research scientist
- Yiyu Tan, RIKEN AICS, research scientist

Abstract:
Deep learning neural networks currently provide the best solution to many complex problems, such as image recognition and natural language processing. The use of such networks usually involves two steps, namely *training* and *inference*. The inference step obtains a prediction for a given input sample according to the network weights. The training is carried out iteratively to refine the network weights until the desired accuracy is achieved by applying massive training data at the input layer, propagating it through all hidden layers, generating the prediction, and feeding back the prediction error in order to update the weights. Therefore, the training step is computationally intensive since both massive training data and a large number of arithmetic operations are involved. Current mainstream state-of-the-art deep learning algorithms, such as CNN, DNN, heavily rely on dense matrix multiplication operations. Although both GPUs and FPGAs can provide acceleration of matrix operations, GPUs provide a much higher bandwidth to off-chip memory, more float-point computing units, and higher operating clock frequency. Therefore today they are a natural choice for network training. However, an emerging trend in deep learning is to adopt low precision data types to improve system efficiency. As an extreme example, binary convolutional neural networks in which all weights are binary are gaining popularity. Such custom data types are difficult to work with on GPUs. In contrast, FPGAs are designed for extreme customizability and can support any data types. In this research, we will be investigating a heterogeneous platform for deep learning in which GPUs will be applied for training and FPGAs will be initially used for inference, and later on will be adapted for training with customizable data types. Ultimately, we will be looking at how to utilize both GPUs and FPGAs for training and inference in a tightly coupled system.

Contributions:
- An OpenCL-based framework for a heterogeneous platform for deep learning algorithms, such as CNN, DNN.
- GPU and FPGA-specific methodology for optimizations of the OpenCL training and inference kernels.
- Design of an FPGA boosting solution for training in deep learning coupled with a study of the impact of variable data types on training and inference quality.

Timeline:
- September 2017: information gathering and survey
- October 2017 – January 2017: development of OpenCL prototype
- February 2018 – March 2018: platform-specific optimizations and prototype evaluation

Computer resource needs:
- none

Expected results:
- Prototype of a framework for a heterogeneous platform for deep learning
- Papers and/or internal reports
- Joint proposal applications