AMD Proposal, RIPS2017

Title: Exploration of Numerical Precision in Deep Neural Networks towards an Efficient Processor Implementation

Sponsor: Advanced Micro Devices

Industry Mentors: Allen Rush, Ph.D., AMD Fellow, *and* Nicholas Malaya, Ph.D.

Introduction:

Deep Neural Networks (DNN) provide near-human performance on a number of learning tasks, including image classification, speech recognition and language understanding. The task of determining the values of parameters for a deep neural network, called "training" the network and, similarly, the task of utilizing a trained network to perform a specific recognition task, called "inference", are both computationally demanding. This project explores numerical precision requirements of this computation with the goal of utilizing reduced precision computation thus enabling more efficient hardware utilization achieving the desired results sooner, at lower hardware and power requirements.

Technical Background:

A key algorithm for training DNNs is the Stochastic Gradient Descent algorithm [1]. The computational demands of this algorithms are a good match for highly parallel Graphics Processing Units, as well as processors equipped with efficient vector engines. The computational load on such devices includes a cost in hardware architecture, energy consumption during operation, and time to solution. Deep neural networks exhibit a natural resiliency to errors and imprecise computation, having been demonstrated to tolerate computation at reduced precision and computational range. This property can, and has been, exploited to generate more efficient computational algorithms and hardware architectures to achieve efficient training and inference in DNNs by exploiting reduced precision computation, either using lower-precision floating-point or fixed-point arithmetic.

However, work in evaluating the resiliency of DNNs to this date has primarily been empirical [3, 4, 5], and there are few studies that formally examine topics such as algorithm resiliency [2] or numerical stability and convergence [6,7], especially with an eye towards efficient architectural and microarchitectural implementations. This project will examine state-of-the-art algorithms and frameworks for DNN training and inference to explore the opportunities for hardware-efficient reduced-precision computation, including novel approaches such as compensated arithmetic[8]. As an additional stretch goal, students will apply the results of the analysis to a framework on a theoretical basis. This will attempt to relate the outcomes of observations of numerical precision and efficiency to underlying multi-layer perceptron, and Restricted Boltzmann Machine theory for discriminative or generative models.

Expectations:

Student will learn about state-of-the-art deep neural networks, computer arithmetic and numerical representations, and analysis of optimization algorithms using open-source frameworks. Students will characterize numerical precision tolerance of specific machine learning algorithms, and relate their findings towards efficient implementations. This is expected to relate to arithmetic computation available in state-of-the-art hardware, such various flavors of floating-point computation, fixed point and compensation arithmetic, as well as a stretch goal of identifying potential extensions to existing architectures.

This project will teach students about the basic structure of computational units, tradeoffs in numerical representation, and rounding methods. Open source frameworks, Python simulations, and numerical analysis will be the primary tools for this study.

Recommended Reading and References:

[1] Q.V.Le et al, On Optimization Methods for Deep Learning, Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.

[2]Bousquet, Olivier, and Léon Bottou. "The tradeoffs of large scale learning." Advances in neural information processing systems. 2008.

[3]Gupta, Suyog, et al. "Deep Learning with Limited Numerical Precision." ICML. 2015.

[4]Courbariaux, Matthieu, Yoshua Bengio, and Jean-Pierre David. "Training deep neural networks with low precision multiplications." arXiv preprint arXiv:1412.7024 (2014).

[5] Ngiam, Jiquan, et al. "On optimization methods for deep learning." *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011.

[6] Higham, Nicholas J. *Accuracy and stability of numerical algorithms*. Society for industrial and applied mathematics, 2002.

[7] Trefethen, Lloyd N., and David Bau III. *Numerical linear algebra*. Vol. 50. Siam, 1997.

[8] Yamanaka, Naoya, et al. "A parallel algorithm for accurate dot product." *Parallel Computing* 34.6 (2008): 392-410.

Software Packages and Special Requirements:

Numerical Analysis, Applied Mathematics, Programming (C, Python), Computer Architecture, Statistics, Information Theory

Tensorflow, CAFFE, Theano