Hardware Neural Network Accelerators

O. Temam
INRIA Saclay, France
olivier.temam@inria.fr

1. TOWARDS ACCELERATORS

Because of increasingly stringent energy constraints (e.g., Dark Silicon, there is a growing consensus in the community that we may be moving towards heterogeneous multi-core architectures, composed of a mix of cores and accelerators. Because our community is traditionally focused on general-purpose computing, we have been especially considering accelerator approaches such as GPUs and reconfigurable circuits. An attractive alternative is to investigate accelerators which are focused on a few key algorithms: key algorithms still mean broad application scope, but few algorithms enable energy efficient and cost-effective accelerators.

Assuming we want to go down the path of multi-purpose accelerators for energy reasons, the main question becomes: which applications should be considered ? The PARSEC benchmarks have been introduced to highlight a trend towards a new kind of high-performance applications (e.g., voice recognition, image analysis, navigation, etc). Interestingly, many of the core tasks of these benchmarks turn out to correspond to inherently stochastic algorithms/tasks, such as clustering, classification, optimization, filtering, approximation algorithms, i.e., tasks which are inherently tolerant to a certain degree of inaccuracy.

What accelerator design would then be appropriate ? When considering the need for energy efficiency and faults/defects tolerance, as well as the nature of the emerging high-performance applications, *hardware neural networks* come across as an attractive alternative accelerator design. While Neural Networks (NNs) are often considered as a niche application, one can consider all the aforementioned applications: many of the emerging high-performance applications are based on machine-learning techniques, and there are competitive alternatives based on neural networks for all five aforementioned core algorithms (clustering, classification, optimization, filtering, approximation). As a result, NNs are much more a *kernel* based on which many applications can be developed rather than a niche algorithm. At the same time, an NN circuit is much closer to an ASIC than a processor, so an NN accelerator can potentially achieve the energy efficiency of an ASIC while still having the broad aforementioned application scope. Finally, one of the most attractive properties of neural networks are their inherent robustness to faults. Thanks to its learning algorithm, an NN can automatically silence out faulty parts through retraining, without having to identify or disable these faults, an attractive feature.

2. WHY HARDWARE NEURAL NETWORKS AGAIN ?

However, NNs are certainly not a new concept. They have been hyped in the 1980s and 1990s, and then fell into oblivion. Why should they be considered now ? In the 1980s and 1990s, neural networks (NNs) became very popular, both in the machine learning community, and in the hardware community as well, with Intel even manufacturing a hardware neural network chip called ETANN. The reason for the surge of interest in neural networks, especially Multi-Layer Perceptron (MLP), stemmed from the fact the concept was loosely derived from an interpretation of how biological neurons operate in the brain, and from the promising capabilities of neural networks, especially for classification tasks.

They fell out of favor in the 1990s for three reasons: (1) from a mathematical standpoint, ANNs were outperformed by other machine-learning algorithms such as Support Vector Machines (SVMs) with better classification capabilities, (2) hardware NNs could not keep pace with the the speed of software versions run on microprocessors with rapidly progressing clock frequency, the so-called “killer micro”, exactly like massively parallel architectures, and, (3) at a time when scientific computing formed the bulk of high-performance applications, the application scope of NNs was inappropriate.

However, all three situations have drastically changed: (1) Deep Networks, i.e., NNs made of a large number of wide layers, have recently been shown to outperform SVMs on a broad range of tasks, becoming again state-of-the-art machine-learning algorithms, (2) the processor clock frequency has mostly stalled, so that a hardware accelerator will now retain an advantage of several orders of magnitude in energy and performance over a software model run on a processor, and, (3) as already mentioned, NNs are almost ideally suited for many of the emerging high-performance applications.

3. A NICE CONCEPT OR FOR REAL ?

The main goal of our research [1, 2, 4, 3] is to show that hardware neural network accelerators are not just a nice concept, but a proposition with real value in the current technology and application context.

4. REFERENCES

[1] B. Belhadj, A. Joubert, Z. Li, R. Heliot, and O. Temam, “Continuous Real-World Inputs Can Open Up Alternative Accelerator Designs,” in *International Symposium on Computer Architecture*, 2013.

[2] T. Chen, Y. Chen, M. Duranton, Q. Guo, A. Hashmi, M. Lipasti, A. Nere, S. Qiu, M. Sebag, and O. Temam, “BenchNN: On the Broad Potential Application Scope of Hardware Neural Network Accelerators,” in *International Symposium on Workload Characterization*, 2012.

[3] A. Hashmi, H. Berry, O. Temam, and M. Lipasti, “Automatic Abstraction and Fault Tolerance in Cortical Microachitectures,” in *International Symposium on Computer architecture*. New York, NY: ACM, 2011.

[4] O. Temam, “A Defect-Tolerant Accelerator for Emerging High-Performance Applications,” in *International Symposium on Computer Architecture*, Portland, Oregon, 2012.

978-1-4799-1417-3/13/$31.00 ©2013 IEEE