2016 IEEE 24th Annual International Symposium on Field-Programmable Custom Computing Machines

Evaluating Embedded FPGA Accelerators for Deep Learning Applications

Gopalakrishna Hegde, Siddhartha, Nachiappan Ramasamy, Vamsi Buddha, Nachiket Kapre
School of Computer Engineering
Nanyang Technological University
Singapore, 639798
nachiket@ieee.org

***Abstract*—**
FPGA-based embedded soft vector processors can exceed the performance and energy-efficiency of embedded GPUs and DSPs for lightweight deep learning applications. For low complexity deep neural networks targeting resource constrained platforms, we develop optimized Caffe-compatible deep learning library routines that target a range of embedded accelerator-based systems between 4–8 W power budgets such as the Xilinx Zedboard (with MXP soft vector processor), NVIDIA Jetson TK1 (GPU), InForce 6410 (DSP), TI EVM5432 (DSP) as well as the Adapteva Parallella board (custom multi-core with NoC). For MNIST (28×28 images) and CIFAR10 (32×32 images), the deep layer structure is amenable to MXP-enhanced FPGA mappings to deliver 1.4–5× higher energy efficiency than all other platforms. Not surprisingly, embedded GPU works better for complex networks with large image resolutions.

I. Idea

As deep neural networks are becoming popular choice for computer vision and artificial intelligence, their adaptation to embedded scenarios is an interesting challenge. Modern embedded SoCs are typically augmented with accelerators such as GPUs, DSPs or FPGAs. These accelerators can support high processing throughput for specific parallel tasks with low power consumption. However, each platform presents a unique mapping challenge with specific computational, and memory access constraints. For FPGA acceleration, we consider Vectorblox MXP[2], [1], a 16 lane soft vector processor, configured on top of the FPGA fabric. Furthermore, we develop Caffe API support libraries optimized for different embedded SoC platforms.

Optimizations Each embedded platform we consider has unique memory hierarchy, communication bandwidth, memory capacity and processing resources. We express key deep learning routines such as 2D convolution, pooling, activation and inner products that impose varying requirements on these different resources. We use vendor libraries to implement these routines for the DSP and GPU platforms while exposing optimization hooks to enable performance tuning. For FPGA and Parallella implementations, we develop C-code descriptions of the APIs from bottom-up while still exposing optimization hooks to the tuning framework. We then use these hooks in conjunction with high-level specifications of Caffe prototxt network architectures to generate optimized code for the various platforms.

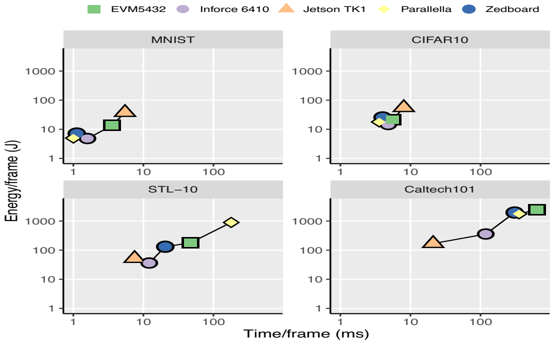

Fig. 1: Comparing Throughput and Energy Efficiency.

In Figure 1 we present combined trends for performance and energy efficiency across platforms for networks on different datasets. On one hand, simpler CNNs which use small images as in MNIST and CIFAR-10 datasets, perform better on the MXP and Parallella (by 2-3×) when compared to GPU and DSP platforms. On the other hand, Jetson TK1 GPU outperforms all other platforms for larger resolutions with complex CNN architectures. Even for complex CNNs, MXP performs 2× better compared to TI DSP and Parallella but is unable to beat Hexagon DSP due to its larger memory, faster clock and 8b fixed point computations used in vendor library. The scalability, flexibility and programmability of FPGA based accelerator such as MXP makes it an attractive solution for diversified deep learning applications.

References

[1] G. Hegde and N. Kapre. Energy-Efficient Acceleration of OpenCV Saliency Computation Using Soft Vector Processors. In *Field-Programmable Custom Computing Machines (FCCM), 2015 IEEE 23rd Annual International Symposium on*, pages 76–83, May 2015.

[2] A. Severance and G. G. Lemieux. Embedded supercomputing in FPGAs with the VectorBlox MXP Matrix Processor. In *Hardware/Software Codesign and System Synthesis (CODES+ ISSS), 2013 International Conference on*, pages 1–10. IEEE, 2013.

978-1-5090-2356-1/16 $31.00 © 2016 IEEE
DOI 10.1109/FCCM.2016.14