CNN Face Detection Application on an Energy-Optimized Accelerator For Neural Networks

Alexandre Carbon, Renaud Schmit, Jean-Marc Philippe

CEA, LIST

Computing and Design Environment Laboratory

F-91191 Gif sur Yvette, France

Email: alexandre.carbon@cea.fr

Abstract—This demonstration shows the efficiency of a new energy-optimized hardware accelerator for deep neural networks compared to two multicore processors typically found in smartphones. On the same face detection application running on an image database, our architecture prototyped on an FPGA is able to be up to 10 times more efficient. This work is the result of a collaboration between an SME (GlobalSensing Technologies) and a research institute (CEA Tech) [1] in a joint laboratory in the NeuroDSP project [2].

I. Introduction and technical details

Deep Neural Networks (e.g. Convolutional Neural Networks - CNN) are a promising approach to design smart machines for a wide range of application domains (automotive, home automation, industry, etc.). They are compute intensive and difficult to embed into low power systems. To tackle this challenge, LIST (a CEA Tech institute) investigated an energy-efficient hardware accelerator IP, able to be embedded into FPGA- or ASIC-based systems. Providing the system with a dramatic performance/watt ratio improvement, the IP can sustain 450GMACS/W in FDSOI 28nm technology, meeting the requirements of high-end embedded applications. This cluster-based architecture (see Figure 1) is generic enough to also support data processing algorithms (pre/post-processing). It has been designed to be scalable in a lot of architectural parameters to be useful in systems ranging from IoT (Internet of Things) platforms to computing servers.
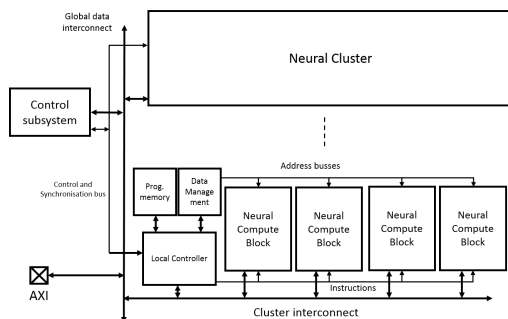

Fig. 1. Architecture of our accelerator.

II. Demonstration and results

The proposed demonstration (see Figure 2) features a comparison between three implementations of the same CNN processing chain used to detect faces in a database containing 18.000 images. It shows that a single cluster FPGA-based implementation of the accelerator IP at 100MHz is able to outperform both a Raspberry Pi 2 and an Odroid-XU3 boards by a factor of respectively 10 and 6 in performance.
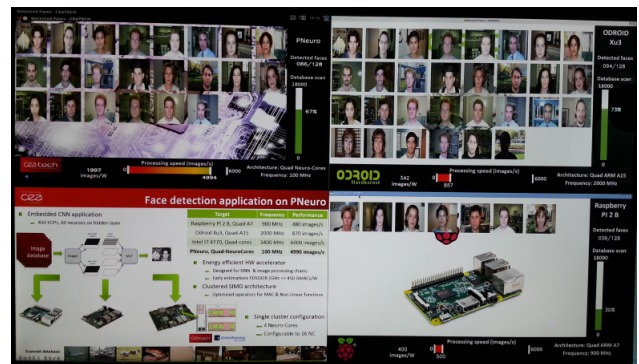

Fig. 2. Overview of the demonstration: top left, our accelerator - top right, Odroid-XU3 platform - bottom right, Raspberry Pi 2 platform.

TABLE I. Performance and energy efficiency of different platforms with respect to the CNN-based application.

Platform	Performance Images/s	Energy efficiency Images/W
Our accelerator (FPGA)	4995	1998
Odroi-XU3	870	350
Raspberry Pi 2	480	400

The topology of the CNN processing chain is the following: an input layer of 48x48 pixels processed by four convolution filters (two 3x3 and two 5x5 coefficients) followed by a max pooling layer and finally sent to a fully connected classifier consisting of 60 neurons in the hidden layer (coefficients and synaptic weights learned using backpropagation of errors). With roughly 1000 lines of instructions for the entire CNN, one cluster of four Neural Compute Blocks of our accelerator IP can simultaneously categorize four images from Caltech-256 database in almost 70.000 clock cycles. The same processing chain was executed on a Raspberry Pi 2 B (900MHz quad-core ARM A7 CPU) in OpenMP and on an Odroid-XU3 (2GHz quad-core ARM A15 CPU) to perform a comparison on this particular application. Publicly available numbers as well as profiling results were used for execution time and power consumption estimates. Table I gives the performance and energy efficiency of the different platforms for the execution of the application. It clearly shows the global efficiency of our accelerator, even on an FPGA.

Based on these promising results, studies have begun to derive a manycore chip including our accelerator in FDSOI 28nm technology, to be integrated into the product line of GlobalSensing Technologies.