Embedded Acceleration of Image Classification Applications for Stereo Vision Systems

Mohammad Loni, Carl Ahlberg, Masoud Daneshtalab, Mikael Ekström, Mikael Sjödin
School of Innovation, Design and Engineering, Mälardalen University,
Västerås, Sweden
{mohammad.loni, carl.ahlberg, masoud.daneshtalab, mikael.ekstrom, mikael.sjodin}.mdh.se

I. Introduction

Autonomous robots use various sensors such as RADAR, LIDAR, cameras and stereo vision cameras for dynamically navigating and exploring unknown environments. What make stereo vison systems attractive is the multimodal sensing, allowing for extraction of three-dimensional (3-D) information, luminance, color, distance, and shape. Current stereo cameras produce high-resolution images which require massive computational resources leading to considerable energy consumption, which is an obstacle for embedded system implementation. To overcome these problems, we use GIMME2 [1], an FPGA-based stereo vision system as a high-throughput and power efficient embedded device which is developed at MDH. The GIMME2 hardware block diagram is illustrated in Figure 1. GIMME2 features two 10 Megapixel cameras and a Xilinx Zynq 7020 SoC, which is equipped with a dual-core Cortex-A9 ARM processor and Artix-7 85K FPGA-fabric.

Approximation computing can be applied to image processing to achieve better performance and/or higher energy utilization. To benefit from this, we introduce an approximation accelerator compatible with GIMME2. Our proposed solution aims to map Deep Neural Network (DNN) based image-classification algorithms to an FPGA by employing DeepMaker, which is an evolutionary based framework embed in our accelerator. There are other NN-based approximation accelerators [2, 3], but they fail to generate an efficient NN architecture. The architecture of the DNN has a great impact not only on the output quality, but also the performance. Previous work employed a simple exploration method to restrict the search space. DeepMaker is an automated framework in which the frontend layer is responsible for generating a robust DNN, and the backend maps the generated network to an FPGA. DeepMaker must be able to efficiently search the vast exploration space, to find the Pareto-optimal surface. Hence, it utilizes a multi-objective genetic programming approach, NSGA-II [4], to discover near-optimal NN architectures regarding network size and accuracy. We employed a template-based DNN accelerator, DNNWeaver [5], to efficiently map DNNs to the FPGA. To evaluate DeepMaker, we used the MNIST dataset, which is for recognition of hand written digits. Figure 2 plots five Pareto points modeled by a multi-layer DNN architecture for the first and the tenth generations. The Pareto-optimal curves are shifted towards left after ten generations which means the set of points on the left curve features improved network architectures in the terms of network accuracy and the network size. In this demo we propose an FPGA-based accelerator for hand written digit classification implemented on the GIMME2 embedded system. In our proposed solution, we first generate an efficient DNN on a HDL-level using DeepMaker, then integrate the generated accelerator module in the processing pipeline of GIMME2, as shown in Figure 1.

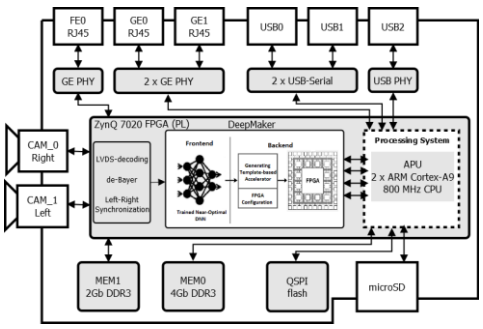

Figure 1. The hardware block diagram of GIMME2

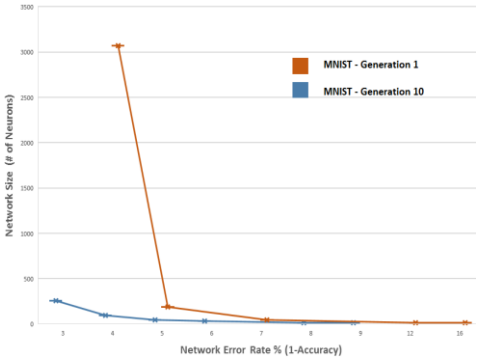

Figure 2. MNIST Pareto frontier curves for generations 1 and 10

References

[1] C. Ahlberg, F. Ekstrand, M. Ekstrom, G. Spampinato, and L. Asplund, "GIMME2 - An embedded system for stereo vision and processing of megapixel images with FPGA-acceleration," in *2015 International Conference on ReConFigurable Computing and FPGAs, ReConFig 2015*,

[2] T. Moreau, M. Wyse, J. Nelson, A. Sampson, H. Esmaeilzadeh, L. Ceze, and M. Oskin, "SNNAP: Approximate computing on programmable SoCs via neural acceleration," in *2015 IEEE 21st International Symposium on High Performance Computer Architecture, HPCA 2015*, 2015, pp. 603–614.

[3] A. Yazdanbakhsh, J. Park, H. Sharma, P. Lotfi-Kamran, and H. Esmaeilzadeh, "Neural acceleration for GPU throughput processors," in *Proceedings of the 48th International Symposium on Microarchitecture - MICRO-48*, 2015, pp. 482–493.

[4] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, 2002.

[5] H. Sharma Jongse Park Emmanuel Amaro Bradley Thwaites Praneetha Kotha Anmol Gupta Joon Kyung Kim Asit Mishra Hadi Esmaeilzadeh, "DNNWEAVER: From High-Level Deep Network Models to FPGA Acceleration," *IEEE Int. Conf. Mechatronics, Electron. Automot. Eng.*, no. 2, pp. 76–80, 2015.