EASpiNN: Effective Automated Spiking Neural Network Evaluation on FPGA

Sathish Panchapakesan*, Zhenman Fang*, Nitin Chandrachoodan†,
* School of Engineering Science, Simon Fraser University, Canada
† Department of Electrical Engineering, Indian Institute of Technology - Madras, India
{sathishp, zhenman}@sfu.ca nitin@ee.iitm.ac.in

Neural networks (NNs) have been widely used in many machine learning algorithms and have been deployed for various industrial applications like image classification, speech recognition, and automated control. Spiking neural network (SNN), known as the third-generation neural network, incorporates timing information in the network and is more biologically plausible [1]. Compared to today's artificial and convolutional neural networks (ANN and CNN) where all neurons in each layer will always be activated and computed, SNN only activates those neurons whose membrane potential exceed the threshold potential [2]. As a result, SNN requires fewer computation resources and less data communication between network layers due to its event-driven nature. Although SNN has been blamed for the relatively lower accuracy, recent studies on converted SNNs have improved its accuracy to a similar level of ANN and CNN for smaller network models like MNIST and CIFAR-10, and have demonstrated the great potential of SNN in future deep learning systems [2].

For this work, we focus on SNNs that have been obtained through the conversion of ANNs (fully-connected layers) and leave convolutional layers for future work. Although this limits the accuracy of larger networks, our current focus is on the aspects of hardware implementation. A major issue with the converted SNNs is that their network topology is based on the underlying ANN model, which results in a large number of memory accesses. This is especially significant in resource constrained devices that are used for edge inference, where low-latency low-power memory is at a premium.

In this paper, we design and implement a framework called EASpiNN, with the goal to enable fast and effective evaluation of various SNN-based network models for inference on edge devices. EASpiNN implements the widely used integrate-and-fire (IF) SNN model [2] on Xilinx ARM-FPGA System-on-Chips (SoCs) using high-level synthesis (HLS) C++. EASpiNN can automatically run any MNIST and CIFAR-10 based networks without rebuilding the hardware design on the same ARM-FPGA SoC. Moreover, it supports the automatic selection of the optimal design point across a range of Xilinx ARM-FPGA SoCs including the ZedBoard, Zynq ZC706, Zynq UltraScale+ ZCU102 and ZCU104 boards.

Within EASpiNN, we optimize the performance of the SNN implementation by customizing both its computation and memory access. For computation optimization, we explore the loop pipelining and parallelization techniques for major computing engines [3]. For the memory optimization, the biggest challenge is to buffer the large weight matrix using on-chip BRAM and/or UltraRAM (URAM) resource: the weight matrix dominates the storage requirement for ANNs like the MNIST and CIFAR-10 networks and its size exceeds the total size of on-chip BRAM and/or URAM on an embedded FPGA. EASpiNN automatically decides the optimal cutoff point to partially buffer the maximum amount of weight matrix, based on the network model parameters and the size of the available BRAM and/or URAM resource on a given FPGA. Finally, it also enables burst access for all off-chip DRAM accesses.

Our EASpiNN framework is built using Xilinx SDSoC 2019.1 and the FPGA accelerator runs at a frequency of 100MHz across all the four aforementioned Xilinx ARM-FPGA SoCs. In our preliminary implementation, no model compression has been applied yet, and the weights and membrane potential are full 32-bit floating point numbers without any data quantization. For three different network models MNIST (610 neurons), MNIST (2,410 neurons), and CIFAR-10 (2,410 neurons), EASpiNN achieves 9.1x, 4.3x, and 4.2x speedups over the ARM CPU on the Xilinx ZCU104 FPGA board that has the largest amount of on-chip memory among the four boards we studied. Even on the ZedBoard that has the smallest amount of on-chip memory, EASpiNN achieves 2.31x, 1.76x, and 2.95x speedups over the ARM CPU.

ACKNOWLEDGMENTS

We acknowledge the support from Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant RGPIN-2019-04613 and DGECR-2019-00120), Simon Fraser University New Faculty Start-up Grant, Huawei Canada and Xilinx.

REFERENCES

[1] M. Pfeiffer and T. Pfeil, "Deep learning with spiking neurons: Opportunities and challenges," *Frontiers in Neuroscience*, vol. 12, p. 774, 2018.

[2] P. U. Diehl, D. Neil, J. Binas, M. Cook, S. Liu, and M. Pfeiffer, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–8.

[3] J. Cong, Z. Fang, M. Lo, H. Wang, J. Xu, and S. Zhang, "Understanding performance differences of fpgas and gpus," in *2018 IEEE 26th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, April 2018, pp. 93–96.