Automatic Generation of FPGA Kernels From Open Format CNN Models

Dimitrios Danopoulos
Department of Electrical
and Computer Engineering
NTUA, Athens, Greece
dimdano@microlab.ntua.gr

Christoforos Kachris
Democritus University of Thrace
& ICCS-NTUA
Athens, Greece
kachris@microlab.ntua.gr

Dimitrios Soudris
Department of Electrical
and Computer Engineering
NTUA, Athens, Greece
dsoudris@microlab.ntua.gr

***Abstract*—The continuing exponential increase of deep learning applications like image classification or object detection requires faster and faster processing speeds while keeping the development time small. Specifically, there is a broad interest for unifying machine learning models into a universal ecosystem so that developers can benefit from framework interoperability and seamless device-specific acceleration. This is a more challenging task for FPGAs which are promising platforms but need extra effort in order to be part of this ecosystem. This work is based on an early development stage open-source project which is called HLS4ML originally created for particle physics applications via the automatic translation of neural networks on embedded Xilinx FPGAs. Our proposed solution involves a generalized optimization scheme on top of HLS4ML that automatically converts open format AI models called ONNX for cloud FPGAs. Our design also achieved in a demonstrated inference 102× over single-core CPU and 6.6× over GPU with a good trade-off between accuracy.**

The field of Machine Learning has dramatically improved in recent years with new emerging applications that achieve impressive results in various areas [1]. Convolutional Neural Networks (CNNs) have gained significant traction due to their high accuracy and performance on visual recognition algorithms [2]. FPGA implementations on the other hand have seen great advancement as is it shown that they have been extremely effective on CNN tasks due to their massive parallelism and reconfigurability on the bit level. However, it is still hard for AI developers to move models between state-of-the-art tools and even harder to optimize them using hardware-specific acceleration. We present to the research community new architectures and optimization techniques for automatic translation of neural networks to cloud FPGAs all running fully in the reconfigurable hardware. Consequently, the development process of ONNX-based deep learning applications on FPGAs is faster and facilitates neural network compilation and acceleration in general. In summary, the main contributions of the paper are as follows:

- A new scheme for automatic translation of neural networks, specifically open format ONNX models on cloud FPGAs using as base design the HLS4ML project [3].
- Novel optimizations on kernel, memory and host level for cloud FPGAs like the Xilinx Alveo U200.
- A Neural Network (NN) inference for demonstration, from the hardware oriented training to the automatic HLS generation and last the inference on an Alveo U200.

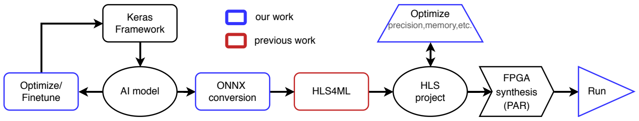

Fig. 1. Design steps for ONNX model deployment to FPGAs.

As for *performance*, our hardware engine uses a flexible heterogeneous streaming architecture with different precision between NN layers. Every layer is parallelized acting as a seperate module and feeds its output to the next layer which accepts the previous output layout as input overlapping the previous layer's operation. Thus, we did not need to store each layer's intermediate results in off-chip memory since they are immediately passed down the stream. Also, we used the most efficient multipliers for the computations (two 8-bit weights multiplied in a single DSP) while keeping the accuracy error small. In terms of *modularity*, we created an heterogeneous NN with the appropriate accuracy in each layer depending on the activations needed which can be changed seamlessly to fit the needs of the application. Last, as for the *re-configurability* of our design, our HLS project can be synthesized for a wide range of OpenCL FPGA devices with multiple copies of the same kernel for batch processing scenarios. Our results on a MNIST-based clothing dataset showed that the proposed architecture can achieve inference of $6.3\mu s$ latency surpassing the performance of a Xeon 2.4GHz CPU with 102.3× speed-up and the performance of a K80 GPU with 6.6× speed-up.

This project was funded from the Xilinx University program and the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No 2212- Hardware Acceleration of Machine Learning Applications in the Cloud.

References

[1] D. Danopoulos, C. Kachris, and D. Soudris, *Approximate Similarity Search with FAISS Framework Using FPGAs on the Cloud*. 08 2019.

[2] D. Danopoulos, C. Kachris, and D. Soudris, "Acceleration of image classification with caffe framework using fpga," in *2018 7th International Conference on Modern Circuits and Systems Technologies (MOCAST)*.

[3] J. Duarte, S. Han, P. Harris, S. Jindariani, E. Kreinar, B. Kreis, J. Ngadiuba, M. Pierini, R. Rivera, N. Tran, and Z. Wu, "Fast inference of deep neural networks in FPGAs for particle physics," *Journal of Instrumentation*, vol. 13, pp. P07027–P07027, jul 2018.

2576-2621/20/$31.00 ©2020 IEEE
DOI 10.1109/FCCM48280.2020.00070