The HLSinf AI hardware accelerator for Safety-related Applications in SELENE

Laura Medina *, Jose Flich *,
Carles Hernàndez *

* *Universitat Politècnica de València (Spain)*

ABSTRACT

Machine Learning (ML) applications have improved rapidly over the past decade. This improvement implies an increase in the amount of processed data and requires devices capable of satisfying the computing requirements imposed by the ML algorithms. Furthermore, safety-related autonomous applications demand computing platforms capable of executing efficient NN inference processes while simultaneously meeting the safety requirements. This paper introduces the HLSinf accelerator and describes the approach followed in the SELENE project to accelerate inference processes.
KEYWORDS: Neural Networks; Artificial Intelligence; inference; real-time

1 Introduction

Artificial Intelligence (AI) is becoming increasingly popular in applications that require processing a large amount of data. Machine Learning (ML) algorithms can be seen behind many applications. It is not only common to see AI applications behind daily-use technologies like social media or phone face unlocking but also in medical areas where ML models are being developed for disease detection [RR+19].

The goal of the H2020 SELENE project is the development of a flexible computing platform for autonomous applications. One of the requirements of SELENE is to achieve real-time Neural Network (NN) inference. The SELENE Acceleration Framework (SAF) is used to ease the utilization of the AI accelerators in the SELENE platform. The SAF consists of several building blocks: a deep learning library, an application programming interface (API), and the AI hardware accelerator.

As a deep learning library, we have chosen The European Distributed Deep Learning Library (EDDL) [pro20]. EDDL is an open-source platform that enables the definition, training, and inference of NN models. The EDDL provides native support for FPGA acceleration which also eases HW/SW co-design. The API developed for Linux provides end-users an easy way to offload computations from the cores to the accelerators. As an AI accelerator, we have developed the HLSinf accelerator. HLSinf is an open-source platform used to accelerate inference processes of NN models. One of the main characteristics of this accelerator is its

[1]E-mail: {laumecha, jflich, carherlu}@upv.es

flexibility. HLSinf provides several parameters that allow the creation of customized implementations. This accelerator is specially well suited for SoCs, including embedded FPGAs or systems including FPGA devices.

2 The SELENE Acceleration Framework

The SELENE System-on-Chip (SoC) is formed by a multi-core NOEL-V RISC-V system divided into General Purpose Processing (GPP) elements, L2 cache, an AI accelerator, memory controllers, and IO elements.

The SELENE Acceleration Framework (SAF) is an Application Programming Interface (API) for interfacing hardware (HW) memory-mapped accelerators implemented in the SELENE SoC with the upper software (SW) layers, the inference library, and the user application. The SAF provides the interface between the HLSinf AI accelerator implemented in HW and the EDDL deep learning inference library used by the user application executing the ML model, as shown in Figure 1.

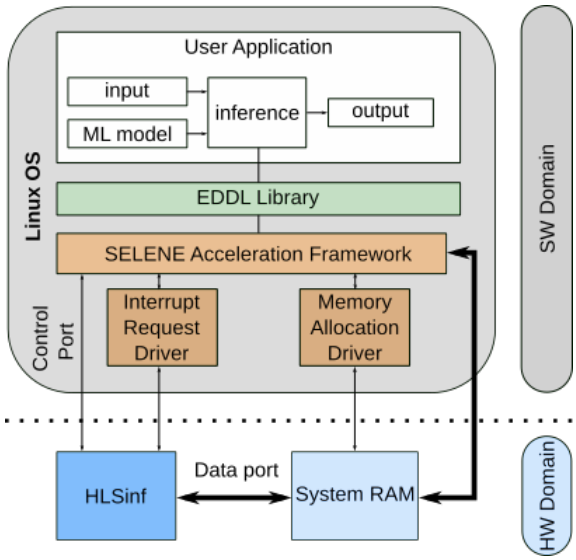

Figure 1: SELENE Acceleration Framework

3 HLSinf AI Harwdare Accelerator

HLSinf is an accelerator for inference processes of NN models based on convolutions. This platform is a High-Level Synthesis open-source project [UPV21]. This accelerator is designed using the channel slicing concept where a set of input channels are taken as an input in parallel, and a set of output channels are produced in parallel. The set of input channels defines the input speed up, and the set of output channels defines the output speed up. The HLSinf design is shown in Figure 2. The accelerator is implemented around the dataflow model where streams interconnect modules. Each module can perform one or more operations needed in NN. This design allows us to pipeline convolutions with additional functions.

One of the main goals of this accelerator is its flexibility. The HLSinf allows us to define several characteristics: the type of operation to support, the convolution operation, the data

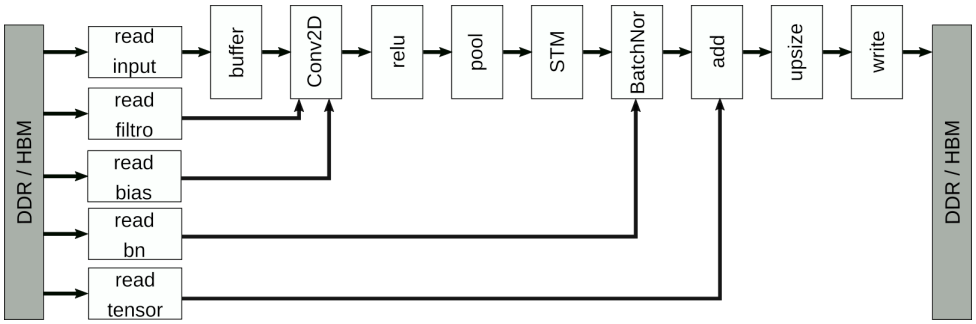

Figure 2: HLSinf accelerator overall design.

type of each module, and the input speed up and the output speed up. This flexibility allows us to implement a specific HW accelerator targeted to a particular use case, NN model, or FPGA device.

3.1 HW/SW co-design

HLSinf has been designed to run in the EDDL library providing the needed support to run offloaded AI model layers on the FPGA. HLSinf can be used to configure and compile a given subset of network layers for its use in an inference process running with EDDL. HLSinf and EDDL allow a perfectly coupled HW/SW co-design approach where some parts of the model run in the FPGA, whereas the rest run in the CPU or GPU when available.

To allow for an effective HW/SW co-design with the SELENE platform, we have designed a new layer in EDDL called HLSinf. This new layer contains all the needed information for running HLSinf-supported NN algorithms in the HLSinf accelerator implemented in the FPGA. Then, the EDDL will run each layer using the proper device. CPU and GPU for non-HLSinf layers and FPGA for HLSinf layers. For the compatibility between devices, we have created a new transforming model functionality in EDDL. This functionality merges layers into a single HLSinf layer, adapts layers parameters for the HLSinf accelerator, and performs the data transformation when a CPU/GPU layer feeds an FPGA layer. For the data transformation, a new Transform layer has been created.

4 Integration Results

Figure 3 shows the average inference time of 10 images in the EDDL library using the Stacked Hourglass network model [N$^+$]. These results compare the inference time for the Intel Core i7-7800X CPU (represented as Intel CPU) and several HLSinf implementations on the Alveo U200 board. More specifically, for the FPGA, three different implementations can be seen: a CPI and CPO of 4 and 32-bits floating-point data type implementation, a CPI and CPO of 8 and 16-bits fixed-point data type implementation, and, finally, a CPI and CPO of 16 and 8-bit integer data type implementation. For the CPU results, all the model layers have been executed on a six-core CPU. On the other side, for the FPGA results, the supported layers run on the FPGA while the unsupported layers run on the CPU. If we compare the Intel CPU and the 16x16 implementation, we can obtain a speedup of 4 for the Stacked Hourglass model.

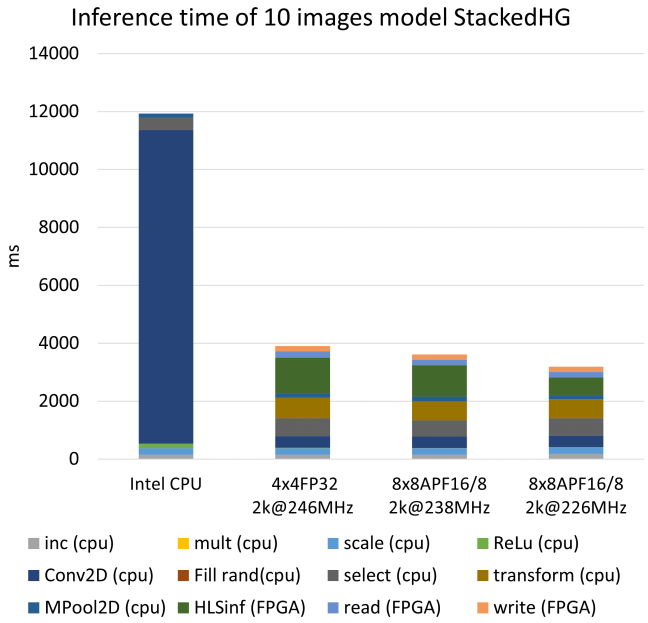

Figure 3: Inference time of the model StackedHG

5 Conclusions

This paper introduces the HLSinf AI hardware accelerator and its integration with the SELENE open-source acceleration framework. In particular, it describes how NN models can be deployed in our platform. Results show that with HLSinf, we can accelerate the inference process by a factor of 4 compared with an Intel Core i7-7800X CPU.

6 Acknowledgements

This work has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement no. 871467

References

[N+] Alejandro Newell et al. Stacked hourglass networks for human pose estimation. In *Computer Vision - ECCV 2016*, pages 483–499.

[pro20] H2020 DEEPHEALTH project. European distributed deep learning (eddl) library, 2020. https://github.com/deephealthproject/eddl.

[RR+19] Rodriguez-Ruiz et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *JNCI: Journal of the National Cancer Institute*, 111(9):916–922, 03 2019.

[UPV21] UPV. Hlsinf neural network inference accelerator for fpgas. `https://github.com/PEAK-UPV/HLSinf`, 2021.