HSLU Hochschule Luzern

Technik & Architektur

A CNN Acceleration Framework for FPGA-SoC
Master Thesis Electrical Engineering

Master Thesis Electrical Engineering

A CNN Acceleration Framework for FPGA-SoC

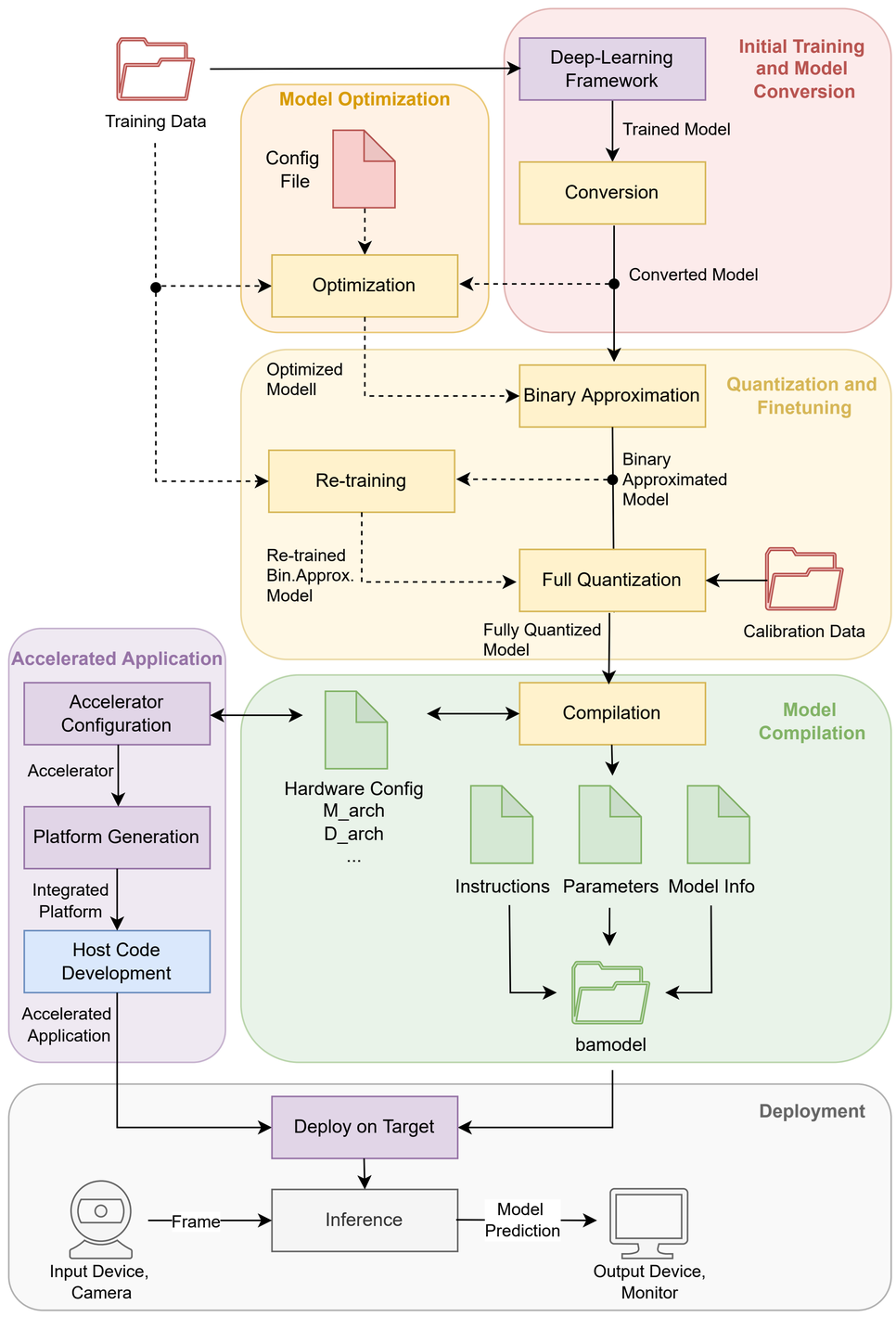

Fig. 1: The proposed BinAccel framework.

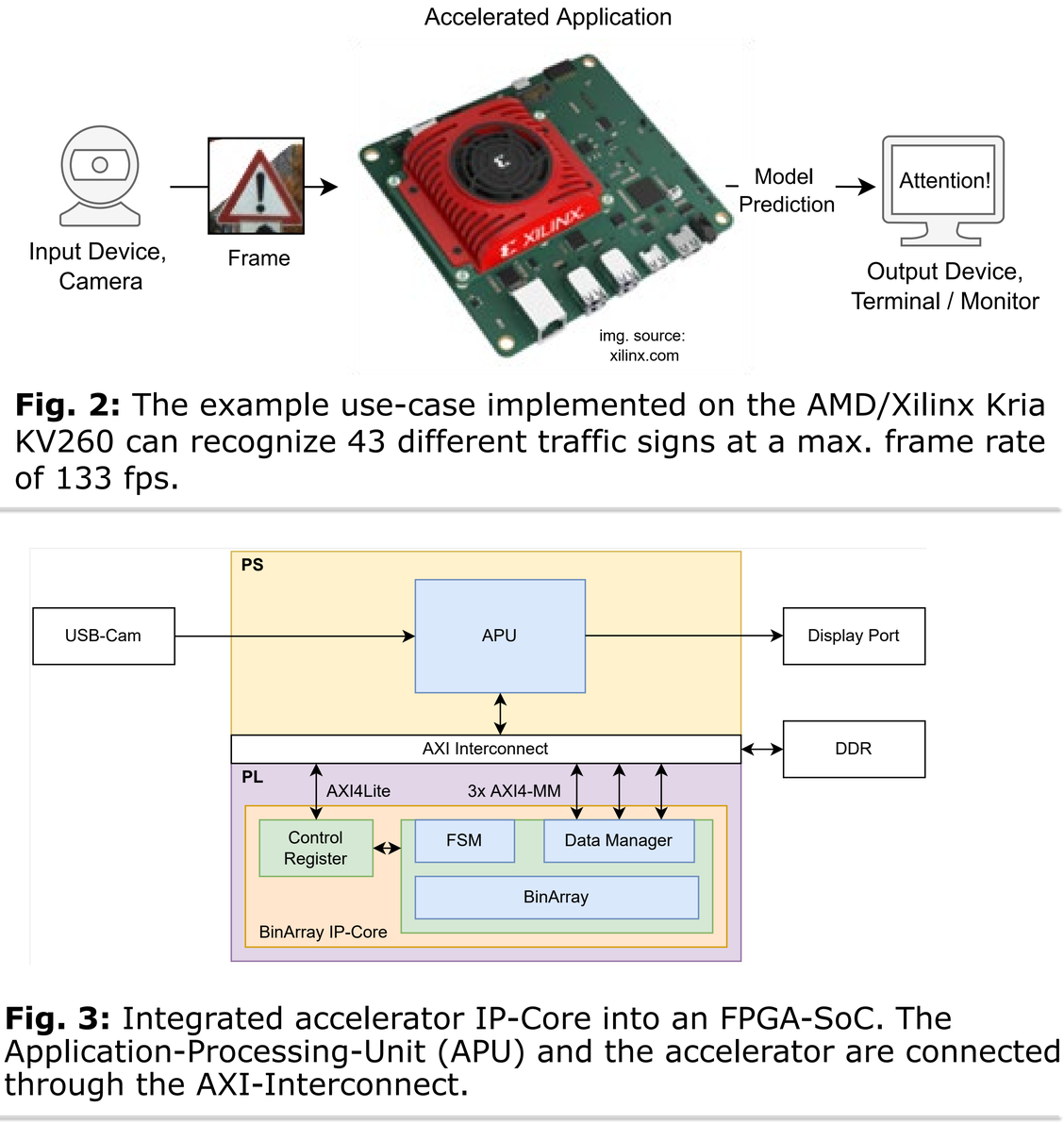

Fig. 2: The example use-case implemented on the AMD/Xilinx Kria KV260 can recognize 43 different traffic signs at a max. frame rate of 133 fps.

Fig. 3: Integrated accelerator IP-Core into an FPGA-SoC. The Application-Processing-Unit (APU) and the accelerator are connected through the AXI-Interconnect.

Task Description

Convolutional Neural Networks (CNNs) are a type of deep learning algorithm that are state-of-the-art to perform tasks such as image classification, object detection, and natural language processing. Because CNNs are computationally intensive, FPGA-SoCs are widely used to accelerate CNNs by providing parallel processing capabilities.

The main objective of this study was to design and implement a flexible CNN acceleration framework **BinAccel** for the efficient proprietary BinArray method, which has been developed in previous work. The acceleration framework concept includes tools to optimize, convert, quantize and compile a CNN into a deployable model format, as well as the steps to build an FPGA-SoC platform with an integrated accelerator for binary approximated CNNs.

Approach

Main parts of the framework have been implemented and applied on an example use-case: A deep-learning based traffic sign recognition application on live input data. The configurability of the acceleration framework has been demonstrated by going through the design steps for multiple points of the design space. Specifically, the CNN model has been quantized and compiled for different hardware accelerator configurations. To improve processing performance, different options have been analyzed and evaluated to increase parallelism in the accelerator, which is crucial for large CNN models that require more computational power.

Results

The results on the example use-case demonstrated how the accuracy, throughput, and resource utilization can be influenced by the user through the configuration options of the framework. A comparison of the processing performance with a non-accelerated version of the application showed that by only accelerating the convolutional layers of the model, the inference can be speeded up by a factor of 1.4. Suggestions were provided to further enhance the overall performance of the application. Furthermore, it was shown that extending the accelerator using input channel parallelization (ICP) makes the BinArray method competitive with commercial frameworks.

Bettina Wyss

Advisor
Prof. Dr. Jürgen Wassner

Expert
Dr. Marc Wegmüller

Research Project
ESA Study Object Tracking

FH Zentralschweiz