Accelerating Convolutional Neural Networks for TensorFlow Lite on Embedded FPGA with Custom Floating-Point Computation

1st Yarib Nevarez
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Convolutional neural networks (CNNs) have become ubiquitous in the field of computer vision and image processing. Given the elevated computational demands of CNNs, dedicated hardware accelerators have been developed to enhance performance and energy efficiency. However, most of commercial deep learning processing units (DPUs) are not targeting compatibility for resource-limited FPGAs. In this publication, we present a dedicated hardware accelerator for TensorFlow (TF) Lite on embedded FPGA for CNN and depthwise CNN. The hardware design is implemented with high-level synthesis (HLS). This accelerator incorporates support for TF Lite quantization for fixed-point and floating-point representations. The proposed hardware optimization decomposes floating-point calculation for the convolution dot-product in order to accelerate computation, reduce energy consumption and resource utilization. To demonstrate the potential of the proposed architecture, we address a design exploration with custom-built CNNs with fixed-point quantization, floating-point single precision, half-precision, brain floating point, NVidia's TensorFloat, and customized reduced formats for approximate computing including logarithmic representation. A single accelerator instance on a Xilinx Zynq-7020 achieves a peak runtime acceleration of $45\times$ on convolution operators compared to the embedded CPU, and $5\times$ compared with the standard Xilinx floating-point LogiCORE IP on MAC operations. With regards to throughput and power efficiency, a single accelerator at 150 MHz yields 152 MFLOP/s and 1.1 TFLOPS/watt, respectively. The entire hardware design and the implemented TF Lite delegate extensions are available as open source project.**

***Index Terms*—Artificial intelligence, convolutional neural networks, hardware accelerator, embedded systems, FPGA, custom floating-point, logarithmic, approximate computing**

This work is funded by the *Consejo Nacional de Ciencia y Tecnologia – CONACYT* (the Mexican National Council for Science and Technology).

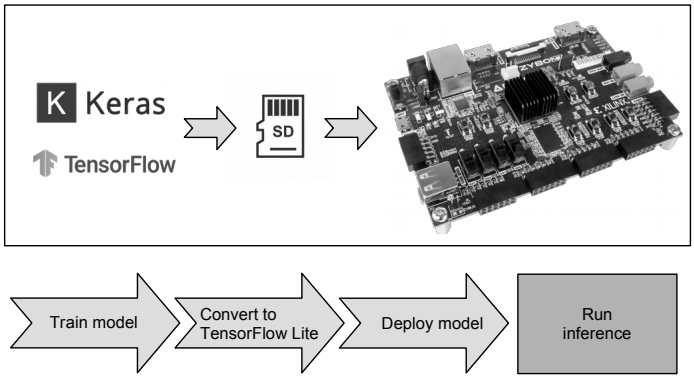

Fig. 1. Workflow.

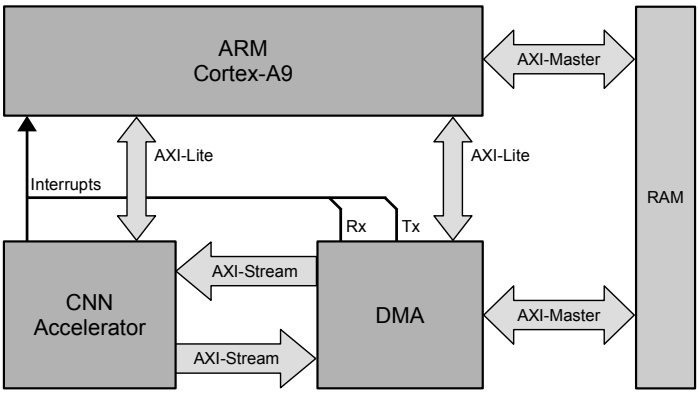

Fig. 2. System-level architecture of the proposed embedded platform.

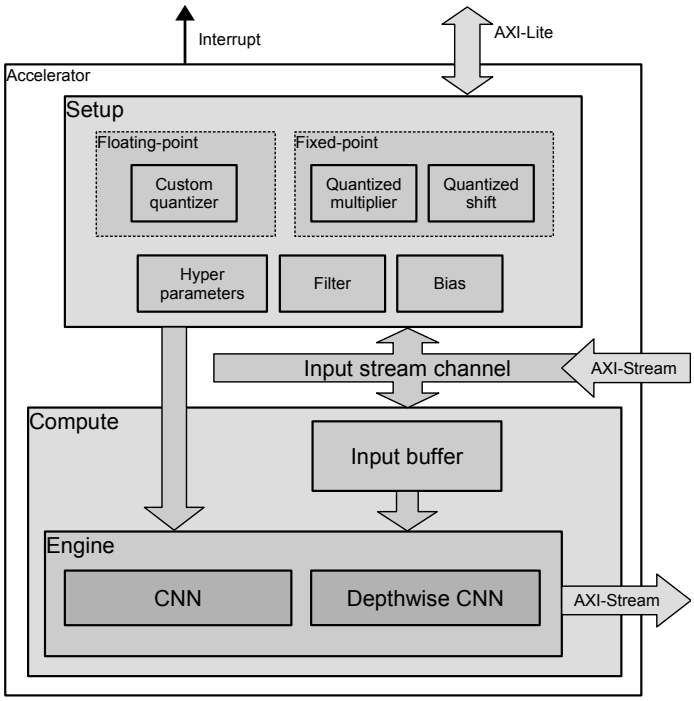

Fig. 3. Hardware architecture of the proposed accelerator.

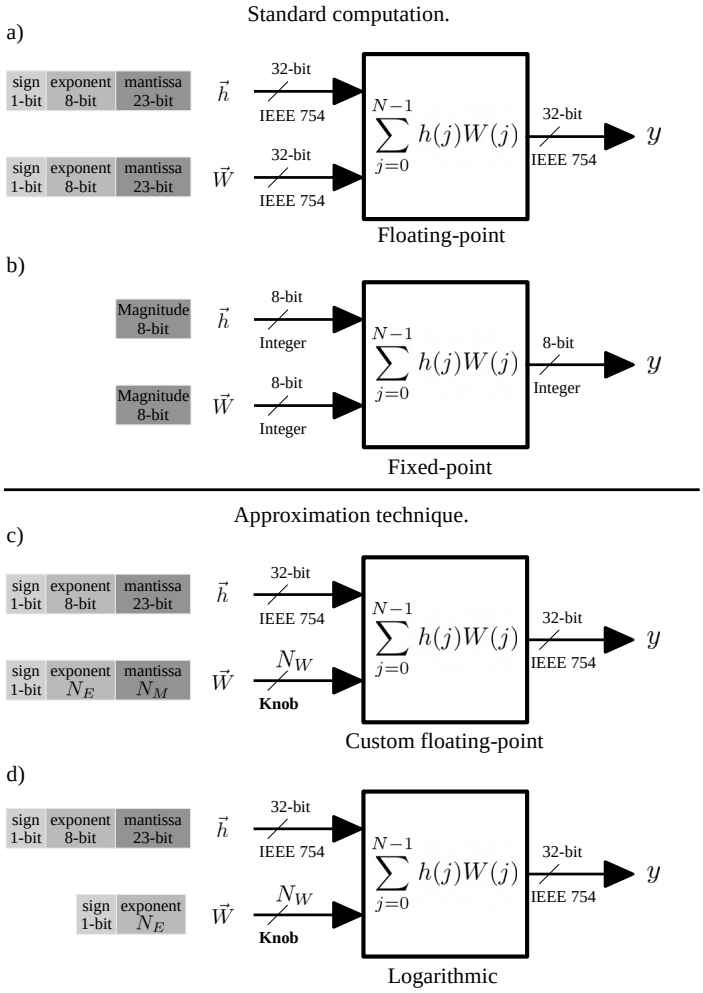

Fig. 4. Proposed hardware modules for vector dot-product.

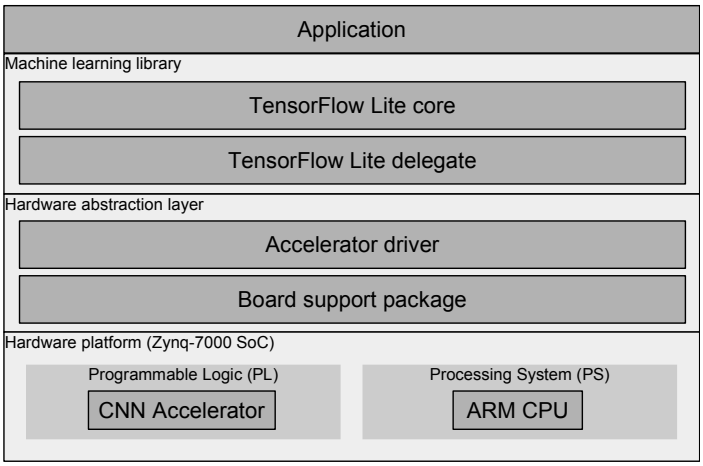

Fig. 5. System-level overview of the embedded software architecture.