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 high computational demands of CNNs, dedicated hardware accelerators have been developed to improve performance and energy efficiency in FPGAs. However, most commercial general-purpose deep learning processing units (DPUs) struggle with support for low-power, resource-limited devices. In this publication, we present a dedicated hardware accelerator for TensorFlow (TF) Lite on embedded FPGA to delegate convolution and depthwise convolution tensor operators. The hardware design is implemented with high-level synthesis (HLS). This accelerator incorporates the support for TF Lite quantization for fixed-point and floating-point. The proposed compute optimization decomposes floating-point calculation for the convolution dot-product. This approach accelerates computation, reduces energy consumption and resource utilization. To demonstrate the potential of the proposed accelerator, we address a design exploration with custom-built CNNs covering fixed-point quantization, floating-point single precision, half-precision, brain floating-point, TensorFloat, and custom reduced formats for approximate processing, including logarithmic computation. A single accelerator running at 150 MHz on a Xilinx Zynq-7020 achieves 45X runtime acceleration on convolution tensor operators compared with ARM Cortex-A9 at 666MHz, and 5X compared with the equivalent implementation with Xilinx LogiCORE IP. This accelerator yields a peak performance of 1.1 TFLOPS/watt and 152 MFLOP/s. The entire hardware design and the implemented TF Lite software extensions are available as open-source project.**

Index Terms—**Artificial intelligence, convolutional neural networks, depthwise separable convolution, hardware accelerator, TensorFlow Lite, embedded systems, FPGA, custom floating-point, logarithmic computation, approximate computing**

I. INTRODUCTION

II. RELATED WORK

III. BACKGROUND

IV. SYSTEM DESIGN

V. EXPERIMENTAL RESULTS

VI. CONCLUSIONS

ACKNOWLEDGMENTS
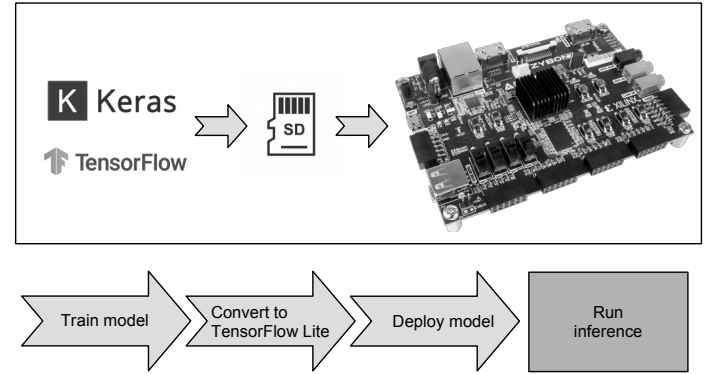
Fig. 1. Workflow.
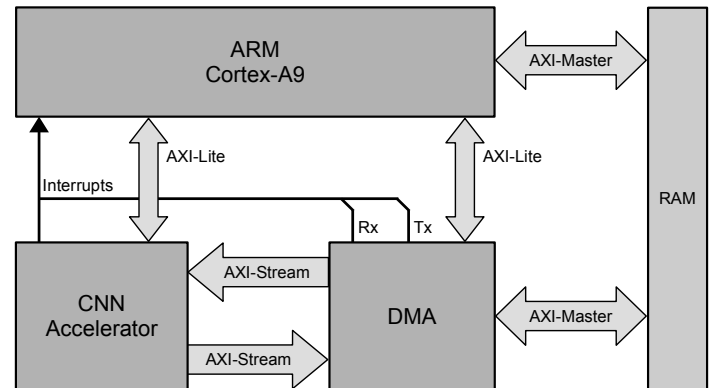
Fig. 2. System-level architecture of the proposed embedded platform.

Standard computation.

a)

Floating-point

b)

Fixed-point

Approximation technique.

c)

Custom floating-point

d)

Logarithmic

Fig. 4. Proposed hardware modules for vector dot-product.
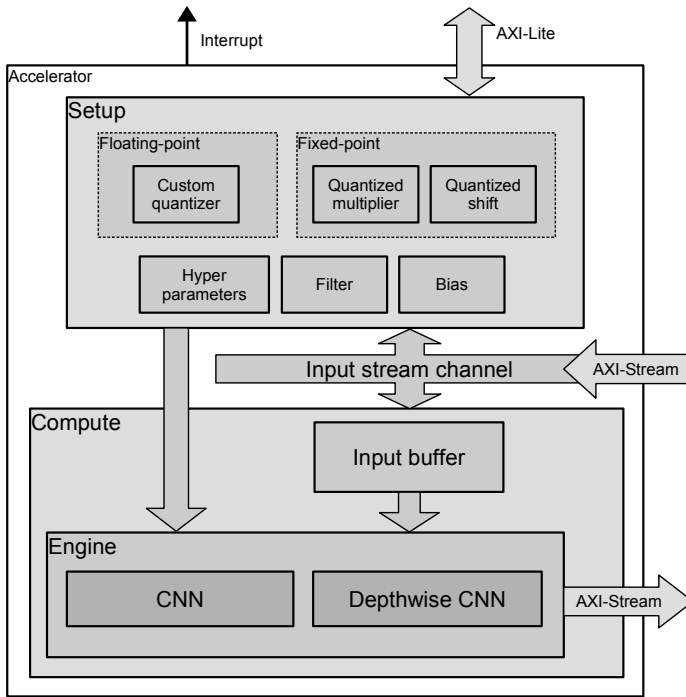
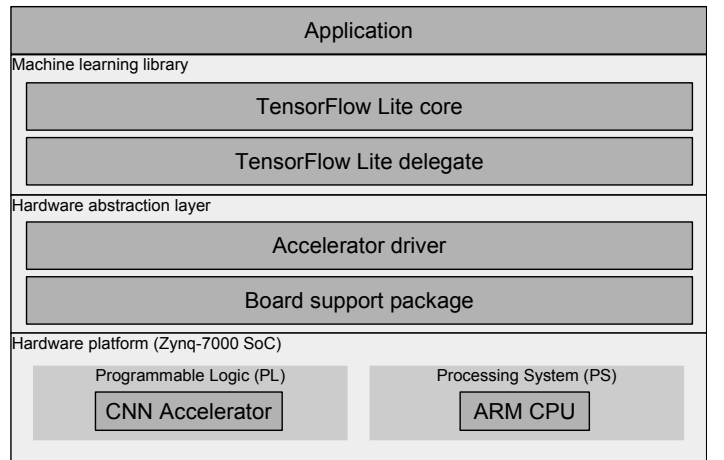Fig. 3. Hardware architecture of the proposed accelerator.

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