BIOGRAPHICAL SKETCH

From 2012 to 2015, Muluken T. Hailesellasie was a Research Assistant with the Institute of Electron Devices and Circuits, Ulm University. He joined Intel as an SoC Design Engineer, in 2018, as Electrical Validation Engineer, in 2017, and as Product Development Engineer, in 2016. Since 2015, he has been a Research Assistant with the Electrical and Computer Engineering Department, Tennessee Tech. He has received funding from Intel Inc. for a Capstone research project at Tennessee Tech in the 2017-2018 academic year. He has reviewed several peer-reviewed IEEE conference papers and journals.

EDUCATION

Addis Ababa University
Addis Ababa, Ethiopia
BS, Electrical & Computer Engineering, 2009

Ulm University
Ulm, Germany
MS, Electrical Engineering, 2015

Tennessee Technological University
Cookeville, Tennessee, USA
PhD, Engineering, August 2019 *(expected)*

College of Engineering

TENNESSEE TECH

The Department of

Electrical & Computer Engineering

Announces the Dissertation Defense

of

Muluken T. Hailesellasie

In Partial Fulfillment of the Requirements

For the degree of

Doctorate of Philosophy

May 13th, 2019

Held in

208 Brown Hall at 3:00 p.m.
115 West 10th Street
Tennessee Tech University

FIELD OF STUDY

Embedded Systems, Digital Design and Signal Processing

DISSERTATION TOPIC

"HARDWARE PROCESSOR DSIGN FOR REGULAR CONVOLUTIONAL NEURAL NETWORKS TARGETING RESOURCE-CONSTRAINED DEVICES WITH AN AUTOMATED FRAMEWORK"

EXAMINING COMMITTEE

Dr. Syed Rafay Hasan, Committee Chair
Professor, Electrical & Computer Engineering

Dr. William Eberle
Professor, Computer Science

Dr. Mohamed Ashiqur Rahman
Electrical & Computer Engineering,
Florida International University

Dr. Mohamed Mahmoud
Professor, Electrical & Computer Engineering

Dr. Ghadir Radman
Professor, Electrical & Computer Engineering

Abstract

The popularity of deep learning has radically increased in the past few years due to its promising results. Convolutional Neural Network (CNN) is one of the most widely used deep learning algorithms in various computer vision applications. While the performance of CNN is impressive, their deployment in the current embedded technology is a challenge since CNN models are computation-intensive and memory-intensive. Hence, there is a growing need for hardware-based solutions in these embedded technologies that can make the dream of real-time computer vision in resource-constrained devices a reality. Due to their reconfigurability, high-performance and low-power features, embedded systems with Field Programmable Gate Arrays (FPGAs) are becoming a hardware platform choice for many deep learning applications. In this work, we explore various hardware architectures and design strategies to improve computation time and minimize the required computing resources of a model when deployed in existing FPGA technology. To alleviate the computation intensiveness of CNN models, first, we propose an efficient convolutional layer architecture with improved computation time per convolution. The proposed architecture finds a trade-off between latency and resource consumption through a technique of distributing the input data into a number of memory blocks. By distributing the input data into parallel memory blocks, where each memory block can be read simultaneously, clock cycle reduction is achieved. Subsequently, we propose an architecture that performs parallel computation of feature maps using a custom designed data flow. The strategy proposed obtained substantial computation speedup compared to the state-of-the-art for the same CNN model. On the other hand, while there is a need for flexible architectures that can be used for various models, the existing architectures are tailored or optimized to a particular CNN architecture. To address this limitation, we propose a novel and highly flexible hardware architecture that can process most regular CNN variants and achieves better resource utilization. We proposed processing cores implemented with multipliers and without multipliers. A fixed-point and power-of-2 quantization schemes are also developed to significantly reduced the on-chip memory space and the logic needed in the targeted device. With substantial on-chip memory reduction and an increase in performance and power efficiency, our results demonstrate that the proposed architecture can be very expedient for resource-constrained devices. To enhance the usability of our proposed architecture for deep learning practitioners and to improve the scalability of the proposed design, a framework that auto-generates a CNN processor in the form of a synthesized hardware intellectual property (IP) is proposed. The proposed framework optimizes the hardware IP based on the model workload and the target device specifications. A memory traffic optimization algorithm that results in higher performance and on-chip fitting optimization that results in higher resource utilization efficiency are employed. Our results demonstrate that the proposed framework is effective in reducing the design time and optimizing the performance and the resource consumption of the hardware IP.