Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Electrical, Electronics and Communications Engineering
(36^{th} cycle)

Accelerating Quantized DNNs with Dedicated Hardware Accelerators and RISC-V Processors Using Precision-Scalable Multipliers

Luca Urbinati

* * * * *

Supervisors
Prof. Casu, Mario R.
Prof. Lavagno, Luciano

Politecnico di Torino
July 1st, 2024

Summary

Mixed-Precision Quantization (MPQ) and Transprecision Computing (TC) represent two valuable techniques used to optimize Deep Neural Networks (DNNs) inference. They aim at minimizing the number of activation and weight bits for each DNN layer during training, and dynamically adjusting the numerical precision during runtime, respectively. Their goal is to find an optimal balance between accuracy, latency, and energy consumption. Implementing MPQ and TC in practice necessitates the use of Precision-Scalable (PS) and reconfigurable hardware. This aspect constitutes the primary topic of this thesis. Given that Deep Learning (DL) algorithms essentially involve scalar multiplications and dot products for executing convolutions and matrix multiplications, our attention is on PS multipliers. Specifically, we focus on two main categories of PS multiplier architectures, *Sum-Apart (SA)* and *Sum-Together (ST)*, and we integrate them into the Multiply-and-Accumulate (MAC) units of DNN accelerators and low-power extreme-edge RISC-V processors. In these multipliers, N multiplications are computed in parallel in a Single Instruction Multiple Data (SIMD) fashion, with operands on 16/N bits, where N = 1, 2, 4. While SA multipliers keep the results separate from each other, ST multipliers accumulate the results of low-precision multiplication internally, eliminating the need for an external adder. Consequently, they enable support for MPQ and TC and, at the same time, accelerate MAC operations by a factor of up to N compared to conventional full-precision 16-bit multipliers.

Our study provides a comprehensive comparison of the main ST multipliers in the literature. We begin with an overview of State-of-the-Art (SoA) ST multiplier architectures. Next, we introduce three new designs: one optimizing the critical path of a Baugh-Wooley (BW) ST multiplier, another derived from High-Level Synthesis (HLS), and the third based the Booth architecture. We evaluate their performance, power, and area (PPA) characteristics across a wide clock frequency range, after normalizing all the architectures to support 16, 8, and 4 bits of precision. The key finding reveals no single winner that satisfies all PPA scenarios, but rather a set of optimal ST multipliers depending on specific PPA constraints.

Our research also contributes to the advancement of ST-based DNN hardware accelerators by proposing implementations for 2D-Convolution (2D-Conv), Depth-wise Convolution (DW-Conv), and Fully-Connected (FC) layers. These

Application-Specific Integrated Circuit (ASIC) are PS and can be reconfigured at runtime to support operands at 16-, 8-, and 4-bit. We explain their working principles and architectures, and illustrate the design flow. Through extensive HLS-driven design space exploration (DSE), we analyze trade-offs between latency, area, and power, exploring various hardware parameters to identify Pareto-optimal accelerators. Furthermore, we demonstrate the benefits of our ST-based accelerators over those equipped with fixed-precision 16-bit multipliers, i.e., *standard* accelerators. The results of executing the four Machine Learning Performance (MLPerf) Tiny networks quantized in mixed-precision (MP), using SoCs integrating ST-based accelerators tailored to different PPA scenarios (i.e., low-area, low-power, and low-latency), show: an average inference latency speedup, across the four models, of 1.46x, 1.33x, and 1.29x, respectively; a reduced average energy reduction in most of the cases; and a marginal area overhead of 0.9%, 2.5%, and 8.0%, compared to SoCs equipped with standard accelerators. In conclusion, our study offers a complete analysis of ST-based accelerators within the context of SoCs, while highlighting future improvements to address identified inefficiencies.

Considering that SA and ST multipliers have typically been proposed separately in the literature, in this study we introduce a novel class of PS multipliers named *Sum-Together/Apart Reconfigurable (STAR)*, capable of working in both SA and ST modes within a single design. We develop four STAR multiplier architectures, including designs based on established *Divide-and-Conquer (D&C)* and *Sub-word Parallel (SWP)* families, as well as those based on three mutually exclusive datapaths (*3-way*) and separate SA and ST multipliers with multiplexed outputs. Comparative analysis in terms of PPA, conducted in a 28-nm technology, identifies STAR SWP as optimal for low-power and low-area requirements, STAR 3-way as the most suitable for high-performance scenarios, and STAR D&C as competitive for mid-range PPA requirements. These results offer valuable insights for designers aiming to implement efficient multipliers tailored to specific design targets.

Lastly, we integrate a STAR multiplier into the MAC unit of a low-power extreme-edge RISC-V processor for the first time, enabling support for MP-quantized DNNs. Specifically, we replace the default 16-bit multiplier inside the Multiplier/Divider unit of the *Ibex* processor with a 16-bit STAR SWP BW multiplier and introduce new MAC instructions, including standard 32-bit MAC and 16/8/4-bit MAC operations in ST/SA mode. The comparison between our modified Ibex processor and the original one unveils an acceleration up to 5.8× in FC, 3.7× in 2D-Conv, and 2.8× in DW-Conv quantized layers. Additionally, in a 28-nm technology with target clock frequencies of 200 and 600 MHz, the area and power consumption of the proposed solution are 0.015 and 0.017 mm^2, and 1.5 and 4.3 mW, respectively, with a limited overhead within 10% and 3% with respect to the original Ibex. In summary, with notable acceleration gains for typical quantized DNN layers and minimal overhead in terms of area and power consumption, our STAR MAC unit presents a viable option for efficient DL inference in resource-constrained devices.