Low-Power Neural Network Accelerators: Advancements in Custom Floating-Point Techniques.

Abstract

The expansion of Artificial Intelligence (AI) is addressing a new era characterized by omnipresent connected devices. To ensure the sustainability of this transformation, it is imperative to adopt design strategies that harmonize precise computational results with economically viable system architectures. Consequently, refining the efficiency and quality of AI hardware engines stands as a critical consideration in this evolution. This necessitates a balanced approach that prioritizes energy-efficient computations, precise and reliable results, and integration across various platforms and devices.

Machine Learning (ML) algorithms are serving as the foundational enabler for the integration of AI into Internet-of-Things (IoT) devices, particularly in the context of Industry 4.0. These advancements are shaping applications to be more intelligent and economically rewarding. This transformation improves numerous domains, from scientific research to industrial processes and everyday living. However, this technological evolution also brings its own set of challenges. ML algorithms pose significant computational and energy demands. Consequently, a central objective of this dissertation is to explore innovative methods for enhancing the hardware efficiency of computing engines.

Approximate computing techniques, such as quantization, exploit the inherent error resilience of ML algorithms to address key design concerns in computer systems: energy efficiency, performance, and chip area. Quantization, which involves reducing the number of bits used to represent numbers, can significantly reduce power consumption and data movement, thereby enhancing energy efficiency by employing compact arithmetic units that save chip area. These techniques often yield computation acceleration due to reduced data sizes, which promotes faster, more parallel, and pipelined processing, particularly in neural network computation. However, this approach introduces a trade-off between number precision and model accuracy, necessitating proper hardware design methodologies. While state-of-the-art methods are advancing, significant research opportunities remain, especially for accelerators with custom Floating-Point (FP) computation.

In this dissertation, a hardware design methodology is presented for low-power inference of Spike-by-Spike (SbS) neural networks for embedded applications, within the field of Spiking Neural Networks (SNNs). Compared to conventional SNNs employing the Leaky Integrate-and-Fire (LIF) mechanism, SbS neural networks are highlighted for their reduced model complexity and exceptional noise robustness. However, despite their advantages, SbS networks inherently possess a memory footprint and computational cost that makes them challenging for deployment in constrained embedded systems. To solve this issue, this research leverages the intrinsic error resilience of SbS models, aiming to enhance performance and reduce hardware complexity, while avoiding number quantization. Specifically, this research introduces a novel Multiply-Accumulate (MAC) module designed to optimize the balance between computational accuracy and resource efficiency of FP operations. This MAC module features configurable quality through a hybrid approach. It combines standard FP number representations with a custom 8-bit FP format, as well as a 4-bit logarithmic number representation. This design excludes the use of a sign bit, further contributing to the compact and efficient representation of numbers. This design enables the MAC module to be tailored to the specific resource constraints and performance requirements of a given application, making SbS neural networks possible for deployment in resource-constrained environments.

In the field of Convolutional Neural Networks (CNNs), this dissertation presents a hardware design methodology for low-power inference, specifically targeting sensor analytics applications. Central to this work is the proposal of the Hybrid-Float6 (HF6) quantization scheme and its dedicated hardware accelerator, designed to function as a Conv2D Tensor Processor (TP). This quantization strategy employs a hybrid number representation, combining standard FP and a 6-bit FP format. This strategy allows for a highly optimized FP MAC, reducing mantissa multiplication into a multiplexer-adder operation. This research introduces a Quantization-Aware Training (QAT) method that, in certain cases, offers beneficial regularization effects. The efficacy of this exploration is demonstrated with a regression model, which improves its precision despite the applied quantization. For ML portability, the custom FP representation is encapsulated within a standard format -- a design characteristic that enables the proposed hardware to process it automatically. To validate the interoperability of this approach, the hardware architecture is integrated with TensorFlow Lite, demonstrating compatibility with industry-standard ML frameworks and affirming the potential for practical deployment in various sensing applications while maintaining compliance with established ML infrastructure.

This dissertation addresses an essential challenge in the current technological landscape: the harmonization of computational accuracy with energy efficiency and compatibility of hardware solutions. This dissertation stands as a significant contribution towards the development of a sustainable next-generation of neural network processors, essential to empower the increasingly connected and intelligent world of tomorrow.