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

Abstract

This dissertation presents an investigation into design techniques involving custom Floating-Point (FP) computation for low-power neural network accelerators in resource-constrained embedded systems. It focuses on the sustainability of the future omnipresence of Artificial Intelligence (AI) through the development of efficient hardware engines, emphasizing the balance between energy-efficient computations, inference quality, application versatility, and cross-platform compatibility.

The study emphasizes the role of Machine Learning (ML) in advancing low-power Internet-of-Things (IoT), particularly in Industry 4.0, and acknowledges the computational and energy challenges posed by ML algorithms. The dissertation aims to enhance hardware efficiency, primarily through approximate computing techniques like quantization. This approach, while improving energy efficiency and accelerating computations, requires careful design to balance numerical precision and model accuracy.

The research presents a hardware design methodology for low-power inference of Spike-by-Spike (SbS) neural networks. Despite the reduced complexity and noise robustness of SbS networks, their deployment in constrained embedded devices is challenging due to high memory and computational costs. The dissertation proposes a novel Multiply-Accumulate (MAC) hardware module that optimizes the balance between computational accuracy and resource efficiency in FP operations. This module employs a hybrid approach, combining standard FP with custom 8-bit FP and 4-bit logarithmic numerical representations, enabling customization based on application-specific constraints and implementing acceleration for the first time in embedded systems.

Additionally, the study introduces a hardware design for low-power inference in Convolutional Neural Networks (CNNs), targeting sensor analytics applications. This proposes a Hybrid-Float6 (HF6) quantization scheme and a dedicated hardware accelerator. The proposed Quantization-Aware Training (QAT) method demonstrates improved quality despite the numerical quantization. The design ensures compatibility with standard ML frameworks as TensorFlow Lite, highlighting its potential for practical deployment in real world applications.

In summary, this dissertation addresses the critical challenge of harmonizing computational accuracy with energy efficiency in AI hardware engines with inference quality, application versatility, and cross-platform compatibility as a design philosophy. It contributes significantly to the development of sustainable neural network processors, crucial for the increasingly connected and intelligent world.