RADemics

Optimized Hardware Acceleration of Deep Learning Algorithms Using Xilinx Zynq SoCs

Saravanan D, S. Pragadeeswaran
KARPAGA VINAYAGA COLLEGE OF ENGINEERING AND TECHNOLOGY, VEL TECH RANGARAJAN DR. SAGUNTHALA R&D INSTITUTE OF SCIENCE AND TECHNOLOGY

Optimized Hardware Acceleration of Deep Learning Algorithms Using Xilinx Zynq SoCs

[1]Saravanan D, Professor of Mathematics, Department of Science and Humanities, Karpaga Vinayaga College of Engineering and Technology, Chengalpattu, Tamilnadu, India saravanan.danapal@gmail.com

[2]S. Pragadeeswaran, Assistant professor, School of Computing, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology. pragadeesresearch@gmail.com

Abstract

The growing demand for intelligent edge computing has intensified the need for deploying deep learning models on energy-efficient and performance-constrained platforms. Xilinx Zynq System-on-Chip (SoC) devices offer a unique architectural blend of programmable logic and embedded processors, enabling customized hardware acceleration for artificial intelligence applications. This book chapter explores advanced techniques in model quantization and structured pruning to optimize neural networks for inference on Zynq-based systems. It provides an in-depth analysis of co-design strategies, layer sensitivity-aware compression, and the implementation of automation pipelines for quantization-aware training. By leveraging toolchains such as Vitis AI and PYNQ, the discussion covers complete deployment workflows from model training to hardware execution, supported by practical insights into latency, power consumption, and resource utilization. Through detailed benchmarking and performance evaluation, this work highlights how high-accuracy inference can be maintained while significantly reducing memory and compute overhead. The methodologies presented in this chapter contribute to scalable and sustainable AI acceleration solutions tailored for edge environments. The comprehensive integration of model compression, tool support, and hardware mapping provides a roadmap for efficient deployment of deep learning on reconfigurable SoCs.

Keywords: Quantization, Structured Pruning, Xilinx Zynq SoC, Edge AI, Hardware Acceleration, Model Optimization

Introduction

The rapid expansion of edge computing has redefined the performance expectations of embedded systems, especially with the integration of artificial intelligence (AI) capabilities in real-time and resource-constrained environments [1]. Edge devices are increasingly expected to execute computationally intensive deep learning models locally, without relying on cloud infrastructure [2]. This shift introduces significant challenges related to energy efficiency, latency, hardware limitations, and thermal management [3]. To address these issues, reconfigurable hardware platforms, particularly Xilinx Zynq SoCs, offer a compelling solution due to their hybrid architecture [4]. These systems combine ARM-based processing systems (PS) with FPGA-based programmable logic (PL), enabling developers to offload and accelerate critical components of AI workloads in a power-efficient and customizable manner [5].

Xilinx Zynq SoCs are uniquely positioned to support deep learning inference at the edge due to their ability to tightly couple software control with parallelized hardware execution [6]. The PS typically handles task scheduling, data input, and control logic, while the PL is tailored for acceleration of compute-intensive layers such as convolutions, matrix multiplications, and activation functions [7]. The AXI interconnect facilitates high-speed communication between the two domains, ensuring that data flows are optimized without introducing significant overhead [8]. Unlike conventional embedded processors, the reconfigurable fabric of Zynq SoCs allows for the implementation of task-specific pipelines and hardware accelerators, which can be dynamically modified to suit the requirements of varying AI models [9]. This versatility makes Zynq SoCs particularly suitable for dynamic environments where workload characteristics and system constraints may evolve during operation [10].