Title: OPTIMIZING NEURAL NETWORKS PERFORMANCE ON PARALLEL ARCHITECTURES

Abstract:

The Ph.D. thesis presented here delves into optimizing the inferencing phase of Neural Networks (NNs) for edge devices with limited computational resources and stringent energy constraints. With the burgeoning growth of NN applications in various domains, the need for energy-efficient solutions becomes paramount for edge devices such as smartphones and wearables.

The study encompasses three key classes of NNs: Self-organizing maps (SOMs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). Each class poses distinct challenges, prompting the development of tailored optimization strategies. The research introduces a comprehensive overview of NN models, emphasizing the importance of optimizing the inferencing phase in developing and deploying NN applications.

Within the realm of CNNs, the focus is optimizing data reuse through innovative partitioning and scheduling schemes. An analytical framework is devised to estimate off-chip memory accesses, enabling the comparison of various approaches. This framework facilitates the identification of optimal solutions to enhance energy efficiency and throughput across different CNN layers.

Addressing the unique challenges of RNNs, a novel data reuse approach is proposed, specifically designed to handle dependencies in consecutive time-step computations. FPGA implementations of Long-Short Term Memory Network (LSTM) accelerators showcase the effectiveness of our approach in improving energy efficiency and throughput for popular LSTM models.

For SOMs, the study explores the impact of quantization techniques on accuracy and energy efficiency. A custom semi-systolic array design is introduced, analyzing the trade-offs between NN accuracy and energy consumption for various bit-width implementations. This research provides insights into the benefits and trade-offs associated with different bit resolutions, which are crucial for energy-constrained systems.

In summary, this Ph.D. thesis advances the field by contributing novel data reuse techniques and analytical frameworks, enhancing the energy-efficient acceleration of modern NNs tailored for edge devices. The research improves the inferencing phase and lays the foundation for future exploration in NN optimization and edge AI applications in energy-constrained environments.