Politecnico di Torino

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

Doctoral Program in Electrical, Electronics and Communication Engineering (37th Cycle)

Efficient End-to-End Deployment of DNNs
Architectural, Model, and Algorithmic Optimizations

By
Pierpaolo Mori

Supervisor
Prof. Claudio Passerone

Summary

The adoption of Deep Neural Networks (DNNs) has led to transformative advancements in a variety of domains such as autonomous driving, industrial automation, and edge medical diagnostics. Despite their high predictive accuracy and generalization capabilities, modern DNNs are characterized by considerable computational and memory demands. These demands often exceed the capabilities of edge computing platforms, which are constrained by limited energy budgets, compute resources, and memory bandwidth. Consequently, a central challenge in edge AI deployment is achieving efficient inference while preserving the accuracy and robustness of the model.

This thesis proposes a comprehensive, end-to-end approach to the efficient deployment of DNNs on edge devices, grounded in the co-optimization of architecture, model, and algorithmic components. It introduces a systematic exploration of deployment bottlenecks and presents targeted solutions to mitigate them through a combination of hardware-aware design and learning-based adaptation techniques. The optimization strategies are grouped into three main categories:architectural optimizations, model-level transformations, and algorithmic acceleration, each contributing synergistically to the overall efficiency of deployment.

At the architectural level, this work proposes techniques to better align neural network execution with hardware constraints by optimizing memory access, data reuse, and parallelism. A flexible accelerator is developed that dynamically adjusts tiling, unrolling, and memory scheduling strategies based on model and hardware characteristics. Structured pruning, guided by a hardware latency model, ensures that compression directly translates into practical performance gains. At the model level, it introduces training strategies that increase robustness to quantization, allowing consistent performance across a range of bit-widths without requiring retraining. A novel multi-quantization-aware training method allows models to adapt flexibly to both uniform and mixed-precision scenarios, supported by automated search techniques to identify optimal deployment configurations. At the algorithmic level, it overcomes the numerical instability of the fast Winograd algorithm in low-precision regimes by explicitly modeling its effects during training, avoiding prediction quality degradation while preserving the computational advantages that enable faster execution on hardware. These contributions are integrated into a unified deployment methodology that bridges model design and hardware execution, enabling accurate and low-latency inference on resource-constrained platforms.

Overall, this thesis advances the state of the art in edge AI by demonstrating that coordinated optimization between model design and the underlying deployment hardware is essential to realizing the full potential of DNNs in real-world, constrained environments.