Politecnico di Torino

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

Doctoral Dissertation
Doctoral Program in Computer Engineering (34^{th}cycle)

Vertical Optimizations of Convolutional Neural Networks for Embedded Systems

By

Antonio Cipolletta

******

Supervisor(s):
Prof. Enrico Macii
Prof. Andrea Calimera

Doctoral Examination Committee:
Prof. Ibrahim Elfadel, Referee, Khalifa University
Prof. Fatih Ugurdag, Referee, Ozyegin University
Prof. Andrea Acquaviva, Università di Bologna
Prof. Alberto Bosio, École Centrale de Lyon
Prof. Andrea Bottino, Politecnico di Torino

Politecnico di Torino
2022

Summary

Deep Convolutional Neural Networks (CNNs) have recently achieved remarkable progress in many data-intensive fields, like computer vision and natural language processing. The eagerness to adopt these powerful algorithms in various applications has recently required moving the CNN inference process from the cloud, powered by near-infinite resources, to "the edge," that is, on lightweight resource-constrained embedded systems. The need for high user privacy, low latency, and low cost are the main driving factors of such a paradigm shift.

Processing the CNN inference directly on-device offers several advantages. It guarantees higher levels of user privacy and higher quality of service, as data stay local and latency is much more deterministic. Moreover, it lowers cost and energy consumption, reducing the pressure on the network infrastructure. At the same time, it also creates a high technical challenge: filling the gap between the computational and memory requirements of modern CNNs and the limited hardware and energy resources of embedded systems.

Achieving the goal of bringing intelligence to the end systems, therefore, relies on the availability of *small, fast, and energy-efficient* CNNs. This dissertation stems from the idea that the key to small, fast, and energy-efficient CNNs is *vertical and automated* optimizations across the entire software stack. Optimizations must act vertically across the different layers to holistically combine the benefits of specialization at the algorithmic, the compiler, and the computational level, achieving remarkable gains. Optimizations must be automated to free embedded designers from the burden of manually dealing with the wide variety of CNN architectures and with the diversity of embedded platforms employed at the edge. To this end, the contribution of this dissertation is threefold. Novel automated optimization techniques are presented to improve the efficiency of state-of-the-art CNNs with minimal to no accuracy loss. New dynamic knobs are introduced to extend the achievable accuracy-complexity trade-off at run time. Finally, the proposed optimizations show that working across multiple levels of the optimization stack pushes further the boundary of accurate CNNs that can be deployed on tiny embedded devices.

This dissertation is organized into three main parts. In the first part, it focuses on how to build small CNNs. It first reviews the memory allocation problem in CNN compilers, quantitatively surveying the most adopted problem formulations. Second, it introduces dataflow restructuring, a novel functionality-preserving, automated method for minimizing the memory footprint of the intermediate activations. Finally, it presents a new compression pipeline, which combines weight pruning with dataflow restructuring to deploy more accurate CNNs on tiny MCU devices. The second part focuses on how to build fast CNNs suitable for tackling challenging inference tasks on low-power devices. This part presents a comprehensive design and optimization framework to accelerate monocular depth estimation on mobile-friendly ARM Cortex-A CPUs and on the tiny MCUs adopted at the edge of the Internet of Things. The last part of the dissertation deals with the deployment of energy-quality scalable CNNs on embedded systems. Specifically, this chapter builds upon the idea that the quality of the result can be gracefully degraded at run time to achieve higher energy efficiency, resorting to a more costly but accurate mode only when strictly needed. To this end, it first describes and characterizes an energy-quality scalable system for monocular depth estimation named EQPyD-Net. Finally, it introduces Nested Sparse ConvNets, a class of low-footprint, dynamic Convolutional Neural Networks purposely built to tackle inference tasks at the edge of the IoT.