Modern Deep Neural Networks (DNNs) - Convolutional Neural Networks (ConvNets) above all - bring computers closer to human intelligence in several tasks, from natural language understanding to computer vision. Motivated by those findings, a broad class of Internet-of-Things (IoT) services began to use ConvNets as a tool for sensing the raw data sampled from in-field sensors. In many novel IoT applications, the data cycle is cloud-centric: data are ceaselessly transmitted to servers hosting ConvNets trained to infer decisions that are then sent back to the end nodes. This approach is often impractical due to latency uncertainty and energy consumption for exchanging data from and to the cloud. Such a centralized infrastructure represents a single point of failure and exposes the users' private data.
On the contrary, deploying ConvNets directly on the edge node reduces the energy and infrastructure construction costs. Moreover, it guarantees higher performance and privacy standards since data are locally sampled and processed.

However, state-of-the-art ConvNets have billions of parameters and heavy arithmetic operations. This prevents their deployment on small, resource-constrained end devices. Against this backdrop, this dissertation aims to fill the gap between the computational and memory requirements of modern ConvNets and the severe computational resource and energy budget limitations of a typical embedded system. To tackle this problem, novel ConvNets must simultaneously satisfy different extra-functional constraints, such as latency, memory occupation, and energy consumption, besides accuracy. Therefore, the design and optimization process should be *hardware driven* and *cross-layer*, embracing different levels of the software and hardware stack. Such *cross-layer* optimization approach comprehensively integrates the advantages of specialization at both algorithmic and hardware levels, achieving remarkable results.
To this end, this dissertation: *(i)* presents novel *hardware-aware cross-layer optimizations* to improve the efficiency of modern ConvNet with a minimum accuracy degradation; *(ii)* proposes novel *dynamic knobs* that allow scale accuracy and energy at run time; *(iii)* demonstrates that a vertical approach across multiple levels of the software and hardware stacks opens to new optimization opportunities, pushing further the boundaries of accurate ConvNets that can be deployed on an embedded device.

This dissertation is organized into three main parts. The first part presents novel cross-layer optimization for deploying *tiny and fast* ConvNets on low-power embedded systems. Firstly, it reviews different quantization approaches and their limitations. Then, it describes a new architectural template combining *binarization* and *ensemble theory* to reduce the memory footprint of a ConvNet. Finally, it introduces a technique to combine fixed-point quantization and binarization to accelerate the inference and improve the accuracy of ConvNet on tiny MCU devices. The second part focuses on building *dynamic scalable ConvNets*, which adapt their working point according to input complexity or energy budget at run-time. Specifically, it first introduces *Nested Sparse ConvNets*, a novel class of compressed ConvNets, that can scale energy and accuracy at run-time, leveraging sparsity as a dynamic knob. Then, it describes an efficient implementation of test-time augmentation tailored for embedded platforms. Such a strategy dynamically adapts the inference latency according to the input complexity. Finally, this dissertation deals with the *hardware and software co-design* solutions to build energy-efficient ConvNets. It reviews different hw/sw co-design strategies and proposes a new approach that exploits arithmetic approximation and data reuse to build an energy-efficient inference engine.