


Doctoral Dissertation
Doctoral Program in Computer and Control Enginering (33^{th} cycle)

Hardware-Aware Compression Techniques for Embedded Deep Neural Networks

Matteo Grimaldi
* * * * * *

Supervisors
Prof. Enrico Macii, Supervisor
Prof. Andrea Calimera, Co-supervisor

Doctoral Examination Committee:
Prof. Jose Ayala, Referee, Universidad Complutense de Madrid
Prof. Anupam Chattopadhyay, Referee, Nanyang Technological University
Prof. Andrea Acquaviva, Università di Bologna
Prof. Paolo Garza, Politecnico di Torino
Prof. Eugenio Villar, Universidad de Cantabria

Politecnico di Torino
October 15, 2021

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

. . . . . .

Matteo Grimaldi

Turin, October 15, 2021

Summary

The success of the Internet-of-Things (IoT) is not about the amount of data collected but regards the ability to convert raw data to valuable information. In the last years, data management and interpretation have been brought to a higher level of difficulty, as the number of connected IoT devices has grown dramatically. All the raw data collected locally by the IoT sensors need to be processed and evaluated to extract highly informative data. This process, also known as *sensemaking*, consists of complex data-analysis tasks leveraging artificial intelligence algorithms.

These strategies can predict future states just by learning from past experience, which is represented by features processed on IoT-sensor data collections. Thanks to the rapid advancements in deep learning theory, deep Neural Networks, in particular, machines took a closer step towards human intelligence. These solutions are becoming ubiquitous and scalable to different levels, ranging from natural language processing, speech recognition, computer vision, autonomous driving. Usually, the actual solutions on this topic are deployed offline on centralized high-performance data centers, based on cloud platforms, where the distance between the raw sensor data and the computational hardware is critical. This solution suffers from low scalability.

A wide consensus among the research community is assessing that the overcoming of the IoT computational needs will pass through the development of near-sensor data-analytics accelerators, able to process the collected data at the edge, without introducing time latencies and further power consumption due to cloud computing solutions. Near-sensors data analytics is the key to sustain the IoT ecosystem indeed. Objects that can autonomously extract and process information from the physical world will allow the development of real-time classification solutions able to improve the quality of service in several applications, from remote health care and domotics to smart transportation, smart manufacturing, and smart power delivery. However, deep learning algorithms are extremely complex: the processing requires huge power consumption due to considerable storage, memory bandwidth, and high demand for computational resources. The challenge comes here: make deep learning algorithms fit low-power end-nodes of the IoT.

This dissertation aims to investigate hardware-aware compression techniques to facilitate the process of embedding deep learning solutions on resource-constrained

architectures. The goal is to reduce the energy required to run such large deep neural networks on resource-limited devices. More specifically, the main objective of our research is to find the perfect trade-off between the complexity of the deep learning model and the resources available on-chip, without performance degradation during prediction. This is accomplished through the development of novel software-level optimizations able to address these compelling technological demands.

Various compression techniques have been explored in the last decade to bridge this gap: pruning to remove redundant parameters, quantization to reduce their numerical precision, encoding algorithm with sparse-matrix computation to exploit the approximated parameters, are a subset of them. Several strategies tried to merge these and other compression methods to optimize such deep learning algorithms in terms of storage, memory, latency, and power; however, yet prioritizing one aspect to respect others. Moreover, these techniques were often designed without a proper consciousness of the hardware, limiting the compression effectiveness to a theoretical aspect.

With such a premise, this dissertation is organized into three main parts, each of them focusing on a different objective. The first part focuses on statistically oriented compression of neural networks, with particular concerning on new strategies to exploit the natural over-parametrization of these models. It first illustrates a constrained training to enable an effective approximation of the distribution of weights, with a proper encoding scheme it reaches high compression rates. Then, it presents a novel hybrid methodology capable to discriminate between model layers in terms of significance, with the aim to boost the final compression achievements. In the second part, the focus shifts on the hardware awareness of the compression strategies, a crucial feature to meet the real constraints of the deployment. The dissertation first analyzes the optimality of memory-bounded convolutional neural networks, through a smart heuristic able to explore the memory vs. accuracy solution space. Then, it presents a new technique able to empower the processing of n-ary precision networks on general-purpose microcontroller units. At last, it illustrates an adaptive sparse training designed to maximize the compression of storage-bounded networks. In the third part, the scalability of the deep learning models is addressed with innovative solutions to explore the latency vs. accuracy space. In particular, it presents a novel training and compression pipeline for building nested sparse networks: a set of sub-networks enclosed in a unique model able to run-time scale configuration points, during the inference stage.

The techniques proposed in this thesis provide some useful insights into the edge-driven compression of neural networks. For each of these three topics, results show that the aforementioned demand to balance the trade-off between model complexity and available resource can be effectively addressed. We hope that this work may contribute, with other research in this field, to open up more space and help to make artificial intelligence accessible to everyone, improving the quality of life of our next future.