Doctoral Program in Electrical, Electronics and Communication Engineering (36th Cycle)

Hardware/Neural Network Codesign for Energy-Efficient Inference on Edge Devices with Optimal Mapping and Compression

By

Emanuele Valpreda

Supervisors

Prof. Maurizio Martina, Supervisor

Prof. Guido Masera, Co-Supervisor

Summary

The diffusion of artificial intelligence (AI) applications in daily life has motivated the development of hardware and software techniques to optimize their execution on edge devices. Contrary to cloud computing, edge devices enable private and secure data processing, personalized algorithms, and lower latency. However, they have limited computational resources and strict power budgets, which makes the deployment of AI algorithms challenging. As modern neural network architectures, which are the backbone of AI, become more complex, it is necessary to develop or adapt the optimization strategies used to reduce computation energy and latency. A common approach is to remove unnecessary neurons and connections or to decrease the numeric precision of the data, leveraging the redundancy and intrinsic error resilience of neural networks. Moreover, the design of specialized hardware accelerators that leverage the computation patterns of neural networks can reduce the cost of data movement and energy consumption. While a significant research effort has been devoted to this topic, the joint optimization of hardware and neural networks is still an open problem. This doctoral thesis aims to investigate hardware/neural network codesign to optimize the performance of neural networks on edge devices. In particular, the problem of optimal hardware mapping with and without compression is addressed, focusing on reducing the cost of data movement during the inference. The hardware mapping performed across multiple layers is also discussed. It is achieved by shaping the communication and computation patterns so that several layers can reuse the same data. Moreover, the thesis presents a methodology for achieving robust and low-power neural network inference on approximated hardware, leveraging a reconfigurable multiplier architecture seamlessly embedded in a microcontroller. The topic of error resilience in safety-critical applications is also addressed. An algorithm to detect and correct errors in object detection is proposed, mitigating the accuracy degradation in case of logic transients. The techniques presented in this thesis are evaluated on popular datasets or hardware platforms, the latter supported by custom simulation tools, showing the effectiveness of the proposed methodologies.