PhD in Electrical, Electronics and Communications Engineering

Research Title: Embedded Machine Learning

SESSION: SPRING 2020

Funded by	Dipartimento di Elettronica e Telecomunicazioni (DET)

Supervisor	Prof. Mario Roberto Casu mario.casu@polito.it

Contact	http://www.det.polito.it/research/research_groups/electronics/vlsilab_group

Context of the research activity	The recent development in the field of Machine Learning (ML), and particularly in the area of ML that deals with Deep Neural Networks (DNN), has made it possible to solve many complex problems with innovative solutions. Due to the wide applicability of DNNs to diverse problems, an extensive and widespread adoption of these technologies is expected in the near future. Many applications, however, will have to function autonomously, without the support of a continuous connection to the electrical and communication networks. Consider for example the huge amount of raw data that environment or industrial sensors would need to send to a centralized processing system for neural network inference. It is clear that to avoid the explosion of communication-related costs we need to relocate part of the processing power to the **edge**, whereas operations like incremental training, testing of new neural networks, storage of high-quality and high-valued data are bound to stay in the central processing system. The challenge is that edge devices are typically small **embedded systems**, with limited processing power and precision, small memory size, and severe constraints on low power consumption. To deploy DNNs and other ML algorithms on such devices require optimizing and tailoring the application on the specific embedded system, as well as developing specialized hardware accelerators implemented in dedicated System-on-Chips (SoCs) with dedicated logic (either fixed or programmable), possibly in conjunction with low-power GPUs. In

	this way it will be possible to provide the best performance for the least amount of energy. The research activity carried out by the selected PhD will be in this context, and will focus especially in 1) the co-design of embedded hardware accelerators and ML algorithms given tight performance/power/cost constraints, and 2) the development of specific design methodologies using the highest abstraction level to explore a vast and complex Hw/Sw design space.

Objectives	The Embedded Machine Learning (EML) domain encloses applications for which the typical design goal is to reach the minimum amount of energy per operation for a given performance constraint (fixed latency or fixed throughput) and a given accuracy requirement. Examples of these applications include object detection/classification in videos acquired from remote cameras or aerial vehicles (fixed throughput) and speech recognition (fixed latency). Energy minimization in the related literature has often been limited to the digital logic of the hardware accelerators and to the optimization of the ML algorithm (e.g., weights pruning and quantization, etc.), without considering that the data access to the memory hierarchy and the I/O system is often the limiter in terms of energy optimization. For this reason, the PhD candidate will take a holistic approach that aims to minimize the energy of the entire system. This requires an accurate power/energy model (either analytic or based on look-up tables) of the system under consideration (e.g. an SoC implemented either in an ASIC or a FPGA technology, its memory system, the board in which the SoC is mounted including the power supply system and the power management, if available), which will be one of the first objectives of the research activity carried out by the PhD student. This model will be used to assess the feasibility and the effectiveness (in terms of optimization of the energy efficiency) of hardware accelerators that the PhD student will develop. Another objective will be the design of configurable hardware accelerators that can adapt and evolve. This because first ML algorithms change rapidly, and second because the limited capacity of embedded systems calls for configurable accelerators that can adapt to different applications (for example, the same computing kernel could be reconfigured and reused for object detection and speech recognition) rather than for multiple concurrent accelerators, which require more hardware resources. Since the design space of these ML hardware accelerators is large, determining the optimal solution or a Pareto front of solutions (in a multi-objective scenario) requires on the one hand a high level of abstraction in the design, and on the other hand it requires that optimization methods and heuristics are developed. This is also an important objective that the PhD candidate will be pursue in

	his/her research activity.

Skills and competencies for the development of the activity	The candidate is required to have skills and competences such as those typically obtained with a Master of Science degree in areas like Electronic Engineering or Computer Engineering. In addition, the candidate should possess the following competences: - Experience with hardware design of accelerators to be used in Systems-on-Chip (developed in RTL using VHDL/Verilog or with synthesizable code like C/C++ for High-Level Synthesis) implemented in ASIC or FPGA technology. - Knowledge of the main ML and Deep Learning Libraries (for example, Tensorflow, Caffe, Pytorch) and Python libraries for ML development (for example Keras, scikit-learn, hyperopt);