ELECTRICAL, ELECTRONICS AND COMMUNICATIONS ENGINEERING

Ateneo/DET - Reduction of complexity of Neural Networks by exploiting innovative neuron architecture for Tiny Machine Learning

Funded By	Politecnico di TORINO [P.iva/CF:00518460019] Dipartimento DET
Supervisor	PARESCHI FABIO - fabio.pareschi@polito.it
Contact	PRONO LUCIANO - luciano.prono@polito.it SETTI GIANLUCA - gianluca.setti@polito.it
Context of the research activity	Aim of the project is to investigate the possibility to reduce the complexity of Neural Network by exploiting known techniques (structural/unstructural pruning, quantization, etc.) and at the same time to develop innovative techniques, such as the definition of new hardware architectures for neurons. The investigation will be pursued both at hardware (i.e., low) level and system (i.e., high) level, and will be focused to devices with reduced computational capacity for TinyML or Edge-AI.
Objectives	The aim of this project is to develop circuits, systems and algorithms for the implementing Artificial Intelligence techniques in devices with reduced computational capacity for TinyML or Edge-AI. Of paramount importance for the success of the project will be a good knowledge of the grounding mechanisms of the most used algorithms for Artificial Intelligence applications, necessary for the optimization of the algorithms themselves in order to implement them on hardware devices with limited resources. An additional requirement is a deep knowledge of the statistical method to analyze the activity of the neurons in order to identify which neurons are of fundamental importance, and which one can be pruned, even assuming an unconventional architecture. Possible areas for applying such optimization processes, generally considered fundamental for implementation on low resource devices, may include (but will not necessarily be limited to) i) reduction of the number of parameters of an algorithm (fundamental given the memory footprint devices considered) and ii) quantization of the weights and/or parameters of an algorithm (necessary given the probable presence of only low-precision arithmetic units), preventing at the same time the loss of accuracy of the algorithm. The capability to estimate the reduction in energy consumption will also be fundamental.

Skills and competencies for the development of the activity	* Good knowledge of Python and Pythorch programming languages * Acquaintance of multi-GPU programming methodologies