POLITECNICO DI MILANO

PhD in INGEGNERIA DELL'INFORMAZIONE / INFORMATION TECHNOLOGY - 39th cycle

Research Area n. 2 - Electronics

PNRR 117 Research Field: DESIGN METHODOLOGIES FOR IN-MEMORY COMPUTING TO EXPLOIT SPARSITY AND QUANTIZATION FOR DEEP LEARNING

Monthly net income of PhDscholarship (max 36 months)
€ 1400.0 In case of a change of the welfare rates during the three-year period, the amount could be modified.

Context of the research activity	
Motivation and objectives of the research in this field	Artificial intelligence has become an essential computing task for edge computing, where complex neural networks must be operated locally, efficiently, securely and in real time. To this purpose, one of the key bottleneck of existing accelerators of deep learning for AI is the massive use of external memory, that critically affects the energy consumption and performance. In this context, in-memory computing allows computation within the data, thus minimizing data movement and enhancing parallelism of data processing. In-memory computing has been demonstrated with virtually all memory technologies, including volatile memories, such as static random access memory (SRAM) and nonvolatile memories, such as embedded phase change memory (PCM). Developing efficient in-memory computing circuits require a detailed methodology covering several aspects, such as device technology, analog/digital circuit design, neural information processing and system architecture. The codesign of device, network, circuit and overall system is essential to develop accurate and efficient deep learning accelerators. In particular, limited-precision multilevel operation and weight sparsity are among the key parameters characterizing in-memory computing and affecting the performance and accuracy of the deep learning task. A ‘structured sparsity’ can be introduced at

	the training stage to minimize the number of weights and decrease the energy consumption. Quantization is dictated by the available number of conductance levels in the memory device technology and/or by the number of bit cells available for each weight. Since quantization and sparsity can provide both constraints and boosters of deep learning accelerators, they need to be carefully considered and optimized. In this context, the PhD thesis will aim at the design of novel deep learning accelerators based on in-memory computing in the presence of significant quantization and sparsity. The thesis will first address the problem from a high level of abstraction to identify the main figures of merits and the theoretical correlation linking sparsity, quantization, accuracy, memory/area occupation and energy consumption. Various mapping strategies will be explored to identify the best tradeoff between area consumption and parallelism. In the second phase, integrated circuits blocks will be developed to allow a realistic simulation of the various figures of merit of the in-memory computing accelerator. The system level architecture will then be developed by including converters and digital processors. Various device technologies (SRAM, PCM) and computing approaches (digital, analog, mixed) will be considered in this phase to provide a comprehensive benchmark of the main figures of merit. Finally, real-life applications in the automotive, industrial and biomedical sectors will be explored to identify a path to commercial exploitation.
Methods and techniques that will be developed and used to carry out the research	- Integrated circuit design to develop the analog/digital in-memory computing circuits - Device engineering to control the number of conductance levels, their precision and readout time - Neural processing to tailor quantization and sparsity in the deep network - Circuit simulation to assess the impact of sparsity and quantization on performance, efficiency and accuracy
Educational objectives	Acquire and/or consolidate knowledge and/or practical

	skills around: - Circuit simulation and design for application specific integrated circuits - Neural modeling for the deep learning - Device, circuit, algorithm codesign for optimized neural networks
Job opportunities	The research proposal addresses an output profile that responds to the needs of the edge-computing industry for technical experts in the design and the development of next generation computing systems to accelerate Deep Learning applications.
Composition of the research group	1 Full Professors 0 Associated Professors 3 Assistant Professors 8 PhD Students
Name of the research directors	Daniele Ielmini

Contacts
E-mail: daniele.ielmini@polimi.it Phone: 02 2399 6120 https://ielmini.faculty.polimi.it

Additional support - Financial aid per PhD student per year (gross amount)	
Housing - Foreign Students	--
Housing - Out-of-town residents (more than 80Km out of Milano)	--

Scholarship Increase for a period abroad	
Amount monthly	700.0 €
By number of months	6

National Operational Program for Research and Innovation	
Company where the candidate will attend the stage (name and brief description)	STMICROELECTRONICS S.R.L., Agrate Brianza
By number of months at the company	6
Institution or company where the candidate will spend the period abroad (name and brief description)	STMICROELECTRONICS S.R.L., Crolles, Francia
By number of months abroad	6

Additional information: educational activity, teaching assistantship, computer availability, desk availability, any other information
EDUCATIONAL ACTIVITIES (purchase of study books and material, including computers, funding for participation in courses, summer schools, workshops and conferences): financial aid per PhD student. TEACHING ASSISTANTSHIP: availability of funding in recognition of supporting teaching activities by the PhD student. There are various forms of financial aid for activities of support to the teaching practice. The PhD student is encouraged to take part in these activities, within the limits allowed by the regulations. COMPUTER AVAILABILITY: individual use. DESK AVAILABILITY: individual use