UNIVERSITY OF SALERNO

DEPARTMENT OF INDUSTRIAL ENGINEERING

Ph.D. Course in Industrial Engineering
Curriculum in Electronic Engineering - XXXVII Cycle

DESIGNING TINY ULTRA-LOW-POWER NEURAL NETWORK HARDWARE ACCELERATORS FOR IN-SENSOR COMPUTING

Supervisors
Prof. Alfredo Rubino
Prof. Gian Domenico Licciardo

Ph.D. Student
Paola Vitolo

Scientific Referees
Prof. Maurizio Valle
Prof. Rajkumar Kubendran

Ph.D. Course Coordinator
Prof. Massimo De Santo

Abstract (Italiano)

La rapida diffusione dei dispositivi IoT pone la necessità di integrare capacità di intelligenza direttamente a livello di sensore, garantendo al contempo elevata accuratezza e consumi energetici estremamente ridotti. Questa tesi di dottorato analizza il paradigma dell'In-Sensor Computing (ISC) attraverso una metodologia di co-progettazione hardware–software guidata dai vincoli hardware, finalizzata alla realizzazione di acceleratori neurali miniaturizzati e a bassissima potenza, strettamente integrati con sensori MEMS. L'approccio proposto definisce fin dalle prime fasi di progetto la topologia delle reti neurali, le strategie di quantizzazione e l'aritmetica a punto fisso, al fine di rispettare stringenti vincoli di area, potenza e memoria. Le soluzioni sviluppate sono validate mediante prototipazione su FPGA e sintesi CMOS. La validità della metodologia è dimostrata attraverso tre casi di studio applicativi, sviluppati in collaborazione con STMicroelectronics. Il primo riguarda l'elaborazione audio per applicazioni di keyword spotting, in cui una rete neurale convoluzionale 1D sostituisce la tradizionale catena CIC+FIR per la conversione PDM–PCM, integrando filtraggio e decimazione in un unico blocco neurale. La soluzione produce segnali PCM a 8 bit e 16 kHz con un rapporto segnale-rumore di 48 dB, mantenendo un'accuratezza complessiva dell'89%. Il core sintetizzato in tecnologia CMOS a 130 nm raggiunge una potenza di 128,7 µW/MHz in un'area inferiore a 1 mm^2. Il secondo caso di studio affronta la manutenzione predittiva basata su segnali vibrazionali mediante una pipeline ibrida ed event-driven, che combina un autoencoder in-sensor sempre attivo, parzialmente binarizzato, per il rilevamento di anomalie, con un classificatore su microcontrollore attivato su richiesta. L'approccio consente di ottenere prestazioni di rilevamento di anomalie prossime allo stato dell'arte (AUC pari a 0,99) e un'accuratezza di classificazione fino al 94,83%, sostenendo rate di dati in uscita dal sensore fino a 365 kHz. I risultati di sintesi riportano un'area di 0,49 mm^2 in tecnologia CMOS a 65 nm e una potenza dinamica di 138,6 µW/MHz. Il terzo caso di studio è dedicato alla compensazione dello stress termico nei sensori di pressione MEMS. Viene proposta un'unità di compensazione riconfigurabile basata su intelligenza artificiale (AI-ReSCU), che combina un meccanismo di trigger adattivo con uno stimatore neurale iterativo dell'errore, caratterizzato da pesi binarizzati e attivazioni a punto fisso. La soluzione consente di ripristinare l'accuratezza del sensore entro ±0,5 hPa, recuperando fino a 1,6 hPa, con una potenza dinamica dell'ordine dei nanowatt e un'area di 0,55 mm^2. Nel loro insieme, i risultati dimostrano la generalità e l'efficacia del flusso di progettazione proposto, mostrando come, nonostante la diversità dei domini applicativi e dei requisiti prestazionali, sia possibile ottenere soluzioni ad alta efficienza energetica e accuratezza competitiva. La tesi distilla in-

oltre principi di progettazione ISC generalizzabili, quali la propagazione precoce dei vincoli, la condivisione aggressiva delle risorse con calcolo serializzato, la quantizzazione a pochi bit e la binarizzazione selettiva, nonché l'adozione di meccanismi event-triggered con modalità di deep sleep. Infine, durante un periodo di ricerca di sei mesi presso la Johns Hopkins University, è stato condotto uno studio esplorativo sull'impiego di Large Language Models (LLM) per la generazione automatica di descrizioni hardware, includendo codice Verilog sintetizzabile, testbench e documentazione, applicati alla progettazione di una rete neurale spiking ricorrente validata su FPGA e implementata tramite un flusso open-source in tecnologia SkyWater a 130 nm. Tale studio fornisce una prospettiva complementare su come il workflow di co-progettazione proposto possa essere ulteriormente accelerato.

Abstract (English)

The explosive growth of IoT devices demands on-sensor intelligence that is accurate and radically energy-efficient. This dissertation investigates In-Sensor Computing (ISC) through a constraints-first hardware–software co-design methodology to realize tiny, ultra-low-power neural accelerators tightly coupled to MEMS sensors. The approach shapes network topology, quantization, and fixed-point arithmetic from the outset to meet stringent limits in area, power, and memory, and validates designs through FPGA prototyping and CMOS synthesis. Three application-driven case studies, developed in collaboration with STMicroelectronics, substantiate the methodology. The first addresses audio processing for keyword spotting, where a learned 1D-CNN replaces the conventional CIC+FIR PDM-to-PCM chain, fusing filtering and decimation and delivering 8-bit/16 kHz PCM with 48 dB SNR while preserving downstream accuracy of 89%. The synthesized core in 130 nm CMOS achieves 128.7 µW/MHz within less than 1 mm^2. The second focuses on vibration-based predictive maintenance, employing a hybrid, event-driven pipeline that combines an always-on, partially binarized in-sensor autoencoder for anomaly detection (AUC = 0.99; 99.61% accuracy) with an on-demand MCU classifier (up to 94.83%). The in-sensor accelerator sustains sensor output data rates up to 365 kHz and exhibits 333 µW/MHz dynamic power on FPGA, while standard-cell synthesis in 65 nm reports 0.49 mm^2 and 138.6 µW/MHz dynamic power. The third case concerns thermal-stress compensation for MEMS pressure sensors: the proposed AI-based Reconfigurable Sensor Compensation Unit (AI-ReSCU) couples a reconfigurable trigger with an iterative neural error estimator with binarized weights and fixed-point activations to restore accuracy within ±0.5 hPa, recovering up to 1.6 hPa, with 4.46 nW dynamic power in 0.55 mm^2. Taken together, these diverse studies confirm the general applicability of the proposed design flow: despite their different sensing domains and performance targets, each achieves state-of-the-art accuracy and efficiency. The dissertation distills generalizable ISC design principles—early constraint propagation, aggressive resource sharing with serialized compute, selective binarization and low-bit quantization, and event-triggered operation with deep sleep—showing that competitive machine-learning accuracy and real-time throughput can be achieved at milliwatt-to-nanowatt power and sub-mm^2 area, enabling practical in-sensor AI. Finally, during a 6-month research period at Johns Hopkins University, an exploratory study investigated Large Language Models (LLMs)-assisted hardware-description generation, including synthesizable Verilog, testbenches, and documentation for a recurrent spiking neural network validated on FPGA and implemented with an open-source SkyWater 130 nm flow, as a complementary perspective on how the proposed co-design workflow could be accelerated.