Progettazione di una architettura per inferenza di CNN basata su acceleratori in-memory computing

L'attività mira alla progettazione di un'architettura avanzata per l'inferenza di reti neurali convoluzionali (CNN), sfruttando acceleratori basati su computing in-memory (IMC). L'obiettivo principale è ottimizzare sia le parti della rete neurale profonda (DNN) non direttamente mappabili sull'acceleratore IMC, sia l'interazione tra l'acceleratore stesso e i componenti digitali ausiliari attraverso estensioni dell'ISA, architetturali e microarchitetturali che permettano di controllare l'acceleratore IMC in maniera flessibile. Questa integrazione punta a garantire un'elaborazione efficiente e coordinata, migliorando le prestazioni complessive in termini di latenza, efficienza energetica e scalabilità per applicazioni di intelligenza artificiale avanzate. L'attività userà il background sia di ST Microelectronics che di UNIBO, ed Il codice sviluppato non sarà rilasciato come open source. L'avanzamento delle attività sarà monitorato tramite meeting bimensili online e meeting trimestrali in presenza, a cui saranno invitati i responsabili aziendali e universitari. I risultati della ricerca saranno brevettati oppure pubblicati in conferenze e riviste di settore di alto livello, come DATE, DAC, TVLSI e TCAD.

Design of an Architecture for CNN Inference Based on In-Memory Computing Accelerators

The activity aims at designing an advanced architecture for convolutional neural network (CNN) inference, leveraging accelerators based on in-memory computing (IMC). The main objective is to optimize both the parts of the deep neural network (DNN) that are not directly mappable onto the IMC accelerator and the interaction between the accelerator itself and auxiliary digital components through ISA, architectural, and microarchitectural extensions that allow flexible control of the IMC accelerator. This integration aims to ensure efficient and coordinated processing, improving overall performance in terms of latency, energy efficiency, and scalability for advanced artificial intelligence applications. The activity will use the expertise from both ST Microelectronics and UNIBO, and the developed code will not be released as open source. Progress will be monitored through bi-monthly online meetings and quarterly in-person meetings, where company and university managers will be invited. The research results will be either patented or published in high-level conferences and journals, such as DATE, DAC, TVLSI, and TCAD.