Abstract

L'attività di ricerca descritta in questa tesi ha lo scopo di dimostrare la fattibilità di integrare feature di Intelligenza Artificiale (AI) a bordo di dispositivi wearable e portatili tramite l'implementazione e l'esecuzione di modelli di reti neurali (Neural Networks – NNs) in prossimità dell'elemento sensibile. Tra i vari modelli di AI, il Deep Learning (DL) e le Deep Neural Networks raggiungono le performance più elevate in diversi task, come la classificazione di immagini, il riconoscimento automatico di attività umane, e così via. Tuttavia, i modelli di DL richiedono solitamente un'enorme quantità di risorse in termini di memoria e la loro esecuzione necessita di architetture digitali ad elevate performance. Queste specifiche sono nettamente in contrasto con le caratteristiche tipiche dei dispostivi wearable e portable, che devono essere quanto più compatti possibile e garantire una durata della batteria soddisfacente. Per questo motivo, spesso viene utilizzata la strategia del cloud computing. Tuttavia, questa introduce delle latenze più elevate, che in talune applicazioni possono risultare inaccettabili; si pensi ad esempio alla guida autonoma dei veicoli o alla microchirurgia assistita. Inoltre, il trasferimento dei dati consuma banda ed energia. In questo contesto risulta quindi altamente desiderabile spostare la computazione in prossimità dell'elemento sensibile. Questo paradigma è chiamato edge computing. Tuttavia, l'implementazione di modelli di DL sui cosiddetti dispositivi edge è tuttora una sfida. Le tipiche piattaforme general purpose, come le CPU o le GPU, non rappresentano una soluzione ottimale in termini di efficienza energetica, specialmente nel caso di dispositivi wearable e alimentati a batteria, dove la durata è un aspetto critico. Di conseguenza, si sta facendo molta ricerca su come progettare acceleratori HW dedicati per il DL e spostare la circuiteria richiesta per implementare la computazione in prossimità dell'elemento sensibile. Quello che si ottiene in questo caso è uno smart sensor. In questa tesi è proposto un modello innovativo chiamato Hybrid Binary Neural Network (HBN), che sfrutta i vantaggi delle Binarized Neural Network (BNN). Come caso studio è stata scelta la Human Activity Recognition (HAR) basata su sensori inerziali. Inoltre, è stato sviluppato un algoritmo di pre-processing che permette di risolvere il problema della device-orientation negli accelerometri triassiali. Combinando le operazioni di pre-processing alla classificazione tramite HBN si può ottenere un miglioramento dell'accuracy in alcune condizioni. I risultati mostrano un'accuracy che arriva fino al 99% nel riconoscimento di 5 diverse attività umane. Dopo aver sviluppato il modello, è stato progettato un acceleratore HW dedicato ultra-low power ed implementato sia su FPGA che su standard cell CMOS. Considerando la frequenza operativa molto bassa associata alle applicazioni HAR, il consumo di potenza è stato ridotto riducendo il numero di risorse. Il design permette di implementare sia le operazioni di pre-processing che il modello HBN. I risultati di sintesi in tecnologia CMOS 65 nm Low-Power (LP) High Voltage Threshold (HVY) mostrano che l'acceleratore HW ha un consumo di potenza di 6.3 μW e un'area di 0.20 mm^2. Il design proposto ha un consumo di potenza che è di almeno 7.3 volte inferiore rispetto allo stato dell'arte. Inoltre, è stata sviluppata una board dimostrativa basata su FPGA per dimostrare il funzionamento in real-time del sistema.