Diss. ETH No. 28937

Near-Sensor Analytics and Machine Learning for Long-Term Wearable Biomedical Systems

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
XIAYING WANG
MSc ETH BME, ETH Zurich
born on 03.10.1994
citizen of Zhejiang, China

accepted on the recommendation of
Prof. Dr. Luca Benini, examiner
Dr. Benjamín Béjar Haro, co-examiner
Prof. Dr. Maurizio Valle, co-examiner

2022

Abstract

Wearable devices for biomedical applications have become increasingly pervasive. In a field where privacy is a major concern and latency is not well-tolerated, low-power and small-sized edge devices are of central importance. An emerging trend is to embed the processing algorithms near the sensors on the edge device to preserve privacy, reduce latency, and increase battery life. A new generation of wearable Internet of things and smart sensing systems should not only provide continuous data monitoring and acquisition but are also expected to process and make sense of the acquired data in similar ways as human experts do.

Over the years, machine learning has achieved impressive results in many applications, including the biomedical field. However, the limited resources available on battery-operated devices pose enormous challenges in executing machine learning algorithms that are generally resource-demanding.

In this thesis, we first evaluate and assess the ability of two representative and leading-edge ultra-low-power microcontrollers to execute machine learning models for wearable applications. We then focus on the challenging task of brain–machine interface based on the motor imagery paradigm, which allows direct communication between the human brain and external machines by merely thinking of a body part movement. We identify the state-of-the-art classification algorithms in this domain and introduce methods to reduce their computational

complexity and model size allowing efficient implementation on edge devices. We further propose optimized and energy-efficient deployment techniques by exploiting hardware extensions and parallel computing. Finally, we design a new model architecture that requires significantly less memory footprint and fewer computations while at the same time keeping state-of-the-art accuracy. We additionally limit the resource requirements by proposing an effective method to reduce the dimensionality of the input data, significantly lowering the overall system's power consumption without significantly degrading the model accuracy.

With this thesis, we demonstrate for the first time that it is feasible to execute real-time inference at the edge for a brain–machine interface based on motor imagery and reach the state-of-the-art trade-off among accuracy, resource demands, and power consumption necessary for a next-generation smart wearable device.

Sintesi

I dispositivi indossabili per applicazioni biomediche sono diventati sempre più pervasivi. In un campo in cui la privacy è una delle preoccupazioni maggiori e la latenza non è ben tollerata, i dispositivi edge a bassa potenza e di piccole dimensioni sono di fondamentale importanza. Una tendenza emergente è quella di incorporare gli algoritmi di elaborazione vicino ai sensori sul dispositivo edge per preservare la privacy, ridurre la latenza e aumentare la durata della batteria. Una nuova generazione di Internet delle cose indossabili e sistemi di rilevamento intelligenti non solo dovrebbe fornire un monitoraggio e un'acquisizione continui di dati, ma anche elaborare e dare un senso ai dati acquisiti in modo simile a quanto fanno gli esperti umani.

Nel corso degli anni, l'apprendimento automatico ha ottenuto risultati impressionanti in molte applicazioni, incluso il campo biomedico. Tuttavia, le limitate risorse disponibili sui dispositivi alimentati a batteria pongono enormi sfide nell'esecuzione degli algoritmi di apprendimento automatico che sono generalmente esigenti in termini di risorse.

In questa tesi, valutiamo innanzitutto la capacità di due microcontrollori rappresentativi e d'avanguardia di bassissima potenza di eseguire modelli di apprendimento automatico per applicazioni indossabili. Ci concentriamo di seguito sulla difficile applicazione dell'interfaccia neurale basata sul paradigma dell'immaginazione motoria, che consente la comunicazione diretta tra il cervello umano e le

macchine esterne semplicemente pensando al movimento di una parte del corpo. Identifichiamo gli algoritmi di classificazione d'avanguardia in questo dominio e introduciamo metodi per ridurre la loro complessità computazionale e le dimensioni del modello, consentendo un'implementazione efficiente sui dispositivi edge. Proponiamo inoltre tecniche di implementazione ottimizzate ed efficienti dal punto di vista energetico sfruttando le estensioni hardware e il calcolo parallelo. Infine, progettiamo una nuova architettura del modello che richiede un utilizzo di memoria significativamente inferiore e un minor numero di calcoli, mantenendo allo stesso tempo un'accuratezza all'avanguardia. Limitiamo inoltre le esigenze di risorse proponendo un metodo efficace per ridurre la dimensionalità dei dati di ingresso, riducendo in modo significativo il consumo energetico complessivo del sistema senza degradare significativamente l'accuratezza del modello.

Con questa tesi, dimostriamo per la prima volta che è possibile eseguire in tempo reale l'inferenza al margine per l'interfaccia neurale basata sull'immaginazione motoria e raggiungere il miglior compromesso tra accuratezza, risorse richieste e consumo di energia necessario per un dispositivo indossabile intelligente di prossima generazione.