ENGLISH

Borsa di Ricerca

Title: Design and Evaluation of Custom RISC-V Matrix Extensions for Accelerating LoRa-based Floating-Point AI Algorithms

Reference Project: Axelera AI

Duration: 12 Months

Contract Start Date: 1 April 2025

Amount: €19,367

Financial Coverage: Axelera AI

Candidate Requirements: Master's degree in Electronic Engineering, microprocessor architectures, digital hardware design

Selection Committee: Angelo Garofalo, Davide Rossi, Francesco Conti; substitute Giuseppe Tagliavini

Description of activities

Design and evaluation of custom RISC-V Matrix Extensions for Accelerating LoRa-based Floating-Point AI Algorithms.

The rapid evolution of AI algorithms, and the pervasive influence of AI-enhanced applications across many edge application-domains like computer vision, robotics, automotive, space, IoT, call for a paradigm shift from simple micro-controllers towards powerful and AI-accelerated heterogeneous edge computers.

Recent research has focused on developing highly specialized AI accelerators to infer various Deep Neural Network models in a variety of application use-cases. Such accelerators leverage
cutting-edge technologies such as RISC-V processors, In-Memory Compute, and quantization techniques, providing a cost-effective and energy-efficient to be integrated in complex end-to-end computing pipelines typical of edge scenarios.

However, while many neural networks have demonstrated resilience to quantization during inference, training phases typically require higher precision data formats, often relying on floating-point representations. Despite growing research into optimizing training in lower precision, production-ready training solutions still demand robust support for floating-point operations, and few cost-effective and energy-efficient hardware solutions have been proposed in literature.

The aim of this activity is to fill this research gap by developing and evaluating a RISC-V matrix extension for edge on-device training and fine-tuning tasks, including LoRA-style algorithms with a focus on solutions which consider floating-point quantization techniques.

The activities will be articulated as follows:

1) Study of state-of-the-art real-world workloads for training and fine-tuning, and selection of a representative subset;
2) Identification of mini-benchmarks based on workloads studied at point 1;
3) Bare-metal C implementation of these mini-benchmarks on existing scalar and vector RISC-V processors developed by UniBo;
4) Development of an FPGA platform for a heterogeneous compute cluster, including scalar, vector and benchmarking it with kernels developed at point 3 to assess a performance baseline;
5) Adapt an existing tensor processor to the requirements of the RISC-V AME Instruction Set Architecture (ISA) extension and extend the support to these instructions in LLVM;
6) Optimization and evaluation of mini-benchmarks on the RISC-V AME accelerator running on FPGA;
7) End-to-end training/finetuning application demonstration

ITALIAN

Borsa di Ricerca

Titolo: Design and Evaluation of Custom RISC-V Matrix Extensions for Accelerating LoRa-based Floating-Point AI Algorithms

Progetto di riferimento: Axelera AI

Durata: 12 Mesi

Data di decorrenza contratto: 1 Aprile 2025

Importo: €19.367

Copertura finanziaria: Axelera AI

Requisiti richiesti ai candidati: Laurea magistrale in Ingegneria Elettronica, architetture dei microprocessori, progettazione hardware digitale

Commissione Giudicatrice: Angelo Garofalo, Davide Rossi, Francesco Conti; supplente Giuseppe Tagliavini

Descrizione delle attività

Design and Evaluation of Custom RISC-V Matrix Extensions for Accelerating LoRa-based Floating-Point AI Algorithms

La rapida evoluzione degli algoritmi di intelligenza artificiale (IA) e l'influenza pervasiva delle applicazioni potenziate dall'IA in numerosi domini applicativi edge, come visione artificiale, robotica, automotive, spazio e IoT, richiedono un cambiamento di paradigma dai semplici microcontrollori verso potenti e eterogenei computer edge accelerati dall'IA.

Le ricerche recenti si sono concentrate sullo sviluppo di acceleratori IA altamente specializzati per inferire vari modelli di reti neurali profonde in una varietà di casi d'uso applicativi. Tali acceleratori sfruttano tecnologie all'avanguardia come processori RISC-V, calcolo in memoria e tecniche di quantizzazione, fornendo soluzioni economiche ed efficienti dal punto di vista energetico da integrare in complessi pipeline di calcolo end-to-end tipici degli scenari edge.

Tuttavia, mentre molte reti neurali hanno dimostrato resilienza alla quantizzazione durante l'inferenza, le fasi di addestramento tipicamente richiedono formati di dati a maggiore precisione, spesso basati su rappresentazioni in virgola mobile. Nonostante la crescente ricerca nell'ottimizzazione dell'addestramento a bassa precisione, le soluzioni di addestramento pronte per la produzione richiedono ancora un robusto supporto per le operazioni in virgola mobile, e poche soluzioni hardware economiche ed efficienti dal punto di vista energetico sono state proposte in letteratura.

L'obiettivo di questa attività è colmare questa lacuna di ricerca sviluppando e valutando un'estensione matriciale RISC-V per attività di addestramento e fine-tuning on-device edge, inclusi algoritmi in stile LoRA con un focus su soluzioni che considerano tecniche di quantizzazione in virgola mobile.

Le attività saranno articolate come segue:

1. Studio dei carichi di lavoro reali allo stato dell'arte per l'addestramento e il fine-tuning, e selezione di un sottoinsieme rappresentativo;
2. Identificazione di mini-benchmark basati sui carichi di lavoro studiati al punto 1;
3. Implementazione in C bare-metal di questi mini-benchmark su processori RISC-V scalari e vettoriali esistenti sviluppati da UniBo;
4. Sviluppo di una piattaforma FPGA per un cluster di calcolo eterogeneo, includendo unità scalari e vettoriali, e benchmarking con i kernel sviluppati al punto 3 per valutare una baseline delle prestazioni;
5. Adattamento di un processore tensoriale esistente ai requisiti dell'estensione dell'Instruction Set Architecture (ISA) AME RISC-V ed estensione del supporto a queste istruzioni in LLVM;
6. Ottimizzazione e valutazione dei mini-benchmark sull'acceleratore AME RISC-V in esecuzione su FPGA;
7. Dimostrazione applicativa di un'applicazione end-to-end di addestramento/fine-tuning utilizzando l'acceleratore AME RISC-V sviluppato.