Progetto di ricerca

L'attività di assistenza alla ricerca si inserisce nel contesto del progetto HAL4SDV e si concentra sull'integrazione delle tecniche di Intelligenza Artificiale (IA) negli stack software per il settore automotive, con un focus particolare sulle architetture RISC-V. L'obiettivo principale è quello di sviluppare e ottimizzare algoritmi avanzati di IA in grado di migliorare l'affidabilità, la sicurezza e l'efficienza degli stack software. Attraverso l'applicazione di tecnologie quali deep learning, analisi predittiva, elaborazione dei segnali e Large Language Model (LLM), il progetto mira a realizzare software intelligenti capaci di interagire in tempo reale con sensori avanzati e sistemi di controllo del veicolo.

Un focus specifico è dedicato al dispiegamento di funzionalità basate su LLM su piattaforme embedded ed edge, affrontando i vincoli legati alle risorse computazionali, all'occupazione di memoria, alla latenza e al consumo energetico. Questo include lo studio di tecniche di compressione dei modelli, quantizzazione, inferenza efficiente e strategie di co-progettazione hardware–software adattate a sistemi basati su RISC-V. Gli LLM embedded vengono esplorati come abilitatori di funzionalità avanzate a bordo veicolo, quali supporto decisionale contestuale, diagnostica intelligente, interfacce uomo–macchina adattive e miglioramento della manutenibilità del software.

L'integrazione di tecnologie di IA e LLM embedded mira ad aumentare la flessibilità, la scalabilità e l'autonomia dell'intera infrastruttura software, favorendo l'introduzione di sistemi automotive di nuova generazione che siano al contempo intelligenti e affidabili.

Piano delle attività

Il piano delle attività include le seguenti fasi:

- Analisi dei requisiti e dello stato dell'arte relativi all'architettura RISC-V di riferimento (Carfield), con particolare attenzione ai vincoli tipici dei sistemi embedded ed edge, agli acceleratori hardware e alle soluzioni esistenti per l'inferenza di modelli di IA e LLM on-device.
- Sviluppo di algoritmi di Intelligenza Artificiale, inclusi modelli di deep learning e Large Language Model a ridotto consumo di risorse, progettati per l'operatività in tempo reale, contesti safety-critical e piattaforme automotive embedded.
- Ottimizzazione di modelli di IA e LLM per il deployment embedded, mediante tecniche quali compressione dei modelli, quantizzazione, pruning e co-progettazione hardware–software su architetture RISC-V.
- Integrazione di tecniche di IA e LLM nell'infrastruttura software, abilitando un'interazione intelligente con sensori, sistemi di controllo del veicolo e componenti software di livello superiore.
- Testing e validazione, con particolare attenzione alla correttezza funzionale, alle prestazioni real-time, alla robustezza, all'efficienza energetica e alla conformità ai requisiti di sicurezza automotive.
- Disseminazione dei risultati attraverso pubblicazioni scientifiche, contributi open-source e partecipazione a workshop e conferenze sui temi dell'IA embedded, RISC-V e software automotive.

Research project

The research activity is part of the HAL4SDV project and focuses on the integration of Artificial Intelligence (AI) techniques into software stacks for the automotive sector, with a particular emphasis on RISC-V architectures. The main goal is to develop and optimize advanced AI algorithms capable of improving the reliability, safety, and efficiency of the software stacks. Through the application of technologies such as deep learning, predictive analytics, signal processing, and Large Language Models (LLMs), the project aims to create intelligent software that can interact in real time with advanced sensors and vehicle control systems.

A specific focus is placed on the deployment of LLM-based capabilities on embedded and edge platforms, addressing constraints related to computational resources, memory footprint, latency, and energy consumption. This includes the investigation of model compression, quantization, efficient inference techniques, and hardware–software co-design strategies tailored to RISC-V-based systems. Embedded LLMs are explored as enablers for advanced in-vehicle functionalities such as context-aware decision support, intelligent diagnostics, adaptive human–machine interfaces, and enhanced software maintainability.

The integration of AI and embedded LLM technologies aims to increase the flexibility, scalability, and autonomy of the entire software infrastructure, supporting next-generation automotive systems that are both intelligent and dependable.

Activity plan

The activity plan includes the following phases:

- Analysis of requirements and state-of-the-art regarding the reference RISC-V architecture (Carfield), with particular attention to embedded and edge constraints, hardware accelerators, and current solutions for on-device AI and LLM inference.
- Development of Artificial Intelligence algorithms, including deep learning models and resource-efficient Large Language Models, tailored for real-time operation, safety-critical contexts, and embedded automotive platforms.
- Optimization of AI and LLM models for embedded deployment, through techniques such as model compression, quantization, pruning, and hardware–software co-design on RISC-V architectures.
- Integration of AI and LLM techniques into the software infrastructure, enabling intelligent interaction with sensors, vehicle control systems, and higher-level software components.
- Testing and validation, focusing on functional correctness, real-time performance, robustness, energy efficiency, and compliance with automotive safety requirements.
- Dissemination of results through scientific publications, open-source contributions, and participation in workshops and conferences related to embedded AI, RISC-V, and automotive software.