Integrated Methodologies for Hardware dimensioning and AI algorithms configuration

The main purpose of the research project consists of studying and developing methods to automatize the process of matching the configuration of Artificial Intelligence (AI) algorithms with and available hardware resources under user-defined constraints and to solve different tasks from various domain. This topic is in line with the goal of the StairwAI EU Horizon2020 project, I.e., to develop an optimization engine for matching target AI applications with the HW platforms (from edge to cloud and on a wide range of architectures).

The problem of determining what is the right hardware (HW on premise or on the cloud) architecture and its dimensioning for AI algorithms is still crucial. Searching for the optimal solution is often challenging, as it is not trivial to anticipate the behaviour of an algorithm on diverse architectures. This is especially true if the AI application must respect quality-of-service constraints or budgets. In this scenario, having an automated decision support tool to match algorithms, user constraints and HW resources would be a great advantage for companies and practitioners working with AI applications. Often, a key requirement is recommending users the proper technique for the application (e.g., suggesting the right algorithm) and also finding the proper hardware (HW) resources to run it, together with the optimal configuration.

Activities:
The project will consist of two main phases. At first, the definition and analysis of different use cases and benchmark to perform multiple Machine Learning (ML) models to learn the relationship between AI algorithms (under various configurations) and their performance (e.g., runtime, memory footprint, solution quality, etc.). In the second phase of the project, the ML algorithms will be integrated with an optimization model to automatically recommend a portfolio of solutions (algorithms and their configurations) given user constraints and available HW (and budget). The overall optimization model will be refined with new ad-hoc constraints and optimization techniques based on the different dataset explored during the project.

Metodologie integrate per il dimensionamento hardware e la configurazione degli algoritmi di IA

Lo scopo principale del progetto di ricerca consiste nello studio e nello sviluppo di metodi per automatizzare il processo di matching tra configurazione degli algoritmi di Intelligenza Artificiale (AI) e le risorse hardware disponibili rispettando vincoli e funzioni obiettivo definiti dall'utente in diversi domini applicativi. Le attività sono in linea con l'obiettivo del progetto StairwAI EU Horizon2020, ovvero sviluppare un framework di ottimizzazione per abbinare le applicazioni AI con le piattaforme HW (dall'edge al cloud e su un'ampia gamma di architetture).

Il problema di determinare quale sia la giusta architettura hardware (HW on premise o sul cloud) e il suo dimensionamento per gli algoritmi di intelligenza artificiale è ancora cruciale. La ricerca della soluzione ottimale è spesso impegnativa, in quanto non è banale anticipare il comportamento di un algoritmo su architetture diverse. Ciò è particolarmente vero se l'applicazione di intelligenza artificiale deve rispettare i vincoli o i budget relativi alla qualità del servizio. In questo scenario, disporre di uno strumento automatizzato di supporto alle decisioni per abbinare algoritmi, vincoli utente e risorse HW sarebbe un grande vantaggio per le aziende e i professionisti che lavorano con le applicazioni di intelligenza artificiale. Spesso, un requisito chiave è consigliare agli utenti la tecnica corretta per l'applicazione (ad esempio, suggerendo l'algoritmo giusto) e anche trovare le risorse hardware (HW) appropriate per eseguirla, insieme alla configurazione ottimale.

Attività:

Il progetto si articolerà in due fasi principali. Inizialmente, la definizione e l'analisi di diversi casi d'uso e benchmark per eseguire più modelli di Machine Learning (ML) per apprendere la relazione tra algoritmi di intelligenza artificiale (in varie configurazioni) e le loro prestazioni (ad esempio, runtime, footprint di memoria, qualità della soluzione, ecc.). Nella seconda fase del progetto, gli ML saranno integrati in un modello di ottimizzazione per consigliare automaticamente un portafoglio di soluzioni (algoritmi e loro configurazioni) dati i vincoli dell'utente e l'HW disponibile (e il budget). Il modello di ottimizzazione generale sarà perfezionato con nuovi vincoli ad hoc e tecniche di ottimizzazione basate sul diverso set di dati esplorato durante il progetto.