Politecnico di Torino

ARTIFICIAL INTELLIGENCE

UNIBO - Optimizing the Execution of Deep Neural Networks on Edge SoCs using AI Compilers for Heterogenous Systems

Funded By	ALMA MATER STUDIORUM UNIVERSITA' DI BOLOGNA [P.iva/CF:01131710376]
Supervisor	JAHIER PAGLIARI DANIELE - daniele.jahier@polito.it
Contact	BENINI LUCA - luca.benini@polito.it JAHIER PAGLIARI DANIELE - daniele.jahier@polito.it
Context of the research activity	TVM (Tensor Virtual Machine) is an open-source deep learning compiler and runtime that aims to accelerate the deployment of machine learning models. MATCH is an AI compiler based on TVM designed for heterogeneous system AI deployment by our team. The candidate will optimize the execution of deep neural networks (DNNs) on low-power microcontroller units (MCUs) by exploring the use of both TVM and MATCH. The scope of the also research includes improving the compilers themselves to support more heterogeneity, platforms, and operators.
	The candidate will explore using open-source compilers to improve DNN execution on low-power heterogeneous systems. In particular, the candidate will tailor TVM and MATCH (an academic open-source AI compiler tool) to produce optimized code for RISC-V-based heterogeneous platforms, such as the Greenwave's platforms, GAP8 or GAP9, the Diana platform developed in KU Leuven or more complex systems such as the Occamy multi-core multi-chiplet platform. Objectives: 1. To review the existing literature on the execution of deep neu-ral networks (DNNs) on low-power system-on-chips (SoCs) and identify potential optimization techniques that can be ported inside TVM and MATCH as an extension. 2. Implement and evaluate the selected optimization techniques using a series of DNNs and low-power platforms. The scope of the research is to optimize networks for specific heteroge-neous platforms. In particular, a set of different optimizations will be developed, which would be either “general-purpose”, i.e., applied to every hardware target, “specific”, i.e., applied to a single target, or “tunable”, i.e., applied to every target but with target-specific parameters. 3. To improve the efficiency of DNN execution on low-power devices by using either the TVM, MATCH or a new AI com-piler (that would possibly be

Objectives	developed as an additional con-tribution of the Ph.D. candidate), enabling their broader use in various applications. The benchmark will be the TinyML Perf Suite, which lists four tasks with four different architectures, used to benchmark low-power hardware platforms. Addition-ally, the candidate will benchmark the novel attention-based Foundation models (FMs), which are becoming more and more popular on edge heterogeneous devices. The target hardware platforms will be different heterogeneous hardware platforms: Diana, a Soc that contains a main RISC-V control unit and two DNN hardware accelerators, one 16x16 digital systolic accelerator, and one 1152x512 analog-in-memory computing accelerator; GAP8, characterized by a RISC-V con-trol unit and a small accelerator made of 8 general-purpose RISC-V cores with a dedicated Level 1 scratchpad memory; GAP9, a platform based on GAP8, which extends the acceler-ator capability from 8 to 9 cores and also includes a pro-grammable DNN-specific digital accelerator; Occamy, a multi-core multi-chiplet platform that can scale up to hundreds of RISC-V cores, with a complex memory multi-level intercon-nect, allowing for efficient data transfers. Outline of Work: 1. Review of existing literature on the execution of DNNs on low-power MCUs and identification of potential optimization techniques that can be implemented using TVM and MLIR. This will involve a thorough review of relevant papers, arti-cles, and other sources to identify the most promising optimi-zation techniques that can be implemented using these two open-source deep learning compilers and runtimes. Relevant similar works will also be analyzed to take inspiration on how to implement new optimization for the target hardware plat-forms. 2. Familiarization with the hardware targets. The target platforms will be explored, installed, and configured to be available to be easily programmed. The candidate will familiarize himself with the different Software Development Kits (SDKs) and the different compilation toolchains. 3. Implementation and evaluation of selected optimization tech-niques using chosen DNN/FM benchmarks on low-power platforms (Diana, GAP8, GAP9, Occamy). This will involve implementing the selected optimization techniques using TVM and MATCH, and evaluating their performance. As previously said, different optimization techniques will be de-veloped and tested. Description: Descrivere l’ambito e gli obiettivi dell’attività di ricerca. Se la borsa si riferisce ad una collaborazione industriale indicarlo esplicitamente. Compilare in lingua inglese (max 5000 characters) The candidate will explore the use of open-source compilers to improve DNN execution on low-power heterogeneous systems. In particular, the candidate will tailor TVM and MATCH (an academic open-source AI compiler tool) to produce optimized code for RISC-V-based heterogeneous platforms, such as the Greenwave's platforms, GAP8 or GAP9, the Diana platform developed in KU Leuven or more complex systems such as the Occamy multi-core multi-chiplet platform. Objectives: 1. To review the existing literature on the execution of deep neu-ral networks (DNNs) on low-power system-on-chips (SoCs) and identify potential optimization techniques that can be ported inside TVM and MATCH as an extension. 2. Implement and evaluate the selected optimization techniques using a series of DNNs and low-power platforms. The scope of the research is to optimize networks for specific heteroge-neous platforms. In particular, a set of different optimizations will be developed, which would be either “general-

	purpose", i.e., applied to every hardware target, "specific", i.e., applied to a single target, or "tunable", i.e., applied to every target but with target-specific parameters. 3. To improve the efficiency of DNN execution on low-power devices by using either the TVM, MATCH or a new AI com-piler (that would possibly be developed as an additional con-tribution of the Ph.D. candidate), enabling their broader use in various applications. The benchmark will be the TinyML Perf Suite, which lists four tasks with four different architectures, used to benchmark low-power hardware platforms. Addition-ally, the candidate will benchmark the novel attention-based Foundation models (FMs), which are becoming more and more popular on edge heterogeneous devices. The target hardware platforms will be different heterogeneous hardware platforms: Diana, a Soc that contains a main RISC-V control unit and two DNN hardware accelerators, one 16x16 digital systolic accelerator, and one 1152x512 analog-in-memory computing accelerator; GAP8, characterized by a RISC-V con-trol unit and a small accelerator made of 8 general-purpose RISC-V cores with a dedicated Level 1 scratchpad memory; GAP9, a platform based on GAP8, which extends the acceler-ator capability from 8 to 9 cores and also includes a pro-grammable DNN-specific digital accelerator; Occamy, a multi-core multi-chiplet platform that can scale up to hundreds of RISC-V cores, with a complex memory multi-level intercon-nect, allowing for efficient data transfers. Outline of Work: 1. Review of existing literature on the execution of DNNs on low-power MCUs and identification of potential optimization techniques that can be implemented using TVM and MLIR. This will involve a thorough review of relevant papers, arti-cles, and other sources to identify the most promising optimi-zation techniques that can be implemented using these two open-source deep learning compilers and runtimes. Relevant similar works will also be analyzed to take inspiration on how to implement new optimization for the target hardware plat-forms. 2. Familiarization with the hardware targets. The target platforms will be explored, installed, and configured to be available to be easily programmed. The candidate will familiarize himself with the different Software Development Kits (SDKs) and the different compilation toolchains. 3. Implementation and evaluation of selected optimization tech-niques using chosen DNN/FM benchmarks on low-power platforms (Diana, GAP8, GAP9, Occamy). This will involve implementing the selected optimization techniques using TVM and MATCH, and evaluating their performance. As previously said, different optimization techniques will be de-veloped and tested. 4. Improvement of the efficiency of DNN execution on low-power devices through the use of AI compilers. This will in-volve measuring the performance improvements obtained through these two compilers and runtimes and analyzing the results to determine the most effective optimization tech-niques for improving the efficiency of DNN execution on low-power devices.

Skills and competencies for the development of the activity	1. Experience with neural network architectures and their com-putational workload. 2. Familiarity with programming languages such as C and Py-thon: The project will involve implementing and testing vari-ous algorithms and techniques 3. Familiarity with edge devices: Understanding the constraints and limitations of edge devices would be important in order to optimize the deep learning models for these platforms.