Master Thesis

Memory-Oriented Approaches for Deployment of DNNs on Low-Cost Edge Heterogeneous Systems

The thesis will contribute to the research on memory-centric approaches for the deployment of DNN models on resource-constraint heterogeneous systems and provide insights into the existing challenges in this field.

In recent years, the rapid growth of Artificial Intelligence (AI) and the explosion of hardware devices with AI-specific features have led to a rising demand for tools and frameworks capable of translating Deep Learning models from high-level languages like Python into lower-level code optimized for a particular hardware target, often in C.

This thesis focuses on edge heterogeneous systems, which have limited computational capabilities, low memory, and prioritize energy efficiency. The proliferation of diverse hardware platforms and programming ecosystems makes porting AI models to every device a non-trivial task. An ideal solution would be a universal tool that can translate high-level model representations, e.g., in Python, into low-level code while accommodating various hardware constraints, programming languages, and interfaces. Unfortunately, achieving this goal without compromising performance is still a challenge. For example, the TVM compiler stack is a popular open-source toolchain for deploying networks on many devices, including CPUs, GPUs, or ARM and RISC-V-based Microcontrollers (MCUs) but falls short when generating code for heterogeneous Systems-on-Chip (SoCs) containing different accelerators.

Recent efforts have focused on integrating TVM with memory-oriented deployment frameworks like DORY [1] and ZigZag [2], aiming to address these challenges.

[1] Van Delm, et al. "HTVM: Efficient neural network deployment on heterogeneous TinyML platforms." In 2023 60th ACM/IEEE Design Automation Conference (DAC), pp. 1-6. IEEE, 2023.

[2] Hamdi, Mohamed Amine. "Integrating Design Space Exploration in Modern Compilation Toolchains for Deep Learning." PhD diss., Politecnico di Torino, 2023.

Your work:

1. Conduct a comprehensive literature review of existing works.
2. Compare the references to identify gaps and unresolved challenges.
3. Investigate the integration flow of the references 1 and 2 in TVM.
4. Work on integrating the approaches outlined in references 1 and 2 using the UMA framework within TVM.

Requirements:

- Fundamental understanding of neural networks and embedded systems
- Basic understanding of TVM compiler
- Experience in programming C\C++ and Python
- Self-motivation and ability to work independently

If you are interested in this topic, please contact me at: samira.ahmadifarsani@tum.de.