Low-Power and High-Throughput LUT-based Accelerator Architecture for Distributed CNN Inference at the Edge

Authors: Quentin Dariol, Domenik Helms
German Aerospace Center (DLR), Institute for Systems Engineering and future mobility (SE), Oldenburg, Germany.
Contact: quentin.dariol@dlr.de

I. Motivation

Challenges:

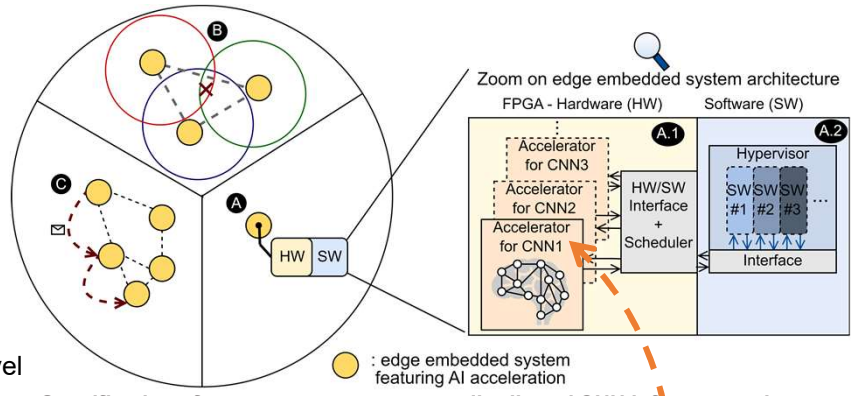
- Implementing Convolutional Neural Networks (CNNs) at the edge due to computational and memory demands versus resource limitations.
- Distributed AI inference at edge key to allow novel applications in multiple domains [1] e.g. future mobility and aerospace.

Key concepts of the proposed approach:

- LUT-Based HW accelerators: Store CNN layers as Look-Up Tables (LUTs) in FPGA, removing the need for external memory and enabling fast, energy efficient inference.
- Specification of an ecosystem to enable distributed CNN inference at edge involving multiple systems.

II. Ecosystem for distributed inference

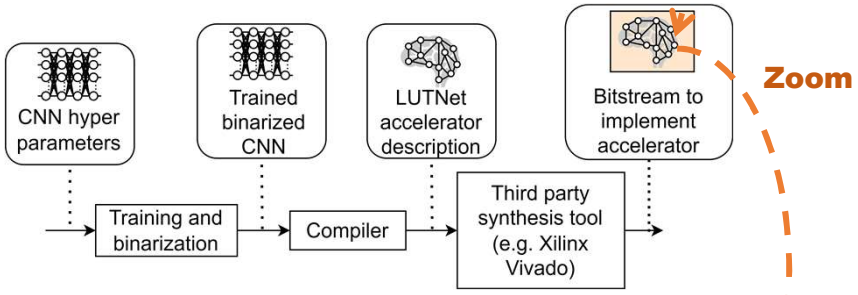
- **A.1. HW:** Ensures high throughput and energy efficiency with FPGA-based AI accelerators near data sensors.
- **A.2. SW:** Hypervisor and scheduler - Manage access of high-level SW applications to HW accelerators, leveraging power management techniques and Dynamic Partial Reconfiguration (DPR).
- **B. AI applications** - Manage access of high-level SW applications to HW accelerators, leveraging power management techniques and Dynamic Partial Reconfiguration (DPR).
- **C. Networking:** The system must also be able to locate and network with nearby systems, and update its network map.



Specification of an ecosystem to support distributed CNN inference at the edge, using HW/SW architectures.

III. LUT-based accelerator architecture using ElasticAI [2] / LUTNet [3] framework

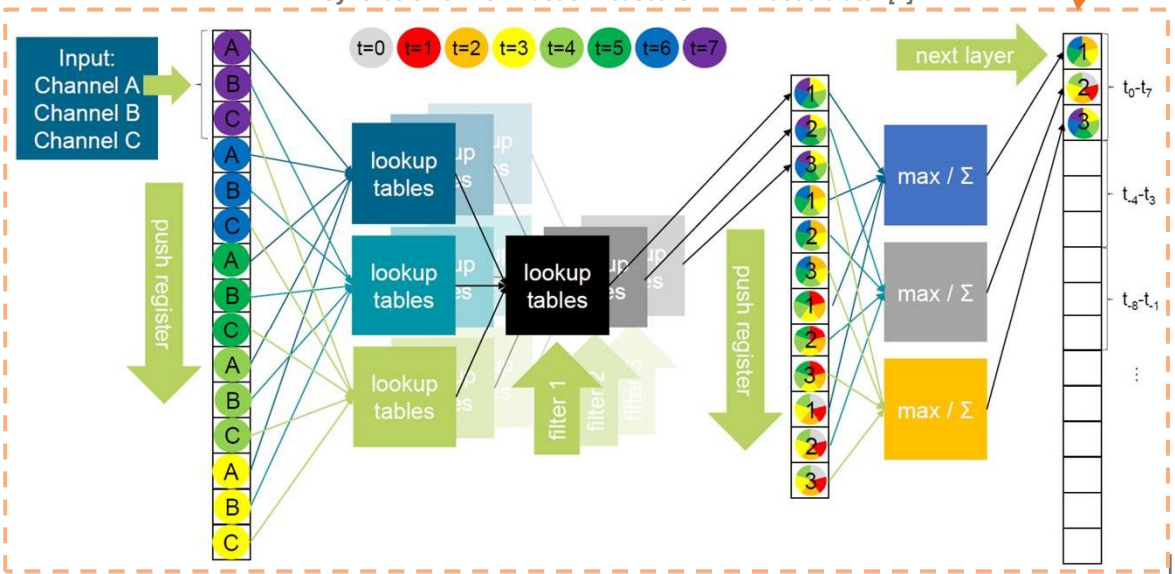
- Originality: pre-computed weights are synthetized as LUTs, allowing to remove external memory access.
- Evaluation results on Conv-1D electrocardiogram recognition on Spartan S15 FPGA. **Resource Usage:** 40% LUTs, no block RAM or DSP. **Performance:** Throughput of 10 million samples/second, energy cost of 26.2 nJ/sample.



Synthesis flow for ElasticAI-based CNN HW accelerator [2]

IV. Prospects

- Enable support for Conv-2D, critical for complex applications.
- Address ressource utilization issues, and low prediction accuracy.



Implementation example of a Conv-1D layer with streaming input, leveraging LUTs for efficient processing [3]

[1] S. Laskaridis et al. The Future of Consumer Edge-AI Computing. 2023. arXiv: 2210.10514 [cs.LG].
[2] C. Qian et al. "ElasticAI: Creating and Deploying Energy Efficient Deep Learning Accelerator for Pervasive Computing". In: IEEE International Conference on Pervasive Computing and Communications. Open source release. 2023. URL: https://github.com/es-ude/elastic-ai.creator
[3] D. Helms et al. "FPGA-Based Low Latency, Low Power Stream Processing AI". In: European Workshop on On-Board Data Processing. 2021.

>> This poster was presented at the GDR SoC Colloquium 2024 in Toulouse, France. 10-12.06.2024