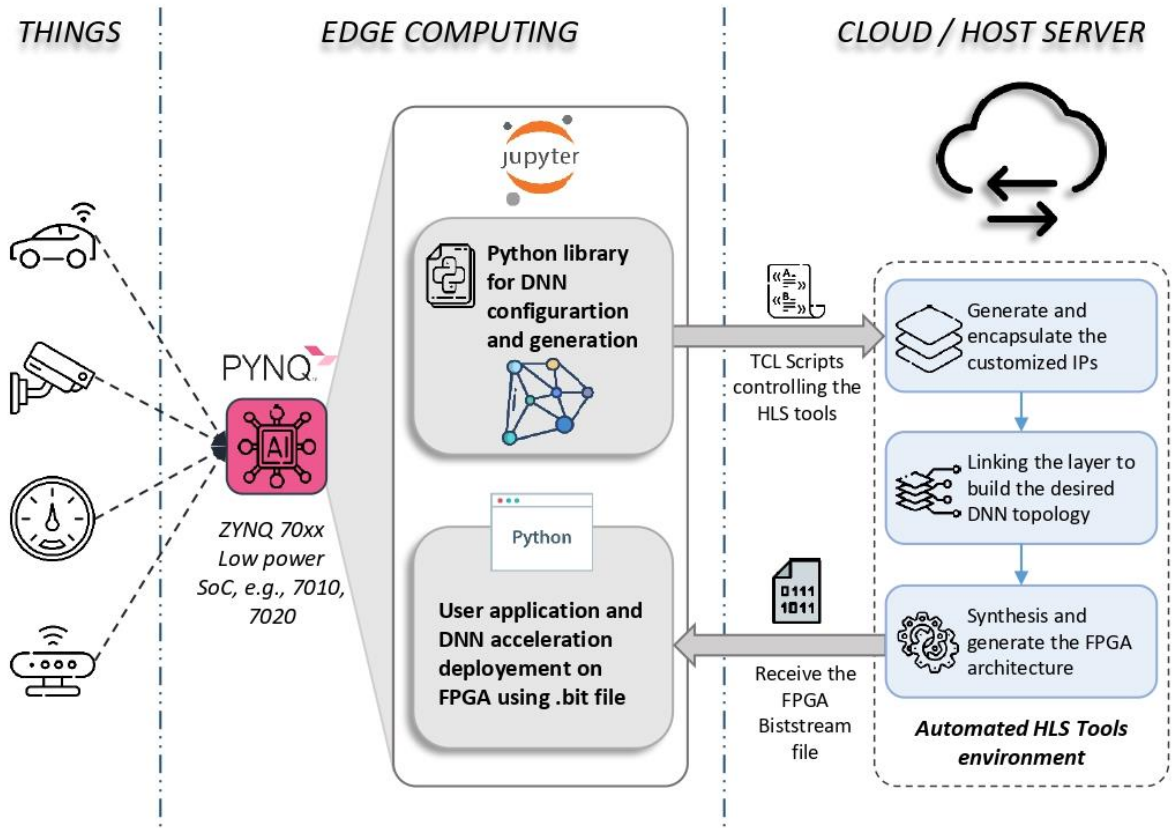
Title: Hardware Acceleration Architectures for Artificial Neural Networks

Abstract:

Deep Learning (DL) techniques have successfully solved many Artificial Intelligence (AI) applications problems. However, owing to topologies with many hidden layers, Deep Neural Networks (DNNs) have high computational complexity, which makes their deployment difficult in contexts highly constrained by requirements such as performance, real-time processing, or energy efficiency. Numerous hardware/software optimization techniques using GPUs, ASICs, and reconfigurable computing (e.g., FPGAs), have been proposed in the literature. With FPGAs, very specialized architectures have been developed to provide an optimal balance between high speed and low power. However, user requirements and hardware constraints must be efficiently met when targeting edge computing and IoT applications. Furthermore, deployment of such FPGAs architecture requires strong skills in both hardware and software domains. In this context, we propose an automated framework for the implementation of hardware-accelerated DNN architectures. Based on a high-level Python interface that mimics the leading DL software frameworks dedicated to GPUs, this framework provides an end-to end solution that facilitates the efficient deployment of topologies on System-on-Chip (SoC) based FPGAs accelerator by combining custom hardware scalability with optimization strategies. To do this, our design methodology covers the three main phases: **(a) customization:** where the user specifies the optimizations needed on each DNN layer through Python interface, **(b) generation:** the framework generates on the Cloud the necessary binaries for both FPGA and software parts, and **(c) deployment:** the SoC on the Edge receives the resulting files serving to program the FPGA and related Python libraries for user applications. Cutting-edge comparisons and experimental results demonstrate that the architectures developed by our framework offer the best compromise between performance, energy consumption, and system costs. For instance, the low power (0.266W) DNN topologies generated for the MNIST database achieved a high throughput of 3,626 FPS.



Nom et prénom: BELABED Tarek

Bibliographie:

TAREK BELABED received a B.Sc. degree in applied science and technologies, and an M.Sc. degree in science and technology at ISSATSo of the University of Sousse, Tunisia in 2014 and 2016 respectively. He received a Phd degree in a sandwich program between University of Mons, Belgium and Univesity of Sousse, Tunisia in 2023. His profile and research interests are oriented toward Edge Computing, re-

configurable hardware, Embedded Systems, EDA, unified hardware Tools, RTL design, and Deep Learning solution approaches.

Jury member

Professeur Pierre Manneback, Université de Mons, Président du jury:
Pierre.Manneback@umons.ac.be

Professeur assistant Sidi Ahmed Mahmoudi, Université de Mons, Secrétaire:
Sidi.MAHMOUDI@umons.ac.be

Professeur Carlos Valderrama, Université de Mons, Directeur de thèse:
carlos.valderrama@umons.ac.be

Externe :

Professeur Ali Douik, Université de Sousse, Examinateur:
ali.douik@eniso.u-sousse.tn

Professeur Mohsen machhout, Université de Monastir, Rapporteur:
mohsen.machhout@fsm.rnu.tn

Professeur Eva Dokladova, Université de Paris, Rapporteur:
eva.dokladalova@esiee.fr

Professeur Chokri Souani, Université de Sousse, Directeur de thèse:
chokri.souani@gmail.com

Logos:

POLYTECH MONS

UMONS
University of Mons


UNIVERSITY OF SOUSSE
جامعة سوسة


NISo
المدرسة الوطنية للمهندسين بسوسة
École Nationale d'Ingénieurs de Sousse