ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

XXXVI Cycle

Hardware and Software Co-design for Accelerating Convolutional Neural Networks on Edge Devices

Emanuele Valpreda

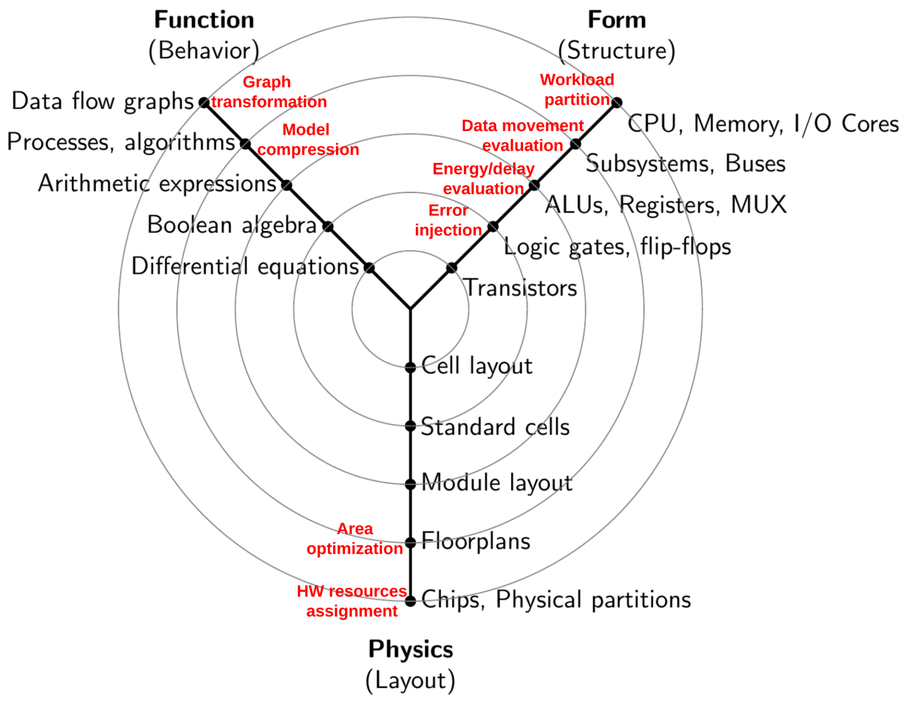
Supervisors: Prof. Maurizio Martina, Prof. Guido Masera

Research context and motivation

- Convolutional Neural Networks (CNNs) are the standard in computer vision tasks, such as image classification, semantic segmentation and object detection.
- State-of-the-art CNNs require billions of multiply-and-accumulate operations to process a single image.
- Due to the high computational demands, CNNs are usually executed on specialized hardware (HW) accelerators. Moreover, to meet hardware constraints such as energy and latency, CNNs are compressed with pruning or quantization techniques.
- Additionally, error resilience and model compression are of paramount importance in safety-critical systems such as the onboard computer of an autonomous electric vehicle.

Addressed research questions/problems

- Co-design improves the trade-off decision between task accuracy and hardware performance. However, the combined search space of model compression and hardware mapping cannot be explored exhaustively due to its complexity. Therefore, it is important to find an optimal search strategy to explore the design space [1,2,4,5,6].
- Model compression techniques such as pruning and quantization can be leveraged to reduce data movement and the energy and delay of each operation, respectively [1,2,6].
- Errors can occur due to logic transient, physical defects or adversarial attacks. Hardware-aware and error-aware training and compression strategies can be used to mitigate faults and improve energy efficiency. Compressed models are more susceptible to misdetections induced by malicious adversarial attacks and to computation errors due to logic transients alike [3].



- To holistically optimize both the CNN and the target hardware accelerator for deployment, several abstraction levels must be considered.

List of attended classes

- 01UMNRV – Advanced deep Learning (15/6/2021, 30)
- 01UJBRV – Adversarial training of neural networks (3/6/2021, 15)
- 01UJRIU – Computing Paradigms for Error-Tolerant Applications (26/7/2021 25)
- 03QTIIU – Mimetic learning (8/3/2021, 20)
- 01DNMIU – Optimized execution of neural networks at the edge (5/9/2022, 25)
- 01DNHRV– System level low power techniques for IoT (15/7/2022, 20)

Submitted and published works

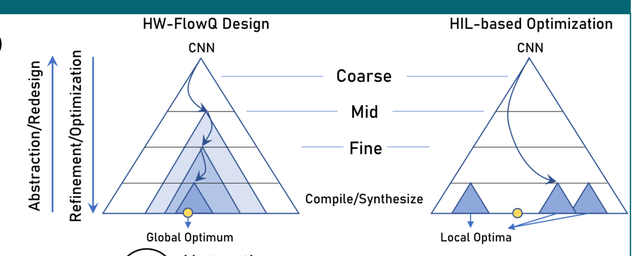
1. Fasfous, N. , Vemparala, M. , Frickenstein A. , Valpreda E. , et al. , "*HW-FlowQ: A Multi-Abstraction Level HW-CNN Co-design Quantization Methodology*", ACM TECS, vol. 20, Issue 5s, Article No. 66, 2021, pp. 1-25
2. Fasfous, N. , Vemparala, M. , Frickenstein A. , Valpreda E. , et al. , "*AnaCoNGA: Analytical HW-CNN Co-Design Using Nested Genetic Algorithms*", DATE , Antwerp, 2022, pp. 238-243
3. Fasfous, N. , Vemparala, M. , Neumer M. , Frickenstein A. , et al. , "*Mind the Scaling Factors: Resilience Analysis of Quantized Adversarially Robust CNNs*", DATE, Antwerp, 2022, pp. 706-711
4. Aiello G. , Bussolino B. , Valpreda E. , Ruo R. , et al. , "*NLCMAP: A Framework for the Efficient Mapping of Non-Linear Convolutional Neural Networks on FPGA Accelerators*", ICIP, Bordeaux, 2022
5. Valpreda E. , Morì P. , Fasfous N. , Vemparala M. , et al. , "*HW-Flow-Fusion: Inter-Layer Scheduling for Convolutional Neural Network Accelerators with Dataflow Architectures*", MDPI Electronics, vol. 11, no. 18, 2022, pp. 2933-2957
6. Vemparala, M. , Fasfous, N. , Frickenstein A. , Valpreda E. , et al. , "*HW-Flow: A Multi-Abstraction Level HW-CNN Codesign Pruning Methodology*", Schloss Dagstuhl LITES, vol. 8, no. 1, 2022, pp. 1-30

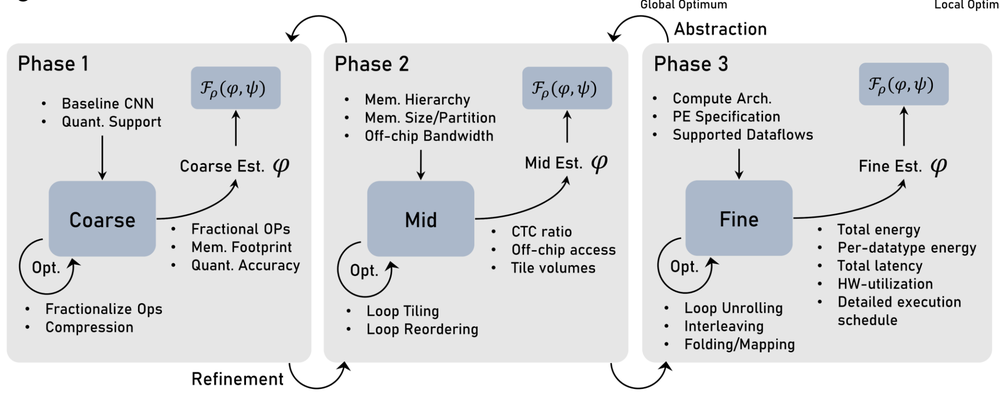
Novel contributions

- A HW-model-in-the-loop compression methodology allowing design space exploration of CNN quantization and pruning strategies and hardware platforms at different design phases [1,2,6]. A genetic agent is used to navigate the quantization search space in [1-2], whereas a reinforcement learning agent is used to select optimal pruning policies in [6].
- A map-space exploration framework that can find the optimal scheduling of CNN workloads executed on re-configurable hardware accelerators based on spatial arrays, which exploits different hardware and CNN abstraction levels to increase/reduce the search space dynamically, to find the schedule that minimizes data movement [1,5,6].
- An error resilient and adversarial robust training strategy for quantized CNNs, based on the observation that models quantized with smaller scaling factors tend to be significantly less susceptible to errors, compared to models with larger ones [3].
- An analytical model of inter-layer scheduling that leverages intermediate data reuse to reduce data movement during the inference on resource-constrained devices, reducing computation energy and latency. This is achieved by fusing the execution of multiple workloads with data dependencies [5].

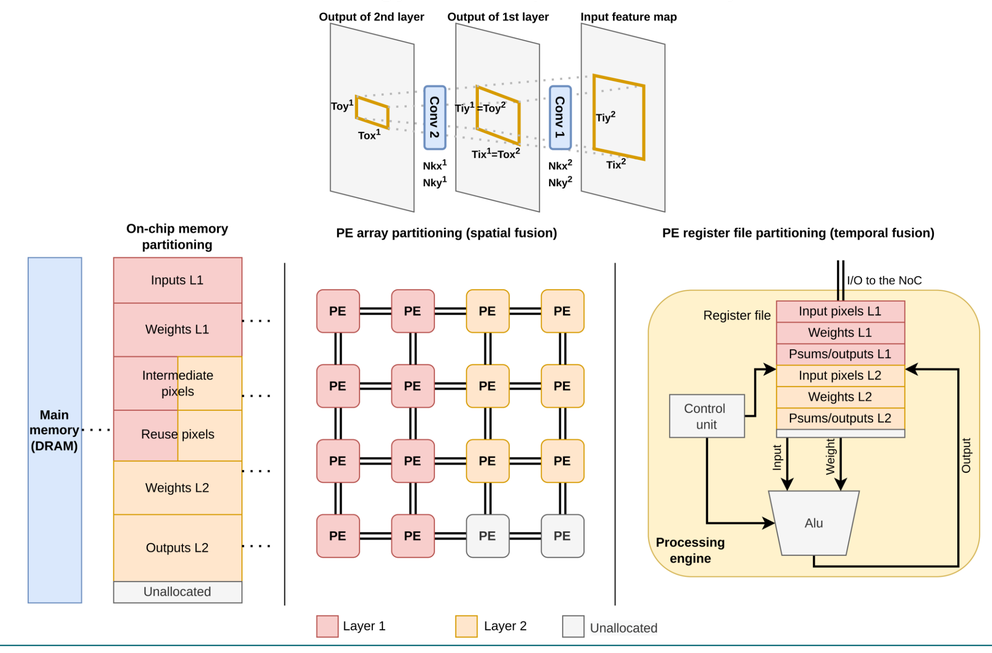
Adopted methodologies

- In HW-Flow and its variants [1,5,6], the map space is explored through iterative steps. Relevant CNN/HW metrics are evaluated to compare the relative performance of the mapping and select the best solution.





- In HW-Flow-Fusion, this three-step approach is adapted to support layer fusion. In phase 1 are searched which layers of a CNN can be fused and scheduled with the available on-chip memory. In phase 2 are selected the fused loop schedules that have high bandwidth efficiency. In phase 3, the remaining schedules are evaluated considering low-level HW details such as NoC/MAC/RF energy and computation latency.



Future work

- Analysis of bit-level pruning techniques to increase 0s within each operand, to reduce dynamic energy of communication and computation and allow compressed data movement. The target hardware architecture is based on a bit-serial accelerator.
- Detection of errors in the output of CNNs for object detection by leveraging the temporal correlation in sequences of frames. Errors are detected by an external agent, the correct prediction is estimated and used for both the current output and future error detection.
- Approximate arithmetic-aware training schemes to reduce the accuracy drop of CNNs due to arithmetic errors (continuation of work [3]). The goal is to trade-off task accuracy with reduced energy, latency, and area.

Politecnico di Torino

PhD program in
Electrical, Electronics and Communications Engineering