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

XXXVII Cycle

Hardware-Software Co-Design for Faster and Stronger CNNs on Edge Devices

Pierpaolo Mori

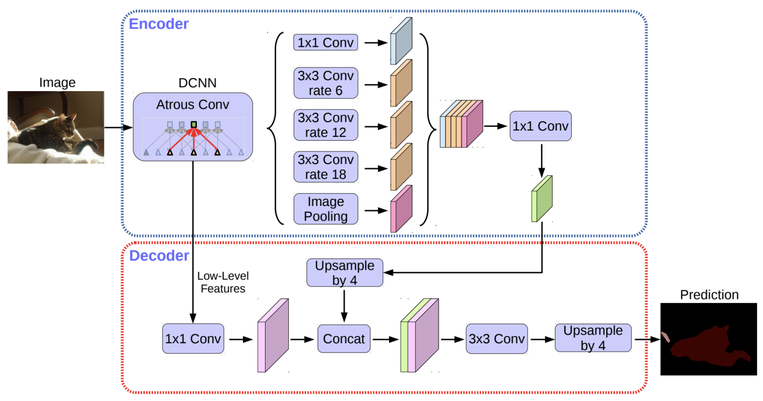
Supervisor: Prof. Claudio Passerone

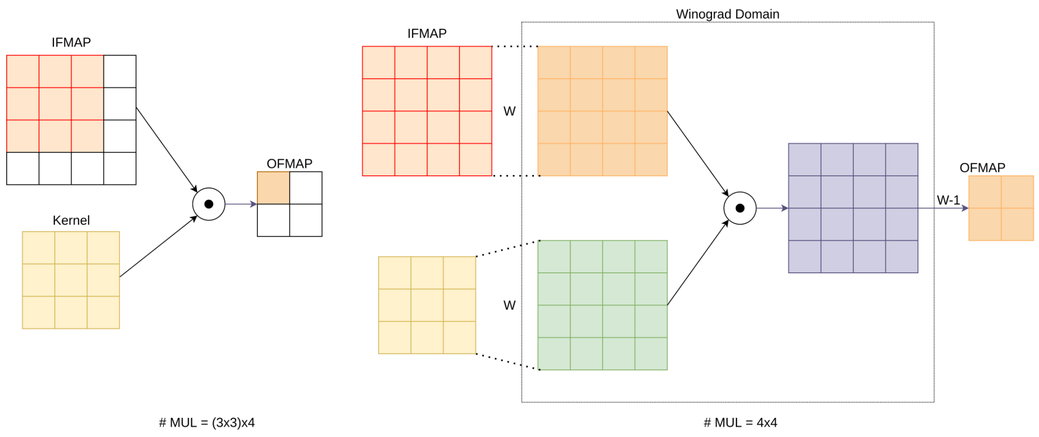
Research context and motivation

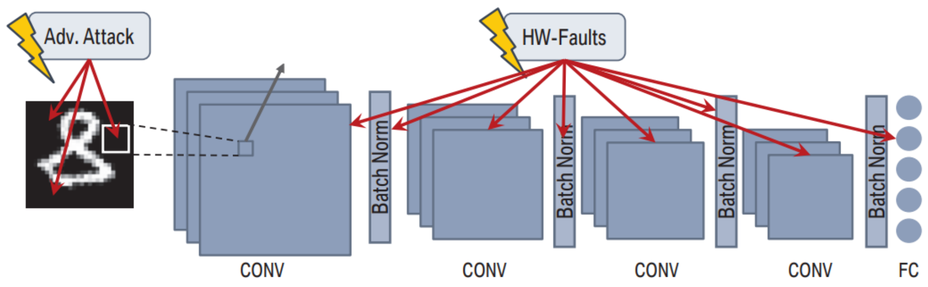
- Convolutional neural networks (CNN) are widely used for solving problems like image classification, object detection, and semantic segmentation.
- The deployment of these networks on resource-limited hardware (HW), such as FPGAs , is challenging due to high memory requirements and energy consumption.
- Modern CNNs models are able to achieve high accuracy.
- HW designs can achieve high performance.
- Hardware and Software have to cooperate in order to design faster, stronger and accurate CNN systems.

Addressed research questions/problems

- Deployment on FPGA of Semantic Segmentation Models is challenging because of model complexity. Model compression techniques such as pruning and quantization are therefore needed. However, a reduction in the number of operations does not always cause a reduction in the inference time.
- Convolution represents the core of CNNs, Winograd transformation allows to turn convolution into a simpler element-wise matrix multiplication, reducing inference time on HW accelerators. However, when Winograd is used with quantization in CNN models, the resulting accuracy is heavily degradeted because of numerical instability.
- Hardware faults, such as bit-flips and stuck at error, can affect the prediction quality of the model.







List of attended classes

- 01UJBRV – Adversarial training of neural networks (5/6/2022, 3)
- 01DNHRV – System level low power techniques for IoT(14/7/2022, 4)
- 01DNMIU – Optimized execution of neural networks at the edge(4/9/2022, 5)

Submitted and published works

- Mori Pierpaolo, Vemparala M. R., Fasfous N., Passerone C., "*Accelerating and Pruning CNNs for Semantic Segmentation on FPGA*", Proceedings of the 59th ACM/IEEE Design Automation Conference, 2022, pp.145-150
- 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.

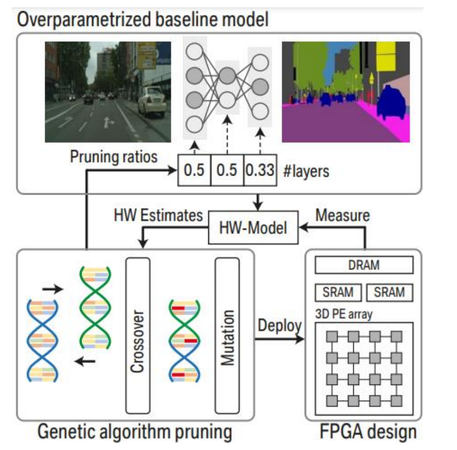
Novel contributions

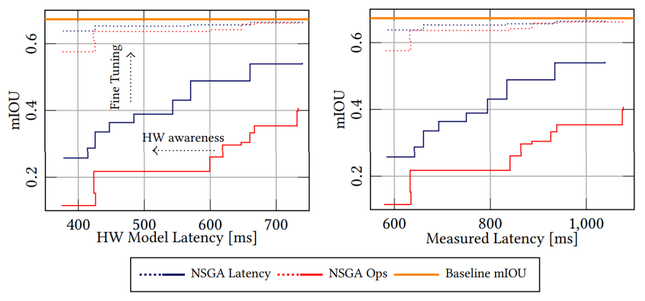
- **FPGA Accelerator Design for segmentation:** able to support 2D dynamic tiling, 3D unrolling, dilated convolutions, and bilinear upsampling. Achieving 90% DSP utilization, 183.3 GOPS throughput, and DRAM accesses reduction by a factor of 2.33×.
- **Genetic Algorithm-based Channel Pruning:** non-dominated sorting genetic algorithm (NSGA-II) is used to determine Pareto optimal pruning configurations, obtaining 2.75× reduction in the number of operations with minimal degradation in prediction quality for DeepLabV3+ model over Cityscapes dataset.
- **Hardware-Aware Pruning:** an analytical HW model of the accelerator is used to steer a latency-driven GA search and outperform pruning configurations based on proxy metrics. The latency-driven GA search provides a performance improvement of 2.44×, with minimal degradation in the prediction quality for DeepLabV3+ model over Cityscapes dataset.
- **End-To-End Winograd-based convolution**: Winograd algorithm with complex number system and quantization are considered at training time.
- **Fault-Aware training:** a fault injection module able to model hardware faults (bitflip, stuck at) is introduced at training time, increasing the robustness of the model.

Adopted methodologies

GA-based pruning

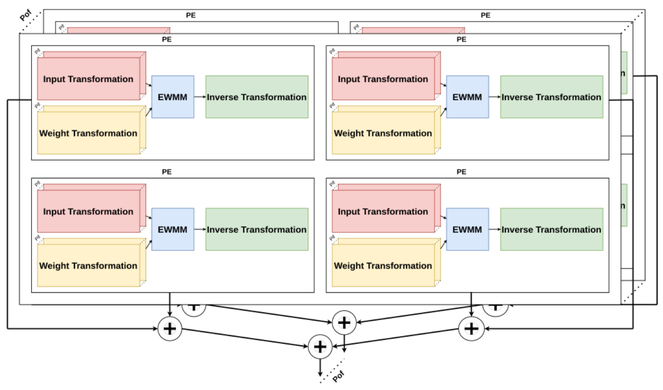
- Channel pruning is formulated as a search problem, where search space is discrete and non-differentiable.
- Single genome represents a potential CNN strategy with as many genetic loci as layers in CNN.
- Loci encapsulate pruning rates
- For each generation, $|\mathcal{P}|$ networks are evaluated to obtain their mIOU and HW estimates.
- HW-related optimization criterion can either be the **number of operations** or the **latency values** calculated using the HW model.





Winograd-based accelerator

- Complex numerical system allows to make the coefficients of Winograd matrices hardware-friendly
- 3D unrolling
- Standard convolution support is no more needed
- Stride-2 is supported by decomposing filters and activations in 4 tiles



Future work

- Analysis and comparison of robustness of Systolic and dataflow architectures (FINN-based).
- New quantization scheme for bit-serial accelerators: trainable scaling factor for each bit of the N-bit value.
- Robustness of Winograd-based accelerators.

Politecnico di Torino

PhD program in
Electrical, Electronics and Communications Engineering