KI DELTA LEARNING
Scalable AI for Automated Driving

Accelerating and Pruning CNNs for Semantic Segmentation on FPGA

Manoj Rohit Vemparala, Alexander Frickenstein, Nael Fasfous, Pierpaolo Mori, Saptarshi Mitra, Lukas Frickenstein

WP 4.2

Networks

Use-Case in Delta Learning

Semantic segmentation is one of the popular tasks in computer vision and autonomous driving, providing pixel-wise annotations for scene understanding.

Technical Solution

We establish an end-end deployment pipeline for semantic segmentation using channel pruning and HW model (Figure 1). We formulate the channel pruning as a search problem using genetic algorithm, where redundant filters are pruned based on layer-wise compression ratios and a magnitude-based heuristic. The pruning rates result in an integer number of remaining channels for each layer. Pruning certain filters leads to large degradation in accuracy (mIOU), highlighting different sensitivities for various pruning choices.

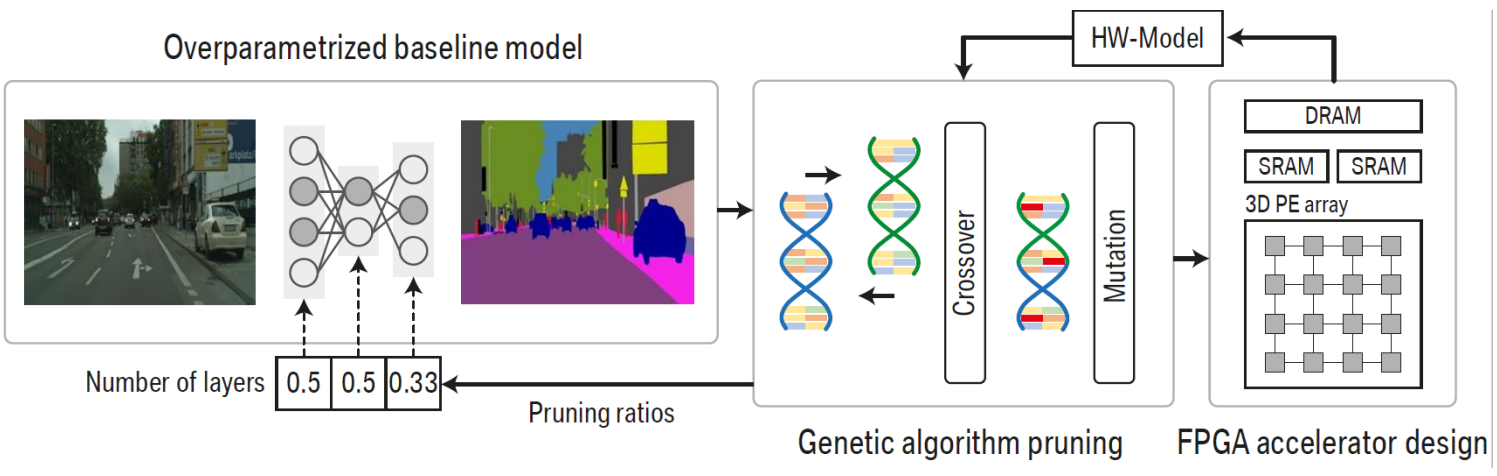

Figure 1: Depiction of the end-to-end CNN deployment pipeline on an embedded platform using genetic search based channel pruning..

Segmentation-based convolutional neural networks require tremendous computational power. The deployment of these networks on resource-limited hardware (HW), such as in-vehicle systems, remains challenging due to high memory requirements and energy consumption. In this work, we focus on the deployment of a state-of-the art semantic segmentation-based CNN architecture, namely the DeepLabV3+ [3] model, on an embedded reconfigurable hardware. We achieve 2.44x lower latency with minimal degradation in semantic prediction quality (-1.98 pp lower mean intersection over union) compared to baseline DeepLabV3+ model, evaluated on Arria-10 FPGA.

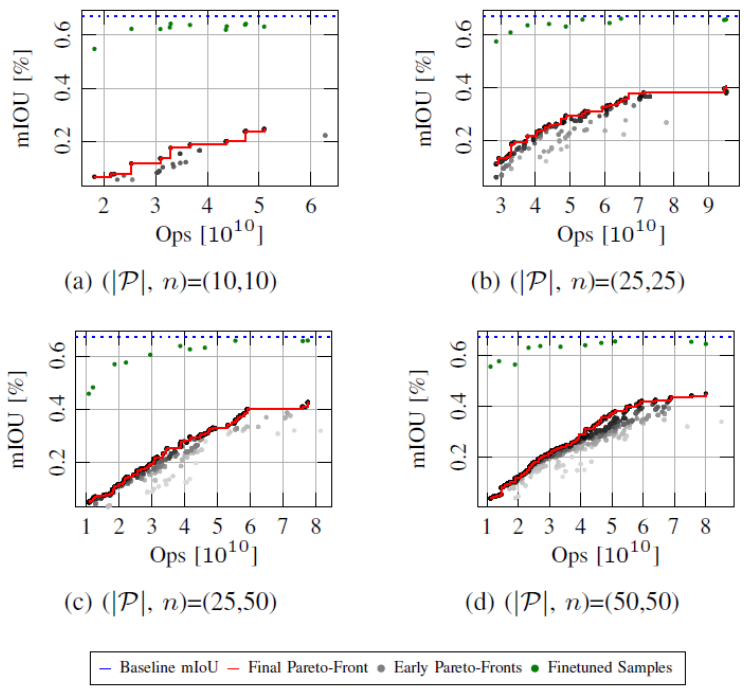

Figure 2: NSGA Search using Ops without compression constraints. Grey to black shades represent Pareto-fronts of older to newer generations, red points belong to the final Pareto-front.

Evaluation

We conduct a hyper-parameter study for NSGA-II based search to understand the characteristics of the pruning search space. In Figure 2, we study the pruning performance by varying the hyper-parameters of the population size P and number of gen n. Increasing the P and n demands higher number of GPU hours required to obtain Pareto optimal pruning configurations. Proxy metrics, such as operation count (OPs), does not always guarantee tangible improvements on measured hardware estimates. Figure 3 illustrates the dominating configurations for the HW aware pruning.

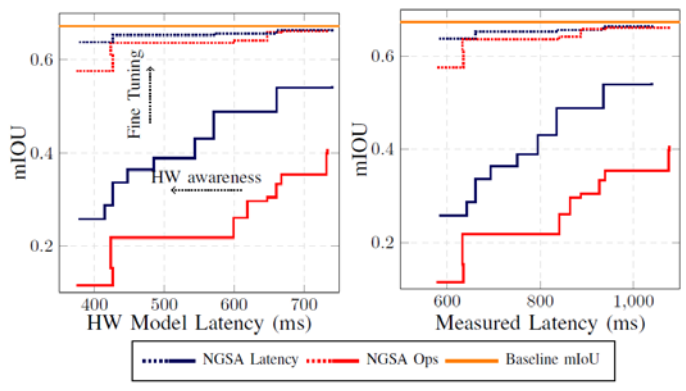

Figure 3: NSGA Search using HW Model Latency with compression constraints.

BMW GROUP

For more information contact:
Manoj-Rohit.Vemparala@bmw.de

Partners

External partners

KI Delta Learning is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Energy.

Supported by: Federal Ministry for Economic Affairs and Energy on the basis of a decision by the German Bundestag

www.ki-deltalearning.de @KI_Familie KI Familie