KI DELTA LEARNING
Scalable AI for Automated Driving

Accelerating and Pruning CNNs for Semantic Segmentation on FPGA

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

* BMW AG | + TUM

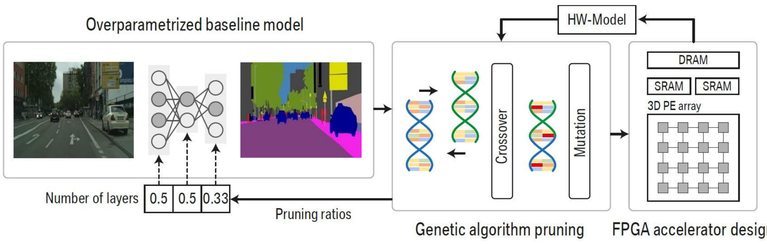

Figure 1: Depiction of the end-to-end CNN deployment pipeline on an embedded platform using genetic search based channel pruning.

Research problem

Semantic segmentation is one of the popular tasks in computer vision and autonomous driving, providing pixel-wise annotations for scene understanding.
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+ model, on an embedded reconfigurable hardware.

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 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. Fine-tuning is performed after the pruning is complete resulting in minimal accuracy degradation with HW benefits.

Evaluation

We conduct a hyper-parameter study for non dominated sorting genetic algorithm (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.

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. Table 1 highlights the benefits of the approach on NVIDIA-GPUs for DeepLab-ResNet 50.

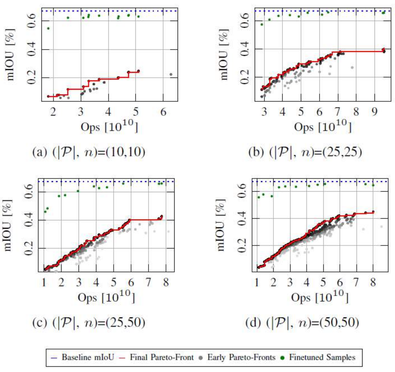

Figure 2: NSGA Search using Ops without compression constraints. Grey to black shades represent Pareto-fronts of older to newer generations, red points belong to the final Pareto-front. Green dots represent fine-tuned pruned solutions.

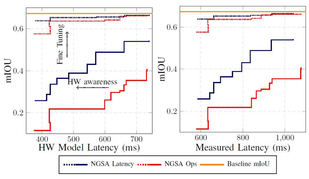

Figure 3: NSGA Search using HW Model Latency with compression constraints.

Typez	Compute Complexity (GFLOPs)	Latency per batch (ms)	mIOU (%)
Baseline	276	114	71.07
Ops based	171	98	69.99
HW-aware	141	**83**	69.99

Table 1: HW-in-loop based pruning search for NVIDIA-2080 GPU for DeepLab-ResNet50 based semantic segmentation.

Embedded Systems

Partners

External partners

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

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 Climate Action.

www.ki-deltalearning.de @KI_Familie KI Familie

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