KI DELTA LEARNING
Scalable AI for Automated Driving

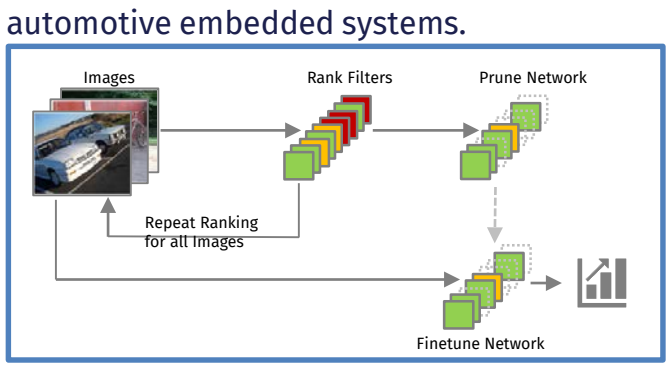
Developing Pruning Methods for Object Detection for Embedded Hardware

Sven Mantowsky, Syed Saqib Bukhari

WP 4.2

Use-Case in Delta Learning

The popularity of deep neural networks in the area of ADAS or AD and their application to embedded systems is increasing rapidly. These embedded systems are limited in their compute resources, as only limited energy can be made available in the car. In order to meet these restrictions and to prevent a failure of software with possible catastrophic consequences during their application, the AI models to be used on this hardware must be optimized properly. Work package 4.2 focuses on the development and implementation of methods for the aforementioned requirements and constrains in order to be able to successfully deploy AI models to automotive embedded systems.



Technical Problem

One of the main problems ZF is dealing with is the difficulty to achieve a minimal inference time of an AI model after its deployment on embedded hardware in order to facilitate a safe execution. However, not only the execution time is significant, but also memory usage, power consumption and other factors. When pruning/optimizing networks, security and its explainability are often not taken into account. ZF is trying to counteract this with a novel pruning method.

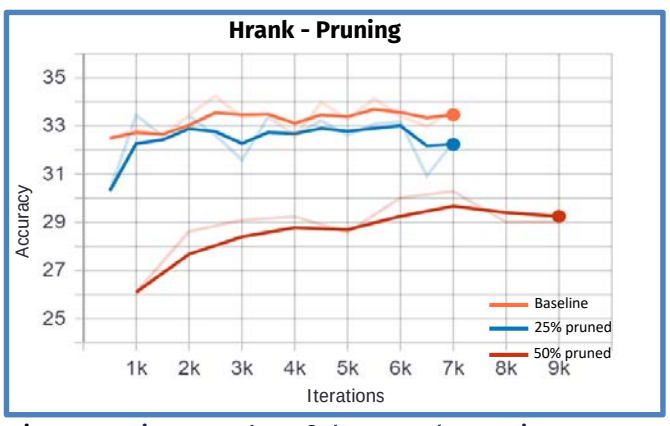


Figure 1: First results of the Hrank Pruning Algorithm, adapted to the SSD300 object detector. Almost the same accuracy was achieved with 25% less parameters than the baseline.

Technical Solution

First we used the HRank pruning algorithm from Lin et. Al [1] and adjusted it to object detection. For this purpose we used the SSD300 [2] network with a VGG/Resnet50 [3][4] backbone. For the pruning process we included the backbone and auxiliary layers of the network. We evaluated different pruning ratios after a finetuning process on ZFs own automotive hardware, the ProAI.
Currently ZF is exploring a new pruning method. The work is under progress, and we will be introduced it in the near future.

Evaluation

We tested the pruned SSD Networks on a workstation PC as well as our embedded system ZF ProAI. The results of the HRank pruning in Fig. 1 show similar accuracy at 25% pruning while minimizing the memory footprint. Fig. 2 shows the inference time per image. The improvement of the pruned network is marginal compared to the baseline. The reason for is the framework Pytorch we used. There are different methods in this framework, like forwar- and backward hooks, which slow down the inference speedup after the pruning process. While using CPU instead of GPU for inferencing, the speedup is great after pruning.

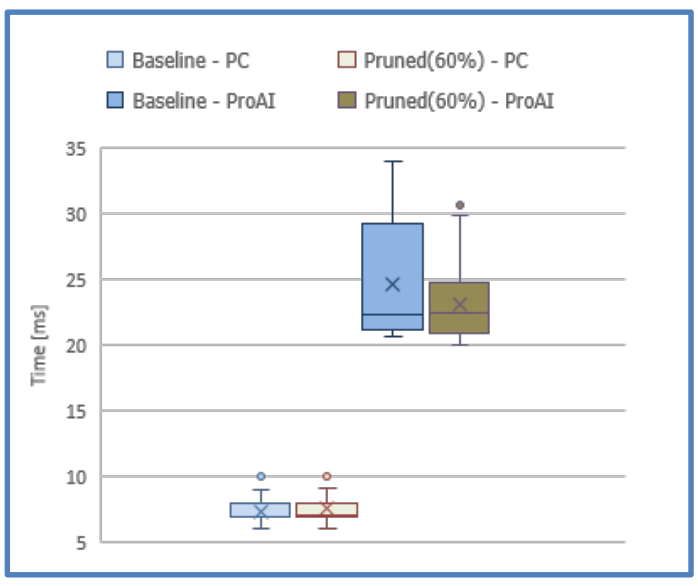


Figure 2: Performance testing of SSD300 pruned with HRank. There is only a marginal improvement due to the peculiarity of the hooks used by the Pytorch framework.

Networks

References:

[1] Lin, Mingbao and Ji, Rongrong and Wang, Yan and Zhang, Yichen and Zhang, Baochang and Tian, Yonghong and Shao, Ling. HRank: Filter Pruning Using High-Rank Feature MapIEEE/CVF Conference on CVPR, 2020.
[2] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, Alexander C. Berg: SSD: Single Shot MultiBox Detector. CoRR abs/1512.02325 (2015)
[3] Simonyan, K. & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556.
[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Deep Residual Learning for Image Recognition. CoRR abs/1512.03385

For more information contact:
Sven.Mantowsky@zf.com

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