KI DELTA LEARNING
Scalable AI for Automated Driving

Interpretable Pruning

Sven Mantowsky, Syed Saqib Bukhari, Georg Schneider

Embedded Systems

Pruning and Explainability

Pruning is a technique to remove less important connections/filters from a model, making it more efficient while preserving or improving performance. This process usually is non-transparent for the user and is barely interpretable and there are hardly any methods that combine both aspects. ZF has faced this problem and developed a method that combines pruning and interpretability, called “Interpretable Pruning”.

HRank Filter Pruning for Object Detection

In the first phases of the project, we extended HRank[1] filter pruning technique:

- extended method from classification to object detection
- tested with SSD[2] and a VGG16[3] backbone (Result shown in Fig. 2).
- tested pruned models on the ZF ProAI (Fig. 1)

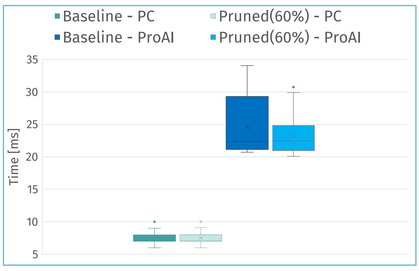

Figure 1: Performance testing of SSD300 model pruned with HRank. There is only a marginal improvement due to the peculiarity of the hooks used by the PyTorch framework. (©ZF Group)

Interpretable Pruning

The process of Interpretable Pruning has the following key points:

- generating heatmaps for each filter for each layer
- heatmaps reflects the influence of the respective filter for one image/prediction
- rank filters for several input images

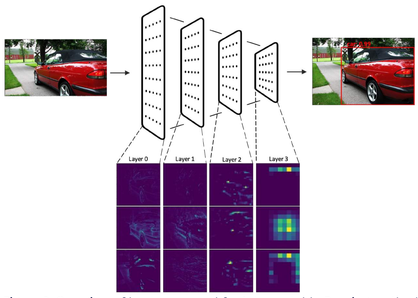

Figure 3: Overview of heatmaps used for Interpretable Pruning method. (©ZF Group)

- prune filter with least activations
- finetune compressed model

→ Method provides explainability and traceability of the pruning process

Results

The results of our Interpretable Pruning method are shown in Fig. 2. Up to a compression rate of 40%, the accuracy remains almost unchanged. We observed a significant drop after increasing the pruning rate by a few percent. Compared to other methods from the literature, however, it was to be expected that the accuracy would continuously decrease as the compression rate increased. It is noticeable that from a compression rate above 55%, the accuracy remais at level of 20%. This behavior could not be confirmed with classification, which indicates an issue with the combination of method and model architecture. This behavior will be further investigated in the future.

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 Map. IEEE/CVF Conference on Computer Vision and Pattern Recognition (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. ECCV (1) , volume 9905 of Lecture Notes in Computer Science, page 21-37. Springer, 2016.
[3] Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556, 2014.

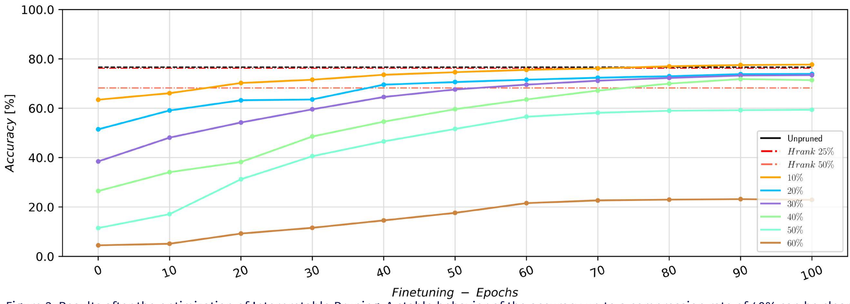

Figure 2: Results after the optimization of Interpretable Pruning. A stable behavior of the accuracy up to a compression rate of 40% can be clearly seen. After that, the accuracy drops very quickly. The comparison to the HRank method shows that our approach needs further optimization. (©ZF Group)

Partners

External partners

For more information contact:
sven.mantowsky@zf.com
Saqib.bukhari@zf.com

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.

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

www.ki-deltalearning.de @KI_Familie KI Familie