Supplementary: Holistic Filter Pruning for Efficient Deep Neural Networks

Lukas Enderich
Robert Bosch GmbH
71229 Leonberg
lukas.enderich@de.bosch.com

Fabian Timm
Robert Bosch GmbH
71272 Renningen
fabian.timm@de.bosch.com

Wolfram Burgard
University of Freiburg
79110 Freiburg im Breisgau
burgard@informatik.uni-freiburg.de

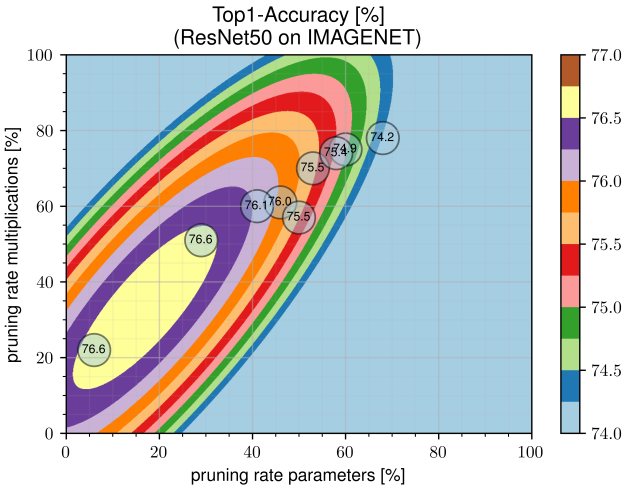

Figure 1. Top-1 accuracies of ResNet-50 on ImageNet with different pruning rates. The performance values are illustrated by colored level curves created by fitting a second-order polynomial.

1. Level curves of pruned ResNet-50

In figure 1, the top-1 accuracies of ResNet-50 on ImageNet are shown for different pruning rates. The performance values are illustrated by colored level curves created by fitting a second-order polynomial. The baseline is 76.15% and taken from the torchvision model zoo.

One can observe that *Holistic Filter Pruning* is able to prune up to 60% of the multiplications and up to 40% of the parameters with no significant loss in die accuracy (76.1% vs 76.15% top-1 accuracy). Furthermore, pruning 50% of the multiplications and 28% of the parameters even improves the top-1 accuracy by approximately 0.5%. This is due to the capability of pruning methods to improve the ability of DNNs to generalize. If more than 50% of the parameters are pruned, the top-1 accuracy drops below 76%. Overall, one can observe that pruning the parameters of ResNet-50 has a greater impact on the performance than pruning the multiplications. Hence, *Holistic Filter Pruning* prunes nearly 80% of the multiplication with less than 2% drop in accuracy.

2. Visualization of ResNet-50

This section analyzes how the overall reduction of parameters and multiplications is proportionally distributed among the individual layers. For example, if 1000 parameters are pruned from the model and the first layer is reduced by 150 parameters, the proportional contribution of the first layer for the parameter pruning is 15%.

Figure 2 shows the proportional pruning rates of ResNet-50 with 55% pruned multiplications and 51% pruned parameters. The first diagram shows the proportional pruning rates for the multiplications while the second diagram shows the proportional pruning rates for the parameters. The pruning rates are shown for different training epochs and refer to the pruning result at that time step (e.g., after 10 epochs 46% of the multiplications were reduced). Additionally, the diagrams show the total number of multiplications and parameters of the unpruned layers (dotted lines).

Initially, a correlation between the layer size and the proportional pruning rates can be observed. As for ResNet-56, the proportional pruning rates of the layers change over the epochs: While the proportional pruning rates for the multiplications of the first 10 layers decrease, the pruning rates with layer index 11, 23, and 41 increases. These are exactly the layers which contribute most to the overall number of multiplications. In case of the pruned parameters, the three basic blocks of the ResNet architecture are visible and marked with A, B, and C. With an increasing layer size, the pruning rates increase to the same degree. Since block C has the highest contribution to the total number of parameters,

it also shows the highest proportional pruning rates.

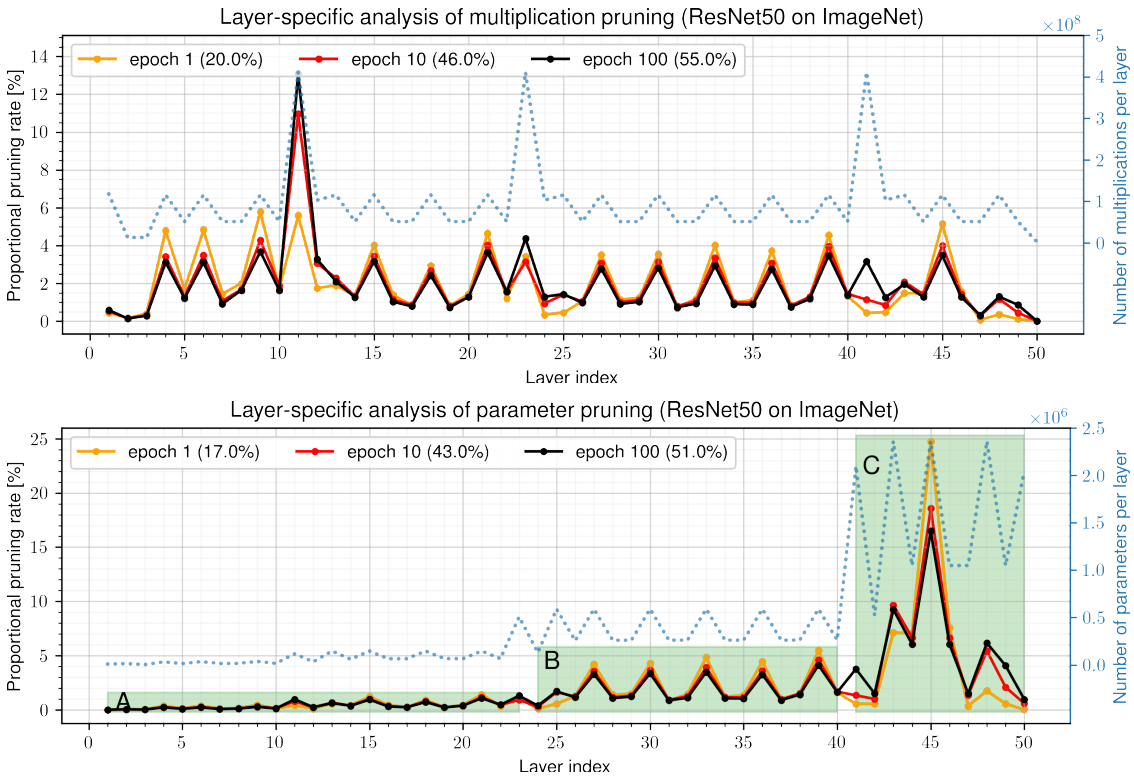

Figure 2. The upper plot shows the proportional pruning rates of the individual layers of ResNet-50 with 55% pruned multiplications and 51% pruned parameters. Proportional pruning rates indicate the contribution of single layers to the overall pruning rate. E.g., if 1000 multiplications are pruned from the model and the first layer is reduced by 150 multiplications, the proportional pruning rate of the first layer is 15%. The lower plot indicates the proportional pruning rates for the number of parameters.