Supplementary Material
Adversarial Robust Model Compression using In-Train Pruning

Manoj-Rohit Vemparala[1], Nael Fasfous[2], Alexander Frickenstein[1], Sreetama Sarkar[1], Qi Zhao[3], Sabine Kuhn[1], Lukas Frickenstein[1], Anmol Singh[1] Christian Unger[1], Naveen-Shankar Nagaraja[1], Christian Wressnegger[3], Walter Stechele[2]

[1]*BMW Autonomous Driving*, [2]*Technical University of Munich*, [3]*Karlsruhe Institute of Technology*

S.1. Hyper parameters for In-Train Pruning

In this section, we discuss the choices for hyper-parameter selection across various experiments in more detail. For the experiments in sections 4.1, 4.2 and 4.3, we initialize the trainable prune masks M at every layer to 1.0. We use SGD optimizer to minimize the prune loss $\mathcal{L}_{\text{prune}}$ with a learning rate of 0.01.

Pruning ResNet20 and ResNet56 on CIFAR-10 For the pruning experiments performed on CIFAR-10, we train for 300 epochs with an initial learning rate of 0.1. We decrease the learning rate by 0.1 after every 80 epochs. We start minimizing the prune loss $\mathcal{L}_{\text{prune}}$ at $E_{\text{Prune, Start}}$ = 20 and freeze the masks at the $E_{\text{Prune, End}}$ = 240. We minimize the $\mathcal{L}_{\text{prune}}$ using the SGD optimizer every epoch. We investigate the effect of scaling factor b in Eq.(3), when the target operation constraint ψ^*=0.2 in Table. 1. We observe that the b=10 doesn't satisfy the target pruning constraints for ResNet20. Thus, for extreme pruning constraints, we use higher value of b=50.

Table 1: Exploring different values of scalar constant b in Prune Loss

Model	b	Ops Reduction Target	Ops Reduction Actual	Acc [%]
ResNet20	10	0.2	0.33	88.65
	50	0.2	**0.17**	88.17
ResNet56	10	0.2	0.21	91.34
	50	0.2	**0.18**	91.57

Regularizing Prune Masks As specified in section 3.2, we incorporate trainable prune masks M in the regularization loss $\mathcal{L}_{reg}$ along with weights W. Regularizing the trainable masks M avoids early bias of binary masks M_b. We demonstrate this effect by studying the training behaviour for the proposed pruning scheme in Fig. 1. We constrain the number of operations to 30% of the baseline model and set b=10 to understand the behaviour of $\mathcal{L}_{HW}$ across training iterations. We observe that regularizing the trainable prune masks M (blue), achieves the target constraint ($\mathcal{L}_{HW} = 0$). We observe that there are no reduction in $\mathcal{L}_{HW}$, when prune mask M is not regularized. This occurs as the initialization of the trainable mask M is set to 1.0. Without regualarizing the prune masks M and using a lower initialization such as $M_{\text{init}} = 0.3$ (green) would result in bias for the pruning decision during the early stages of training process. This would cause longer training time to achieve target constraints.

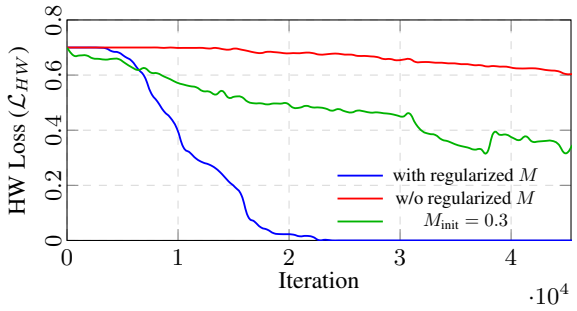

Figure 1: Comparison of HW loss $\mathcal{L}_{HW}$ across several training iterations for different operation constraints ψ^*=0.3 for different settings of continuous masks M.

S.2. Robust Pruning of ImageNet dataset

In Table 2, we analyze the scalability of the in-train pruning approach for larger dataset such as ImageNet [1]. As baseline models, we train ResNet18 and ResNet50 using

Figure 2: Qualitative comparison of In-train pruning (right) with its baseline predictions (left).

FastAT [2] to obtain a robust baseline model. We consider random FGSM with strength ϵ = 2/255, step size α = 2.5/255 and train for 100 epochs. The learning rate is initially 0.1 and scaled down by 10 every 30 epochs. We report adversarial accuracy using the PGD attack with ϵ = 2/255 and α = 0.5/255 for 20 iterations.

Table 2: In-train pruning for various operation constraints for ImageNet dataset

Model	Acc [%]	Adv. Acc [%]	Ops Reduction Target	Ops Reduction Original	Model Size $\times 10^6$
ResNet18-AT	47.22	28.29	1.0	-	11.68
ResNet50-AT	58.35	37.33	1.0	-	25.60
ResNet18-Prune	45.38	29.48	0.7	0.69	10.60
ResNet18-Prune	48.98	27.34	0.5	0.49	10.02
ResNet18-Prune	43.95	19.08	0.3	0.29	8.43
ResNet50-Prune	52.84	31.32	0.7	0.59	19.55
ResNet50-Prune	52.99	27.93	0.5	0.48	17.43
ResNet50-Prune	53.01	29.77	0.3	0.29	14.66

S.3. In Train Pruning on Object Detection

In this section, we constrain the number of operations of CenterNet [3] model to 50% using the in-train pruning approach on the task of object detection. We trained CenterNet with DLA-34 [4] backbone on the Kitti dataset [5]. We use a 75% and 25% split for the training and validation set respectively. We train for 250 epochs using ADAM optimizer and step learning policy. We decrease the learning rate by 0.01 at 60, 90, and 120 epochs. We report the 2D mAP for validation data on easy, medium, and hard constraints of the car class in Table 3.

In Fig. 2, we perform qualitative comparison of the predictions obtained using in-train pruning (right) with the baseline model (left). *Green* boxes indicate *ground truths* and *Blue* boxes indicate *predictions*. In the first row, we observe that the pruned model doesn't predict the bounding box for the car present at the far right corner. In the second row, we observe that the overlap between in-train pruned predictions and ground truth bounding boxes is very high. This also reflects the higher mAP for the hard constraint in Table 3.

Table 3: Kitti validation for in-train pruned CenterNet with 50% constrained operations

Method	mAP (car) [%] Easy	Medium	Hard	Train Time
Baseline	89.24	**80.56**	71.65	23
In-train Prune (Ours)	85.83	78.94	**78.21**	32

References

[1] J. Deng, W. Dong, R. Socher, *et al.*, "ImageNet: A Large-Scale Hierarchical Image Database," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[2] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *International Conference on Learning Representations (ICLR)*, 2020.

[3] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," in *arXiv preprint arXiv:1904.07850*, 2019.

[4] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research (IJRR)*, 2013.