ADVERSARIAL ROBUST MODEL COMPRESSION USING IN-TRAIN PRUNING

*MANOJ ROHIT VEMPARALA, † NAEL FASFOUS, *ALEXANDER FRICKENSTEIN, *SREETAMA SARKAR,
♦ QI ZHAO, *SABINE KUHN, *LUKAS FRICKENSTEIN, *ANMOL SINGH, * CHRISTIAN UNGER,
* NAVEEN-SHANKAR NAGARAJA, ♦ CHRISTIAN WRESSNEGGER, † WALTER STECHELE

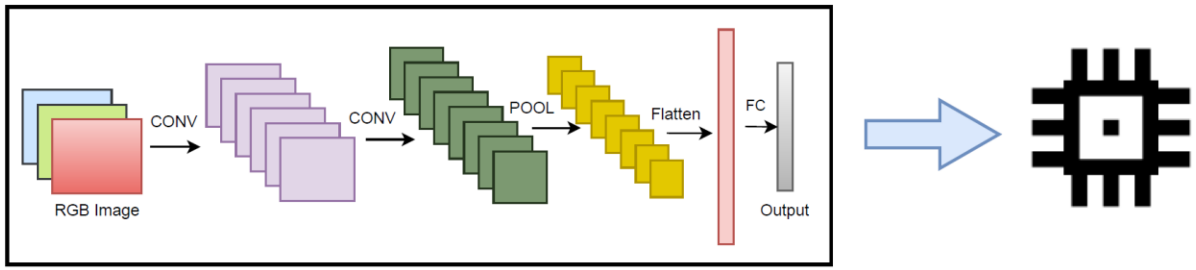
*BMW AG
† TECHNICAL UNIVERSITY OF MUNICH
♦ KARLSRUHE INSTITUTE OF TECHNOLOGY

3RD CVPR WORKSHOP
SAFE ARTIFICIAL INTELLIGENCE FOR AUTOMATED DRIVING

MOTIVATION

- Convolutional Neural Networks (CNNs) have achieved success in **image classification [Deng et al. CVPR 2009]**, **image segmentation [Chen et al. ECCV 2018]** and **object detection [Zhao et al. 2019]**.

- Huge network size consequently increases **latency**, **energy** and **storage requirements.**



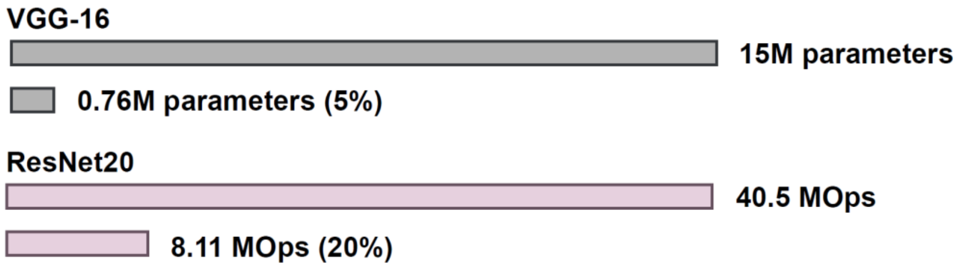
- Compressing CNNs using pruning or quantization techniques is essential for deployment in resource-constrained platforms.

- Robustness of CNNs against Adversarial Attacks [Szegedy et al. ICLR 2014] mandatory for its application in security-critical applications like **Autonomous Driving**, Malware Detection.

- Goal: Efficiently deploy CNNs on secure/robust embedded platforms.

OBJECTIVES

Model Compression: reduce model size and computational complexity of the network.

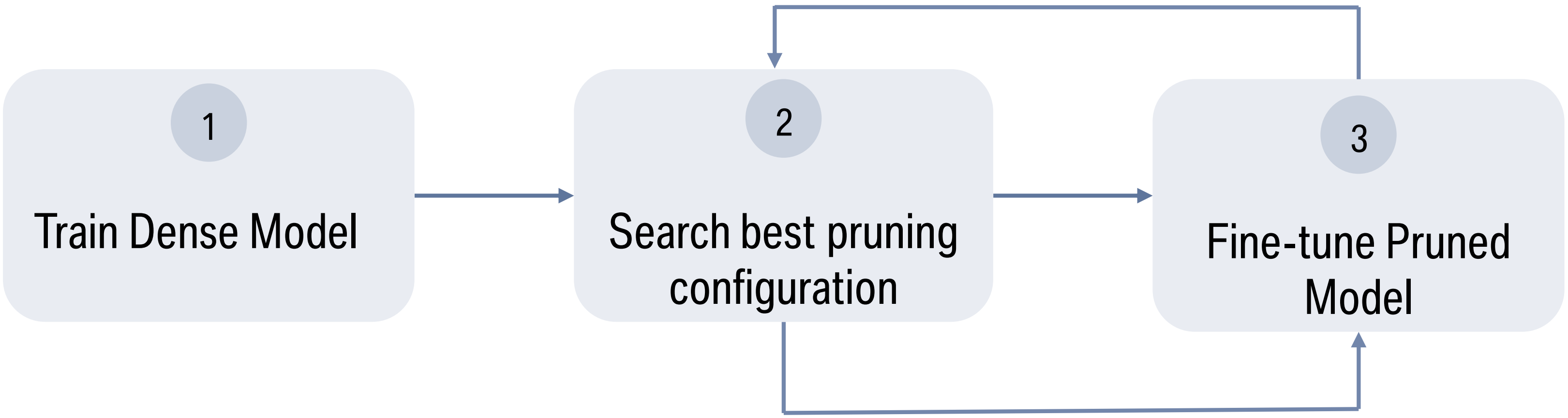


Task-specific Performance: retain accuracy of the model for natural examples.

Adversarial Robustness: correctly classify images generated using adversarial attacks.

Search time Optimization: minimize GPU hours for searching prune configuration

RELATED WORK - POST TRAIN PRUNING



- Three stage pipeline.
- Efficient pruning configuration can be searched using Reinforcement-Learning [He et al. ECCV 2018], [Huang et al. WACV 2018].

Advantages

- automated learning of layerwise sparsities.
- good compression performance with negligible accuracy degradation.

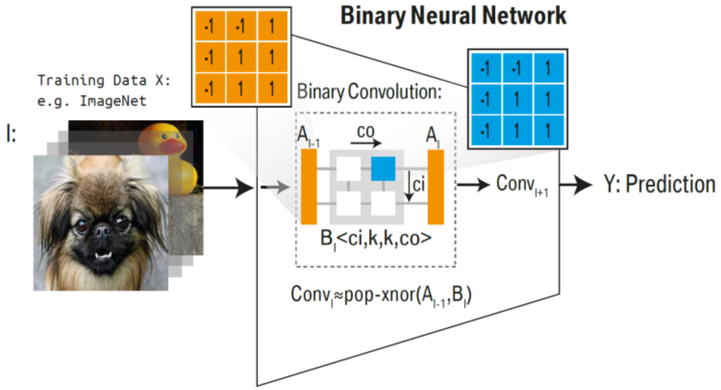
Shortcomings

- Iterative fine-tuning, if required, increases search time manifold.
- leads to sub-optimal performance and high search time when considering adversarial robustness

RELATED WORK - ROBUST MODEL COMPRESSION

Attacking Binary Neural Networks [Galloway et al. ICLR 2018]

- BNNs show inherent improvment of robustness compared to full precision models.
- Discontinuous and approximated gradients of BNNs during the training gives them an advantage over full-precision networks for adversarial attacks.



Backward pass consists of approximated gradient

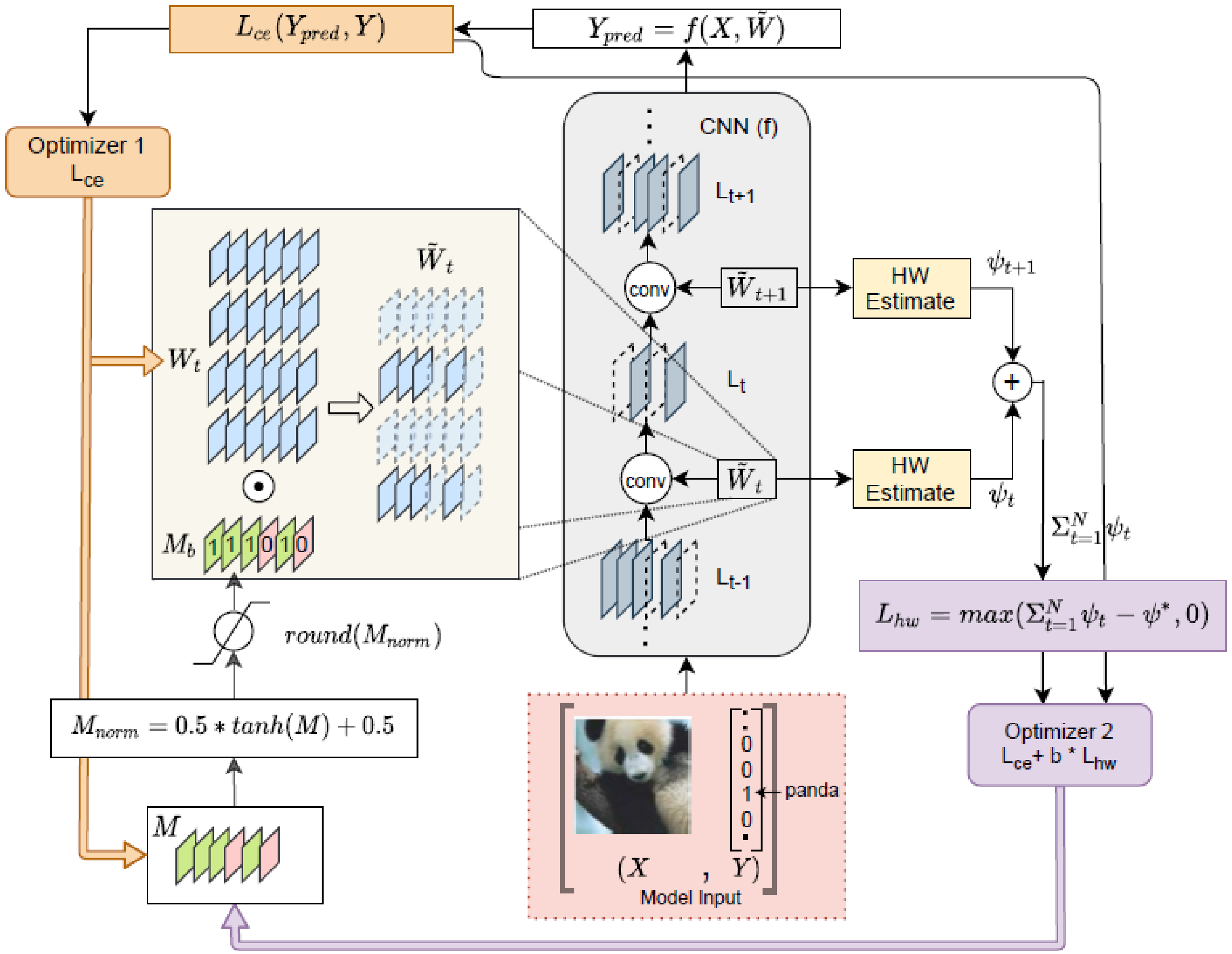
$$g_W = g_B 1_{|w| \leq 1}$$

Robust Pruning

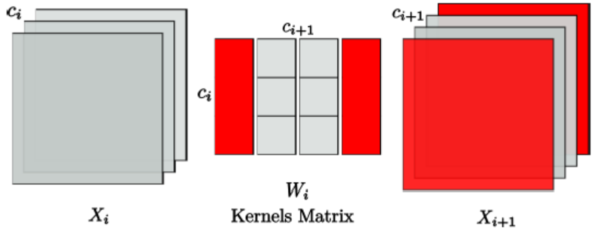
- RobustADMM [Ye et al. ICCV 2019] : concurrently prune and adversarially train an over-parameterized network.
- ATMC Pruning [Gui et al. NeurIPS 2019] : pruning, factorization and quantization.
- Hydra [Sehwag et al. NeurIPS 2020] : gradient based importance score to obtain robust pruned model.

All these approaches require pretrained model.

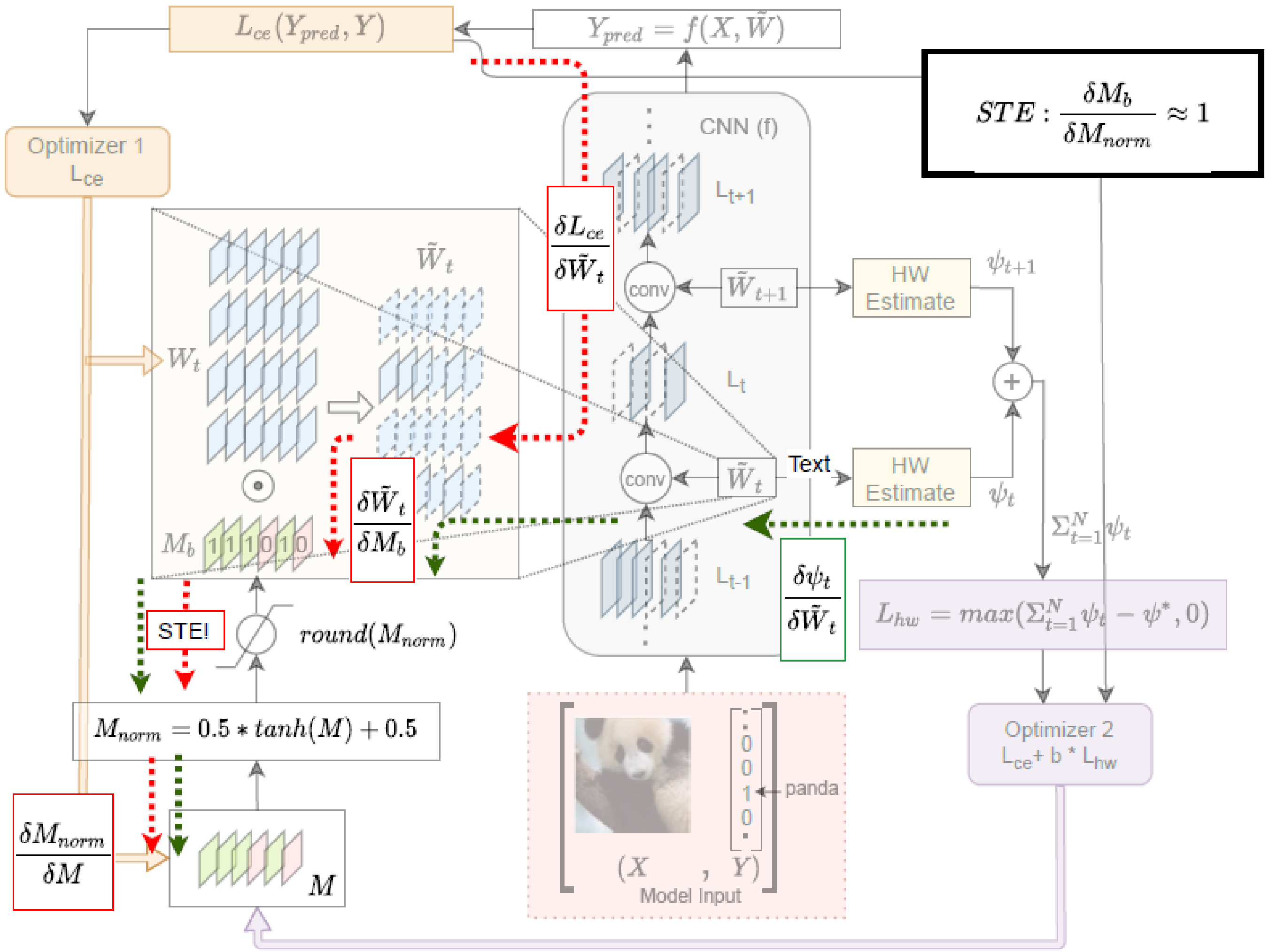
IN-TRAIN PRUNE METHODOLOGY



- Our approach introduces trainable masks (M) for model pruning.
- At every step, the CE entropy loss updates the prune masks capturing the importance scores across the training duration.
- Various pruning regularity such as irregular weight pruning and channel pruning (no specialized HW implementation).

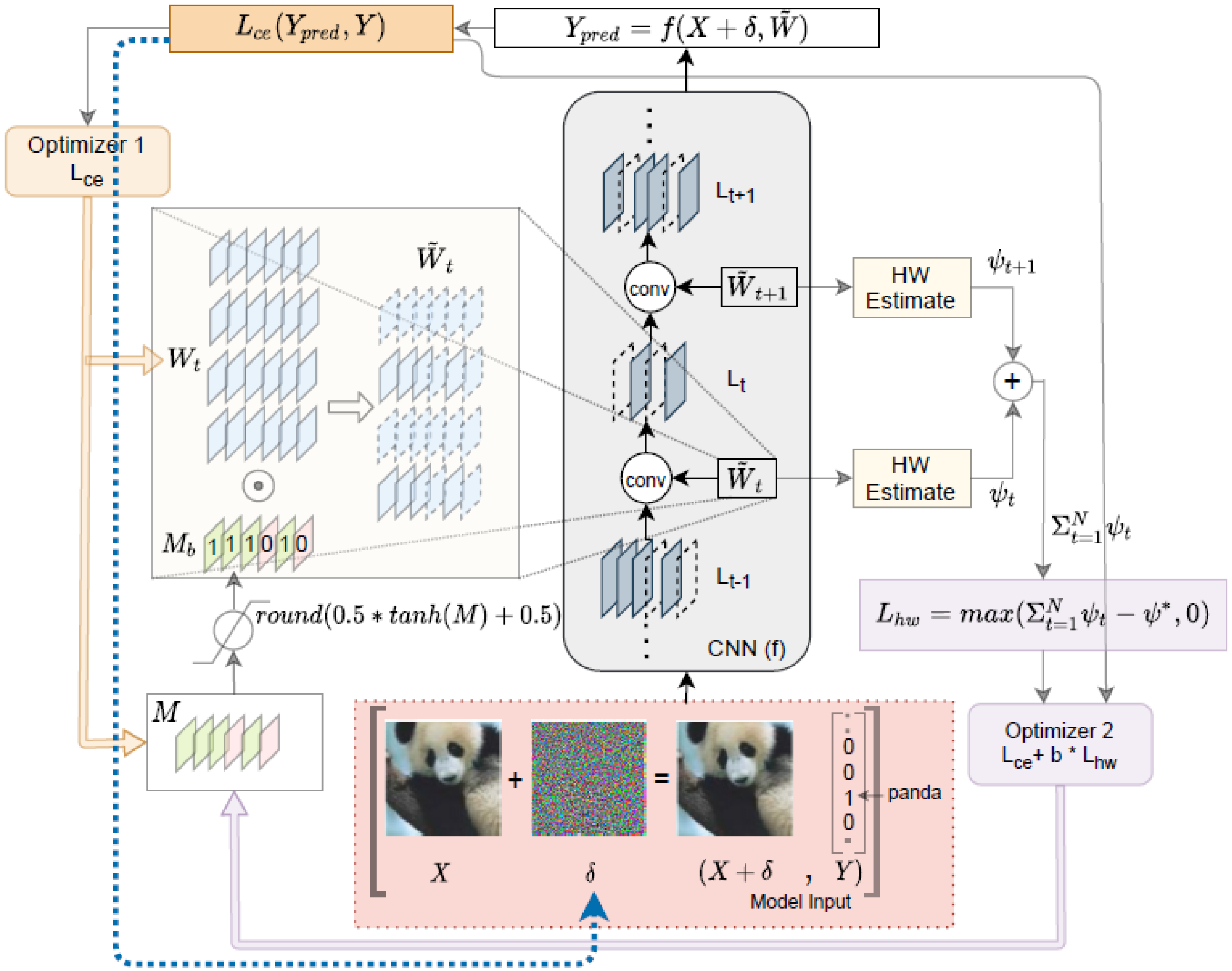


GRADIENT FLOW – UPDATING PRUNE MASKS



- We use tanh, scale,shift and round operations to derive the binary masks $M_b \in \{0, 1\}$
- Any discrete function with a limited range set such as Round () would introduce zero gradients.
- Straight-through Estimator (STE) is used to obtain gradient updates for trainable masks (M) from binary masks (M_b).
- Important to regularize trainable masks along with the CNN weights to ensure frequent updates during the training.

ROBUST IN-TRAIN PRUNE METHODOLOGY

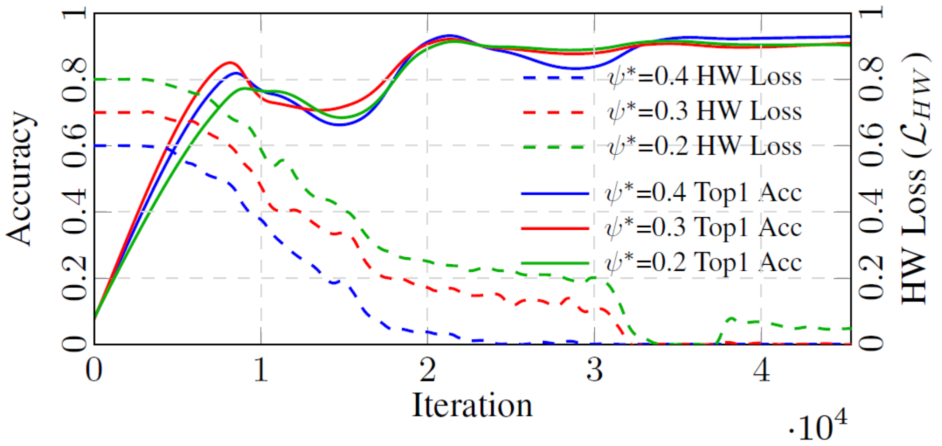


- We integrate the intrain pruning approach with state of the art defense method FastAT [Wong et al. ICLR 2020] to ensure robust compression.
- Fast AT uses single iteration of random FGSM to generate attacked images.

IN-TRAIN PRUNING : CONSTRAINED OPTIMIZATION

Dataset	Model	Accuracy [%]	Ops Reduction Target	Ops Reduction Actual	Param Reduction
CIFAR10	ResNet56	93.56	1.0	-	1.0
		93.03	**0.4**	**0.35**	0.55
		92.38	**0.3**	**0.28**	0.50
		91.57	**0.2**	**0.18**	0.37
ImageNet	ResNet18	68.53	1.0	-	1.0
		67.22	**0.7**	**0.69**	0.88
		65.06	**0.5**	**0.45**	0.78

- Intrain pruning meets target hardware constraints
- Accuracy degradation of 1.99 pp for ResNet56 on CIFAR10 [Krizhevsky et al 2010] for 80% reduction in operations.
- Various HW constraints are met during stages of the training.



ROBUST PRUNING – COMPARISION WITH RL BASED PRUNING APPROACH

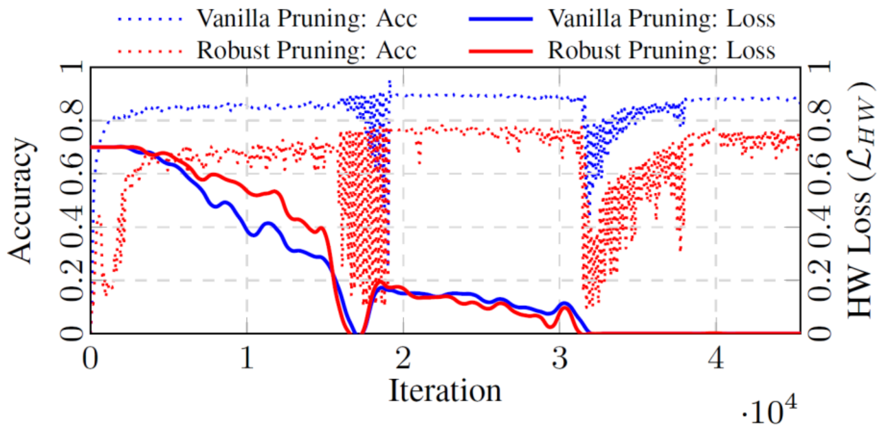
Model	Operation Reduction	Fast AT + RLPrune		Fast AT + Intrain (Our approach)	
		Acc	PGD	Acc	PGD
ResNet20	1.0	81.52	40.65	81.52	40.65
	0.70	78.89	40.39	80.63	39.27
	0.50	77.11	39.65	80.32	40.14
	0.30	66.97	33.89	72.88	34.33
ResNet56	1.0	84.03	38.45	84.03	38.45
	0.70	82.78	42.47	84.52	36.91
	0.50	81.88	41.78	84.56	36.78
	0.30	74.75	36.95	83.40	36.89

Tab: Comparsion of In-train pruning approach with RL based pruning on original images and PGD attacked images.

- Re-implemeted post-train robust pruning uses AMC [He et al. ECCV 2018] approach with KL robustness score in the reward function to make the pruning robustness-aware.

$$R_{acc+kl} = acc_{pruned} \cdot \log_{10}(\psi_{kl}(x))$$

- For 70% reduction in operations, the in-train pruning achieves an improvement of **5.91 pp** and **8.65 pp** in natural accuracy for ResNet20 and ResNet56.



COMPARISON WITH STATE-OF-THE-ART ROBUST PRUNING

Work	Baseline Model	Pretrained Model	Pruning Regularity	PGD iteration	Model Size	Acc [%]	Adv Acc [%]
Robust ADMM [Ye et al. ICCV2019]	ResNet18	✓	channel	10	0.17	73.36	43.17
Ours	**ResNet20**	×	channel	10	**0.16**	**79.67**	**43.22**
Hydra [Sehwag et al. NeurIPS 2020]	VGG-16	✓	weight	50	0.76	78.90	48.70
			channel	50	7.65	52.90	38.00
Ours	**VGG-16**	×	channel	50	**5.51**	**82.54**	**38.36**
			channel	50	0.76	73.40	30.20
ATMC (Prune) [Gui et al. NeurIPS 2019]	ResNet34	✓	weight	7	0.11	84.00	62.00
Ours	**ResNet56**	×	weight	7	0.13	82.68	**68.63**

- Different robust pruning works use different baselines, PGD parameters and adversarial training schemes. Very challenging for comparison.
- RobustADMM considers over parameterized ResNet as a baseline model and prunes it for various parameter constraints.
- Significant improvement for channel pruning configurations compared to Hydra.
- Compared to ATMC-32bit pruned configuration, we achieve 6.63pp higher robustness.

CONCLUSION AND FUTURE WORK

- This work combines **adversarial training** and **model pruning** in a joint formulation of the fundamental learning objective during training.
- Saves the effort of additional post-train pruning and eliminates the **need for a pre-trained model.**
- **Improves natural accuracy** while maintaining same level of adversarial robustness for higher compression rates as compared to **state-of-the-art** approaches.
- Robustness of in-train pruned models needs to be explored on object detection and semantic segmentation tasks.
- As future work, HW-aware robust pruning can be formulated using differentiable loss objective based on real hardware metrics like inference latency.

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. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in European Conference on Computer Vision (ECCV), 2018.
3. X. Zhou, D. Wang, and P. Kr¨ahenb¨uhl, "Objects as points," in arXiv preprint arXiv:1904.07850, 2019.
4. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in International Conference on Learning Representations (ICLR), 2014.
5. Y. He, J. Lin, Z. Liu, H. Wang, L.-J. Li, and S. Han, "AMC: AutoML for model compression and acceleration on mobile devices," in European Conference on Computer Vision (ECCV), 2018.
6. Q. Huang, K. Zhou, S. You, and U. Neumann, "Learning to prune filters in convolutional neural networks," in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 709–718, 2018.
7. APQ
8. A. Galloway, G. W. Taylor, and M. Moussa, "Attacking Binarized Neural Networks," in International Conference on Learning Representations (ICLR), 2018.
9. S. Ye, K. Xu, S. Liu, H. Cheng, J.-H. Lambrechts, H. Zhang, A. Zhou, K. Ma, Y. Wang, and X. Lin, "Adversarial robustness vs. model compression, or both?," in International Conference on Computer Vision (ICCV), October 2019.
10. S. Gui, H. Wang, H. Yang, C. Yu, Z. Wang, and J. Liu, "Model compression with adversarial robustness: A unified optimization framework," in Proceedings of the 33rd Conference on Neural Information Processing Systems, 2019.

REFERENCES

11. E. Wong, L. Rice, and J. Z. Kolter, “Fast is better than free: Revisiting adversarial training,” in International Conference on Learning Representations (ICLR), 2020.
12. A. Krizhevsky, V. Nair, and G. Hinton, “CIFAR-10 (Canadian institute for advanced research),”, 2010.
13. N. Carlini and D. A. Wagner, “Towards Evaluating the Robustness of Neural Networks,” in IEEE Symposium on Security and Privacy (SP), 2017.