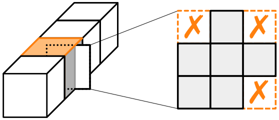
Holistic Adversarially Robust Pruning

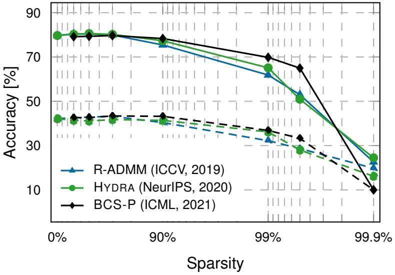
Qi Zhao and Christian Wressnegger

KASTEL Security Research Labs, Karlsruhe Institute of Technology (KIT), Germany

Background

Adversarial Robust Pruning (on VGG16 for CIFAR-10)

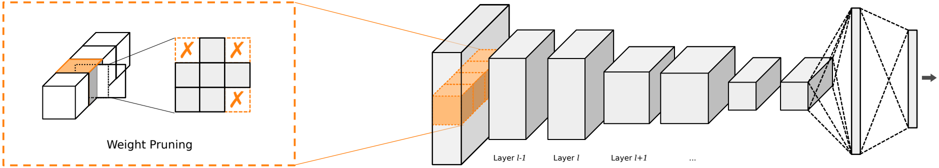




Concern 1: Model pruning inflicts robustness recession (ICML-W, 2021)

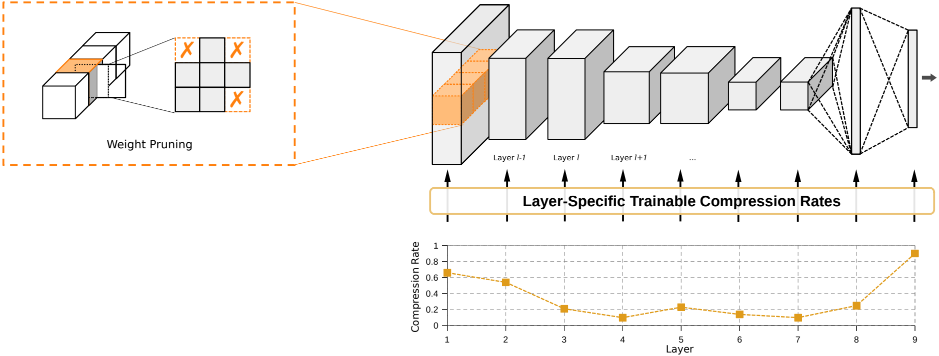
Concern 2: Adversarial pruning has only achieved moderate compression

Motivation



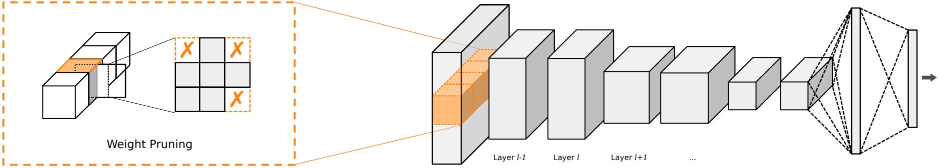
Motivation

Learning on **layer-specific compression rate**



Motivation

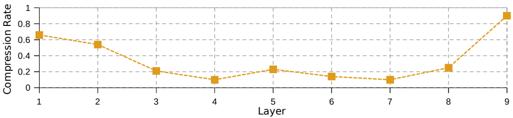
- Learning on **layer-specific compression rate**
- Learning on **prunable weight selection**



Pruning Mask with Trainable Scores

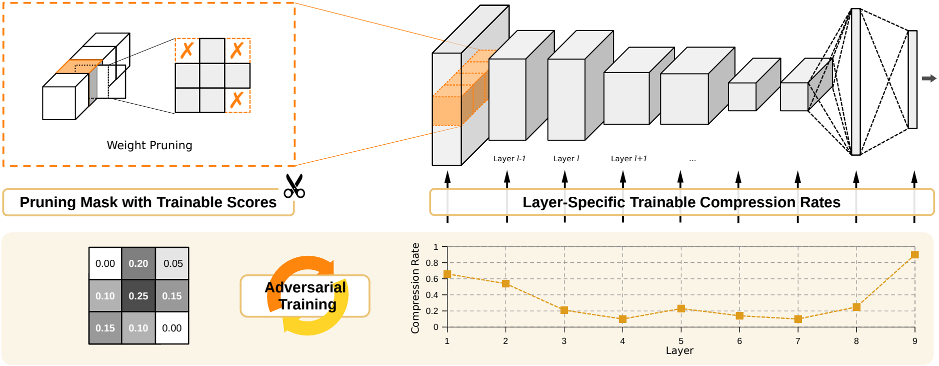
Layer-Specific Trainable Compression Rates

0.00	0.20	0.05
0.10	0.25	0.15
0.15	0.10	0.00



Motivation

- Learning on **layer-specific compression rate**
- Learning on **prunable weight selection**



HARP: Holistic Adversarially Robust Pruning

Global Compression Control for Robust Pruning

$$\min_{\boldsymbol{r},\boldsymbol{S}} \underbrace{\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}} \left[\max_{\delta} \{ \mathcal{L}_{robust}(\boldsymbol{\theta} \odot \boldsymbol{M}, \boldsymbol{x} + \delta, y) \} \right]} + \gamma \cdot \underbrace{\mathcal{L}_{hw}(\boldsymbol{\theta} \odot \boldsymbol{M}, a_t)}$$

global robust training on weight selection & layer-specific compression

global control on model compression

Global Control on Model Compression

$$\mathcal{L}_{hw}(\hat{\boldsymbol{\theta}}, a_t) := \max \left\{ \frac{\Theta_{\neq 0}}{a_t \cdot \Theta} - 1 \; , \; 0 \right\} \text{, where } \hat{\boldsymbol{\theta}}^{(l)} = \boldsymbol{\theta}^{(l)} \odot \boldsymbol{M}^{(l)}$$

HARP: Methodological Implementation

Conduction of Pruning Mask

$$\boldsymbol{M}^{(l)} := \left(\mathbb{1}_{s>P(\alpha^{(l)},\, \boldsymbol{S}^{(l)})}\right)$$

where: $\alpha^{(l)} = 1 - a^{(l)}$ and $a^{(l)} = g(r^{(l)})$ with $g : r \mapsto (1 - a_{min}) \cdot \text{sigmoid}(r^{(l)}) + a_{min}$

$P(\cdot) =$ percentile of $\alpha^{(l)}$ and selection scores $\boldsymbol{S}^{(l)}$

HARP: Methodological Implementation

Conduction of Pruning Mask

$$\boldsymbol{M}^{(l)} := \left(\mathbb{1}_{s > P(\alpha^{(l)},\, \boldsymbol{S}^{(l)})}\right)$$

where: $\alpha^{(l)} = 1 - a^{(l)}$ and $a^{(l)} = g(r^{(l)})$ with $g : r \mapsto (1 - a_{min}) \cdot \text{sigmoid}(r^{(l)}) + a_{min}$

$P(\cdot) =$ percentile of $\alpha^{(l)}$ and selection scores $\boldsymbol{S}^{(l)}$

Learning on Trainable Rates r and Scores S

Back-propagation on non-differentiable operation $\odot$ via “Straight Through Estimation” (STE)

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{S}^{(l)}} = \frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{\theta}}^{(l)}} \cdot \frac{\partial \hat{\boldsymbol{\theta}}^{(l)}}{\partial \boldsymbol{M}^{(l)}} \cdot \frac{\partial \boldsymbol{M}^{(l)}}{\partial \boldsymbol{S}^{(l)}} \overset{\text{STE!}}{=} \frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{\theta}}^{(l)}} \cdot \frac{\partial \hat{\boldsymbol{\theta}}^{(l)}}{\partial \boldsymbol{M}^{(l)}}$$

(NeurIPS, 2016)

$$\frac{\partial \mathcal{L}}{\partial r^{(l)}} = \frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{\theta}}^{(l)}} \cdot \frac{\partial \hat{\boldsymbol{\theta}}^{(l)}}{\partial \boldsymbol{M}^{(l)}} \cdot \frac{\partial \boldsymbol{M}^{(l)}}{\partial g(r^{(l)})} \cdot g'(r^{(l)}) \overset{\text{STE!}}{=} \left\langle \frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{\theta}}^{(l)}} \cdot \frac{\partial \hat{\boldsymbol{\theta}}^{(l)}}{\partial \boldsymbol{M}^{(l)}} \right\rangle \cdot g'(r^{(l)})$$

(ICML, 2020)

HARP: Ablation Study

The Importance of Learning on Rates r and Scores S

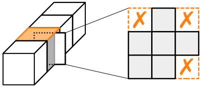
Table: Natural accuracy and PGD-10 adversarial robustness are presented left and right of the / character.

Model	Adv. Training	99 % Sparsity			99.9 % Sparsity		
		HARP-r	HARP-S	HARP	HARP-r	HARP-S	HARP
ResNet18	PGD	76.39 / 46.64	72.05 / 43.69	**80.25** / **50.36**	41.66 / 27.54	57.66 / 35.92	**63.99** / **39.39**
	TRADES	73.31 / 45.14	75.50 / 46.37	**77.78** / **50.16**	73.31 / 45.14	75.50 / 46.37	**77.78** / **50.16**
	MART	70.08 / 48.38	75.27 / 47.11	**75.88** / **50.79**	70.08 / 48.38	75.27 / 47.11	**75.88** / **50.79**
VGG16	PGD	76.17 / 46.74	65.09 / 39.80	**78.50** / **48.71**	36.76 / 28.02	50.33 / 34.03	**59.13** / **37.36**
	TRADES	72.91 / 44.52	66.75 / 41.79	**76.46** / **48.01**	41.63 / 26.95	56.08 / 31.51	**63.43** / **34.64**
	MART	71.63 / 48.64	64.37 / 41.46	**73.04** / **51.09**	37.19 / 30.68	49.51 / 36.29	**55.02** / **39.39**

- HARP-r is beneficial for **moderate compression**
- HARP-S is important in **aggressive compression**
- **Concurrent optimization on r and S** allows HARP to excel

HARP: Experimental Comparison (1)

Comparing Robust Pruning Methods



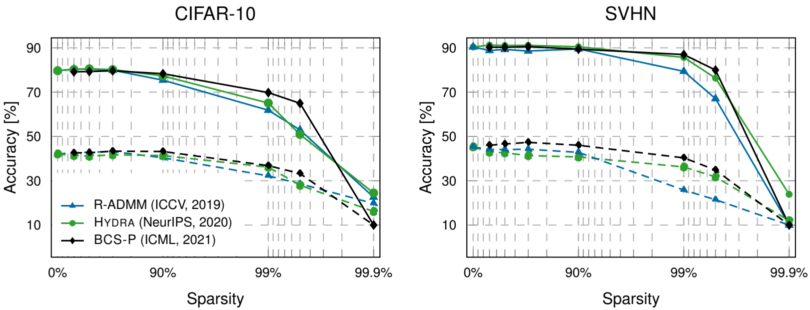


Figure: Overview of pruning weights of a VGG16 model for CIFAR-10 (left) and SVHN (right) with PGD-10 adversarial training. Solid lines show the natural accuracy of all robust pruning methods. Dashed lines represent the robustness against AUTOATTACK.

HARP: Experimental Comparison (1)

Comparing Robust Pruning Methods with HARP

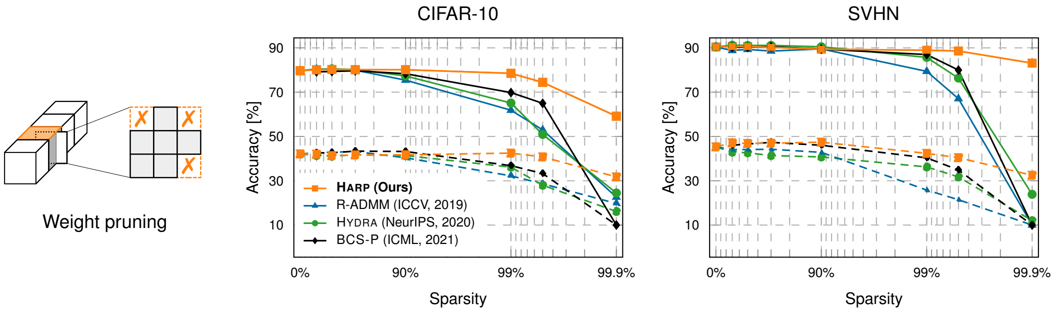


Figure: Overview of pruning weights of a VGG16 model for CIFAR-10 (left) and SVHN (right) with PGD-10 adversarial training. Solid lines show the natural accuracy of all robust pruning methods. Dashed lines represent the robustness against AUTOATTACK.

HARP: Experimental Comparison (2)

Comparing Robust Pruning Methods with HARP on ImageNet

Table: Comparing HARP with R-ADMM and HYDRA on ResNet50 models for ImageNet.

Attack	FREE-AT	90 % Sparsity			99 % Sparsity		
		R-ADMM	HYDRA	HARP	R-ADMM	HYDRA	HARP
–	60.25	35.26±0.46	49.44±0.37	**55.21**±0.36	11.41±0.32	27.00±0.66	**34.62**±0.36
PGD	32.82	14.35±0.41	23.75±0.33	**27.10**±0.41	5.15±0.17	12.23±0.19	**14.67**±0.32
$C\&W_\infty$	30.67	12.35±0.33	21.60±0.27	**24.62**±0.38	4.03±0.22	11.22±0.18	**12.42**±0.33
APGD	31.54	13.53±0.39	23.14±0.27	**25.57**±0.33	4.85±0.31	12.34±0.34	**13.47**±0.34
AA	28.79	11.01±0.25	19.88±0.29	**22.57**±0.41	3.69±0.35	10.09±0.40	**11.24**±0.43

- R-ADMM (ICCV, 2019) suffers a large robustness recession at sparsity of 90 %
- HYDRA (NeurIPS, 2020) significantly benefits from learnable masks
- HARP shows the prominence of concurrent optimization on rates r and scores S

HARP: Impact of Layer-specific Non-uniformity (1)

Table: Comparing performance of R-ADMM and HYDRA by using ERK and LAMP and by HARP on CIFAR-10. Natural accuracy and PGD-10 robustness are presented left and right of the / character.

Model	Sparsity	R-ADMM			HYDRA			HARP
		Original	w/ ERK	w/ LAMP	Original	w/ ERK	w/ LAMP	
ResNet18	99 %	71.42 / 42.31	80.36 / 48.38	**80.64** / 48.28	75.53 / 45.84	79.09 / 49.17	80.16 / 50.07	**80.25** / **50.36**
	99.9 %	26.39 / 20.62	54.51 / 33.06	57.16 / 34.05	34.55 / 26.08	55.73 / 35.09	57.07 / 35.91	**63.99** / **39.39**
VGG16	99 %	62.28 / 37.54	70.33 / 43.30	74.38 / 46.39	67.33 / 41.47	72.19 / 45.05	76.75 / 47.96	**78.58** / **48.71**
	99.9 %	21.28 / 17.46	43.35 / 29.11	48.96 / 32.39	23.41 / 20.99	50.38 / 34.32	57.93 / 36.01	**59.13** / **37.36**

- ERK (ICML, 2020) significantly improves uniform pruning methods
- LAMP (ICLR, 2021) has more promising performance than ERK
- HARP excels in robust pruning, particularly at the sparsity of 99.9 %

HARP: Impact of Layer-specific Non-uniformity (2)

Distribution of layer compression rates

- Non-uniform strategies sacrifice more on middle layers
- HARP favors higher preservation on the front and back layer

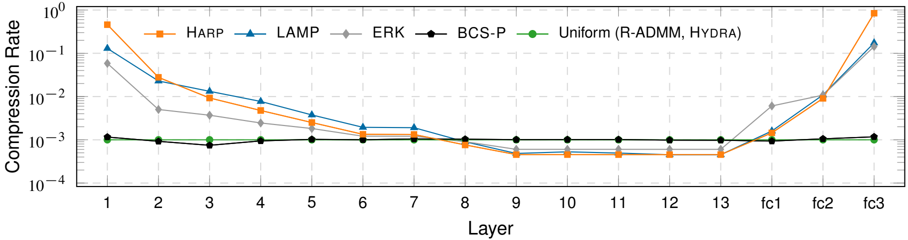


Figure: Layer-wise compression rates of 99.9 % sparsity on VGG16 for CIFAR-10

HARP: Impact of Layer-specific Non-uniformity (3)

Distribution of layer preserved parameters

- Non-uniform strategies result in a close-uniform distribution
- HARP attaches higher importance to front and back layer

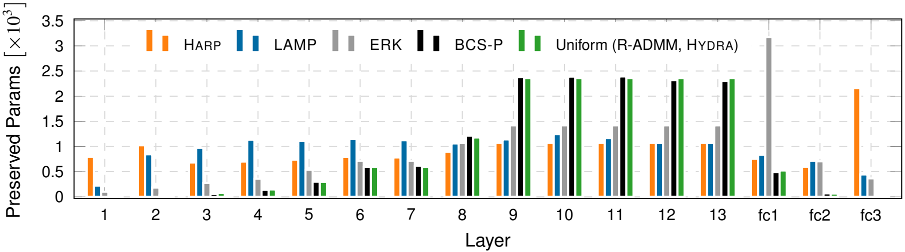


Figure: Layer-wise preserved parameters of 99.9 % sparsity on VGG16 for CIFAR-10

Thank You!

KASTEL Security Research Labs

Karlsruhe Institute of Technology (KIT)

https://intellisec.de/team/qi/

https://github.com/intellisec/harp/

https://intellisec.de/research/harp/

References

Lukas Timpl, Rahim Entezari, Hanie Sedghi, Behnam Neyshabur, and Olga Saukh. Understanding the effect of sparsity on neural networks robustness. In *Proc. of the ICML Workshop on Overparameterization: Pitfalls & Opportunities*, July 2021.

Shaokai Ye, Xue Lin, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, and Yanzhi Wang. Adversarial robustness vs. model compression, or both? In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 111–120, 2019.

Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. HYDRA: Pruning adversarially robust neural networks. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Ozan Özdenizci and Robert Legenstein. Training adversarially robust sparse networks via bayesian connectivity sampling. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 8314–8324, 2021.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2016.

Aditya Kusupati, Vivek Ramanujan, Raghav Somani, Mitchell Wortsman, Prateek Jain, Sham Kakade, and Ali Farhadi. Soft threshold weight reparameterization for learnablesparsity. In *Proc. of the International Conference on Machine Learning (ICML)*, 2020.

Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 2943–2952, 2020.

Jaeho Lee, Sejun Park, Sangwoo Mo, Sungsoo Ahn, and Jinwoo Shin. Layer-adaptive sparsity for the magnitude-based pruning. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.