Training Adversarially Robust Sparse Networks via Bayesian Connectivity Sampling

Ozan Özdenizci [1] [2] and Robert Legenstein [1]

[1] Graz University of Technology, Institute of Theoretical Computer Science, Graz, Austria

[2] Silicon Austria Labs, TU Graz - SAL Dependable Embedded Systems Lab, Graz, Austria

38th International Conference on Machine Learning (ICML) 2021

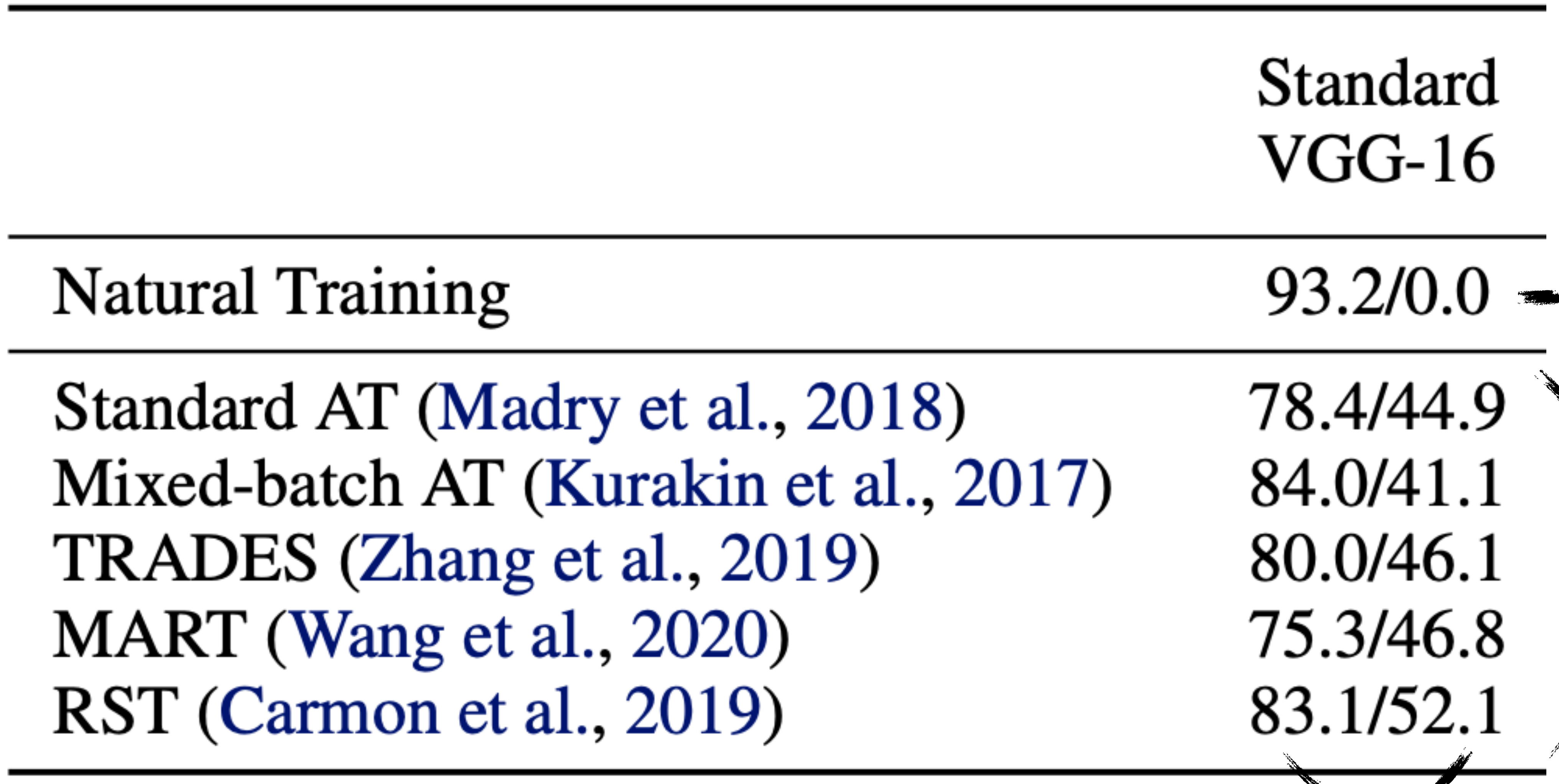
Introduction & Motivation

- Deep neural networks are susceptible to **adversarial attacks**.

Introduction & Motivation

- Deep neural networks are susceptible to **adversarial attacks**.
- Recently successful defenses rely on **robust adversarial training objectives**.

e.g., CIFAR-10 classification



	Standard VGG-16
Natural Training	93.2/0.0
Standard AT (Madry et al., 2018)	78.4/44.9
Mixed-batch AT (Kurakin et al., 2017)	84.0/41.1
TRADES (Zhang et al., 2019)	80.0/46.1
MART (Wang et al., 2020)	75.3/46.8
RST (Carmon et al., 2019)	83.1/52.1

$$\arg\min_{\boldsymbol{\theta}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[-\log p(y|x,\boldsymbol{\theta})]$$

$$\arg\min_{\boldsymbol{\theta}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\tilde{x}\in\mathcal{B}^p_\epsilon(x)} \mathcal{L}_{\text{robust}}(\boldsymbol{\theta},\tilde{x},y)\right]$$

benign acc. / robust acc. (w/PGD^{50} attacks)

Introduction & Motivation

- Deep neural networks are susceptible to **adversarial attacks**.
- Recently successful defenses rely on **robust adversarial training objectives**.
- Better robustness with increasing network width and size was observed.

⇒ Deployment of large models under resource constraints is challenging.

Introduction & Motivation

- Deep neural networks are susceptible to **adversarial attacks**.
- Recently successful defenses rely on **robust adversarial training objectives**.
- Better robustness with increasing network width and size was observed.

  ⇒ Deployment of large models under resource constraints is challenging.
- We highlight the need to consider achieving model compactness and **sparsity** simultaneously with **adversarial robustness** in DNNs.

Introduction & Motivation

- Deep neural networks are susceptible to **adversarial attacks**.
- Recently successful defenses rely on **robust adversarial training objectives**.
- Better robustness with increasing network width and size was observed.

  ⇒ Deployment of large models under resource constraints is challenging.
- We highlight the need to consider achieving model compactness and **sparsity** simultaneously with **adversarial robustness** in DNNs.
- Robustness-aware network pruning methods showed success.

  ⇒ No effective method existed for ***robust end-to-end sparse training***.

Introduction & Motivation

- Deep neural networks are susceptible to **adversarial attacks**.
- Recently successful defenses rely on **robust adversarial training objectives**.
- Better robustness with increasing network width and size was observed.

  ⇒ Deployment of large models under resource constraints is challenging.
- We highlight the need to consider achieving model compactness and **sparsity** simultaneously with **adversarial robustness** in DNNs.
- Robustness-aware network pruning methods showed success.

  ⇒ No effective method existed for ***robust end-to-end sparse training***.

  ⇒ *How can we enable learning with state-of-the-art **robust training objectives** by **end-to-end sparse training** under strict connectivity constraints?*

Robust Training by Connectivity Sampling

- Optimizing the network with a *negative log-posterior loss* which combines a sparse connectivity prior with the robust training objective.

$$p(\boldsymbol{\theta}\,|x,y) \propto \underline{p(\boldsymbol{\theta})} \cdot \underline{p(y|x,\boldsymbol{\theta})}$$

Robust Training by Connectivity Sampling

- Optimizing the network with a *negative log-posterior loss* which combines a sparse connectivity prior with the robust training objective.

$$p(\boldsymbol{\theta}\,|x, y) \propto \underline{p(\boldsymbol{\theta})} \cdot \underline{p(y|x, \boldsymbol{\theta})}$$

- During robust training we update both the **connectivity configuration** and the **weights** such that we are sampling network parameters from the posterior via *stochastic gradient Langevin dynamics.*

$$\Delta\,\boldsymbol{\theta}_k = \eta_t \Big(\underline{\nabla\Omega(\boldsymbol{\theta}_k)} + \underline{\nabla\,\mathbb{E}\left[\,\mathcal{L}_{\text{robust}}(\boldsymbol{\theta}_k, \tilde{x}, y)\right]} \Big) + \zeta_t \qquad\qquad \zeta_t \sim \mathcal{N}(0, \sigma\eta_t)$$

gradient of the
log-prior

gradient of the
data log-likelihood

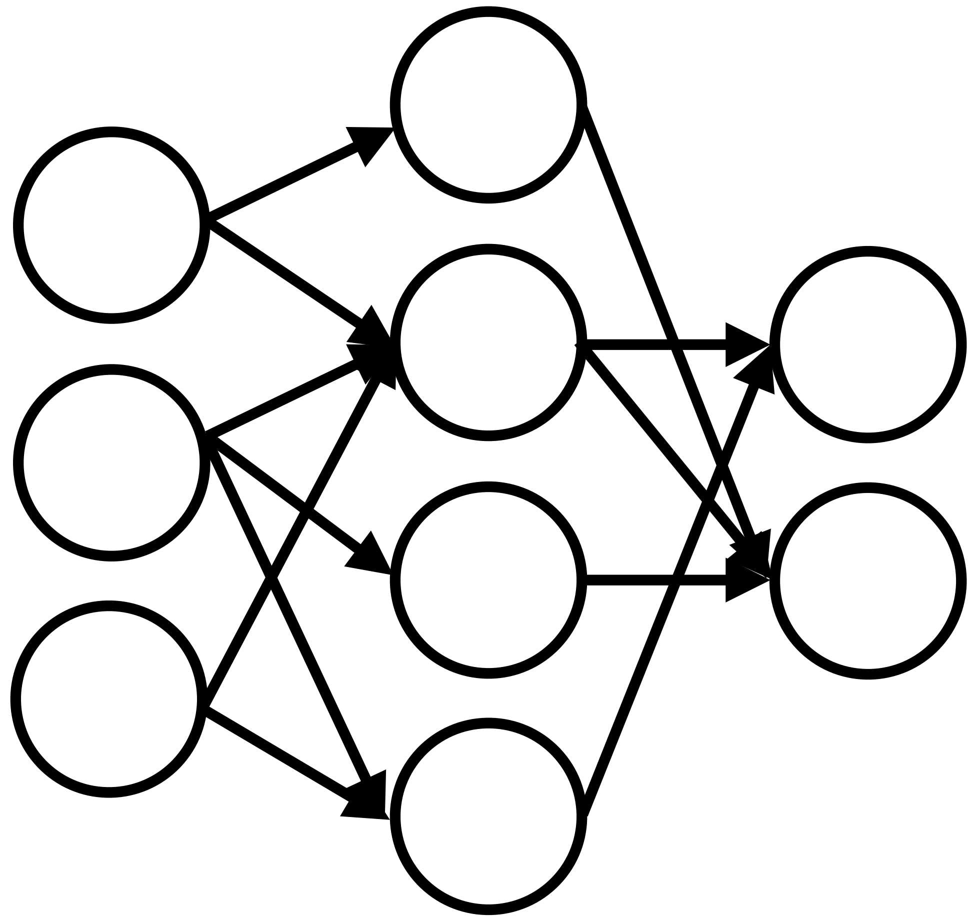
[1] Welling & Teh, "Bayesian learning via stochastic gradient Langevin dynamics", ICML 2011.

[2] Bellec et al., "Deep Rewiring: Training very sparse deep networks", ICLR 2018.

Robust Training by Connectivity Sampling

- Incorporating the sparsity prior by a weight re-parametrization trick.

$$\boldsymbol{w}_k = \underbrace{\gamma_k}_{\substack{\text{sign of the}\\ \text{connection}\\ \text{(fixed)}}} \cdot \underbrace{\max\{0, \boldsymbol{\theta}_k\}}_{\substack{\text{magnitude of}\\ \text{the connection}\\ \text{(optimized)}}} \quad s.t. \quad \gamma_k \in \{-1, 1\}$$

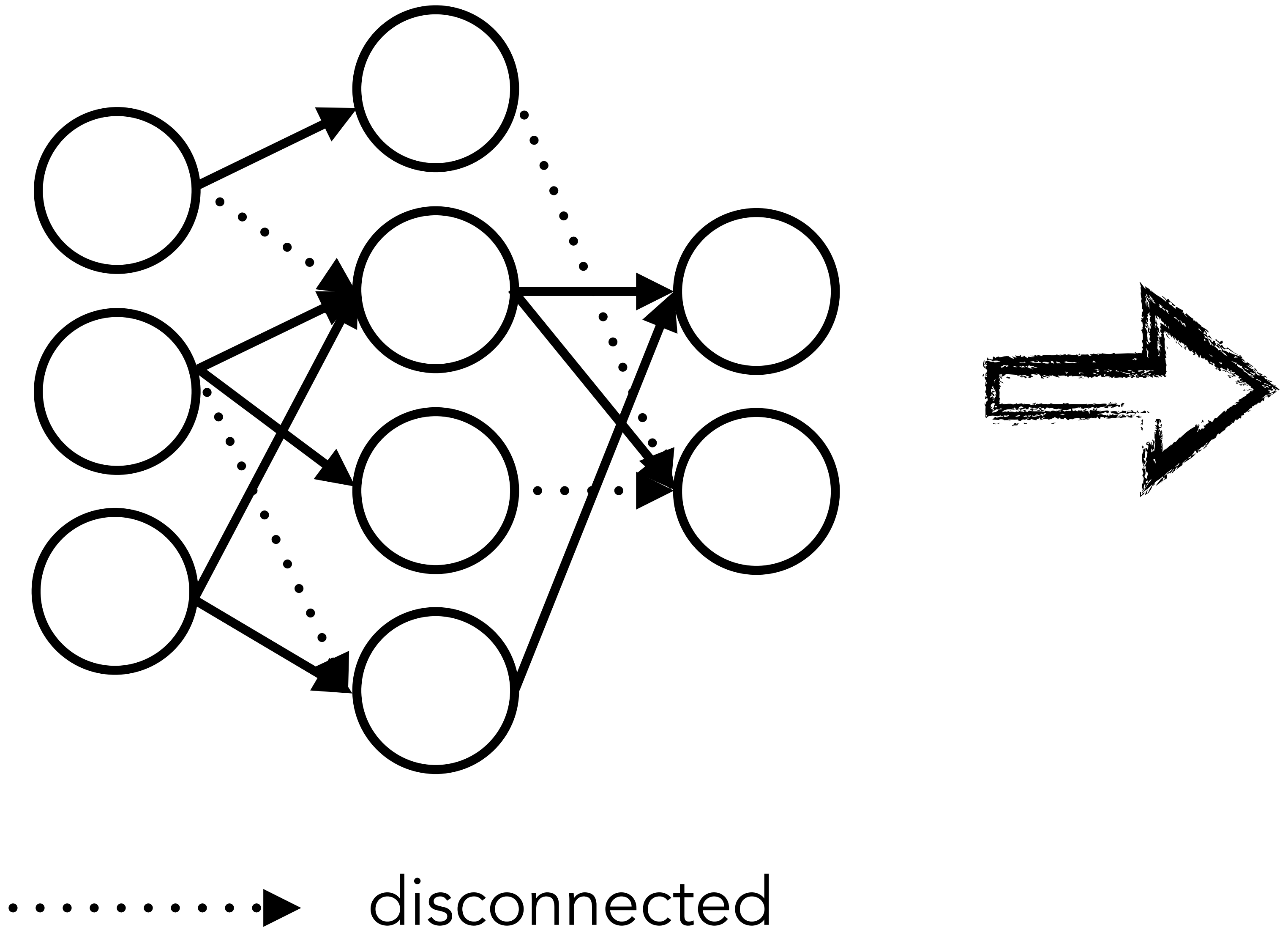


[1] Welling & Teh, “Bayesian learning via stochastic gradient Langevin dynamics”, ICML 2011.
[2] Bellec et al., “Deep Rewiring: Training very sparse deep networks”, ICLR 2018.

Robust Training by Connectivity Sampling

- Incorporating the sparsity prior by a weight re-parametrization trick.

$$\boldsymbol{w}_k = \underbrace{\gamma_k}_{\substack{\text{sign of the}\\\text{connection}\\\text{(fixed)}}} \cdot \underbrace{\max\{0, \boldsymbol{\theta}_k\}}_{\substack{\text{magnitude of}\\\text{the connection}\\\text{(optimized)}}} \quad s.t. \quad \gamma_k \in \{-1, 1\}$$



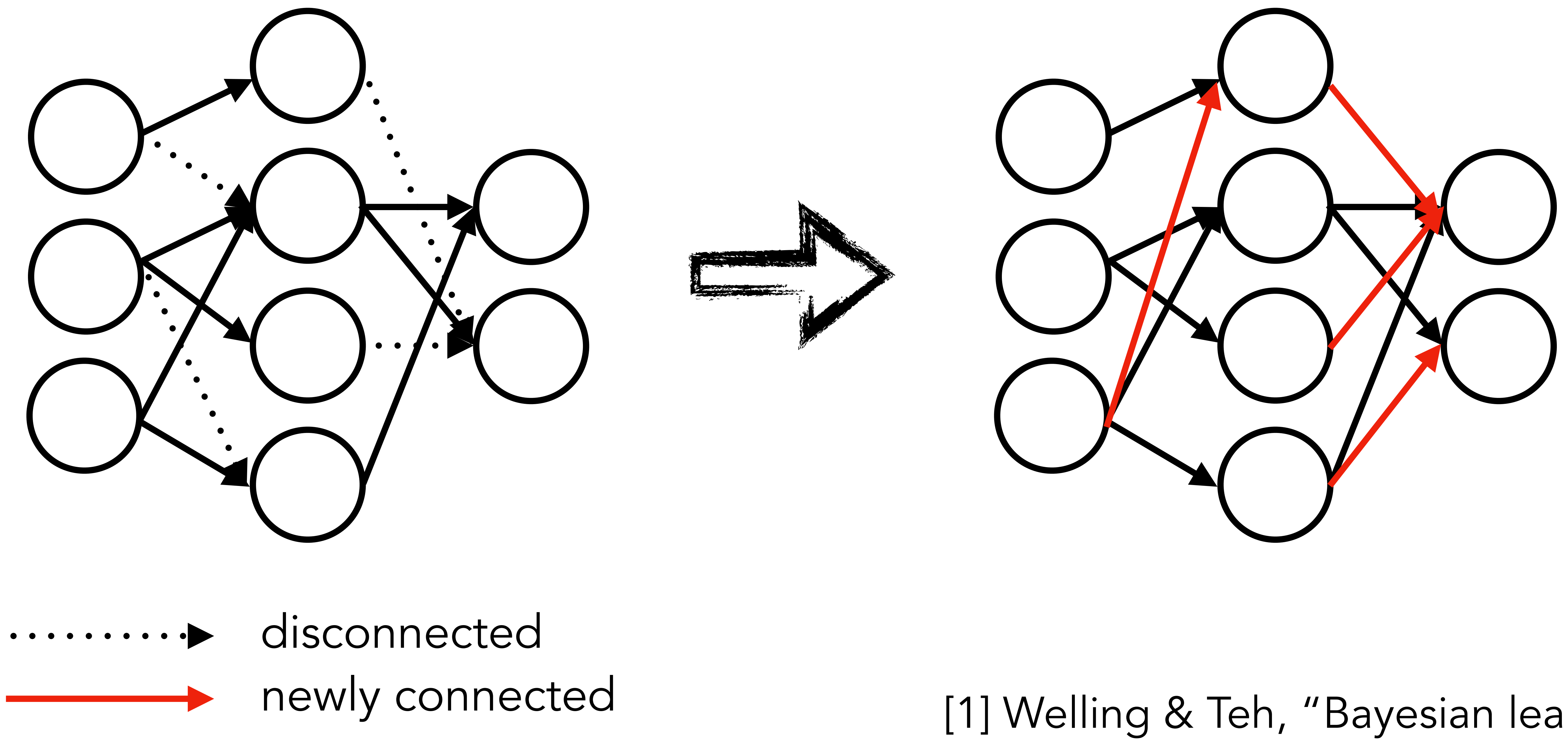
[1] Welling & Teh, “Bayesian learning via stochastic gradient Langevin dynamics”, ICML 2011.

[2] Bellec et al., “Deep Rewiring: Training very sparse deep networks”, ICLR 2018.

Robust Training by Connectivity Sampling

- Incorporating the sparsity prior by a weight re-parametrization trick.

$$\boldsymbol{w}_k = \underbrace{\gamma_k}_{\substack{\text{sign of the}\\ \text{connection}\\ \text{(fixed)}}} \cdot \underbrace{\max\{0, \boldsymbol{\theta}_k\}}_{\substack{\text{magnitude of}\\ \text{the connection}\\ \text{(optimized)}}} \quad s.t. \quad \gamma_k \in \{-1, 1\}$$



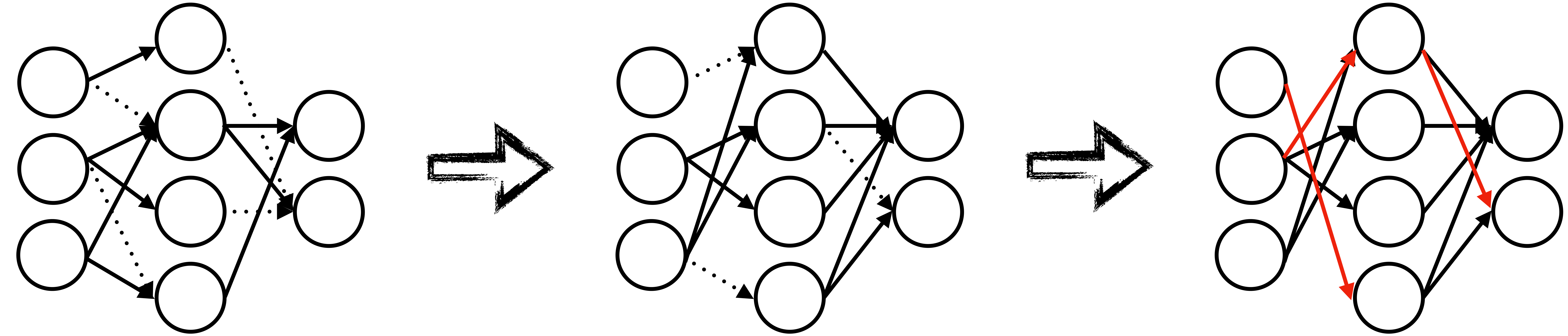
[1] Welling & Teh, "Bayesian learning via stochastic gradient Langevin dynamics", ICML 2011.

[2] Bellec et al., "Deep Rewiring: Training very sparse deep networks", ICLR 2018.

Robust Training by Connectivity Sampling

- Incorporating the sparsity prior by a weight re-parametrization trick.

$$\boldsymbol{w}_k = \underbrace{\gamma_k}_{\substack{\text{sign of the}\\ \text{connection}\\ \text{(fixed)}}} \cdot \underbrace{\max\{0, \boldsymbol{\theta}_k\}}_{\substack{\text{magnitude of}\\ \text{the connection}\\ \text{(optimized)}}} \quad s.t. \quad \gamma_k \in \{-1, 1\}$$



········▶ disconnected

⟶ newly connected

[1] Welling & Teh, "Bayesian learning via stochastic gradient Langevin dynamics", ICML 2011.

[2] Bellec et al., "Deep Rewiring: Training very sparse deep networks", ICLR 2018.

Enabling Robust End-to-End Sparse Training

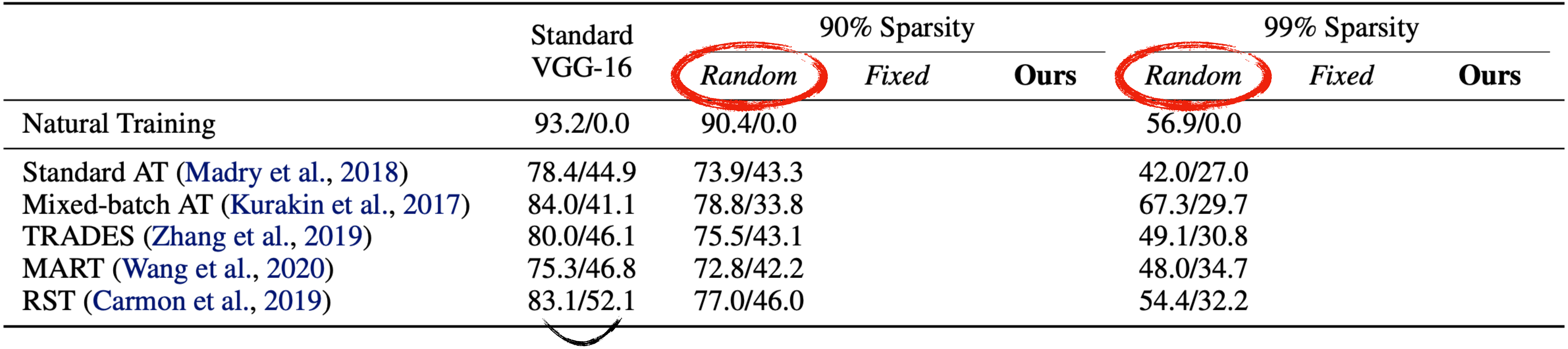
- ***Random:*** Sparse networks trained from scratch with a randomly initialized connectivity that is kept static during optimization.

Equivalent to previous "end-to-end robust and sparse training" baselines [3,4].

[3] Ye et al., "Adversarial robustness vs. model compression, or both?", ICCV 2019.

[4] Sehwag et al., "HYDRA: Pruning adversarially robust neural networks", NeurIPS 2020.

***e.g.*, CIFAR-10 classification**



	Standard VGG-16	90% Sparsity			99% Sparsity		
		Random	*Fixed*	**Ours**	*Random*	*Fixed*	**Ours**
Natural Training	93.2/0.0	90.4/0.0			56.9/0.0		
Standard AT (Madry et al., 2018)	78.4/44.9	73.9/43.3			42.0/27.0		
Mixed-batch AT (Kurakin et al., 2017)	84.0/41.1	78.8/33.8			67.3/29.7		
TRADES (Zhang et al., 2019)	80.0/46.1	75.5/43.1			49.1/30.8		
MART (Wang et al., 2020)	75.3/46.8	72.8/42.2			48.0/34.7		
RST (Carmon et al., 2019)	83.1/52.1	77.0/46.0			54.4/32.2		

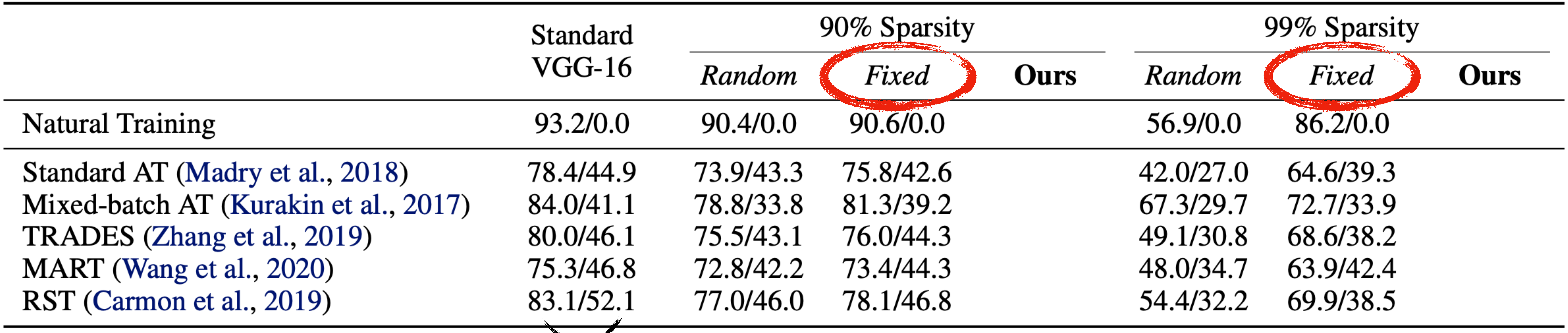
benign acc. / robust acc. (w/PGD^{50} attacks)

Enabling Robust End-to-End Sparse Training

- ***Random:*** Sparse networks trained from scratch with a randomly initialized connectivity that is kept static during optimization.

- ***Fixed:*** Sparse networks trained from scratch with a fixed and static connectivity, *where the layer-wise #connections are chosen equal to the #connections that our method was found to converge at.*

***e.g.*, CIFAR-10 classification**



	Standard VGG-16	90% Sparsity			99% Sparsity		
		Random	*Fixed*	**Ours**	*Random*	*Fixed*	**Ours**
Natural Training	93.2/0.0	90.4/0.0	90.6/0.0		56.9/0.0	86.2/0.0	
Standard AT (Madry et al., 2018)	78.4/44.9	73.9/43.3	75.8/42.6		42.0/27.0	64.6/39.3	
Mixed-batch AT (Kurakin et al., 2017)	84.0/41.1	78.8/33.8	81.3/39.2		67.3/29.7	72.7/33.9	
TRADES (Zhang et al., 2019)	80.0/46.1	75.5/43.1	76.0/44.3		49.1/30.8	68.6/38.2	
MART (Wang et al., 2020)	75.3/46.8	72.8/42.2	73.4/44.3		48.0/34.7	63.9/42.4	
RST (Carmon et al., 2019)	83.1/52.1	77.0/46.0	78.1/46.8		54.4/32.2	69.9/38.5	

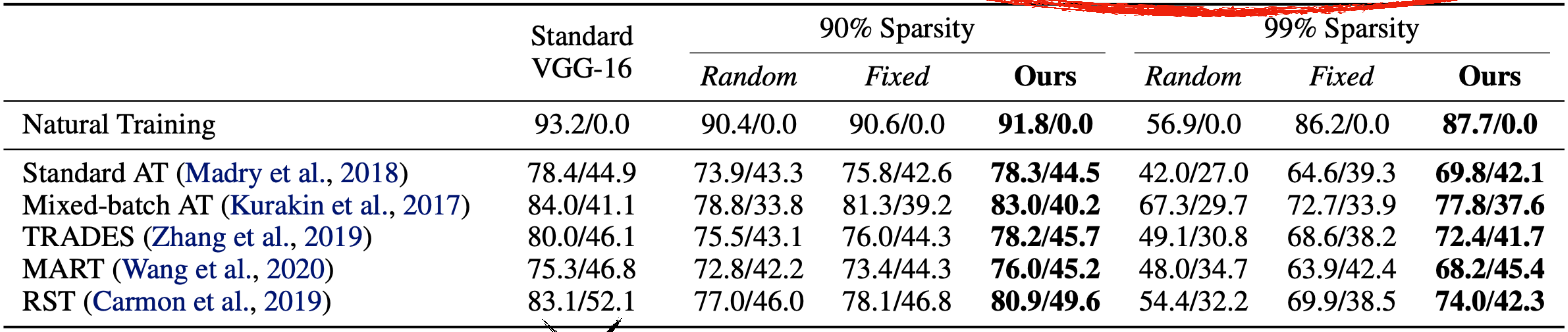
benign acc. / robust acc. (w/PGD^{50} attacks)

Enabling Robust End-to-End Sparse Training

- ***Random:*** Sparse networks trained from scratch with a randomly initialized connectivity that is kept static during optimization.
- ***Fixed:*** Sparse networks trained from scratch with a fixed and static connectivity, *where the layer-wise #connections are chosen equal to the #connections that our method was found to converge at.*

Rearranging connectivities enables robust end-to-end sparse training.

e.g., CIFAR-10 classification



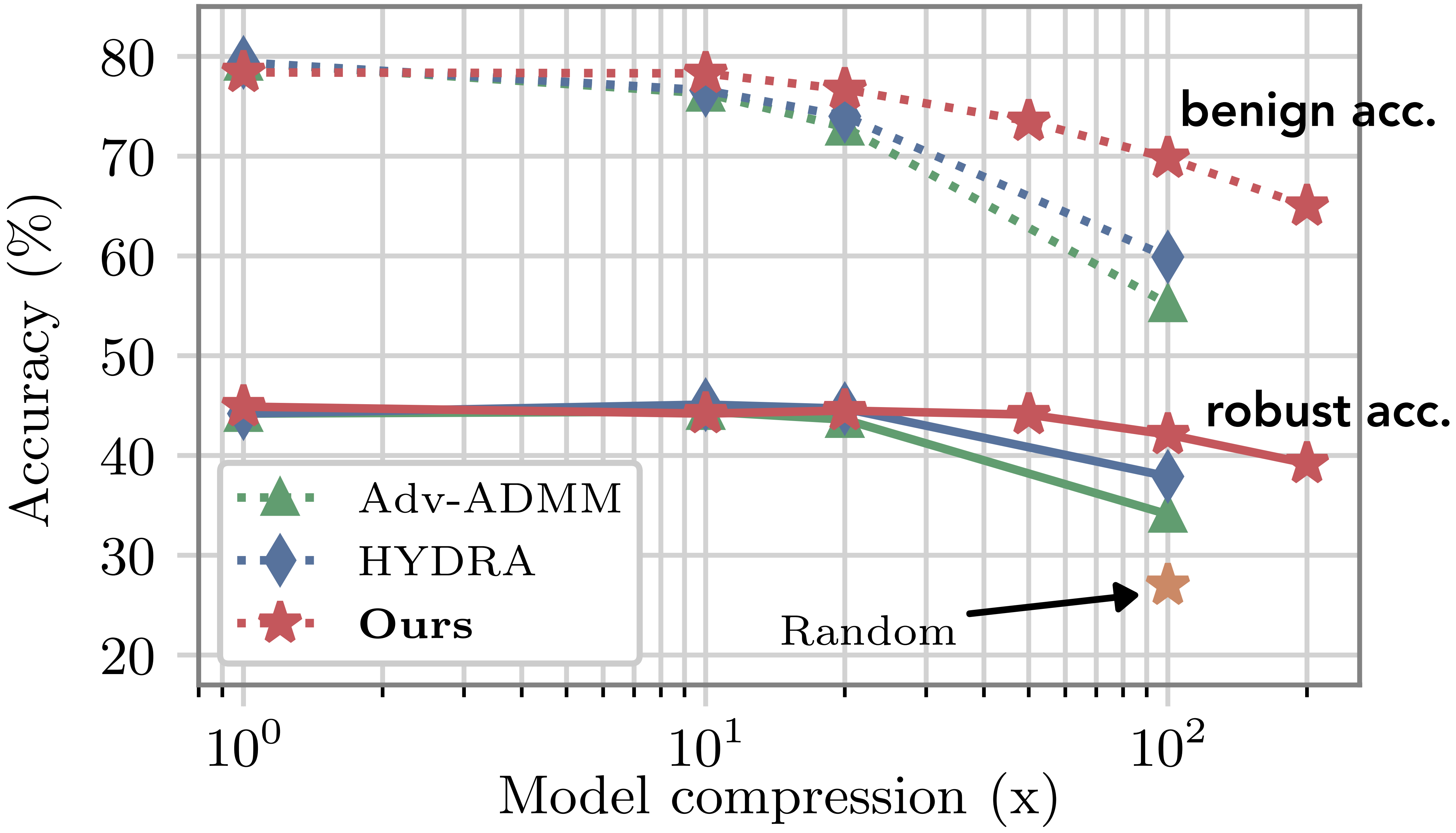
	Standard VGG-16	90% Sparsity			99% Sparsity		
		Random	*Fixed*	**Ours**	*Random*	*Fixed*	**Ours**
Natural Training	93.2/0.0	90.4/0.0	90.6/0.0	**91.8/0.0**	56.9/0.0	86.2/0.0	**87.7/0.0**
Standard AT (Madry et al., 2018)	78.4/44.9	73.9/43.3	75.8/42.6	**78.3/44.5**	42.0/27.0	64.6/39.3	**69.8/42.1**
Mixed-batch AT (Kurakin et al., 2017)	84.0/41.1	78.8/33.8	81.3/39.2	**83.0/40.2**	67.3/29.7	72.7/33.9	**77.8/37.6**
TRADES (Zhang et al., 2019)	80.0/46.1	75.5/43.1	76.0/44.3	**78.2/45.7**	49.1/30.8	68.6/38.2	**72.4/41.7**
MART (Wang et al., 2020)	75.3/46.8	72.8/42.2	73.4/44.3	**76.0/45.2**	48.0/34.7	63.9/42.4	**68.2/45.4**
RST (Carmon et al., 2019)	83.1/52.1	77.0/46.0	78.1/46.8	**80.9/49.6**	54.4/32.2	69.9/38.5	**74.0/42.3**

benign acc. / robust acc. (w/PGD^{50} attacks)

Comparisons with Robustness-Aware Pruning

- Outperforming recent methods for combining **standard AT** and model sparsity.

CIFAR-10 classification with VGG-16



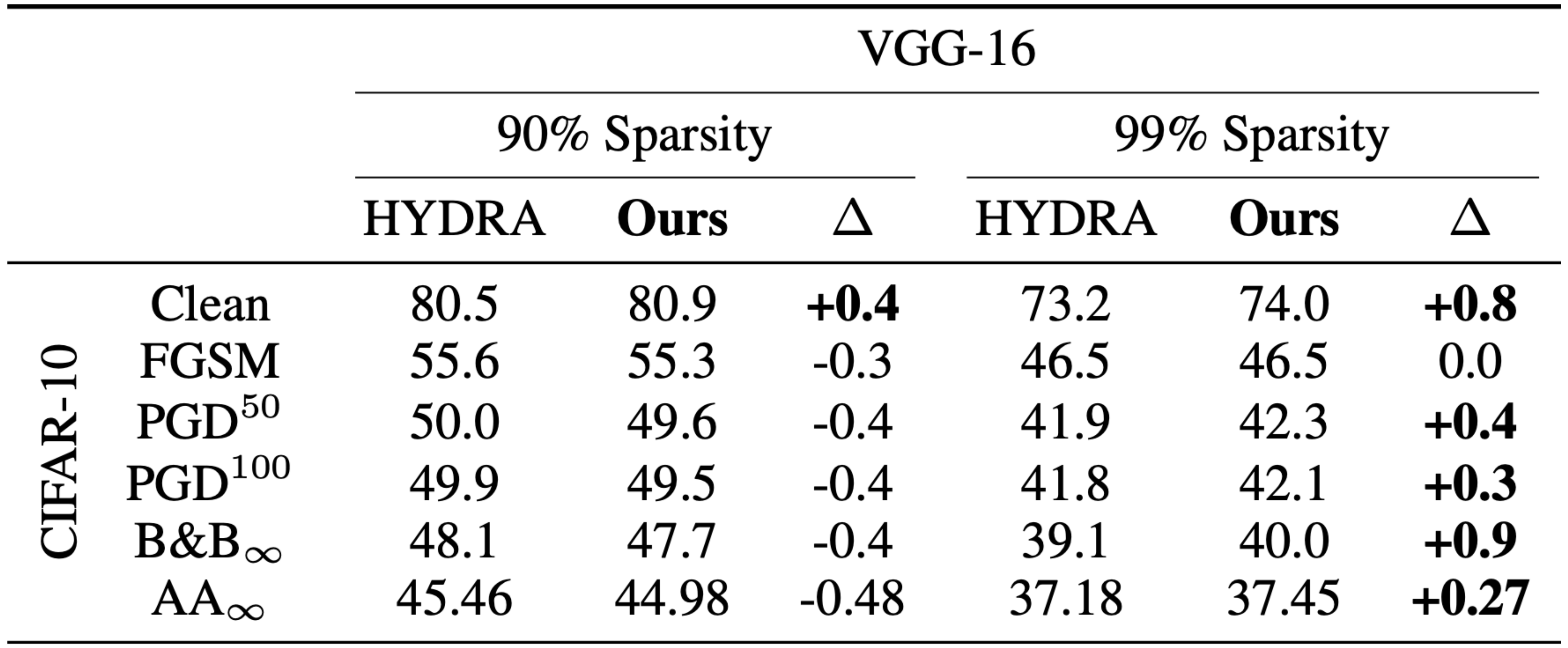
[3] Ye et al., “Adversarial robustness vs. model compression, or both?”, ICCV 2019.

[4] Sehwag et al., “HYDRA: Pruning adversarially robust neural networks”, NeurIPS 2020.

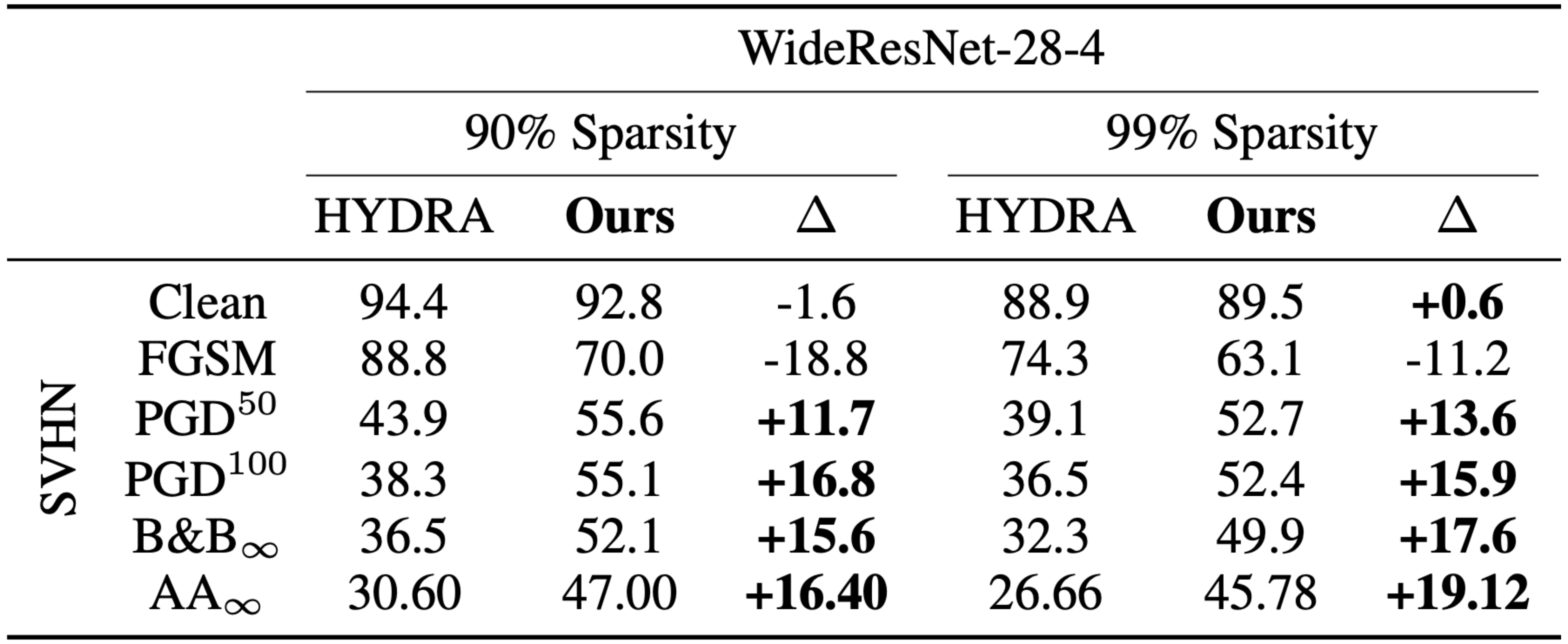
Comparisons with Robustness-Aware Pruning

- Outperforming recent methods for combining **standard AT** and model sparsity.

- State-of-the-art performance against pruning methods based on robust pre-training of densely connected networks under different robust training objectives (e.g., **TRADES**, **RST).**



		VGG-16					
		90% Sparsity			99% Sparsity		
		HYDRA	**Ours**	Δ	HYDRA	**Ours**	Δ
CIFAR-10	Clean	80.5	80.9	**+0.4**	73.2	74.0	**+0.8**
	FGSM	55.6	55.3	-0.3	46.5	46.5	0.0
	PGD^{50}	50.0	49.6	-0.4	41.9	42.3	**+0.4**
	PGD^{100}	49.9	49.5	-0.4	41.8	42.1	**+0.3**
	$\text{B\&B}_\infty$	48.1	47.7	-0.4	39.1	40.0	**+0.9**
	AA_∞	45.46	44.98	-0.48	37.18	37.45	**+0.27**



		WideResNet-28-4					
		90% Sparsity			99% Sparsity		
		HYDRA	**Ours**	Δ	HYDRA	**Ours**	Δ
SVHN	Clean	94.4	92.8	-1.6	88.9	89.5	**+0.6**
	FGSM	88.8	70.0	-18.8	74.3	63.1	-11.2
	PGD^{50}	43.9	55.6	**+11.7**	39.1	52.7	**+13.6**
	PGD^{100}	38.3	55.1	**+16.8**	36.5	52.4	**+15.9**
	$\text{B\&B}_\infty$	36.5	52.1	**+15.6**	32.3	49.9	**+17.6**
	AA_∞	30.60	47.00	**+16.40**	26.66	45.78	**+19.12**

[3] Ye et al., "Adversarial robustness vs. model compression, or both?", ICCV 2019.

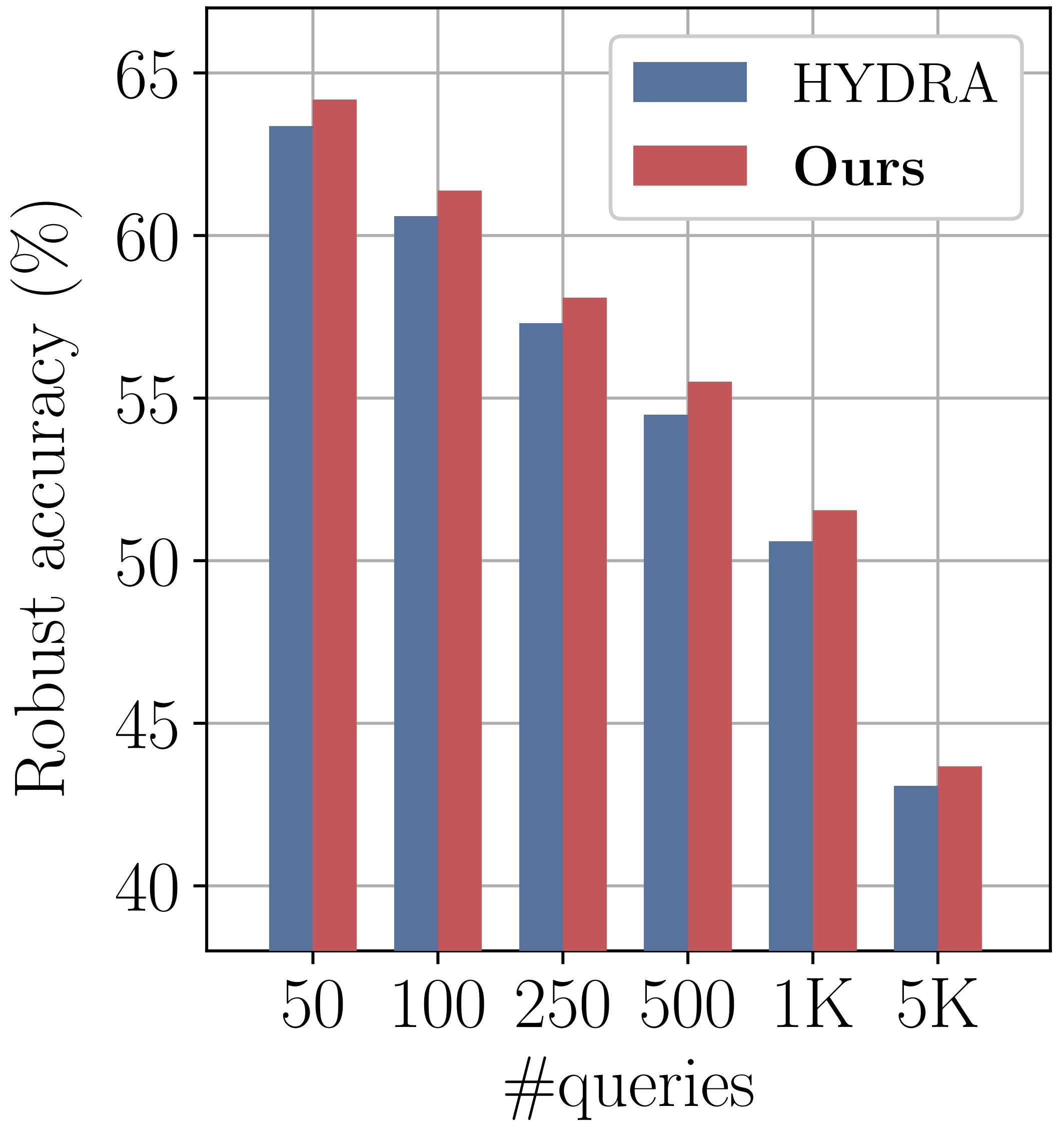
[4] Sehwag et al., "HYDRA: Pruning adversarially robust neural networks", NeurIPS 2020.

Comparisons with Robustness-Aware Pruning

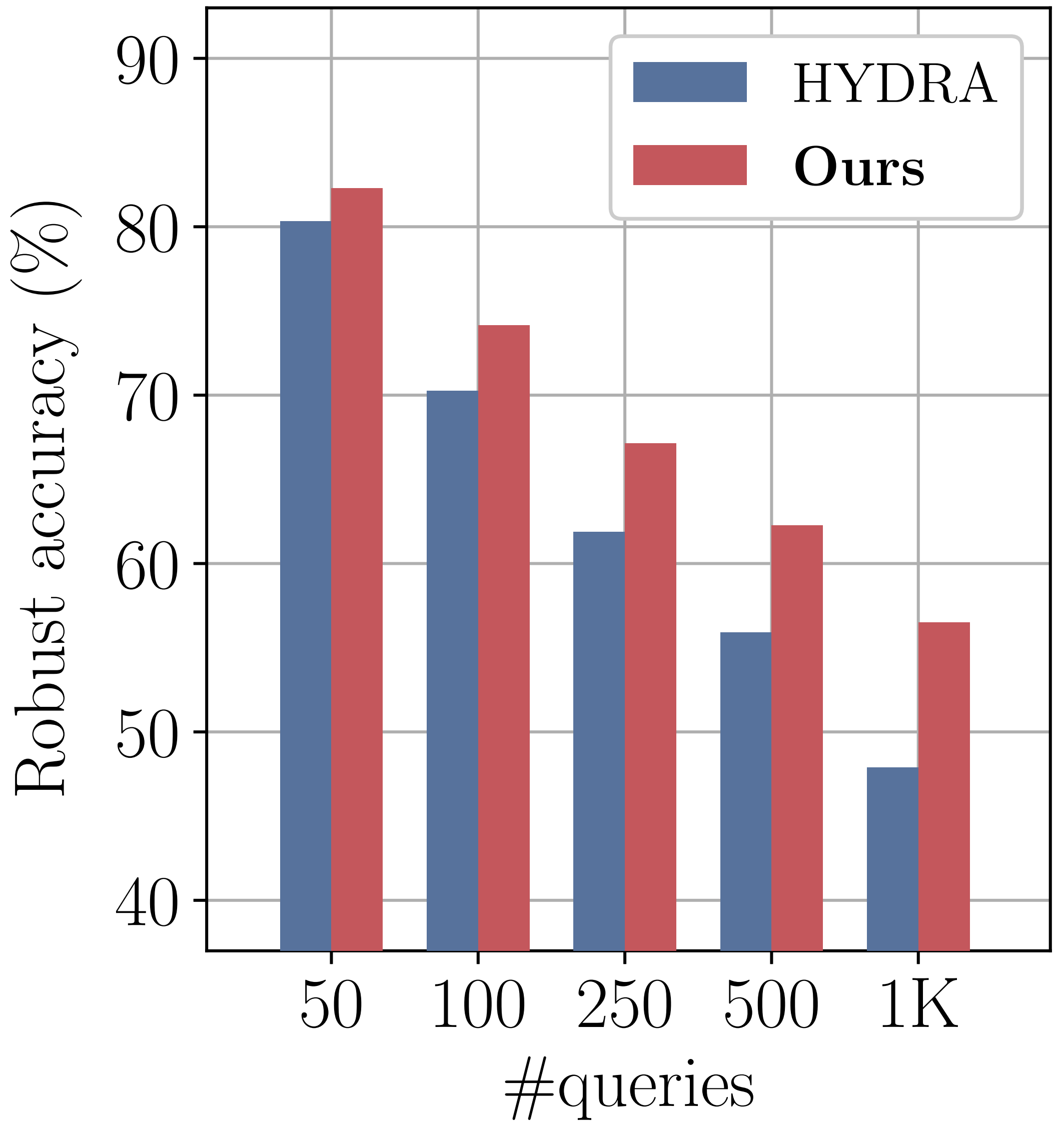
- Outperforming recent methods for combining **standard AT** and model sparsity.

- State-of-the-art performance against pruning methods based on robust pre-training of densely connected networks under different robust training objectives (e.g., **TRADES**, **RST).**

- Enabling robust and strictly-sparse training on-hardware and shows robustness under **query-based black box attacks**.



CIFAR-10 classification with VGG-16 at 99% sparsity



SVHN classification with WideResNet-28-4 at 99% sparsity

[3] Ye et al., "Adversarial robustness vs. model compression, or both?", ICCV 2019.
[4] Sehwag et al., "HYDRA: Pruning adversarially robust neural networks", NeurIPS 2020.

Thank you for your attention!

Code: https://github.com/IGITUGraz/SparseAdversarialTraining

TU Graz

SAL
SILICON AUSTRIA LABS

FWF
Der Wissenschaftsfonds.

Acknowledgements: This work has been supported by the "University SAL Labs" initiative of Silicon Austria Labs (SAL) and its Austrian partner universities for applied fundamental research for electronic based systems. This work is also partially supported by the Austrian Science Fund (FWF) within the ERA-NET CHIST-ERA programme (project SMALL, project number I 4670-N).

Training Adversarially Robust Sparse Networks via Bayesian Connectivity Sampling

Ozan Özdenizci [1 2] **Robert Legenstein** [1]

Abstract

Deep neural networks have been shown to be susceptible to adversarial attacks. This lack of adversarial robustness is even more pronounced when models are compressed in order to meet hardware limitations. Hence, if adversarial robustness is an issue, training of sparsely connected networks necessitates considering *adversarially robust sparse learning*. Motivated by the efficient and stable computational function of the brain in the presence of a highly dynamic synaptic connectivity structure, we propose an intrinsically sparse rewiring approach to train neural networks with state-of-the-art robust learning objectives under high sparsity. Importantly, in contrast to previously proposed pruning techniques, our approach satisfies global connectivity constraints throughout robust optimization, i.e., it does not require dense pre-training followed by pruning. Based on a Bayesian posterior sampling principle, a network rewiring process simultaneously learns the sparse connectivity structure and the robustness-accuracy trade-off based on the adversarial learning objective. Although our networks are sparsely connected throughout the whole training process, our experimental benchmark evaluations show that their performance is superior to recently proposed robustness-aware network pruning methods which start from densely connected networks.

1. Introduction

Despite their widely-acknowledged success and deployment in various application fields, deep neural networks (DNNs) are known to be highly susceptible to intentionally crafted adversarial examples that cause incorrect decision making. Seminal work by (Szegedy et al., 2013) showed that such adversarial examples can be created via perturbations that are hardly perceptible to humans, which exposed important weaknesses of standard deep learning algorithms. Numerous studies explored adversarial defense methods to such threats. Notably successful approaches rely on harnessing adversarial examples during model training (Goodfellow et al., 2015; Madry et al., 2018), and its immediate extensions with robust training losses using regularization schemes to diminish the generalization gap based on an inherent robustness-accuracy trade-off (Tsipras et al., 2019; Zhang et al., 2019; Wang et al., 2020).

Recent work further suggests better robustness with increasing network width and complexity (Madry et al., 2018; Nakkiran, 2019; Wu et al., 2020). Deployment of such large models, however, is challenging in resource-constrained settings. Thus, under consideration of memory and computational demand concerns, this highlights a need to consider achieving model compactness and sparsity simultaneously with adversarial robustness in DNNs.

There has been a growing interest in tackling the problem of achieving robustness against adversarial attacks with very sparsely connected neural networks (cf. Section 2). Success was so far demonstrated by robustness-aware pruning of adversarially trained dense networks (Sehwag et al., 2019; 2020). Importantly these studies only considered naive "end-to-end sparse learning" baseline comparisons with a random and static sparse network initialization. Subsequently, these intrinsically sparse models were found to yield inferior robustness than compressed models obtained with robustness-aware pruning methods. However pruning an adversarially trained DNN does not allow robust training under strict sparsity constraints. To date, no effective method existed for robust end-to-end sparse training to meet such limitations, where the challenge is to enable sparse network connections to rearrange during training such that a well-performing robust and sparse model can be configured.

In this paper we present a method for end-to-end sparse training of neural networks with robust adversarial training objectives. Our approach is motivated by the dynamic synaptic connectivity structure in the brain, which maintains its stable computational function in the presence of an under-

[1] Graz University of Technology, Institute of Theoretical Computer Science, Graz, Austria [2] Silicon Austria Labs, TU Graz - SAL Dependable Embedded Systems Lab, Graz, Austria. Correspondence to: Ozan Özdenizci <ozan.ozdenizci@igi.tugraz.at>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).