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

Introduction & Motivation

- Our work focuses on the interaction of two key challenges in deep learning: achieving **model compactness and sparsity** simultaneously with **adversarial robustness**.
- Robustness aware network pruning methods showed recent success in this domain. Nevertheless, no effective method existed for robust end-to-end sparse training.
- Motivating question: 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 via Bayesian Connectivity Sampling

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

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

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

$$\Delta\,\boldsymbol{\theta}_k = \eta_t\Big(\underbrace{\nabla\Omega(\boldsymbol{\theta}_k)}_{\text{gradient of the log-prior}} + \underbrace{\nabla\,\mathbb{E}\left[\mathcal{L}_{\text{robust}}(\boldsymbol{\theta}_k,\tilde{x},y)\right]}_{\text{gradient of the data log-likelihood}}\Big) + \zeta_t \qquad \zeta_t \sim \mathcal{N}(0,\sigma\eta_t)$$

- Incorporating the sparsity prior by a weight re-parametrization trick: $\boldsymbol{w}_k = \gamma_k \cdot \max\{0,\boldsymbol{\theta}_k\}$ $\;s.t.\quad \gamma_k \in \{-1,1\}$

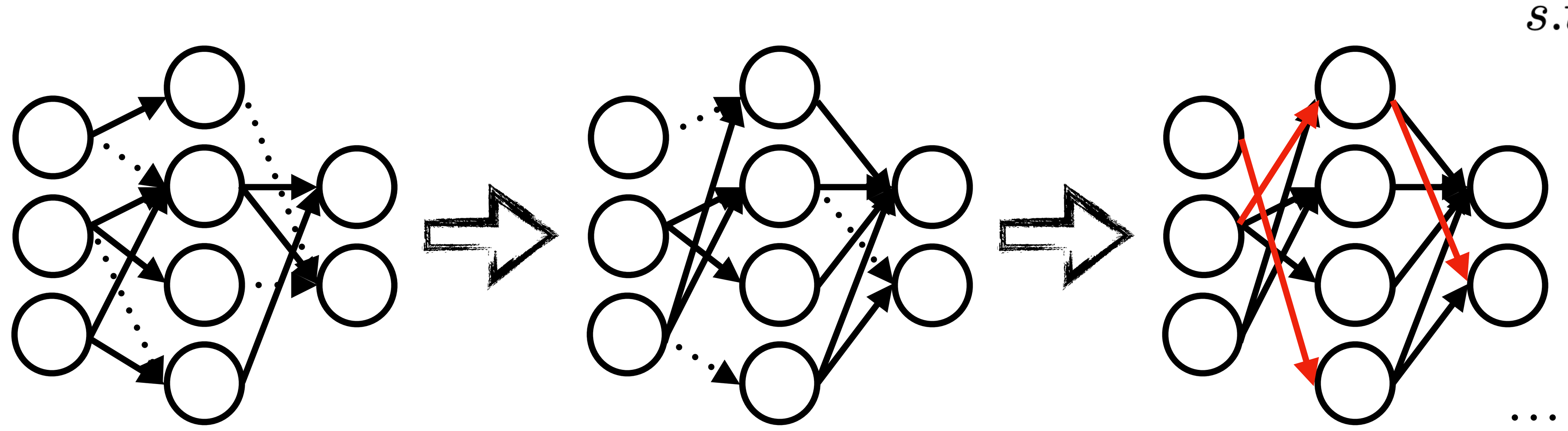

Figure 1: Illustration of the changing connectivity configurations throughout iterations.

Experimental Results

- *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**

- *State-of-the-art performance against pruning methods based on robust pre-training of densely connected networks.*

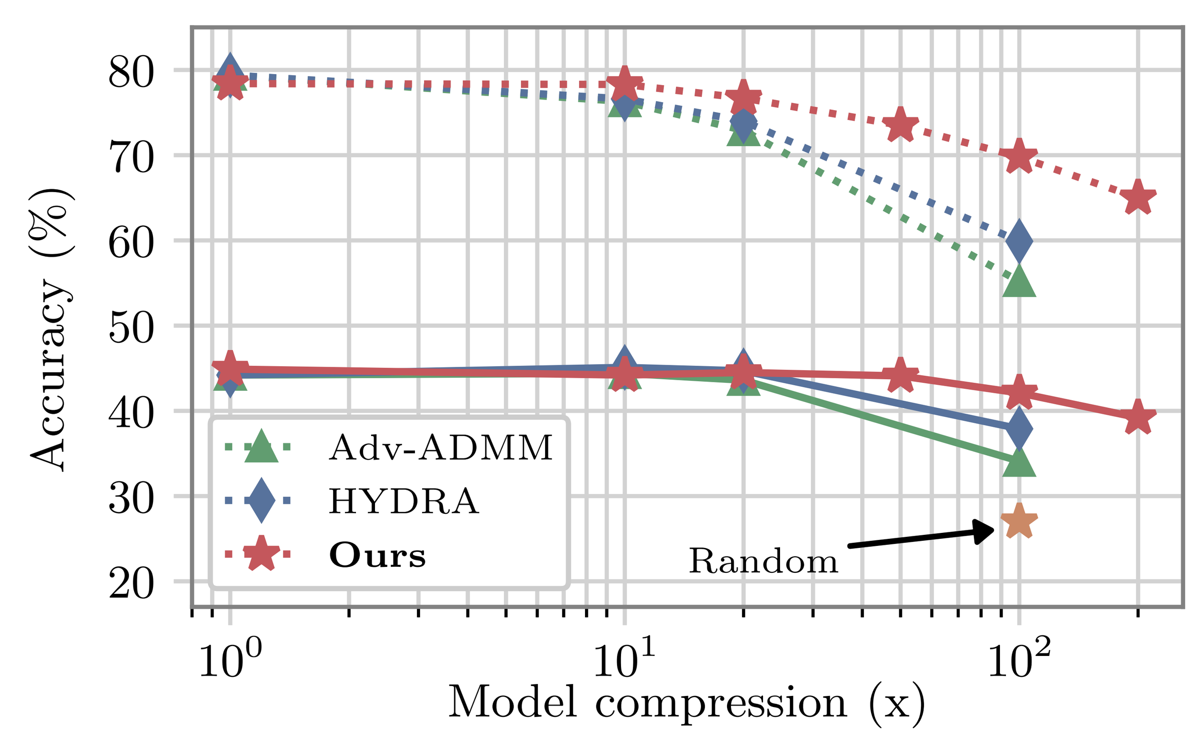

Figure 2: Standard AT on CIFAR-10 with varying VGG-16 model compression. Dotted lines show clean, and solid lines show robust accuracies.

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**
	B&B$_\infty$	48.1	47.7	-0.4	39.1	40.0	**+0.9**
	AA$_\infty$	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**
	B&B$_\infty$	36.5	52.1	**+15.6**	32.3	49.9	**+17.6**
	AA$_\infty$	30.60	47.00	**+16.40**	26.66	45.78	**+19.12**

References

[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.
[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.

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).