ICML
International Conference
On Machine Learning

Removing Batch Normalization Boosts Adversarial Training

Haotao Wang[1] Aston Zhang[2] Shuai Zheng[2] Xingjian Shi[2] Mu Li[2] Zhangyang Wang[1]

[1]University of Texas at Austin, [2]Amazon Web Services

TEXAS
The University of Texas at Austin

amazon

Overview

Method	Clean accuracy	PGD robustness
Normal training (with BN)	76.06%	0%
Adversarial training (with BN)	59.28%	13.57%
NoFrost (without normalizer)	74.06%	22.45%

→ vulnerable to adversarial attacks

Overview

Method	Clean accuracy	PGD robustness
Normal training (with BN)	76.06%	0%
Adversarial training (with BN)	59.28%	13.57%
NoFrost (without normalizer)	74.06%	22.45%

→ vulnerable to adversarial attacks

→ improve robustness but significantly decrease clean accuracy

TEXAS The University of Texas at Austin amazon

Overview

Method	Clean accuracy	PGD robustness
Normal training (with BN)	76.06%	0%
Adversarial training (with BN)	59.28%	13.57%
NoFrost (without normalizer)	74.06%	22.45%

→ vulnerable to adversarial attacks

→ improve robustness but significantly decrease clean accuracy

→ high adversarial robustness almost without sacrificing clean accuracy

TEXAS The University of Texas at Austin amazon

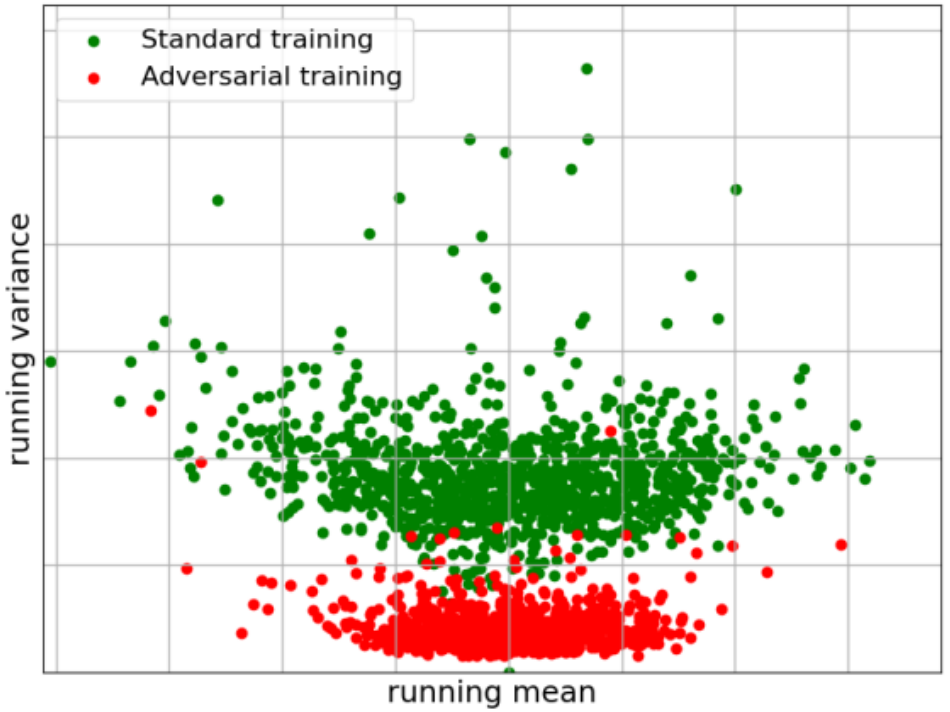
Intuition

$$\min_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}} (1-\lambda)\mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}),y) + \lambda\mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}^*),y),$$

Clean sample

Adversarial sample

Batch normalization (BN) struggles to fit the mixture distribution of clean and adversarial training samples [1].



The channel-wise BN statistics obtained by standard training and adversarial training, respectively. Each dot represents the running mean and variance of a channel in the BN layer.

TEXAS The University of Texas at Austin amazon

[1] Intriguing properties of adversarial training. In ICLR 2020.

Method

Our solution: normalizer-free adversarial training (NoFrost) uses normalizer-free (NF) networks [2] as f_θ.

NoFrost:

$$\min_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}} (1-\lambda)\mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}),y) + \lambda\mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}^*),y),$$

Clean sample

Adversarial sample

We further combine other robust data augmentation methods into NoFrost to achieve comprehensive robustness against multiple distribution shifts:

NoFrost*:

$$\min_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim\mathcal{D}} (\mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}),y)+ \mathcal{L}(f_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}),y) + \mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}^*),y))/3,$$

DeepAugment [3], Texture-debiased augmentation [4], etc.

TEXAS The University of Texas at Austin amazon

[2] Characterizing signal propagation to close the performance gap in unnormalized ResNets. In ICLR, 2021.
[3] The many faces of robustness: A critical analysis of out-of-distribution generalization. In ICCV, 2021.
[4] The origins and prevalence of texture bias in convolutional neural networks. In NeurIPS, 2020.

Results

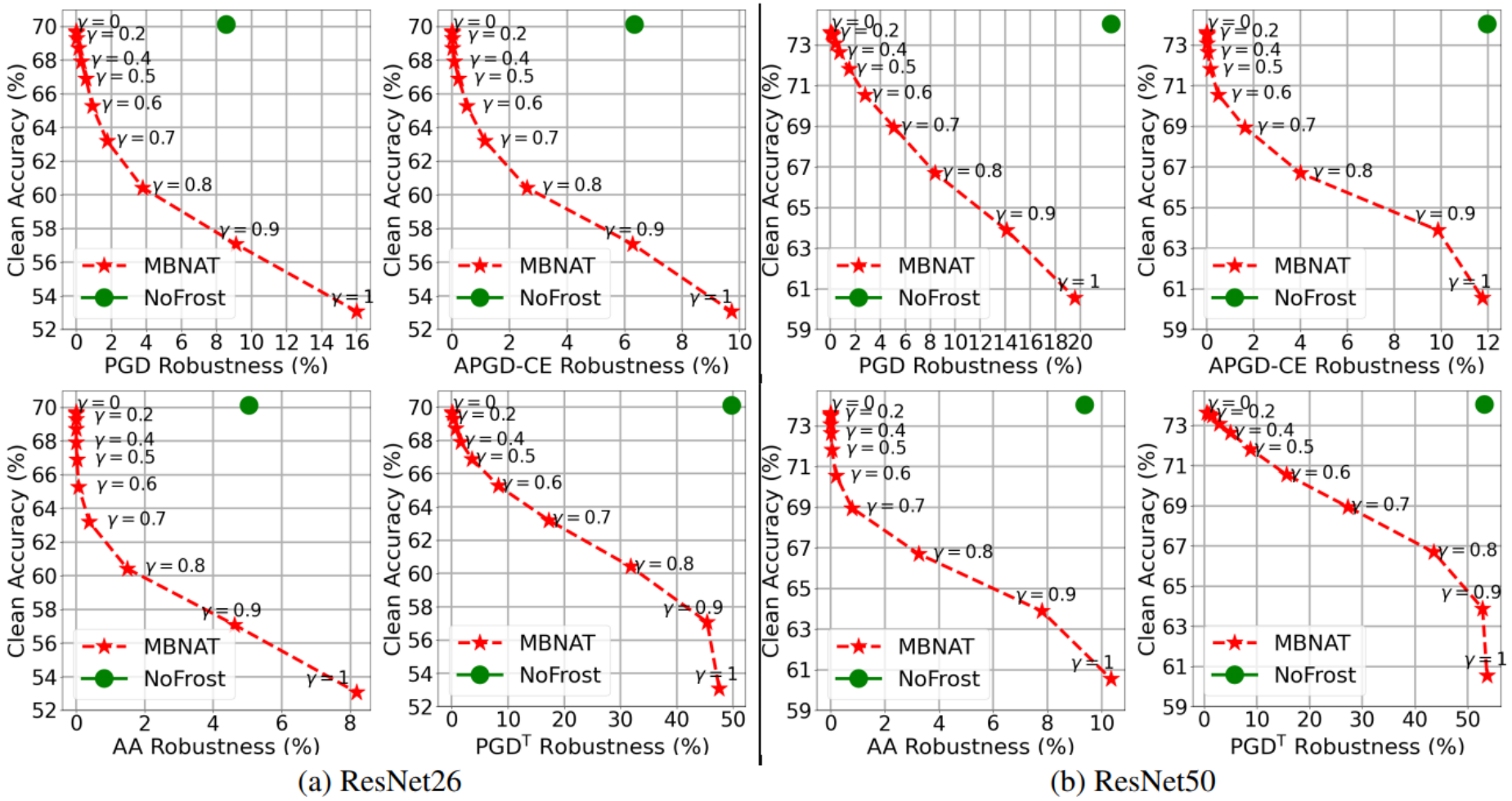


Figure 2. Trade-off between robustness and accuracy on ResNet26 (sub-figure (a)) and ResNet50 (sub-figure (b)) trained by MBNAT and NoFrost. Adversarial robustness is evaluated on PGD (the top-left panel in each sub-figure), APGD-CE (the top-right panel in each sub-figure), AutoAttack (AA, the bottom-left panel in each sub-figure) and targeted PGD (denoted as PGD^T; the bottom-right panel in each sub-figure) attacks. γ is the weight for interpolation between the two BNs in MBNAT.

TEXAS The University of Texas at Austin

amazon

Results

Table 1. Adversarial robustness of ResNet26 under perturbation magnitude $\epsilon = 8$. Classification accuracy on clean images and under different adversarial attacks are reported. The best and second to the best numbers are shown in bold and underlined, respectively.

Method	Clean	White-box Attacks					Black-box Attacks		AA
		PGD	APGD-CE	APGD-DLR	MIA	CW	RayS	Square	
ST	**72.68**	0.01	0.00	0.00	0.00	0.00	18.2	27.5	0.00
SAT	52.65	10.55	5.02	5.30	8.84	9.18	30.5	44.7	3.78
TRADES	39.64	9.94	<u>6.24</u>	4.02	8.33	6.37	20.7	32.8	3.54
FAT	58.72	6.97	2.35	2.68	6.59	6.37	<u>33.6</u>	**50.8**	1.70
TRADES-FAT	55.65	<u>11.91</u>	5.79	<u>6.14</u>	<u>10.83</u>	**10.81**	31.1	46.7	<u>4.63</u>
NoFrost	<u>70.13</u>	**12.24**	**6.34**	**6.60**	**21.83**	<u>10.18</u>	**34.5**	<u>48.3</u>	**5.04**

Table 2. Adversarial robustness of ResNet50 under perturbation magnitude $\epsilon = 8$. Classification accuracy on clean images and under different adversarial attacks are reported. The best and second to the best numbers are shown in bold and underlined, respectively.

Method	Clean	White-box attacks					Black-box attacks		AA
		PGD	APGD-CE	APGD-DLR	MIA	CW	RayS	Square	
ST	**76.06**	0.04	0.00	0.00	0.00	0.00	22.5	31.4	0.00
SAT	59.28	13.57	7.80	<u>8.46</u>	10.28	11.02	27.4	40.2	6.23
TRADES	49.25	<u>14.80</u>	<u>9.20</u>	8.19	<u>12.97</u>	11.80	32.6	39.5	<u>6.66</u>
FAT	58.94	12.45	5.48	7.16	12.56	<u>12.24</u>	<u>35.9</u>	**51.4**	4.73
TRADES-FAT	60.52	11.67	4.71	5.90	11.28	10.29	34.5	<u>48.6</u>	3.87
NoFrost	<u>74.06</u>	**22.45**	**11.96**	**13.37**	**27.02**	**19.17**	**36.1**	43.1	**9.36**

TEXAS The University of Texas at Austin amazon

Results: NoFrost* achieves comprehensive robustness

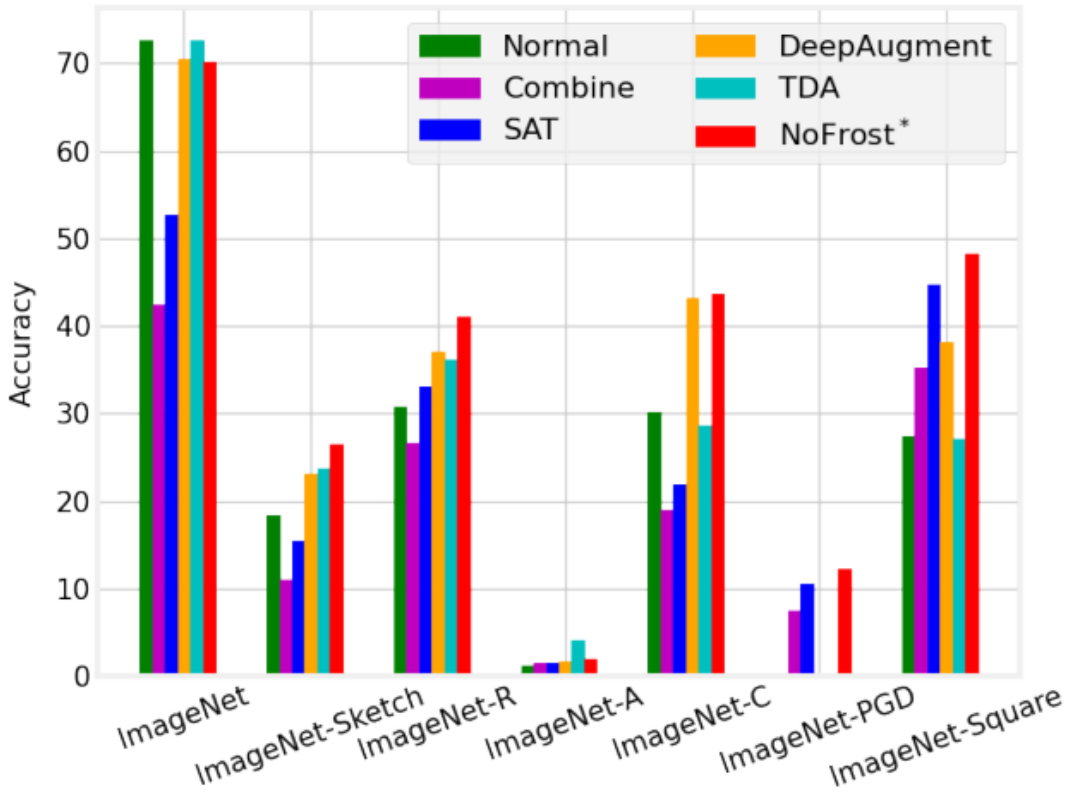


Figure 3. Model performance (accuracy in percentage) on different benchmark datasets or adversarial attacks. All methods are trained on ImageNet with ResNet26.

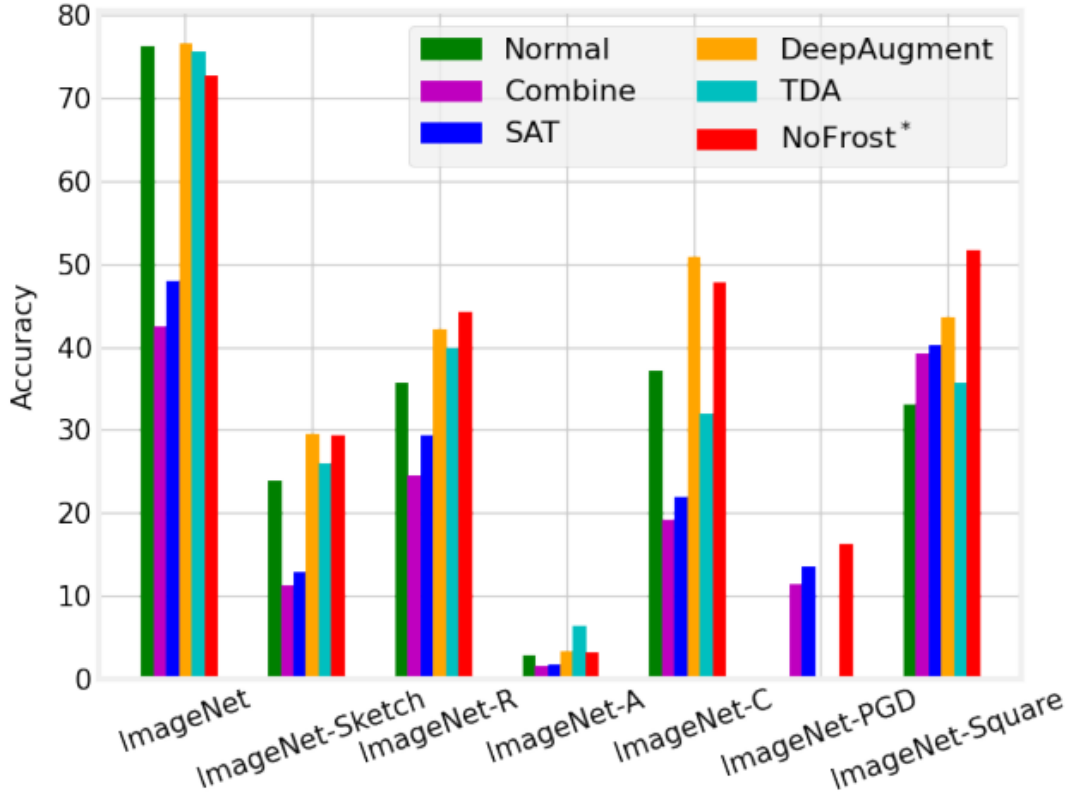


Figure 4. Model performance (accuracy in percentage) on different benchmark datasets or adversarial attacks. All methods are trained on ImageNet with ResNet50.

TEXAS The University of Texas at Austin amazon

Results: Replacing BN with IN doesn't work

Table 3. Standard adversarial training (SAT) with IN-based networks yields worse robustness than NoFrost. Experiments conducted on ResNet26 with different normalizers.

	Clean	**PGD**
SAT w/ BN	52.65	10.55
SAT w/ IN	56.78	11.06
NoFrost	**70.13**	**12.24**

Results: NoFrost Leads to More Robust Model Properties

Table 5. Decision margin, boundary thickness, and model smoothness of adversarially trained (under $\epsilon = 8$) ResNet26 models with different normalization strategies. The best and second-best values are bolded and underlined, respectively.

Normalization strategy (Method)	**Decision margin** $M(\boldsymbol{x})(\uparrow)$	**Boundary thickness** $T(\boldsymbol{x})(\uparrow)$	**Model smoothness** $D(\boldsymbol{x})(\downarrow)$
BN (SAT)	<u>0.3241</u>	<u>17.51</u>	4.927
MBN (MBNAT)	0.3143	13.78	**1.119**
NF (NoFrost)	**0.4700**	**31.49**	<u>2.996</u>

Thank You!

Our code and pretrained models are available on GitHub:

https://github.com/amazon-research/normalizer-free-robust-training

TEXAS
The University of Texas at Austin
amazon