Supplementary Material: Efficient Training Methods for Achieving Adversarial Robustness Against Sparse Attacks

Sravanti Addepalli*, Dhruv Behl*, Gaurang Sriramanan, R.Venkatesh Babu
Video Analytics Lab, Indian Institute of Science, Bangalore, India

1. Details on Training Dataset

In this work, we present detailed evaluations on benchmark datasets such as CIFAR-10 [8], ImageNet-100 (a 100-class subset of ImageNet [4, 11]) and the German Traffic Sign Recognition Benchmark (GTSRB) [6]. The CIFAR-10 dataset consists of 32×32 RGB images of ten distinct classes, and has become a benchmark dataset for the analyses and evaluation of adversarial attacks and defenses. ImageNet [4] is a 1000-class dataset consisting of 224×224 colour images, and is known to be particularly challenging for the setting of robust classification. In this work, we focus on a 100-class random subset [11], to alleviate computational space and time requirements. The GTSRB dataset consists of road signage boards, and is thus immensely important towards the practical deployment of robust networks in autonomous navigation vehicles. Following Rao *et al.* [10], we use a subset of GTSRB, consisting of 32×32 colour images, with 35,600 training images and 1,273 test images across 43 classes. We split each of the training datasets into class-balanced training and validation sets. We use a 2% validation split for CIFAR-10 and GTSRB, and 20% split for ImageNet-100.

2. Training Details

We train ResNet-20 [5] models from scratch using Adam [7] with initial learning rate of 0.01 and weight decay 0.0001 for 125 epochs with batch size 64. We reduce the learning rate by a factor of 10, three times during the training process. On CIFAR-10, we train with a patch size of 5×5 for the patch defense and with $k = 20$ for the ℓ_0 defense. On GTSRB, we we train with a patch size of 7×7. On ImageNet-100, we train with a patch size covering 3% of the image. For CIFAR-10 training, we set α (Eq.3 of the main paper) to 0.2 and λ (Eq.2 of the main paper) to 1. We multiply the value of λ with 9 three times during training. In the FRC-GLA defense, we use the BPFC regularizer [1] with $p = 4$ and $\lambda = 1$ along with a multiplicative factor of 5. We use early stopping on the validation split based on robust accuracy against the ROA PGD-150 attack [12] for the patch defenses and ℓ_0-RS (1k) attack [2] for the ℓ_0 defenses in the last 25 epochs.

Table 1. **ImageNet-100**: Performance (%) of the proposed methods FCR-RL, FCR-GL and FCR-GLA compared to baselines, against PGD 30-step all location patch attack (stride=4) of different sizes (1%, 2% and 3%) with 10 random restarts (RR).
FP : Forward pass, F+BP: Forward and Backward passes

Method	# steps location	# steps attack	Clean Acc	PGD-30 10RR		
				1%	2%	3%
AT-ROA (DOA) [12]	1444 FP	20	71.8	14.7	10.5	7.6
AT-FullLO [10]	80 FP	20	75.1	18.4	15.8	12.0
FCR-RL (Ours)	0 FP	0	**75.5**	13.9	9.6	5.8
FCR-GL (Ours)	0.5 F+BP	0	75.2	18.8	15.7	11.6
FCR-GLA (Ours)	0.5 F+BP	0.5	74.9	**23.1**	**19.8**	**17.1**

Table 2. **GTSRB Stop Sign dataset**: Performance (%) of the proposed methods FCR-RL, FCR-GL and FCR-GLA compared to baselines, against Stop Sign attack [12] with multiple random restarts (RR). FP : Forward pass, F+BP: Forward and Backward passes

Method	# steps location	# steps attack	Clean Acc	1 RR	10 RR	100 RR
AT-ROA (DOA) [12]	676 FP	30	94.3	**85.5**	75.3	74.1
AT-FullLO [10]	200 FP	50	93.2	82.9	68.8	68.0
FCR-RL (**Ours**)	0	0	**94.7**	81.3	71.9	69.5
FCR-GL (**Ours**)	0.5 F+BP	0	93.9	84.1	76.2	75.8
FCR-GLA (**Ours**)	0.5 F+BP	0.5	92.6	85.0	**79.3**	**78.6**

Table 3. **CIFAR-10, ℓ_0 threat model**: Performance (%) of the proposed methods FCR-RL, FCR-GL and FCR-GLA trained using ℓ_0 norm bound perturbations, against l_0-RS attack [2] with 25000 queries and ℓ_0 perturbation bound k. The proposed method achieves robustness comparable to the multi-step defense PGD$_0$-AT [3] at around 6× lower computational cost.

Method	# steps attack	Clean Acc	l_0-RS (25k)		Time/ epoch (s)	No. of epochs	Time (hrs)
			k=10	k=15			
PGD$_0$-AT [3]	40	87.1	**43.2**	**32.8**	390	100	10.8
FCR-RL (**Ours**)	0	**88.6**	31.1	20.7	**35**	125	**1.2**
FCR-GL (**Ours**)	0	86.5	36.3	24.9	43	125	1.5
FCR-GLA (**Ours**)	0.5	85.2	38.7	27.0	50	125	1.7

*Equal contribution.

Correspondence to: Sravanti Addepalli, sravantia@iisc.ac.in

3. Details on Results

To ensure the absence of gradient masking [9], we evaluate the defenses with a black-box gradient-free patch attack. The main candidates for such an attack are Texture-based Patch Attack (TPA) [13], which makes use of reinforcement learning to select the patch location and texture, and Patch-RS [2], which uses a framework based on score-based random search to generate the patch attack. We run these two attacks on our FCR-GLA model on ImageNet-100 with the patch covering 3% of the image. The maximum number of queries allowed is 10,000. The robust accuracy obtained with TPA is 59%, while the accuracy obtained with Patch-RS is 36%. Since Patch-RS is a stronger and more query-efficient attack, we use this to test black-box robustness of our defenses in Tables 1 and 2 of the main paper. From Table 1 in the supplementary, the all-location PGD-30 attack yields the strongest evaluation of robustness against patch attacks on ImageNet-100 with a robust accuracy of 17%. Thus, we report numbers against the all-location PGD attack in Table 2 of the main paper and Tables 1 and 2 in the supplementary.

All the experiments for measuring computational complexity in Table 1 of the main paper and Table 3 of the supplementary are done on a single Nvidia GeForce GTX 1080 Ti GPU card.

References

[1] Sravanti Addepalli, B S Vivek, Arya Baburaj, Gaurang Sriramanan, and R Venkatesh Babu. Towards Achieving Adversarial Robustness by Enforcing Feature Consistency Across Bit Planes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[2] Francesco Croce, Maksym Andriushchenko, Naman D Singh, Nicolas Flammarion, and Matthias Hein. Sparse-RS: a versatile framework for query-efficient sparse black-box adversarial attacks. *ECCV Workshop on Adversarial Robustness in the Real World*, 2020.

[3] Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks (IJCNN)*, 2013.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[9] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the ACM Asia Conference on Computer and Communications Security (ACM ASIACCS)*, 2017.

[10] Sukrut Rao, David Stutz, and Bernt Schiele. Adversarial Training against Location-Optimized Adversarial Patches. *ECCV Workshop (CV-COPS)*, 2020.

[11] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R Venkatesh Babu. Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[12] Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. Defending Against Physically Realizable Attacks on Image Classification. *International Conference on Learning Representations (ICLR)*, 2020.

[13] Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhi Cao, and Alan Yuille. PatchAttack: A Black-Box Texture-Based Attack with Reinforcement Learning. In *The European Conference on Computer Vision (ECCV)*, 2020.