Efficient Training Methods for Achieving Adversarial Robustness against Sparse Attacks

Sravanti Addepalli*, Dhruv Behl*, Gaurang Sriramanan, R. Venkatesh Babu
Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India

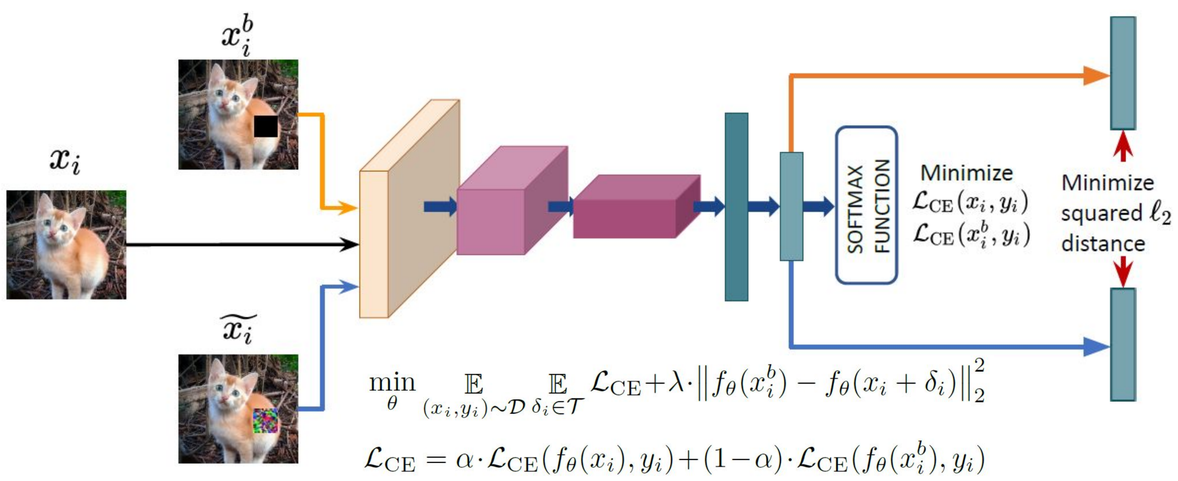
Background and Motivation

- **Goal :** Achieving adversarial robustness against Sparse Attacks (patch attacks, L0 norm bound attacks) efficiently
- There has been steady progress in defending against L-inf and L2 norm bound attacks effectively and efficiently
- **Challenges in defending Real World Attacks**
  - Sparse attacks are easier to implement in the real-world
  - Adversarial attack generation is computationally more expensive under sparse threat models (requires 10-50 attack steps), making standard adversarial training more expensive
  - Adversarial attacks in the real world need not be restricted to a single threat model - generalization to unseen attacks is important

Contributions

- We propose Feature Consistency Regularizer (FCR) based training that uses :
  - [**FCR-RL**] Random augmentations at random locations (RL)
  - [**FCR-GL**] Random augmentations at (single-step) gradient locations (GL)
  - [**FCR-GLA**] Single-step gradients for attack generation and training in alternate training iterations
- Generalizes better that existing empirical and certified defenses to unseen sparse attacks such as multi-patch attacks and L0 norm bound attacks
- Achieves a large boost in robustness when combined with the state-of-the-art certified patch defense, BagCert
- The proposed algorithm can be extended to other threat models as well. We demonstrate improvements on an L0 norm based threat model as well

Feature Consistency Regularizer (FCR)



Results on CIFAR-10

CIFAR-10: Performance (%) of the proposed methods FCR-RL, FCR-GL and FCR-GLA against PGD 150-step all location attack with multiple random restarts (RR) and Patch-RS (P-RS) attack [4] with 10000 queries (Q). FP: Forward pass, F+BP: Forward and Backward passes

Method	# steps location	# steps attack	Clean Acc	PGD 10 RR	PGD 100 RR	P-RS 10k Q
AT-ROA (DOA) [1]	784 FP	30	83.6	30.2	29.8	41.3
AT-FullLO [2]	200 FP	50	**88.7**	33.4	32.9	40.8
FCR-RL (**Ours**)	0 FP	0	87.9	30.6	26.1	40.2
FCR-GL (**Ours**)	0.5 F+BP	0	84.9	38.8	34.3	50.1
FCR-GLA (**Ours**)	0.5 F+BP	0.5	85.3	**42.8**	**41.1**	**56.4**

Results on ImageNet-100

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) [1]	1444 FP	20	71.8	14.7	10.5	7.6
AT-FullLO [2]	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**

Generalization to unseen attacks

Generalization to unseen attacks: Performance (%) of the proposed methods FCR-RL, FCR-GL and FCR-GLA compared to baselines, against patch attacks, ℓ_0 and ℓ_1 norm bound attacks on the CIFAR-10 dataset. All defenses are trained to be robust to a single square patch attack of size 5×5. We evaluate these defenses against various attacks that are unseen during training, such as the square multi-patch attack, rectangular single-patch attack, and ℓ_0, ℓ_1 norm bound attacks. Patch-RS [4] with 10000 queries is used for evaluating robustness to patch attacks. Square attack [5, 6] with 1000 queries and l_0-RS [4] attack with 5000 queries are used for evaluation of ℓ_1 and ℓ_0 attacks respectively. The first two partitions use ResNet-20 architecture and the third partition uses BagNet [7] architecture.

Method	Clean Acc	Patch attack (Total budget ~25 pixels) 1 square 5x5	2 squares 4x4, 3x3	3 squares {3x3}^3	4 squares {3x3, 2x2}^2	5 squares 3x3, {2x2}^4	6 squares {2x2}^6	1 rectangle 3x8/ 2x12/ 1x25	ℓ_1 ($\varepsilon = 5$)	ℓ_0 ($\varepsilon = 7$)	Avg (unseen threat models)	No. of epochs	Time / epoch (sec)	Total time (hrs)
DS (Certified 56.2%)	83.9	70.5	59.2	50.5	43.2	41.9	39.7	40.2	45.1	58.5	49.7	350	42	4.1
Mask-DS (Certified 58.1%)	84.5	**73.1**	**60.5**	**51.3**	44.0	42.6	40.4	**40.7**	43.0	59.1	49.5	350	42	4.1
AT-ROA [1]	83.6	41.3	35.1	32.2	30.6	29.3	28.2	29.4	61.9	65.3	52.6	120	370	12.3
AT-FullLO [2]	**88.7**	40.8	36.3	34.1	31.5	29.7	29.1	28.0	63.2	62.4	52.3	200	400	22.2
FCR-RL (Ours)	87.9	40.2	35.8	33.2	30.4	29.5	28.6	26.3	65.8	61.1	52.5	125	**30**	**1.0**
FCR-GL (Ours)	84.9	50.1	44.4	41.7	40.5	39.5	39.1	33.0	69.6	67.4	58.9	125	38	1.3
FCR-GLA (Ours)	85.3	56.4	50.4	46.9	**44.9**	**44.1**	**43.4**	**40.4**	**70.1**	**69.5**	**61.5**	125	45	1.5
BagCert (Certified 60%) [3]	**85.0**	**76.3**	46.7	42.6	37.8	35.3	34.6	44.2	55.5	49.1	48.2	350	75	7.2
FCR-GLA (Ours)	84.4	64.8	58.5	53.1	**49.4**	**47.5**	**44.1**	**45.3**	**74.1**	**62.7**	**62.1**	200	**70**	**3.9**
FCR-GLA (Ours+BagCert)	84.1	75.2	**61.4**	**54.2**	47.9	43.6	42.8	44.1	65.3	56.5	56.9	350	90	8.7

Results on GTSRB Stop Sign dataset

GTSRB Stop Sign dataset: Performance (%) of the proposed methods FCR-RL, FCR-GL and FCR-GLA compared to baselines, against Stop Sign attack [1] 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) [1]	676 FP	30	94.3	**85.5**	75.3	74.1
AT-FullLO [2]	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**

Robustness against L0 norm bound attacks

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 [4] with 25000 queries and ℓ_0 perturbation bound k. The proposed method achieves robustness comparable to the multi-step defense PGD_0-AT [8] at around $6\times$ lower computational cost.

Method	# steps attack	Clean Acc	l_0-RS (25k) k=10	k=15	Time/ epoch (s)	No. of epochs	Time (hrs)
PGD_0-AT [8]	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

References

[1] Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. Defending Against Physically Realizable Attacks on Image Classification. (ICLR), 2020.
[2] Sukrut Rao, David Stutz, and Bernt Schiele. Adversarial Training against Location-Optimized Adversarial Patches. ECCV Workshop (CV-COPS), 2020.
[3] Jan Hendrik Metzen and Maksym Yatsura. Efficient Certified Defenses Against Patch Attacks on Image Classifiers. ICLR 2021
[4] 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
[5] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square Attack: a query-efficient black-box adversarial attack via random search. ECCV, 2020.
[6] Francesco Croce and Matthias Hein. Mind the box: L1-apgd for sparse adversarial attacks on image classifiers. ICML 2021
[7] Wieland Brendel and Matthias Bethge. Approximating CNNs with Bag-of-local-features models works surprisingly well on ImageNet. ICLR 19
[8] Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. ICCV, 2019.

Acknowledgements: This work was supported by Uchhatar Avishkar Yojana (UAY) project (IISC 10), MHRD, Govt. of India. Sravanti Addepalli is supported by a Google PhD Fellowship in Machine Learning