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

Abstract

The vulnerability of Deep Neural Networks to adversarial attacks poses a serious threat in critical applications such as autonomous navigation systems and surveillance systems. While most of the existing research is focused on defending attacks within the ℓ_∞ and ℓ_2 threat models, real-world attacks are often sparse, as they need to be physically realizable. In this work, we aim to improve the efficiency of defenses against sparse adversaries such as patch-attacks and ℓ_0 norm bound attacks. We achieve this by enforcing the network to learn consistent feature representations between a clean image and a corresponding randomly augmented image that is specific to the considered threat model. The proposed method achieves robustness at a significant speed-up when compared to existing methods. We achieve a further boost in robustness by using single-step gradients for attack generation and location optimization.

1. Introduction

Deep Networks are being widely adopted in many security critical applications such as self-driving cars and face-recognition systems. However, they are susceptible to adversarial attacks [41], which are crafted perturbations to the input image [17] that can lead models to flip their predictions to completely unrelated classes, resulting in potentially dangerous outcomes [38, 16]. This has spurred interest in building adversarial attacks [8, 3, 42] to expose the vulnerabilities of models, and adversarial defenses [30, 50, 40] to improve their robustness. To formalize research in the area of adversarial attacks and defenses, an attack is typically confined within a well-defined threat model [7, 18], such as an ℓ_p norm bound of radius ε, that ensures imperceptibility. The most common threat models in literature are the ℓ_∞ and ℓ_2 norm bounds, which tend to produce low magnitude perturbations on a large number of pixels.

While a systematic study with these settings has led to significant progress over the past few years [18, 30, 50, 20], real-world attacks differ in a few aspects. Firstly, an attacker is not restricted to generate attacks under a single threat model, which makes it important to build defenses that can generalize well to unseen attacks as well. Secondly, the attack needs to be physically realizable. Sparse attacks such as adversarial patches [5, 9, 48], and those constrained within the ℓ_0 norm bound [35, 32, 9] are easier to implement in the real-world, since it is easier to corrupt a few pixels with large perturbations, when compared to mildly perturbing a large number of pixels, as is the case with ℓ_∞ and ℓ_2 norm bound attacks. In a patch attack, the adversary is allowed to perturb a patch of a fixed shape and size (such as a 5×5 square patch) in the given image. The patch size is selected to be roughly 1% to 3% of the image size [45, 37], so that the perturbation is not conspicuous.

In this work, we propose efficient adversarial defenses to achieve robustness against sparse attacks by enforcing consistent feature representations between a clean image, and an augmented image that is specific to the defined threat model. While existing empirical defenses [45, 37] typically use multi-step adversarial attacks (10 to 50 steps) during training, we demonstrate robustness to sparse attacks without the use of adversarial samples during training. Further, we show that with an overhead of one additional gradient computation, we can find a better location for the random perturbation, thereby boosting the robustness significantly. Finally, by using the computed gradients to also generate single-step attacks, we obtain improved results when compared to existing multi-step adversarial training methods.

2. Contributions

Our contributions have been listed below:

- **FCR-RL:** We propose Feature Consistency Regularizer (FCR) based training that uses random patch augmentations at random locations (RL) to achieve robustness to patch attacks, at a computational cost that is comparable to standard training.
- **FCR-GL:** We further propose to use single-step gradients for location optimization in alternate training iterations to improve the strength of the attack, leading to significantly better robustness.

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

- **FCR-GLA:** We propose to use the gradients computed in the alternate training iterations for attack generation as well, to achieve improved results compared to existing multi-step adversarial training methods at a significantly lower computational cost. We demonstrate improved results on the CIFAR-10 [26], ImageNet-100 [13] and GTSRB [24] datasets.
- The proposed method generalizes better that existing empirical and certified defenses to unseen sparse attacks such as multi-patch attacks and ℓ_0 norm bound attacks, and achieves a large boost in robustness when combined with the state-of-the-art certified patch defense BagCert [31].
- We extend the proposed method to the ℓ_0 norm threat model, where we achieve results comparable to adversarial training methods at significantly lesser compute.

3. Related Works

Following the discovery of adversarial examples by Szegedy *et al.* [41], a broad range of defenses [18, 30, 50] have been proposed to improve the worst-case performance of deep networks. Amongst the earliest defenses specific to the ℓ_∞ threat model was Fast Gradient Sign Method based adversarial training (FGSM-AT) [18], wherein the training set is augmented with adversaries generated using single-step optimization. This was later found to be susceptible to gradient masking [34], and was thus not robust against multi-step attacks. This was followed by methods that attempted to use input transformations [21] or other gradient-obfuscation based training methods [6, 29, 15, 47, 39] to build robust models. However, they were broken in the work by Athalye *et al.* [3], where the authors proposed adaptive attacks to circumvent such defenses. Madry *et al.* [30] and Zhang *et al.* [50] proposed robust training techniques using strong, multi-step adversaries, and have withstood the test of time against more sophisticated attacks [11, 40]. However, due to their large computational overhead, efficient training techniques have been proposed that either eliminate the generation of adversaries entirely [1], or make use of only single-step attacks [44, 40]. While such defenses are largely developed for the setting of ℓ_2 or ℓ_∞ based threat models, in this work we seek to build efficient defenses against physically realisable adversaries, as is subsequently expounded.

Patch Attacks: Brown *et al.* [5] first demonstrated the vulnerability of image classification models to adversarial patch attacks, wherein carefully crafted universal adversarial stickers could be printed and placed on any scene to induce misclassification. Subsequently, image-specific patch attacks such as LaVAN [25] were also proposed. While physical adversarial attacks are often modeled as adversarial patches [28, 36, 43], a related variant, which is localized adversarial perturbations of different shapes, have been shown to fool safety-critical systems like Face Recognition Systems [38] and Road Sign Classifiers [16]. Several black-box attacks have been developed for patch attacks, including Texture-based Patch Attack (TPA) [48], which uses a reinforcement learning agent to optimize the patch location and texture, and Patch-RS [9], which is based on random search to efficiently generate the patch attack.

Defenses against Patch Attacks: Input pre-processing based defenses such as Digital Watermarking (DW) [22] and Local Gradient Smoothing (LGS) [33] were first proposed to detect and mask out the adversarial patch in the image before passing it to the classifier. However, such defenses are vulnerable to adaptive adversaries such as BPDA [3]. Adversarial training [30] has been adapted to build robust defenses against patch-attacks [37, 45]. Although these models show significantly improved robustness to patch attacks, they are computationally expensive to train.

Chiang *et al.* [49] proposed the first certified defense against adversarial patch-attacks using Interval Bound Propagation (IBP) [19]. Later, Clipped BagNet (CBN) [51], De-randomized Smoothing (DS) [27], and PatchGuard [46] were introduced to achieve significantly higher certified accuracy on CIFAR-10 and ImageNet. The large inference time incurred by the DS based methods limits their practicality. BagCert [31] uses a modification of the BagNet [4] architecture to limit the receptive field of the network. This coupled with their certification framework and margin based training loss yields high certified robustness and efficient inference. While these defenses improve robustness within the specified threat model, we show that they do not generalize well to other unseen sparse attacks.

ℓ_0-attacks and defenses: Croce *et al.* [10] introduced a score-based ℓ_0-norm attack called CornerSearch (CS), and an ℓ_0-norm variant of the PGD attack PGD_0, which were stronger than prior attacks such as JSMA [35] and Sparse-Fool [32]. Using this, the authors propose an adversarial training method to improve robustness against ℓ_0 norm bound attacks. Croce *et al.* [9] propose a black-box framework for score-based sparse attacks using random search. Their ℓ_0-RS attack achieves state-of-the-art success rate and query efficiency in the ℓ_0 threat model.

4. Notation

In this work, we aim to improve the adversarial robustness of Deep Neural Network based image classifiers. We denote a data sample from a distribution $\mathcal{D}$ as (x_i, y_i) where x_i denotes the input image and y_i denotes its corresponding ground truth label. We denote a Deep Neural Network based classifier using f_θ whose weights are denoted by $\theta \in \Theta$. The network takes an image x_i as input and outputs the pre-softmax output $f(x_i)$. We denote an augmentation that is specific to a threat model $\mathcal{T}$ as $\widetilde{x_i}$. We denote the Cross-Entropy loss by $\mathcal{L}_{\text{CE}}$.

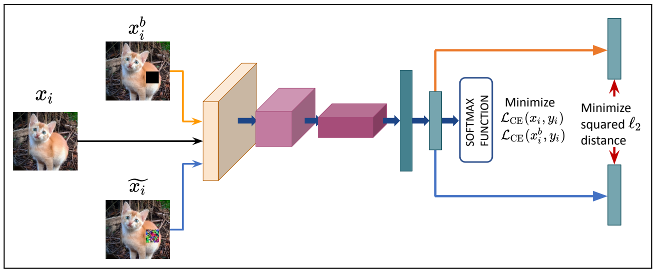

Figure 1. Schematic diagram of the proposed Feature Consistency Regularizer (FCR) based training to defend against patch attacks

5. Proposed Method

5.1. Adversarial Training on Sparse Threat Models

Adversarial Training is the most widely used defense strategy for obtaining models that are robust to adversarial attacks in a well-defined threat model. Projected Gradient Descent based Adversarial Training (PGD-AT) [30] is one of the earliest and most successful defenses that works well on a wide range of threat models. This attempts to solve the minimax optimization problem of firstly maximizing the Cross-Entropy loss for generating strong adversarial attacks, followed by minimizing the worst-case loss for training. The loss formulation of PGD-AT is shown below:

$$\min_{\theta} \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}} \max_{\varepsilon\in\mathcal{T}} \mathcal{L}_{\mathrm{CE}}\left(f_\theta(x+\varepsilon), y\right) \tag{1}$$

While this loss formulation achieves the best results across various threat models, the inner maximization step requires the use of multiple iterations, leading to significantly higher computational cost for training. This makes it hard to scale to large datasets which are common in real-world applications. Further, the complexity of attack generation depends on the threat model considered. While attacks constrained within the ℓ_2 and ℓ_∞ norm bounds typically have lower magnitude perturbations per pixel, and therefore can achieve good attack strength using lesser optimization steps (10 steps) [30, 50], sparse threat models such as patch and ℓ_0 norm produce larger magnitude perturbations per pixel, and hence require more steps (20-50 steps) [37, 45, 10] for the generation of strong attacks. This increases the computational cost of sparse adversarial training further. The generation of adversaries with lesser number of optimization steps leads to the phenomenon of gradient-masking [34] where the network learns to obfuscate the gradients around the data samples to prevent the generation of strong attacks, leading to catastrophic failure of adversarial training [44].

5.2. Feature Consistency Regularizer

FCR-RL (Random Location): We propose to replace the inner optimization step in Eq.1 with an expectation over random augmentations and show using extensive experiments that smoothing the loss surface over random directions can indeed serve as a promising alternative to the minimization of classification loss over strong multi-step adversaries in sparse threat models. The optimization problem used for training models using the proposed Feature Consistency Regularizer (FCR) is shown in Eq.2 and 3.

$$\min_{\theta} \underset{(x_i,y_i)\sim\mathcal{D}}{\mathbb{E}} \underset{\delta_i\in\mathcal{T}}{\mathbb{E}} \mathcal{L}_{\mathrm{CE}} + \lambda \cdot \left\| f_\theta(x_i^b) - f_\theta(x_i+\delta_i) \right\|_2^2 \tag{2}$$

$$\mathcal{L}_{\mathrm{CE}} = \alpha \cdot \mathcal{L}_{\mathrm{CE}}(f_\theta(x_i), y_i) + (1-\alpha) \cdot \mathcal{L}_{\mathrm{CE}}(f_\theta(x_i^b), y_i) \tag{3}$$

For every input image x_i, we generate a randomly augmented image $\widetilde{x}_i = x_i + \delta_i$ by adding a random patch of a fixed size at a random location. The random patch is constructed by sampling the perturbation for every pixel from the distribution $\mathcal{U}(0, 255)$ and clipping the obtained value to the range $[0, 255]$. We further generate a corresponding Cutout [14] based augmentation x_i^b by blanking out the pixel values corresponding to the location of the random patch, or setting them to 0 as shown in Fig.1. As shown in Eq.3, we minimize a convex combination of cross-entropy loss on x_i and x_i^b with a coefficient α. We use the squared ℓ_2 norm based regularizer [1] to impose local smoothness between the pre-softmax outputs of the perturbed image $\widetilde{x}_i$ and its Cutout augmentation x_i^b, in addition to the minimization of cross-entropy loss.

FCR-GL (Gradient based Location): We further explore the use of single-step gradients for location optimization of the random attack in alternate training iterations. This is done by first adding random noise Δ_i of magnitude $8/255$ sampled from a Bernoulli distribution to every pixel, and further computing gradient of the cross-entropy loss of the image $x_i + \Delta_i$. We randomly select one location among the top-5 average gradient locations, where the average is found across a local $p \times p$ patch, that corresponds to the considered threat model.

FCR-GLA (Gradient based Location and attack): In this method, we utilize single-step gradients for both location optimization and attack generation in alternate training iterations. We generate a random patch perturbation as obtained earlier, and multiply this with the sign of gradients at the patch location before adding it to the image. Therefore the perturbation utilizes very weak support from the gradients. The use of random location and attack in alternate training iterations prevents the issue of gradient masking that is common with single-step training methods [34]. We additionally use the BPFC regularizer proposed by Addepalli *et al.* [1] to improve the efficacy of single-step gradients by enforcing local smoothness of the loss surface.

6. Experiments and Results

We present detailed evaluations on benchmark datasets such as CIFAR-10 [26], ImageNet-100 (a 100-class subset

Table 1. **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 [9] with 10000 queries is used for evaluating robustness to patch attacks. Square attack [2, 12] with 1000 queries and l_0-RS [9] attack with 5000 queries are used for evaluation of ℓ_1 and ℓ_0 attacks respectively. The first two partitions use ResNet-20 [23] architecture and the third partition uses BagNet [4] 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%) [27]	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%) [46]	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 [45]	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 [37]	**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%) [31]	**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

Table 2. **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 [9] 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) [45]	784 FP	30	83.6	30.2	29.8	41.3
AT-FullLO [37]	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**

of ImageNet [13, 40]) and the German Traffic Sign Recognition Benchmark (GTSRB) [24]. We include details on datasets and training in the Supplementary.

Results: We present evaluations of the proposed approach along with existing adversarial training methods AT-ROA [45] and AT-FullLO [37] against patch attacks on the CIFAR-10 dataset in Table-2. The results on ImageNet-100 and GTSRB datasets are presented in Tables-1 and 2 of the Supplementary. We evaluate all defenses against an all-location PGD 150-step attack with multiple random restarts (10-100 RR) which is much stronger and exhaustive when compared to evaluations in prior work [45, 37]. We additionally evaluate against the gradient-free attack Patch-RS [9] using 10000 queries on the CIFAR-10 dataset, to ensure the absence of gradient masking. FCR-RL achieves robustness to a significant extent across all three datasets at a budget comparable to standard training, and achieves results comparable to the multi-step (30-50) adversarial training methods on the CIFAR-10 and the GTSRB datasets. We further note that by using merely one additional back-propagation in alternate training iterations for location optimization, we achieve a significant boost in robustness across all three datasets in FCR-GL. Reusing the same gradients for attack generation as well leads to a further boost in robustness across all three datasets. Overall, we achieve gains of 8.2%, 5.1% and 4.5% on CIFAR-10, ImageNet-100 and GTSRB datasets respectively, when compared to the multi-step adversarial training approaches [37, 45].

We compare performance of the proposed patch defense with existing empirical and certified defenses against attacks constrained within various threat models in Table-1. While some of the certified defenses [27, 46, 31] achieve better robustness against the main threat model considered (5×5 square patches), they are computationally more expensive either during training [31] or inference [27, 46]. Further, our proposed defenses generalize very well to other unseen threat models such as multi-patch attacks, rectangular attacks and attacks within the ℓ_0 and ℓ_1 norm bound, while certified defenses are specific to the threat model considered. We obtain the highest average accuracy against unseen threat models, which is computed as an equally weighted average of unseen patch attacks, ℓ_0 and ℓ_1 norm bound attacks. By using the BagNet architecture [4, 31] with the proposed approach, we achieve significant gains in results and obtain additional gains by combining the proposed approach with the BagCert defense [31].

We use the proposed algorithm to defend against ℓ_0 norm bound attacks and obtain results comparable to the multi-step adversarial training method PGD_0-AT [10] at significantly lower compute (Table-3 of the Supplementary).

7. Conclusions

We propose Feature Consistency Regularizer (FCR) based training to achieve robustness against Patch Attacks without the use of expensive multi-step adversarial attacks during training. The proposed defense achieves improved results when compared to existing multi-step defenses on the main threat model used for training, and generalizes much better to unseen threat models when compared to certified patch defenses, using significantly lower compute during training and inference. We extend the proposed framework to defend against other sparse threat models such as the ℓ_0 norm bound as well.

8. 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. We are thankful for the support.

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] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square Attack: a query-efficient black-box adversarial attack via random search. In *The European Conference on Computer Vision (ECCV)*, 2020.

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018.

[4] Wieland Brendel and Matthias Bethge. Approximating CNNs with Bag-of-local-features models works surprisingly well on ImageNet. In *International Conference on Learning Representations (ICLR)*, 2019.

[5] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[6] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

[7] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017.

[9] 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.

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

[11] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020.

[12] Francesco Croce and Matthias Hein. Mind the box: l_1-apgd for sparse adversarial attacks on image classifiers. *International Conference on Machine Learning (ICML)*, 2021.

[13] 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.

[14] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[15] Guneet S. Dhillon, Kamyar Azizzadenesheli, Jeremy D. Bernstein, Jean Kossaifi, Aran Khanna, Zachary C. Lipton, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations (ICLR)*, 2018.

[16] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] Ian Goodfellow and Nicolas Papernot. Is attacking machine learning easier than defending it? Blog post on Feb 15, 2017.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[19] Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[20] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

[21] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*, 2018.

[22] Jamie Hayes. On Visible Adversarial Perturbations & Digital Watermarking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[23] 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.

[24] 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.

[25] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning (ICML)*, 2018.

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

[27] Alexander Levine and Soheil Feizi. (De)Randomized Smoothing for Certifiable Defense against Patch Attacks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[28] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299*, 2018.

[29] Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Michael E. Houle, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations (ICLR)*, 2018.

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Tsipras Dimitris, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[31] Jan Hendrik Metzen and Maksym Yatsura. Efficient Certified Defenses Against Patch Attacks on Image Classifiers. In *International Conference on Learning Representations (ICLR)*, 2021.

[32] Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. SparseFool: A Few Pixels Make a Big Difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] Muzammal Naseer, Salman Khan, and Fatih Porikli. Local Gradients Smoothing: Defense Against Localized Adversarial Attacks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[34] 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.

[35] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS P)*, 2016.

[36] Anurag Ranjan, Joel Janai, Andreas Geiger, and Michael J. Black. Attacking Optical Flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

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

[38] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[39] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

[40] 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.

[41] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2013.

[42] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[43] Rey Wiyatno and Anqi Xu. Physical adversarial textures that fool visual object tracking. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[44] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020.

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

[46] Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. PatchGuard: A Provably Robust Defense against Adversarial Patches via Small Receptive Fields and Masking. In *30th USENIX Security Symposium (USENIX Security 21)*, 2021.

[47] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations (ICLR)*, 2018.

[48] 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.

[49] Ping yeh Chiang*, Renkun Ni*, Ahmed Abdelkader, Chen Zhu, Christoph Studor, and Tom Goldstein. Certified defenses for adversarial patches. In *International Conference on Learning Representations (ICLR)*, 2020.

[50] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.

[51] Zhanyuan Zhang, Benson Yuan, Michael McCoyd, and David Wagner. Clipped BagNet: Defending Against Sticker Attacks with Clipped Bag-of-features. In *IEEE Security and Privacy Workshops (SPW)*, 2020.