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 $\ell_\infty$ and $\ell_2$ 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 $\ell_0$ 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 Neural Networks are susceptible to adversarial attacks [36], which are crafted perturbations to the input image [15] that can lead models to flip their predictions to completely unrelated classes, resulting in potentially dangerous outcomes [14,34]. This has spurred interest in building adversarial attacks [3,7,37] to expose the vulnerabilities of models, and adversarial defenses [26,35,44] 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 [6,16], such as an $\ell_p$ norm bound of radius $\epsilon$, that ensures imperceptibility. The most common threat models in literature are the $\ell_\infty$ and $\ell_2$ 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 [16,18,26,44], 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 [5,8,42], and those constrained within the $\ell_0$ norm bound [8,28,31] 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 $\ell_\infty$ and $\ell_2$ norm bound attacks. In a patch attack, the adversary is allowed to perturb a patch of a fixed shape and size (such as a $5 \times 5$ square patch) in the given image. The patch size is selected to be roughly 1% to 3% of the image size [33,40], 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 [33,40] 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. 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.

- **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 ex-
isting multi-step adversarial training methods at a significantly lower computational cost. We demonstrate improved results on the CIFAR-10 [23], ImageNet-100 [12] and GTSRB [21] datasets.

- The proposed method generalizes better than existing empirical and certified defenses to unseen sparse attacks such as multi-patch attacks and ℓ₀ norm bound attacks, and achieves a large boost in robustness when combined with the state-of-the-art certified patch defense BagCert [27].

- We extend the proposed method to the ℓ₀ norm threat model, where we achieve results comparable to adversarial training methods at significantly lesser compute.

2. Related Works

Following the discovery of adversarial examples by Szegedy et al. [36], a broad range of defenses [16, 26, 44] 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) [16], wherein the training set is augmented with adversaries generated using single-step optimization. This was later found to be susceptible to gradient masking [30], and was thus not robust against multi-step attacks. Madry et al. [26] and Zhang et al. [44] proposed robust training techniques using strong, multi-step adversaries, and have withstood the test of time against more sophisticated attacks [10, 35]. 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 [35, 39]. While such defenses are largely developed for the setting of ℓ₂ or ℓ∞ based threat models, in this work we seek to build efficient defenses against physically realizable 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 [22] were also proposed. While physical adversarial attacks are often modeled as adversarial patches [25, 32, 38], a related variant, which is localized adversarial perturbations of different shapes, have been shown to fool safety-critical systems like Face Recognition Systems [34] and Road Sign Classifiers [14]. Several black-box attacks have been developed for patch attacks, including Texture-based Patch Attack (TPA) [42], which uses a reinforcement learning agent to optimize the patch location and texture, and Patch-RS [8], 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) [19] and Local Gradient Smoothing (LGS) [29] 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 [26] has been adapted to build robust defenses against patch-attacks [33, 40]. Although these models show significantly improved robustness to patch attacks, they are computationally expensive to train.

Chiang et al. [43] proposed the first certified defense against adversarial patch-attacks using Interval Bound Propagation (IBP) [17]. Later, Clipped BagNet (CBN) [45], De-randomized Smoothing (DS) [24], and PatchGuard [41] 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 [27] 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.

**ℓ₀-attacks and defenses:** Croce et al. [9] introduced a score-based ℓ₀-norm attack called CornerSearch (CS), and an ℓ₀-norm variant of the PGD attack PGD₀, which were stronger than prior attacks such as JSMA [31] and Sparse-Fool [28]. Using this, the authors propose an adversarial training method to improve robustness against ℓ₀ norm bound attacks. Croce et al. [8] propose a black-box framework for score-based sparse attacks using random search. Their ℓ₀-RS attack achieves state-of-the-art success rate and query efficiency in the ℓ₀ threat model.

3. 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 \((x, y)\) where \( x \) denotes the input image and \( y \) denotes its corresponding ground truth label. We denote a Deep Neural Network based classifier using \( f_{\theta} \) whose weights are denoted by \( \theta \in \Theta \). The network takes an image \( x_i \) as input and outputs the pre-softmax output \( f_{\theta}(x_i) \). We denote an augmentation that is specific to a threat model \( T \) as \( \tilde{x}_i \). We denote the Cross-Entropy loss by \( L_{CE} \).

4. Proposed Method

4.1. 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) [26] is one of the earliest and most successful defenses that works
ial attacks, followed by minimizing the worst-case loss for generating strong adversarial examples. 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:

\[
\min_{\theta} \mathbb{E}_{(x,y) \sim D} \max_{\varepsilon \in T} \mathcal{L}_{CE} \left( f_{\theta}(x + \varepsilon), y \right) \quad (1)
\]

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 \( \ell_2 \) and \( \ell_{\infty} \) norm bounds typically have lower magnitude perturbations per pixel, and therefore can achieve good attack strength using lesser optimization steps (10 steps) [26, 44], sparse threat models such as patch and \( \ell_0 \) norm produce larger magnitude perturbations per pixel, and hence require more steps (20-50 steps) [9, 33, 40] 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 [30] 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 [39].

### 4.2. 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.

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

\[
\mathcal{L}_{CE} = \alpha \cdot \mathcal{L}_{CE}(f_{\theta}(x_i), y_i) + (1 - \alpha) \cdot \mathcal{L}_{CE}(f_{\theta}(x_i^b), y_i) \quad (3)
\]

For every input image \( x_i \), we generate a randomly augmented image \( x_b = x_i + \delta_i \) 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 \( U(0, 255) \) and clipping the obtained value to the range \([0, 255]\). We further generate a corresponding Cutout [13] 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 \( \alpha \). We use the squared \( \ell_2 \) norm based regularizer [1] to impose local smoothness between the pre-softmax outputs of the perturbed image \( x_b \) 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 \( \delta_i \) 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 [30]. 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.

### 5. Experiments and Results

We present detailed evaluations on benchmark datasets such as CIFAR-10 [23], ImageNet-100 (a 100-class subset of ImageNet [12, 35]) and the German Traffic Sign Recognition Benchmark (GTSRB) [21]. 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 [40] and AT-FullLO [33] against patch attacks on the CIFAR-10 dataset in Table-2. The results on ImageNet-100
Table 1. Generalization to unseen attacks: Performance (%) of the proposed methods FCR-RL, FCR-GL and FCR-GLA compared to baselines, against patch attacks, $\ell_0$ and $\ell_1$ norm bound attacks on the CIFAR-10 dataset. All defenses are trained to be robust to a single square patch attack of size $5 \times 5$. We evaluate these defenses against various attacks that are unseen during training, such as the square multi-patch attack, rectangular single-patch attack, and $\ell_0$, $\ell_1$ norm bound attacks. Patch-RS [8] with 10000 queries is used for evaluating robustness to patch attacks. Square attack [2, 11] with 1000 queries and $\ell_0$-RS [8] attack with 5000 queries are used for evaluation of $\ell_1$ and $\ell_0$ attacks respectively. The first two partitions use ResNet-20 [20] architecture and the third partition uses BagNet [4] architecture.

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 [5] with 10000 queries (Q). FP: Forward pass, F+BP: Forward and Backward passes

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 $\ell_0$ norm bound as well.

6. 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 $\ell_0$ norm bound as well.

7. Acknowledgements

This work was supported by Uchhatar Avishkar Yojana (UAY) project (IISC_10), MHRD, Govt. of India. Srvanti Addepalli is supported by a Google PhD Fellowship.
References


[33] Sukruti Rao, David Stutz, and Bernt Schiele. Adversarial Training against Location-Optimized Adversarial Patches. *ECCV Workshop (CV-COPS)*, 2020. 1, 2, 3, 4


[38] Rey Wiyatno and Anqi Xu. Physical adversarial textures that fool visual object tracking. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2


[40] Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. Defending Against Physically Realizable Attacks on Image Classification. *International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 3, 4


