Local Gradients Smoothing: Defense against localized adversarial attacks

Muzammal Naseer
Australian National University (ANU)
muzammal.naseer@anu.edu.au

Salman H. Khan
Data61, CSIRO
salman.khan@data61.csiro.au

Fatih Porikli
Australian National University (ANU)
fatih.porikli@anu.edu.au

Abstract

Deep neural networks (DNNs) have shown vulnerability to adversarial attacks, i.e., carefully perturbed inputs designed to mislead the network at inference time. Recently introduced localized attacks, Localized and Visible Adversarial Noise (LaVAN) and Adversarial patch, pose a new challenge to deep learning security by adding adversarial noise only within a specific region without affecting the salient objects in an image. Driven by the observation that such attacks introduce concentrated high-frequency changes at a particular image location, we have developed an effective method to estimate noise location in gradient domain and transform those high activation regions caused by adversarial noise in image domain while having minimal effect on the salient object that is important for correct classification. Our proposed Local Gradients Smoothing (LGS) scheme achieves this by regularizing gradients in the estimated noisy region before feeding the image to DNN for inference. We have shown the effectiveness of our method in comparison to other defense methods including Digital Watermarking, JPEG compression, Total Variance Minimization (TVM) and Feature squeezing on ImageNet dataset. In addition, we systematically study the robustness of the proposed defense mechanism against Back Pass Differentiable Approximation (BPDA), a state of the art attack recently developed to break defenses that transform an input sample to minimize the adversarial effect. Compared to other defense mechanisms, LGS is by far the most resistant to BPDA in localized adversarial attack setting.

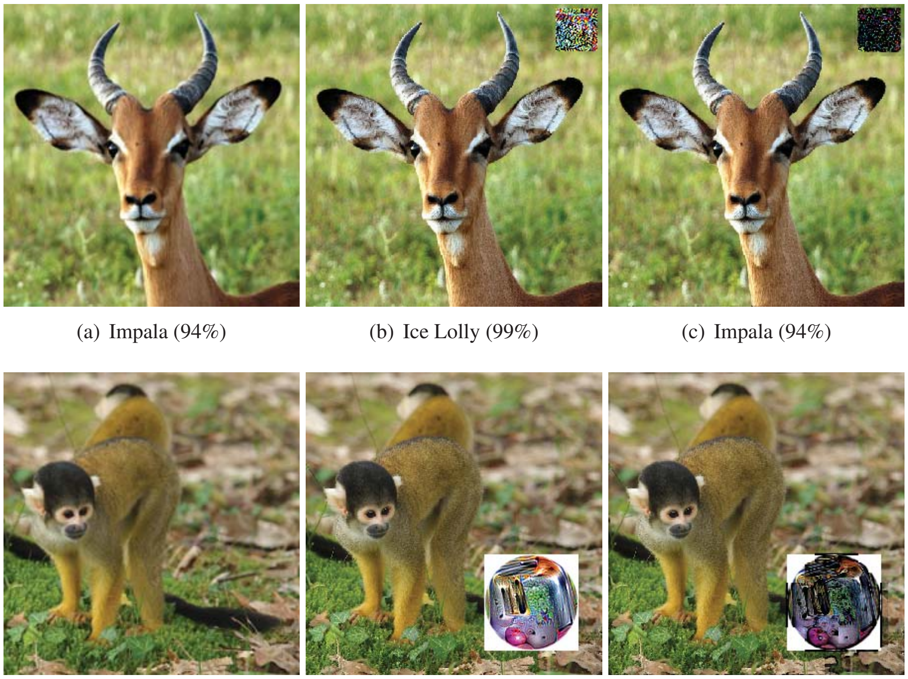
1. Introduction

Deep neural network architectures achieve remarkable performance on critical applications of machine learning including sensitive areas such as face detection [16], malware detection [17] and autonomous driving [11]. However, the vulnerability of DNNs to adversarial examples limit their wide adoption in security critical applications [1]. It has been shown that adversarial examples can be created by minimally modifying the original input samples such that a DNN mis-classifies them with high confidence. DNNs are often criticized as black-box models; adversarial examples raise further concerns by highlighting blind spots of DNNs. At the same time, adversarial phenomena provide an opportunity to understand DNN's behavior to minor perturbations in visual inputs.

Methods that generate adversarial examples either modify each image pixel by a small amount [24, 8, 14, 13] often imperceptible to human vision or few image pixels by a large visible amounts [20, 22, 4, 12, 7]. Pixel attack [22] changes few image pixels, but it requires small images (e.g., 32×32) and does not provide control over noise location. Small noise patches were introduced by [20] in the form of glasses to cover human face to deceive face recognition systems. Similarly, Evtimov *et al*. [7] added noise patches as rectangular patterns on top of traffic signs to cause misclassification. Very recently, localized adversarial attacks, i.e., Adversarial patch [4] and LaVAN [12] have been introduced that can be optimized for triplets (misclassification confidence, target class, perturbed location). These practical attacks have demonstrated high strength and can easily bypass existing defense approaches. Therefore they present a significant challenge for existing deep learning systems.

Contributions: In this work, we study the behavior of localized adversarial attacks and propose an effective mechanism to defend against them (see Fig. 1). LaVAN and Adversarial patch add adversarial noise without affecting the original object in the image, and to some extent, they are complementary to each other. In an effort towards a strong defense against these attacks, this paper contributes as follows:

- Motivated by the observation that localized adversarial attacks introduce high-frequency noise, we pro-



(a) Impala (94%) (b) Ice Lolly (99%) (c) Impala (94%)

(d) Squirrel Monkey (58%) (e) Toaster (91%) (f) Squirrel Monkey (57%)

Figure 1: Inception v3 [23] confidence scores are shown for example images. (a) and (d) represent benign examples from ImageNet [18], (b) and (e) are adversarial examples generated by LaVAN [12] and Adversarial patch [4] respectively, (c) and (f) show transformed adversarial images using our proposed LGS. As illustrated, LGS restores correct class confidences.

pose a transformation called Local Gradient Smoothing (LGS). LGS first estimates region of interest in an image with the highest probability of adversarial noise and then performs gradient smoothing in only those regions.

- We show that by its design, LGS significantly reduces gradient activity in the targeted attack region and thereby showing the most resistance to BPDA [2], an attack specifically designed to bypass transformation based defense mechanisms.
- Our proposed defense outperforms other state-of-the-art methods such as Digital watermarking, TVM, JPEG compression, and Feature squeezing in localized adversarial attacks setting [12, 4].

2. Related Work

Among the recent localized adversarial attacks, the focus of adversarial patch [4] is to create a scene independent physical-world attack that is agnostic to camera angles, lighting conditions and even the type of classifier. The result is an image independent universal noise patch that can be printed and placed in the classifier's field of view in a white box (when deep network model is known) or black box (when deep network model is unknown) setting. However, the size of the adversarial patch should be 10% of the image for the attack to be successful in about 90% cases [12]. This limitation was addressed by Karmoon *et al*. [12], who focused on creating localized attack covering as little as 2% of the image area instead of generating a universal noise patch. In both of these attacks [4, 12], there is no constraint on noise, and it can take any value within image domain, i.e., [0, 255] or [0, 1].

Defense mechanisms against adversarial attacks can be divided into two main categories. (a) Methods that modify DNN by using adversarial training [25] or gradient masking [15] and (b) techniques that modify input sample by using some smoothing function to reduce adversarial effect without changing the DNN [6, 5, 9, 26]. For example, JPEG compression was first presented as a defense by [6] and recently studied extensively by [5, 19]. [26] presented feature squeezing methods including bit depth reduction, median filtering, Gaussian filtering to detect and defend against adversarial attacks. Guo *et al*. [9] considered smoothing input samples by total variance minimization along with JPEG compression and image quilting to reduce the adversarial effect. Our work falls into the second category as we also transform the input sample to defend against localized adversarial attacks. However, as we will demonstrate through experiments, the proposed defense mechanism provides better defense against localized attacks compared to previous techniques.

The paper is organized as follows: Section 3 discusses localized adversarial attacks, LaVAN and Adversarial patch in detail. Section 4 presents our defense approach (LGS) against these attacks. We discuss other related defense methods in Section 5.2. Section 5 demonstrates the effectiveness of the proposed method LGS in comparison to other defense methods against LaVAN and adversarial patch attacks. Section 5.3 discusses BPDA and resilience of different defense methods against it. Section 6 concludes the draft by discussing possible future directions.

3. Adversarial Attacks

In this section, we provide a brief background to adversarial attacks and explain how LaVAN [12] and Adversarial patch [4] are different from traditional attacks.

3.1. Traditional Attacks

The search for adversarial examples can be formulated as a constrained optimization problem. Given a discriminative classifier $\mathcal{F}(\mathbf{y} \mid \mathbf{x})$, an input sample $\mathbf{x} \in \mathbb{R}^n$, a target class $\bar{\mathbf{y}}$ and a perturbation budget ϵ, an attacker seeks to find a modified input $\mathbf{x}' = \mathbf{x} + \delta \in \mathbb{R}^n$ with adversarial noise δ to increase likelihood of the target class $\bar{\mathbf{y}}$ by solving the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{x}'} \ \mathcal{F}(\mathbf{y} = \bar{\mathbf{y}} \mid \mathbf{x}') \\ &\text{subject to: } \ \|\mathbf{x} - \mathbf{x}'\|_p \leq \epsilon \end{aligned} \tag{1}$$

This formulation produces well camouflaged adversarial examples but changes each pixel in the image. Defense methods such as JPEG compression [6, 5], Total variance minimization [9] and Feature squeezing [26] are effective against such attacks especially when the perturbation budget ϵ is not too high.

3.2. LaVAN

LaVAN [12] differs from the formulation presented in Eq. 1 as it confines adversarial noise δ to a small region, usually away from the salient object in an image. It uses the following spatial mask to replace the small area with noise, as opposed to noise addition performed in traditional attacks:

$$\mathbf{x}' = (\mathbf{1} - \mathbf{m}) \odot \mathbf{x} + \mathbf{m} \odot \delta, \quad s.t., \mathbf{m} \in \mathbb{R}^n \text{ and }, \tag{2}$$

where $\odot$ is Hadamard product and δ represents adversarial noise.

They also introduce a new objective function where at each iteration, optimization algorithm takes a step away from the source class and towards the target class simultaneously, as follows:

$$\begin{aligned} &\max_{\mathbf{x}'} \ \mathcal{F}(\bar{\mathbf{y}} \mid \mathbf{x}') - \mathcal{F}(\mathbf{y} \mid \mathbf{x}') \\ &\text{subject to: } \ \|\mathbf{x} - \mathbf{x}'\|_\infty \leq \epsilon, \ \ 0 \leq \epsilon \leq 1, \end{aligned} \tag{3}$$

where $\mathbf{x}'$ is given by Eq. 2.

3.3. Adversarial Patch

Adversarial examples created using the methodology presented in Eq. 1 cannot be used in physical world attacks because adversarial noise loses its effect under different camera angles, rotations and lighting conditions. Athalye *et al.* [3] introduced an Expectation over Transformation (EoT) attack to create robust adversarial examples invariant to chosen set of transformations. Brown *et al.* [4] build upon Athalye's work and used EoT to create a scene independent robust noise patch confined to small region that can be printed and placed in the classifier's field of view to cause misclassification. To generate adversarial patch $\mathbf{p}'$, [4] proposed a patch operator $\mathcal{A}(\mathbf{p}, \mathbf{x}, \mathbf{l}, \mathbf{t})$ for a given image $\mathbf{x}$, patch $\mathbf{p}$, location $\mathbf{l}$ and a set of transformation $\mathbf{t}$. During optimization, patch operator $\mathcal{A}$ applies a set of transformations to the patch $\mathbf{p}$ and then projects it onto the image $\mathbf{x}$ at a location $\mathbf{l}$ to increase likelihood of target class $\bar{\mathbf{y}}$.

$$\mathbf{p}' = \max_{\mathbf{p}} \mathbb{E}_{\mathbf{x} \sim X, \mathbf{t} \sim T, \mathbf{l} \sim L} [\mathcal{F}(\bar{\mathbf{y}} \mid \mathcal{A}(\mathbf{p}, \mathbf{x}, \mathbf{l}, \mathbf{t}))] \tag{4}$$

where X represent training images, T represents distribution over transformations, and L is a distribution over locations in the image.

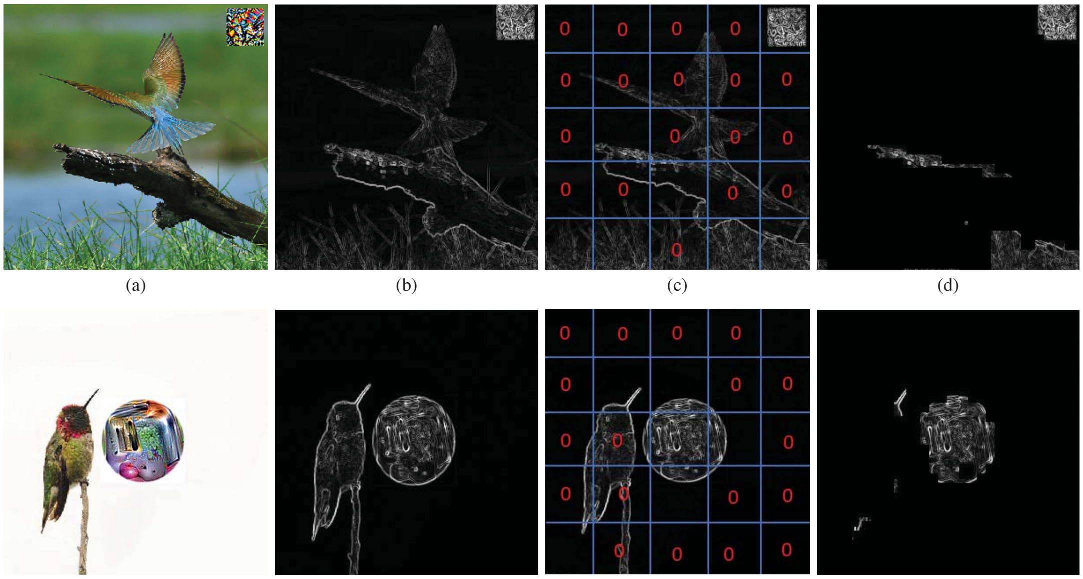
4. Defense: Local Gradients Smoothing

Both of the above discussed attacks [12, 4] introduce high frequency noise concentrated at a particular image location and strength of such a noise becomes very prominent in image gradient domain. We propose that the effect of such adversarial noise can be reduced significantly by suppressing high frequency regions without effecting the low frequency image areas that are important for classification. An efficient way to achieve this is by projecting scaled normalized gradient magnitude map onto the image to directly suppress high activation regions. To this end, we first compute the magnitude of first-order local image gradients as follows:

$$\| \nabla \mathbf{x}(a, b) \| = \sqrt{\left(\frac{\partial \mathbf{x}}{\partial a}\right)^2 + \left(\frac{\partial \mathbf{x}}{\partial b}\right)^2}, \tag{5}$$

where a, b denote the horizontal and vertical directions in the image plane. The range of the gradient magnitude calculated using the above equation is normalized for consistency across an image as follows:

$$g(\mathbf{x}) = \frac{\| \nabla \mathbf{x}(a, b) \| - \| \nabla \mathbf{x}(a, b) \|_{min}}{\| \nabla \mathbf{x}(a, b) \|_{max} - \| \nabla \mathbf{x}(a, b) \|_{min}}. \tag{6}$$



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2: (a) and (e) show adversarial examples generated by LaVAN [12] and adversarial patch [4] respectively, (b) and (f) show normalized gradients magnitude before applying windowing operation to look for highest activation regions, (c) and (g) show concept of window search to estimate noise regions, (d) and (h) show normalized gradients magnitude after applying windowing operation.

The normalized gradient $g(\mathbf{x})$ is projected onto the original image to suppress noisy perturbations in the input data domain. This operation smooths out very high frequency image details. As demonstrated by our evaluations, these regions have high likelihood of being perturbed areas, but they do not provide significant information for final classification. The noise suppression is performed as follows:

$$\mathcal{T}(\mathbf{x}) = \mathbf{x} \odot (1 - \lambda * g(\mathbf{x})), \tag{7}$$

where λ is the smoothing factor for LGS and $\lambda * g(x)$ is clipped between 0 and 1. Applying this operation at a global image level, however, introduces image structural loss that causes a drop in classifier's accuracy on benign examples. To minimize this effect, we design a block-wise approach where gradient intensity is evaluated within a local window. To this end, we first divide the gradient magnitude map into a total of K overlapping blocks of same size (τ) and then filter these blocks based on a threshold (γ) to estimate highest activation regions which also have the highest likelihood of adversarial noise. This step can be represented as follows:

$$\begin{aligned} &\mathbf{g}'_{h,w} = \mathcal{W}(g(\mathbf{x}), h, w, \tau, o) \in \mathbb{R}^{\tau}, \\ &\hat{\mathbf{g}}_{h,w} = \begin{cases} \mathbf{g}'_{h,w}, & \text{if } \frac{1}{|\mathbf{g}'_{h,w}|} \sum_i \sum_j g'_{h,w}(i,j) > \gamma \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{8}$$

where $|\cdot|$ denotes the cardinality of each patch, o denotes the patch overlap, $\mathcal{W}(\cdot)$ represent the windowing operation, h, w denote the vertical and horizontal components of the top left corner of the extracted window, respectively. We set the block size $\tau = 15 \times 15$ with 5×5 overlap and threshold is 0.1 in all of our experiments. The updated gradient blocks represented as $\hat{\mathbf{g}}_{h,w}$ are then collated to recreate the full gradient image: $\bar{\mathbf{g}} = \mathcal{W}^{-1}(\{\hat{\mathbf{g}}_{h,w}\}_1^K)$. Figure 2 shows the effect of windowing search on gradients magnitude maps.

5. Experiments

5.1. Protocol and Results Overview

We used Inception v3 model [23] to experiment with various attack and defense mechanisms in all of our experiments. All attacks are carried out in white-box settings. We consider the validation set available with Imagenet-2012 dataset in our experiments. This set consists of a total of 50k images. We report top-1 accuracy of classifier. Results are summarized in tables 1, 2 and 3.

LaVAN [12] can be optimized for triplets (target, confidence, location) but it is highly sensitive to noise location. Adversary loses its effect with even a small change to the pixel location. To reduce the computational burden and conduct experiments on a large scale, we randomly chose noise location along border areas of the image because they have the least probability to cover the salient object. We ran 1000 iterations of attack optimization per image. We terminate the optimization early if classifier mis-classify with

	No Attack	42x42 noise patch	52x52 noise patch	60x60 noise patch
		covering $\sim$2% of image	covering $\sim$3% of image	covering $\sim$ 4% of image
No Defense	75.61%	11.00%	2.79%	0.78%
LGS [lambda=2.3]	71.05%	**70.90%**	**69.84%**	**69.37%**
LGS [lambda=2.1]	71.50%	70.80%	69.54%	68.56%
LGS [lambda=1.9]	71.84%	70.40%	68.84%	66.98%
LGS [lambda=1.7]	72.30%	69.55%	67.32%	63.38%
LGS [lambda=1.5]	72.72%	67.68%	64.13%	55.67%
DW	52.77%	**67.70%**	**66.19%**	**64.57%**
MF [window=3]	70.59%	**63.90%**	**62.15%**	**59.81%**
GF [window=5]	61.75%	59.52%	57.68%	55.29%
BF [window=5]	65.70%	61.53%	58.70%	55.59%
JPEG [quality=80]	74.35%	18.14%	6.23%	2.06%
JPEG [quality=60]	72.71%	25.69%	11.86%	4.85%
JPEG [quality=40]	71.20%	37.10%	23.26%	12.73%
JPEG [quality=30]	70.04%	45.00%	33.72%	22.04%
JPEG [quality=20]	67.51%	52.84%	46.25%	37.19%
JPEG [quality=10]	60.25%	**53.10%**	**48.73%**	**43.59%**
TMV [weights=10]	70.21%	**14.48%**	**4.64%**	**1.73%**
TMV [weights=20]	72.85%	13.24%	3.78%	1.17%
TMV [weights=30]	73.85%	12.79%	3.53%	1.04%
BR [depth=1]	39.85%	**25.93%**	**15.14%**	**9.73%**
BR [depth=2]	64.61%	16.32%	6.15%	2.68%
BR [depth=3]	72.83%	13.4%	3.89%	1.25%

Table 1: Summary of Inception v3 performance against LaVAN attack on ImageNet validation set with and without defenses including local gradient smoothing (LGS), digital watermarking (DW), median filtering (MF), Gaussian filtering (GF), bilateral filtering (BF), JPEG compression, total variance minimization (TVM) and bit-depth reduction (BR). Bold numbers represent the best accuracy of a certain defense against LAVAN attack.

confidence above than or equal to 99% or we let it run for at max 1000 iterations and attack is considered to be successful if the image label is changed to a random target (not equal to the true object class). Inceptionv3 model accepts 299x299 image as an input. Three adversarial noise masks with size 42x42 ($\sim$2% of the image), 52x52 ($\sim$3% of the image) and 60x60 ($\sim$4% of the image) were applied. Table 1 presents summary of all the results. For the case of adversarial patch [4] attack, placing a patch of size 95x95 (10% of the image) randomly on all Imagenet validation set was not possible because it would cover most of salient objects details in an image. So we carefully created 1000 adversarial examples that model misclassified as a toaster with a confidence score at least 90%. We then applied all the defense techniques and reported results in Table 2. Figure 3 shows runtime of defense methods to process ImageNet [18] validation set. We used optimized python implementations. Specifically, we employed JPEG from Pillow, Total variance minimization (TVM), and Bilateral filtering (BF) from scikit-image, Median filtering (MF) and Gaussian filtering (GF) from scipy, and LGS and Bit Depth Reduction (BR) are written in python 3.6 as well. All experiments were conducted on desktop windows computer equipped with Intel i7-7700k quad-core CPU clocked at 4.20GHz and 32GB RAM.

Defense	None	LGS	DW	MF	JPEG	TVM	BR
Adversarial Patch	0%	**90.5%**	80%	49.10%	45%	1%	0%

Table 2: Accuracy of Inception v3 against adversarial patch attack with and without defense. The size of adversarial noise is 95x95 covering $\sim$10% of image. LGS is used with $\lambda = 2.3$, DW in blind defense scenario, MF with window equal to 3, JPEG compression with quality equal to 30, TVM with weights equal to 10 and BR with depth 3. This hyperparameter choice was made for fair comparison such that the performance on benign examples from ImageNet is approximately the same (first column of Table 1). Results are reported for 1000 adversarial examples misclassified as toaster with confidence above than 90%.

5.2. Comparison with Related Defenses

In this section, we report comparisons of our approach with other recent defense methods that transform the input sample to successfully reduce the adversarial effect. The compared methods include both global and local techniques. Note that our method processes image locally so it has advantage over other defenses like JPEG, MF, TVM and BR that process image globally. First, we provide a brief de-

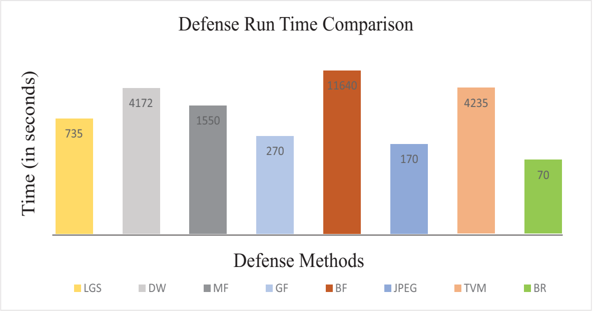


Figure 3: Computational cost comparison of defense methods to process 50k ImageNet validation images. Graph is shown in log scale for better visualization with actual processing times written on the top of each bar in seconds.

scription of the competing defenses which will allow us to elaborate further on the performance trends in Tables 1, 2 and 3.

5.2.1 Digital Watermarking

Hayes et.al [10] presented two, non-blind and blind, defense strategies to tackle the challenge of localized attacks [12, 4]. Non-blind defense considers a scenario, where defender has the knowledge of adversarial mask location. This is unlikely scenario in the context of adversarial attacks because threat is over immediately, once the adversary provides the mask location. Localized attacks have the ability to change the attention of classifier from the original object to adversarial mask. In their blind defense, authors [10] exploited the attention mechanism by first finding the mask location using saliency map and then processing that area before inference. Using saliency map to detect adversarial mask location is the strength of this defense but at the same time its also the weakness of defense because on benign examples, saliency map will give the location of original object and hence processing original object will decrease the performance on clean examples. Authors [10] reported blind defense performance to protect VGG19 [21] on only 400 randomly selected images with 12% accuracy drop on clean images. We have tested this defense on imagenet validation set [18] (50k images). This method has the second best accuracy on adversarial examples after LGS but its accuracy on clean examples expectedly dropped by a large margin (22.8%). Tables 1, 2 and 3 summarizes the performance of digital watermarking [10].

5.2.2 JPEG Compression

[6, 5, 19] extensively studied JPEG compression to defend against adversarial attacks. This way high-frequency components are removed that are less important to human vision by using Discrete Cosine Transform (DCT). JPEG performs compression as follows:

- Convert an RGB image YC_bC_r color space, where Y and C_b, C_r represent luminance and chrominance respectively.
- Down-sample the chrominance channels and apply DCT to 8×8 blocks for each channel.
- Perform quantization of frequency amplitudes by dividing with a constant and rounding off to the nearest integer.

As illustrated in Table 1, image quality decreases as the degree of compression increases which in turn decreases accuracy on benign examples. JPEG compression is not very effective against localized attacks, and its defending ability decreases a lot against BPDA. JPEG performance comparison is shown in Tables 1, 2 and 3 and Figure 4.

5.2.3 Feature Squeezing

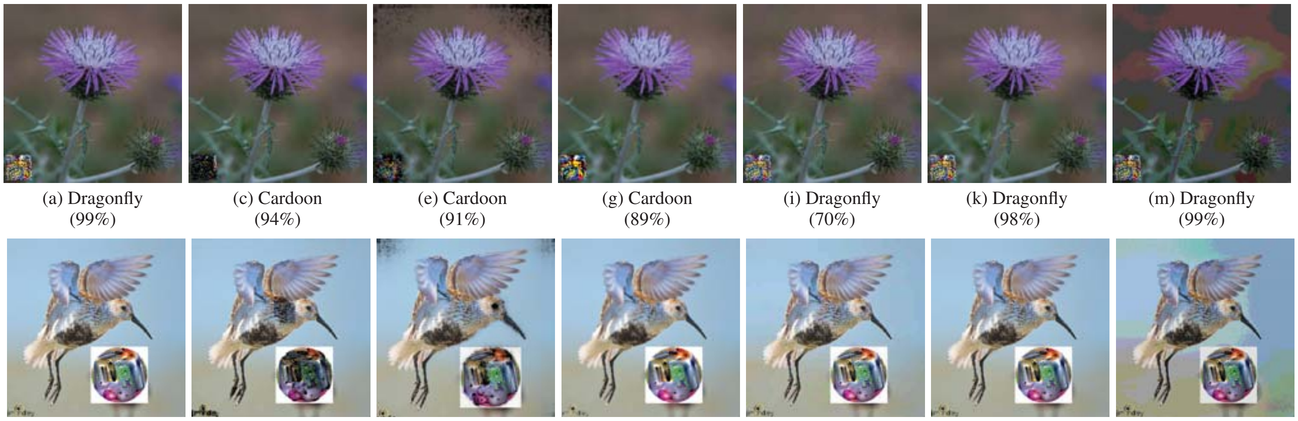
The main idea of feature squeezing [26] is to limit the explorable adversarial space by reducing resolution either by using bit depth reduction or smoothing filters. We found that bit reduction is not effective against localized attacks, however smoothing filter including Gaussian filter, median filter, and bilateral filter reduces localized adversarial effect with reasonable accuracy drop on benign examples. Among smoothing filters, median filter outperforms Gaussian and bilateral filters. Feature squeezing performance is shown in Tables 1, 2 and 3 and Figure 4.

5.2.4 Total Variance Minimization (TVM)

Guo *et al*. [9] considered smoothing adversarial images using TVM along with JPEG compression and image quilting. TVM has the ability to measure small variations in the image, and hence it proved effective in removing small perturbations. As illustrated in Table 1, TVM becomes ineffective against large concentrated variations introduced by the localized attacks. Further comparisons are shown in Tables 2 and 3 and Figure 4.

5.3. Resilience to BPDA

BPDA [2] is built on the intuition that transformed images by JPEG or TVM should look similar to original images, that is, $\mathcal{T}(x) \approx x$. BPDA approximate gradients for non-differentiable operators with combined forward propagation through operator and DNN while ignoring operator during the backward pass. This strategy allows BPDA to approximate true gradients and thus bypassing the defense. In the traditional attack setting like Projected Gradient Descent (PGD) [13], the explorable space available to BPDA



(a) Dragonfly (99%)
(c) Cardoon (94%)
(e) Cardoon (91%)
(g) Cardoon (89%)
(i) Dragonfly (70%)
(k) Dragonfly (98%)
(m) Dragonfly (99%)
(b) Toaster (94%)
(d) Sandpiper (89%)
(f) Sandpiper (45%)
(h) Sandpiper (55%)
(j) Sandpiper (28%)
(l) Toaster (90%)
(n) Toaster (92%)

Figure 4: Inception v3 confidence score is shown on example images. (a,b) represent adversarial examples generated by LaVAN and adversarial patch respectively, (c,d) show transformed adversarial images using LGS with lambda equal to 2.3 respectively, (e,f) show transformed adversarial images using DW processing method respectively, (g,h) show transformed adversarial images using median filter with window size 3 respectively, (i,j) show transformed adversarial images using JPEG with quality 30 respectively, (k,l) show transformed adversarial images using TVM with weights equal to 10 respectively, and (m,n) show transformed adversarial images using BR with depth 3.

is $\mathbb{R}^n$ because it can change each pixel in the image. In localized attack setting explorable space reduces to $\mathbb{R}^{m<<n}$ controlled by the mask size. LGS suppresses the high-frequency noise to near zero thereby significantly reducing gradient activity in the estimated mask area and restricting BPDA to bypass defense. However, as it is the case with all defenses, increasing explorable space, i.e., distance limit in PGD attack [13] and mask size in the case of localized attack [12], protection ability of defense methods decreases. To test performance against BPDA in the localized setting, we randomly selected 1000 examples from Imagenet and attack is optimized against all defenses for the same target, location, mask size and number of iterations. Compared to other defenses methods, LGS significantly reduces the explorable space for localized adversarial attacks within mask size equal to $\sim 2\%$ of the image as discussed in [12]. In the case of DW [10] defense, we tested BPDA against the proposed input processing given the mask location. Summary of attack success rate against defense methods is presented in Table 3.

Defense	None	LGS	DW	MF	JPEG	TVM	BR
LaVAN with BPDA	88%	**18%**	25.6%	75%	73.30%	78.10%	83%

Table 3: Attack success rate against Inception v3 with and without defense (lower is better). The size of adversarial noise 42x42 covering ∼2% of image. LGS is used with $\lambda = 2.3$, DW in blind scenario, MF with window equal to 3, JPEG compression with quality equal to 30, TVM with weights equal to 10 and BR with depth 3. This hyperparameter choice was made for fair comparison such that the performance on benign examples from ImageNet is approximately the same (first column of Table 1). Attack is optimized for 1000 randomly selected images for the same target, location and mask size.

6. Discussion and Conclusion

In this work, we developed a defense against localized adversarial attacks by studying attack properties in gradient domain. Defending against continuously evolving adversarial attacks has proven to be very difficult especially with standalone defenses. We believe that in critical security applications, a classifier should be replaced by a robust classification system with following main decision stages:

- Detection: given the unlimited distortion space, any image can be converted into an adversarial example that can bypass any defense system with 100% success rate [2]; however, this also pushes the adversarial example away from the data manifold, and it would be easier to detect rather than removing the perturbation.
- Projection or Transformation: Adversarial examples within a small distance of original images can be either projected onto the data manifold or transformed to mitigate the adversarial effect.
- Classification: Final stage should be to perform a forward pass through a DNN, whose robustness is increased via adversarial training.

Our method performs a transformation, so it falls into the second stage of robust classification systems. LGS outperforms digital watermarking, JPEG compression, feature squeezing and TVM against localized adversarial attacks with minimal drop in accuracy on benign examples. LGS can be used with a combination of other defense methods, for example, smoothing filters like low pass filter can be applied just on the estimated noisy region to enhance protection for a DNN further.

References

[1] N. Akhtar and A. S. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

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

[3] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning (ICML)*, 2017.

[4] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. Adversarial patch. In *Neural Information Processing Systems (NIPS)*, 2017.

[5] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, S. Li, L. Chen, M. E. Kounavis, and D. H. Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *Knowledge Discovery and Data Mining (KDD)*, 2018.

[6] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy. A study of the effect of jpg compression on adversarial images. In *International Society for Bayesian Analysis (ISBA)*, 2016.

[7] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song. Robust physical-world attacks on machine learning models. In *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICRL)*, 2015.

[9] C. Guo, M. Rana, M. Cissé, and L. van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICRL)*, 2017.

[10] J. Hayes. On visible adversarial perturbations & digital watermarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1597–1604, 2018.

[11] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, et al. An empirical evaluation of deep learning on highway driving. *arXiv preprint arXiv:1504.01716*, 2015.

[12] D. Karmon, D. Zoran, and Y. Goldberg. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning (ICML)*, 2018.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICRL)*, 2017.

[14] S. M. Moosavi Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number EPFL-CONF-218057, 2016.

[15] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016.

[16] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *British Machine Vision Conference(BMVC)*, volume 1, page 6, 2015.

[17] R. Ronen, M. Radu, C. Feuerstein, E. Yom-Tov, and M. Ahmadi. Microsoft malware classification challenge. *arXiv preprint arXiv:1802.10135*, 2018.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[19] U. Shaham, J. Garritano, Y. Yamada, E. Weinberger, A. Cloninger, X. Cheng, K. Stanton, and Y. Kluger. Defending against adversarial images using basis functions transformations. *arXiv preprint arXiv:1803.10840*, 2018.

[20] M. Sharif, S. Bhagavatula, L. Bauer, and M. 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*, pages 1528–1540. ACM, 2016.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[22] J. Su, D. V. Vargas, and S. Kouichi. One pixel attack for fooling deep neural networks. *arXiv preprint arXiv:1710.08864*, 2017.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[24] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICRL)*, 2014.

[25] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICRL)*, 2018.

[26] W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.