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., $32 \times 32$) 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-
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 pro-
vides 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 \( F(y | x) \), an input sample \( x \in \mathbb{R}^n \), a target class \( y \) and a perturbation budget \( \epsilon \), an attacker seeks to find a modified input \( x' = x + \delta \in \mathbb{R}^n \) with adversarial noise \( \delta \) to increase likelihood of the target class \( \bar{y} \) by solving the following optimization problem:

\[
\max_{x'} F(y = \bar{y} | x') \\
\text{subject to: } \|x - x'\|_p \leq \epsilon
\]  

(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 \( \epsilon \) is not too high.

3.2. LaVAN

LaVAN [12] differs from the formulation presented in Eq. 1 as it confines adversarial noise \( \delta \) 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:

\[
x' = (1 - m) \odot x + m \odot \delta, \quad s.t., \; m \in \mathbb{R}^n \quad \text{and},
\]

(2)

where \( \odot \) is Hadamard product and \( \delta \) 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:

\[
\max_{x'} F(\bar{y} | x') - F(y | x') \\
\text{subject to: } \|x - x'\|_\infty \leq \epsilon, \; 0 \leq \epsilon \leq 1,
\]

(3)

where \( x' \) 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 \( p' \), [4] proposed a patch operator \( A(p, x, l, t) \) for a given image \( x \), patch \( p \), location \( l \) and a set of transformation \( t \). During optimization, patch operator \( A \) applies a set of transformations to the patch \( p \) and then projects it onto the image \( x \) at a location \( l \) to increase likelihood of target class \( \bar{y} \).

\[
p' = \max_p E_{x \sim X, t \sim T, l \sim L}[F(\bar{y} | A(p, x, l, t))] \quad (4)
\]

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:

\[
\| \nabla x(a, b) \| = \sqrt{\left( \frac{\partial x}{\partial a} \right)^2 + \left( \frac{\partial x}{\partial b} \right)^2},
\]

(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:

\[
g(x) = \frac{\| \nabla x(a, b) \| - \| \nabla x(a, b) \|_{\min}}{\| \nabla x(a, b) \|_{\max} - \| \nabla x(a, b) \|_{\min}}.
\]

(6)
The normalized gradient $g(x)$ 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:

$$T(x) = x \odot (1 - \lambda \ast g(x)),$$

where $\lambda$ is the smoothing factor for LGS and $\lambda \ast g(x)$ 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 ($\tau$) and then filter these blocks based on a threshold ($\gamma$) to estimate highest activation regions which also have the highest likelihood of adversarial noise. This step can be represented as follows:

$$g'_{h,w} = W(g(x), h, w, \tau, o) \in \mathbb{R}^\tau,$$

$$\hat{g}_{h,w} = \begin{cases} g'_{h,w}, & \frac{1}{|g'_{h,w}|} \sum_{i} \sum_{j} g'_{h,w}(i, j) > \gamma \\ 0, & \text{otherwise}. \end{cases}$$

where $| \cdot |$ denotes the cardinality of each patch, $o$ denotes the patch overlap, $W(\cdot)$ 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 $5 \times 5$ overlap and threshold is 0.1 in all of our experiments. The updated gradient blocks represented as $\hat{g}_{h,w}$ are then collated to recreate the full gradient image: $g = W^{-1}(\{\hat{g}_{h,w}\})$. Figure 2 shows the effect of windowing search on gradients magnitude maps. We further demonstrated LGS efficiency on challenging images in supplementary material.

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
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 LA V AN attack.

Table 2: Accuracy of Inception v3 against adversarial patch attack with and without defense. The size of adversarial noise is 95x95 covering ~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 description 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 \( 8 \times 8 \) 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, \( T(x) \approx x \). BPDA approximate gradients for non-differentiable operators with combined forward propagation through operator and DNN while ignoring operator
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.

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

<table>
<thead>
<tr>
<th>Defense</th>
<th>None</th>
<th>LGS</th>
<th>DW</th>
<th>MF</th>
<th>JPEG</th>
<th>TVM</th>
<th>BR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LaVAN with BPDA</td>
<td>88%</td>
<td>18%</td>
<td>25.6%</td>
<td>75%</td>
<td>73.30%</td>
<td>78.10%</td>
<td>83%</td>
</tr>
</tbody>
</table>

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.
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**


