Title

Exploiting Human Perception for Adversarial Attacks

Author
Quan, Pengrui

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Exploiting Human Perception for Adversarial Attacks

A dissertation submitted in partial satisfaction

of the requirements for the degree

Master of Science in Electrical and Computer Engineering

by

Pengrui Quan

2020

ABSTRACT OF THE DISSERTATION

Exploiting Human Perception for Adversarial Attacks

by

Pengrui Quan
Master of Science in Electrical and Computer Engineering
University of California, Los Angeles, 2020
Professor Mani B. Srivastava, Chair

There has been a significant amount of recent work towards fooling deep-learning-based classifiers, particularly for images, via adversarial inputs that are perceptually similar to benign examples. However, researchers typically use minimization of the L_p-norm as a proxy for imperceptibility, an approach that oversimplifies the complexity of real-world images and human visual perception. We exploit the relationship between image features and human perception to propose a *Perceptual Loss (PL)* metric to better capture human imperceptibly during the generation of adversarial images. By focusing on human perceptible distortion of image features, the metric yields better visual quality adversarial images as our experiments validate. Our results also demonstrate the effectiveness and efficiency of our algorithm.

The dissertation of Pengrui Quan is approved.

Cho-Jui Hsieh

Jonathan Kao

Mani B. Srivastava, Committee Chair

University of California, Los Angeles

2020

To my parents . . .

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my sincere gratitude towards my advisor, Prof. Mani B. Srivastava, without whom this work wouldn't have been possible. I want to express my sincere appreciation to his encouragement and guidance to me so that I am brave enough to explore different challenging research aspects.

Then, I would love to thank the following scholars I have worked with, including Prof. Chih-Jen Lin from National Taiwan University, Prof. Thierry Blu from The Chinese University of Hong Kong, and Prof. Cho-Jui Hsieh at UCLA. My studies and research outcomes wouldn't have been the same without their thoughtful mentorship.

Third, I am grateful to my colleagues and friends who provide me with academic assistance and moral support. I would like to especially thank my labmates at Network and Embedded System Laboratory (NESL) for their inspiration and companion. Besides, my friends Xinkai Zhou, Gongze Cao, Ruiming Guo, Qiming Gu, Ray Lin, Qiuyang Yue, Weixi Feng, and Miaoqing Chen, who have been an integral part of my time at UCLA, also deserve my sincere gratitude. It is my great fortune to share my happiness with them.

Besides, I would like to give my deepest gratitude to my parents. Their unconditional love, guidance, and encouragement have always been the support for my whole life. I couldn't be thankful enough for their love during my entire lifetime.

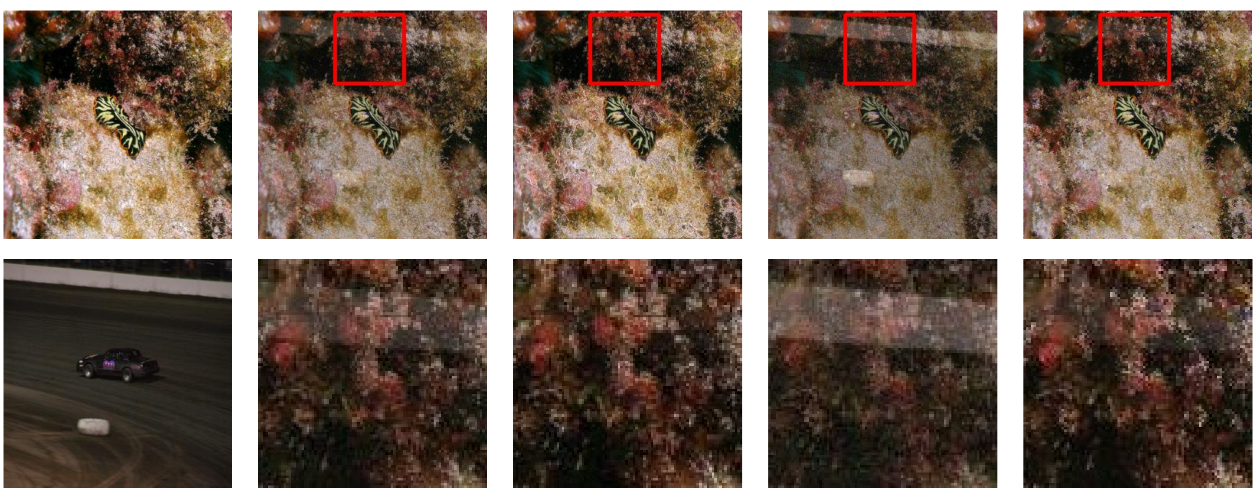
Finally, I would like to thank my collaborator, Ruiming Guo, for his contribution to image feature classification in Chapter 4 and helpful discussions. I also want to acknowledge the funding agencies - DAIS ITA that sponsored this work.

CHAPTER 1

Introduction

1.1 Vulnerability of deep learning model

It has been widely observed that deep neural networks are susceptible to adversarial inputs ([SZS13], [GSS14], [ASC19]). For instance, with a small perturbation added to images, the image classifiers make completely wrong decision ([GSS14], [BRB17]). What is worse, with full structures of deep neural networks exposed to attackers ([CW17]), the attack is easy to perform, and the models can even be forced to make with inconspicuous perturbation.



$\mathbf{x}^{org}$ and $\mathbf{x}^{target}$ | L_2+BA | PL+BA | L_2+SignOPT | PL+SignOPT

Figure 1.1: Visual comparison at 20k iteration. From left to right, first row: original image and adversarial images generated using L_2 + Boundary Attack ([BRB17]), PL (ours) + Boundary Attack, L_2 + Sign-OPT ([CSC19]), and PL (ours) + Sign-OPT. Second row: image in the targeted class and zoomed patches of each adversarial image. Our methods (the third column and the fifth column) can visibly suppress the ghosting effects with the same number of queries.

1.2 Rethinking adversarial attack

In many attack scenarios, if fooling an AI model has been claimed, the adversarial input should be subjected to the following main requirements at the same time: i) Deceptive: the prediction of the perturbed inputs should be modified. ii) Feasible: In many cases or real-world attack situations, the number of queries to the deep learning model is limited and the gradient information and even the output logistic or probability are even not exposed to the attacker. iii) Inconspicuous: Images with unnatural artifacts can be detected by statistical test and can be sent to human inspection ([MGF17], [GMP17], [MC17]). The unnatural property can also be used to defend against adversarial attacks by mapping the adversarial inputs to the natural image space ([GRC17], [TKP17]) and hence, those adversarial attacks can be relatively easy to defend. Therefore, adversarial images with highly perceivable perturbations may not be as destructive as inconspicuous one.

Consider the input image $\mathbf{x} \in \mathbb{R}^N$, where $N = 3 \times W \times H$ is the size of images. The hard-label classifier gives its predicted label y, where $y \in \{1, ..., C\}$. Currently, given the original image $\mathbf{x}^{org}$, its ground-truth label y^{org}, a target class $t \neq y^{org}$ and adversarial image $\mathbf{x}^{adv}$, one of the commonly used methods for generating an adversarial image is to minimize the L_2-distance:

$$\underset{x^{adv} \in \mathbb{R}^N}{\text{minimize}} \qquad \frac{1}{N} \|\mathbf{x}^{adv} - \mathbf{x}^{org}\|_2^2 \tag{1.1a}$$

$$\text{subject to} \qquad f(\mathbf{x}^{adv}) = t \tag{1.1b}$$

By minimizing the L_2-distance, the attackers intend to make the adversarial images inconspicuous from the human perspective such that the harm cannot be easily prevented ([GMP17], [MGF17]). However, we also raise the question about the validity of L_2: is the inconspicuousness automatically satisfied in current adversarial attack tasks by merely minimizing the L_2-norm? To better study the human perception question, we also conduct a subjective test on various adversarial images which will be discussed in Chapter 2.

1.3 Contribution

We summarize the contribution of this thesis as follows:

- Subjective test: We conduct a human perception assessment to study the effectiveness of using L_p-norm in the adversarial attacks and demonstrate certain limitations of these metrics. These data and statistics may serve as the research material for the vision community in the future.

- Better visual quality: with the same amount of resources utilized, we are able to achieve better perceptual quality. In the experiment, we incorporate the *Perceptual Loss (PL)* in the hard-label black-box attack setting, which is a practical attack scenario.

- Novel feature distortion metric: *Perceptual Loss (PL)* is a metric of low-level image feature distortion based on human perception. It is adaptive to image context and does not rely on optimization methods.

- Robust classifier of low-level features: edge, texture, and smooth area. The classification is unsupervised with pixel-level distinction.

CHAPTER 2

Human perception assessment

2.1 Experiment setup

We use Amazon Mechanical Turk to conduct the human perception assessment, where we recruited 165 subjects. We designed a website that serves as the user interface where users are requested to evaluate adversarial images one by one. Some of the presented images are the original ones as captured by a camera, and others have noise added to them to fool an AI algorithm. Subjects were asked to give their opinions by answering whether the image has been perturbed by an adversary to fool an AI algorithm. In this human perception evaluation, we generated adversarial examples of 20 images using several different methods. Firstly, we use L_0, L_{inf}, and L_2 as the objective function and optimizing them with genetic algorithm ([ASC19]). The reason we choose to use the genetic algorithm in [ASC19] is that it is a zero-order optimization approach that is easy to implement and adaptive to various optimization problems. Besides, we also follow the setting in Boundary Attack ([BRB17]) to optimize L_2 objective function. We sampled adversarial images at different iterations, and each time, the subject was asked to evaluate one image by answering whether the image is perturbed. We asked 165 subjects to evaluate benign images and adversarial images at different iterations obtained by the four different combinations of attack methods and attack objectives. Each subject can see 24 to 25 images. There are 800 adversarial examples, and 20 benign images in total and each of them is evaluated by five times.

2.2 Are these attacks really imperceptible?

Denote that $p \in \mathcal{P}$ where p is a specific attack, i.e., GenAttack + L_0 and $\mathcal{P}$ is the set of the attack methods. Explicitly, $\mathcal{P} = \{\text{GenAttack} + L_0, \text{GenAttack} + L_2, \text{GenAttack} + L_{inf}, \text{Boundary Attack} + L_2\}$. The ratio of a method r_p^t at a specific iteration t was calculated by:

$$r_p^t = \frac{1}{5|\mathcal{S}_p|}\sum_{j=1}^{5}\sum_{i\in\mathcal{S}_p}\mathbf{1}\{\mathbf{x}_i^t \text{ is considered as adversarial at iteration t evaluated at time j}\} \tag{2.1}$$

where $\mathbf{1}\{\cdot\}$ is an indicator function, j denotes that the time when the image $\mathbf{x}_i^t$ is evaluated, t denotes the iteration where we sample adversarial images, and $\mathcal{S}_p$ denotes the set of adversarial images generate by attack method p. In the following figures, we will plot how the ratio r_p^t changes with respect to iterations. We treat the ratio r_p^t as a proxy of the human perception on adversarial images.

Typically, we also calculate the ratio $\hat{r}$ denoting how human perceive the benign images:

$$\hat{r} = \frac{1}{5|\hat{\mathcal{S}}|}\sum_{j=1}^{5}\sum_{i\in\hat{\mathcal{S}}}\mathbf{1}\{\mathbf{x}_i \text{ is considered as adversarial evaluated at time j}\} \tag{2.2}$$

where $\hat{\mathcal{S}}$ denotes the set of benign images. Therefore, by comparing how r_p^t changes across different iteration t with $\hat{r}$, we approximate how human perception changes with respect to iterations. Note that $\hat{r}$ will remain constant across every iteration since the benign images do not change during the optimization. Therefore, it plays the role of an indication of whether the adversarial image is indistinguishable from the benign images.

From Figure. 2.5, 2.6, and 2.7, we can have several observations: i) Using L_p-norm as the objective can generally improve the human perception quality. As we can observed from the figures, the ratio r_p^t is generally increasing. ii) However, when we consider the preference of human, L_p-norm cannot always represents it. For instance, from Figure. 2.6 we know that in terms of L_2-norm, GenAttack + L_2 is significantly better than Boundary Attack + L_2. But from the human perspective, they are comparable across iterations.

From the human perception evaluation, we can conclude that commonly use L_p distance is valid but not perfect. Minimizing the L_p-norm used in the above attack methods can generally improve image quality. However, it cannot completely represent the preference of human perceptions. From Fig. 2.6 we can verify that even though GenAttack+L_2 significantly outperforms Boundary Attack + L_2 and other optimization methods in L_2 distance, their perceptual quality does not always give the same results. If we merely treat the L_2 distance as the proxy of human perception, the GenAttack+L_2 should have behaved the best among the four attack methods in human perception. But clearly from Figure. 2.5, this is not the case. Therefore, if the goal of attackers is to make injected noise imperceptible, spending resources in minimizing L_p-norm may not always be the optimal choice. This claim is also supported by the user studies of adversarial images in [SBR18] and [SZM19], where they demonstrate certain mismatches between L_p-norm and human perception. In the following chapters, we will discuss an alternative *Perceptual Loss* metric to better capture the relationship of human perception and perturbation using our novel image feature classification algorithm.

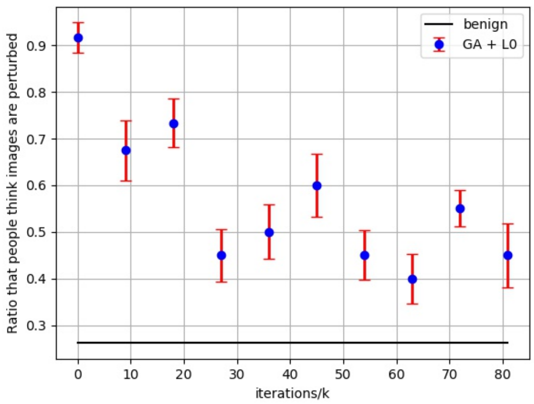


Figure 2.1: GenAttack + L_0

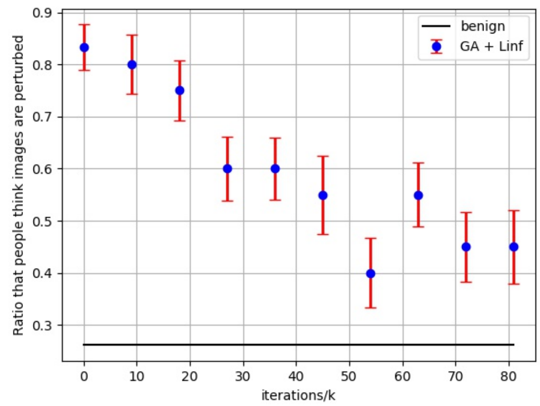


Figure 2.2: GenAttack + L_{inf}

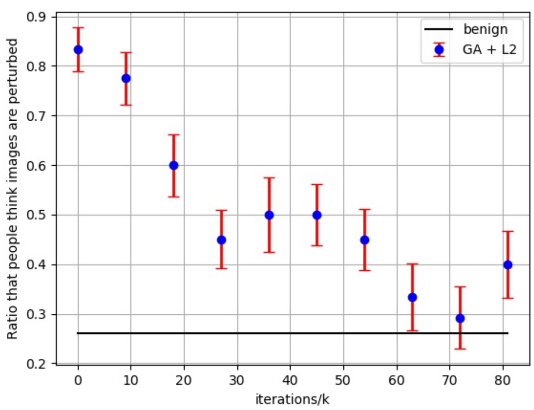


Figure 2.3: GenAttack + L_2

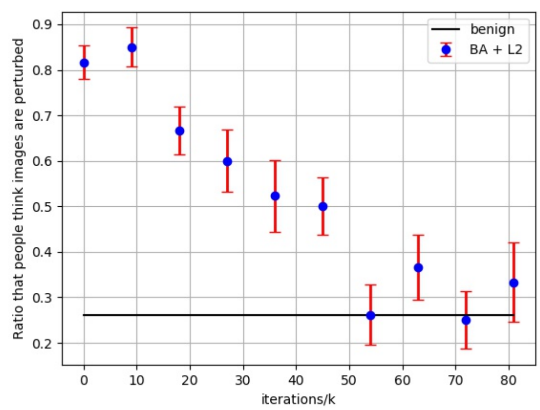


Figure 2.4: Boundary Attack + L_2

Figure 2.5: Results of perception assessment: ratio that human think the given image is adversarial. The black horizontal line denotes the ratio of benign images. Note that from the human perspective, Boundary Attack + L_2 performs the best among the four methods since when iteration becomes larger than 50k, subjects cannot almost distinguish the adversarial images from the benign images (Fig. 2.4).

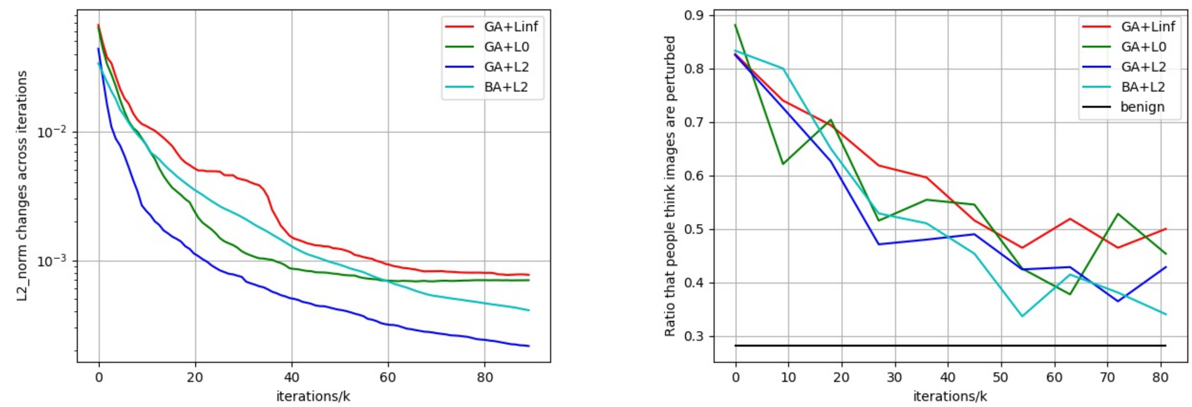


Figure 2.6: Results of perception assessment. **Left**: L_2-norm changes v.s. iterations. **Right**: Ratio that human recognize the image as an adversary; BA: Boundary Attack; GA: GenAttack. According to the L_2 criterion, GenAttack + L_2 is the best, and the Boundary + L_2 and GenAttack + L_0 are comparable, which are not the case of human perception as is reflected in the right.

original original GA + L_2 GA + L_2 BA + L_2 BA + L_2

Figure 2.7: Mismatch between L_2-norm and visual quality. From left to right: the first pair: original image; the second pair: GenAttack (L_2-distance: $2.8e{-}4$); the third pair: Boundary Attack (L_2-distance: $7.8e{-}4$). The second pair contains color distortion at the neck of the squirrel even though its L_2-distance is smaller. Therefore, it implies that L_2 cannot always give the best representation of human perception.

CHAPTER 3

Related work

Adversarial attack Some attack methods consider the white-box setting, where the classifier is completely exposed to the attackers. Among them, C&W attack [CW17] reformulate the objective function into an unconstrained optimization problem using the logistic outputs of classifiers. Besides, [CZS17] and [ASC19] consider the black-box scenario where only the output logistic or classification probability is unknown to the attacker. Furthermore, [IEA18], [BRB17], [CLC18], and [CSC19] consider more extreme cases, hard-label black-box attack, where only the top-1 or top-k hard label is given to the attackers.

Objective function [SBR18] demonstrate that L_p-norm may not be the optimal measurement in adversarial setting by conducting user studies. They study human perception by asking humans to predict the ground-truth label of corrupted images. Moreover, a recent work [SZM19] asked subjects to point out the perturbed image when they think it became just noticeably different from the original image. This claim is also supported by the user studies of adversarial images in [SBR18] and [SZM19], where they demonstrate certain mismatches between L_p-norm and human perception.

The *Perceptual Loss* metrics have been used in many areas, such as audio codec and image processing, to capture the properties of human perception. Leveraging the same ideas, there are some existing works trying to generate adversarial images to improve visual quality. Some methods strive to generate adversarial inputs by shifting the color space ([HP18]), performing geometric transform of the original image ([ETS17]), or using generative models learned from the data manifold ([ZDS17]). These methods may produce adversarial examples with high distance to the original images as denoted by L_p-norm. Furthermore, other methods are trying to improve upon the L_p distance: [CH19] combine L_0 and L_∞ to produce sparser and less perceivable noise. [GMP19] aims to preserve the perceptual quality by maximizing the *Structural Similarity* (SSIM) between original images and adversarial images. [ZAF19] further includes the smoothness penalty into the objective function to smooth noise on the flat areas of the input image using *Laplacian Smoothing*. Besides, [ZLL19] propose an objective function of color distance in CIELCH space to reduce visible artifacts. Among them, [LLW18] is probably the most similar to ours: they computed local variance and tried to perturb pixels at high variance zones. However, they treat features equally across images and only perform attacks in white-box settings. In this work, we propose a more adaptive and accurate metric that is closely connected to human perception and demonstrate an improved visual quality in realistic cases, such as the hard-label black-box attack scenario.

CHAPTER 4

Background & algorithm

In this chapter, we mainly focus on the relationship between human perception and image contexts. We first list the ingredients that influence the human perceptual system and then design a novel pixel-wise Fourier-Argand (FA) classifier to discriminate the image samples based on the perceptual sensitivity adaptively. Then we demonstrate that the optimality and high efficiency of the FA classifier theoretically ensure the reliability and effectiveness of the proposed framework. Exploiting the feature classifier, we finally propose a novel perceptual loss and develop an efficient adversarial attack algorithm.

4.1 Human perceptual system

The human visual evaluation mechanism is a quite sophisticated system related to many aspects, e.g., image resolution, object types, image contrast, etc ([WSL19]). In the image and video processing community, one widely accepted conclusion is that the human evaluation mechanism largely depends on the frequency domain characteristics of the image ([DD90]). For instance, people leverage frequency sensitivity in JPEG image format where $10:1$ compression is achieved with little perceptible loss in image quality ([Hai92], [HLN18]).

However, based on the Parseval's theorem (4.1), we know that minimizing L_2 distance is equivalent to penalizing frequency components $X[k]$ with equal importance:

$$\sum_{n=0}^{N-1} x[n]^2 = \frac{1}{N}\sum_{k=0}^{N-1} |X[k]|^2 \tag{4.1}$$

Therefore, the widely used L_2 norm process the image frequency components indiscriminately, which is far from the truth of visual perception. This motivates us to develop a metric that can better characterize the human perceptual evaluation. Since the perceptual evaluation is subjective and variables affecting visual quality are complicated, we provide a perspective that explicitly characterizes human perception by exploiting low-level image features (e.g., edges, textures, etc.). The motivation behind is that, although the precise characterization of the human perceptual evaluation is infeasible, the link between the visual sensitivity of the human eye and image features is relatively clear and well-studied ([BSD09], [RBM19], [PIM08], [DHL15]). Basically, given an image, people tend to be more sensitive to the low-variation features instead of edges and ridges ([DD90]). Generally speaking, people usually divide the low-level features of the image into three categories: smooth areas, edges, and textures. This actually provides a clue to quantify the perceptual distortion of an adversarial image: compute the feature distortion based on the visual sensitivity, which is the core idea of this work.

4.2 Refined feature classification

As is mentioned above, we divide the low-level features of the image into three categories: smooth areas, edges, and textures. Hence, we propose a novel low-level feature classification (FA classifier) based on the recent Fourier-Argand (FA) filter ([ZB20]). The reason for not using traditional edge detectors or deep-learning based approaches is that general edge detectors can only handle the classification of low-frequency and high-frequency components, and the deep-learning approaches are not robust due to the diversity of the real images. In contrast, we demonstrate that the optimality and high-efficiency of FA classifier guarantee its accuracy and robustness. In the first step, we use the FA filter to distinguish smooth areas and fast-variation features based on the response. Then, we further discriminate between the edges and textures through the feature spatial sparsity and direction.

In essence, low-level features characterize the local directionality of the image, e.g., edges are usually unidirectional, while textures are usually multidirectional. This requires that the edge-detector used should be able to capture all possible directional changes, i.e., $[0, 2\pi)$. However, it is very difficult to balance the accuracy and complexity: either we achieve finer

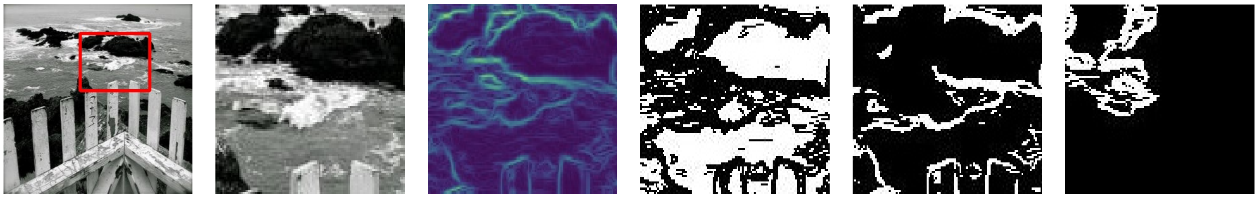


Figure 4.1: Example of FA classifier. From left to right: reef image, a zoomed patch, and the corresponding features generated by the FA classifier. The bright regions are features detected. The difference of edge and texture is the sparsity of the neighboring FA responses.

angle discretization with expensive computational cost, or we keep low complexity by rougher angle discretization, such as Canny edge-detector ([Can86]).

In order to address this issue, people propose steerable filters ([FA91]) whose space of all rotated version is finite as long as the steerability assumption is satisfied. Using a linear combination of basis filters ([FA91]), they can obtain an approximated version of the original filter which is rotation invariant. This rotation-invariance guarantee is still missing in deep-learning based methods. However, this filter approximation in [FA91] is still not perfect. The polynomial representation used in the paper lacks optimality and robustness in the presence of noise. Moreover, it causes numerical stability problems and results in high computational cost, making it difficult to represent fine direction-selective filters.

Fourier-Argand filter Recently, people further develop Fourier-Agrand (FA) filter, which is highly efficient and *optimal* in terms of the approximation error ([ZB20]). The key idea of the FA filter is to find the *optimal* basis consisting of N functions for approximating *all* rotated versions of the given pattern. Here, the pattern is a filter which can accurately capture the image spatial variation along certain direction. Specifically, let ${}^{\alpha}h$ denote the pattern characterized by the direction α ($\alpha \in [0, 2\pi]$), and $\{\phi_0, \phi_1, \cdots, \phi_{N-1}\}$ be an arbitrary basis of N elements. Let $\mathcal{P}\{\cdot\}$ denote the orthogonal projection onto the approximation space $span\{\phi_0, \phi_1, \cdots, \phi_{N-1}\}$. [ZB20] found the optimal basis for all rotated versions of ${}^{\alpha}h$

by minimizing the average approximation error e_N:

$$e_N \stackrel{\text{def}}{=} \int_0^{2\pi} \|^{\alpha}h - \mathcal{P}\{^{\alpha}h\}\|_2^2 \, d\alpha \tag{4.2}$$

Low-level feature classification This minimization automatically leads to the optimal and rotation-invariant Fourier-Argand basis. The optimality ensures that we can use a few basis to approximate the pattern accurately. Furthermore, the rotation-invariance ensures that the Fourier-Argand filter can fully characterize all spatial directions without any angle quantization error. The properties provide a valid and efficient tool to accurately classify the fast-variation image samples with small complexity. We refer to this paper ([ZB20]) for more details if readers are interested.

With the FA response, the next question is how to further discriminate the edge and texture features from the filtered results. The key idea is based on the following observation: intuitively, the spatial sparsity of the texture features in the FA response is much higher than the edge features.

The sparsity criterion provides a valid approach to effectively classify the edge and texture features based on the FA response $\mathcal{FA}(x_{i,j})$. Suppose $\mu_1, \mu_2,$ and μ_3 denote three human perception coefficients corresponding to smooth area, edge, and texture respectively, and $\mathbf{1}_{\{\cdot\}}$ is an indicator function. We first define a sparsity function $g(x_{i,j})$ to characterize the sparsity of FA response within a local patch $B_{i,j}$ centered at pixel $x_{i,j}$:

$$B_{i,j} :\stackrel{\text{def}}{=} \{x_{m,n} \mid i - r_0 \leq m \leq i + r_0 \text{ and } j - r_0 \leq n \leq j + r_0\} \tag{4.3}$$

$$g(x_{i,j}) = \frac{1}{|B_{i,j}|} \sum_{x_{m,n} \in S_{i,j}} \mathbf{1}_{\{\mathcal{FA}(x_{m,n}) \geqslant \sigma\}} \tag{4.4}$$

where $\mathbf{1}_{\{\cdot\}}$ will indicate the FA responses that are above threshold σ. Hence $g(x_{i,j})$ will compute, within the patch $B_{i,j}$, the ratio of the response.

$$M_{i,j} = \begin{cases} \mu_1, & \mathcal{FA}(x_{i,j}) < \sigma \\ \mu_2, & \mathcal{FA}(x_{i,j}) \geqslant \sigma \text{ and } g(x_{i,j}) \leqslant s_0 \\ \mu_3, & otherwise \end{cases} \tag{4.5}$$

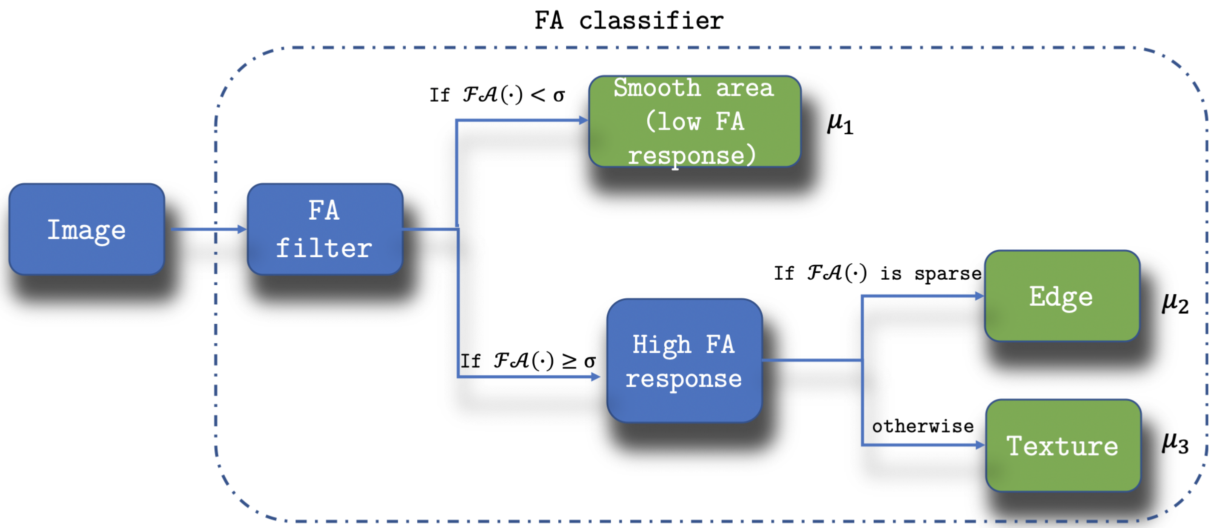


Figure 4.2: FA-filter based feature classification (Eqn. 4.5). For a pixel $x_{i,j}$, it will be first classified by the magnitude of $\mathcal{FA}(\cdot)$. A high-FA-response pixel will be further classified according to the neighboring FA sparsity (Eqn. 4.4)

$M_{i,j}$ indicates the sensitivity coefficient of pixel $x_{i,j}$. Essentially, the procedure of classification is that if the FA response is smaller than a certain threshold, the pixel is classified into the smooth area; If not, a pixel will be classified into texture if the FA response is dense within a local patch. Otherwise will be classified into edge if the response is sparse (as is indicated by Fig. 4.2).

4.3 Perceptual loss formulation & algorithm

Perceptual loss Hence, according to the our discussion in above sections, we propose the the problem formulation and the *Perceptual Loss (PL)* as follows:

$$\underset{\mathbf{x}^{adv} \in \mathbb{R}^N}{\text{minimize}} \quad PL(\mathbf{x}^{org}, \mathbf{x}^{adv}) \tag{4.6a}$$

$$\text{subject to} \quad f(\mathbf{x}^{adv}) = t \tag{4.6b}$$

where

$$PL(\mathbf{x}^{org}, \mathbf{x}^{adv}) \stackrel{\text{def}}{:=} \frac{1}{N} ||\mathbf{x}^{org} \odot \mathbf{M} - \mathbf{x}^{adv} \odot \mathbf{M}||_2^2 \text{ and } t \neq y^{org} \tag{4.7}$$

The notation $\odot$ denotes the Hadamard product of matrices. PL strives to distinguish the significance across image features and assign penalties to them accordingly.

Proposed algorithm We leverage Boundary Attack, a decision-based method ([BRB17]), and give the attack procedures as follows (We also described how to use Sign-OPT Attack [CSC19] to find adversarial samples in the supplementary materials). In essence, Boundary Attack is to perform searches along the decision boundary. In each iteration, the method will sample noise $\boldsymbol{\eta}$ and project $\mathbf{x}^i + \boldsymbol{\eta}$ onto the sphere centered at $\mathbf{x}^{org}$ with radius $d(\mathbf{x}^{org}, \mathbf{x}^i)$ (Eqn. 4.8). And then it makes a small step towards $\mathbf{x}^{org}$ with step size $\beta d(\mathbf{x}^{org}, \mathbf{x}^i)$, (Eqn. 4.9). We refer our readers to [BRB17] for details. In our case, the distance function $d(\mathbf{x}^{org}, \mathbf{x}^i) = \|(\mathbf{x}^{org} - \mathbf{x}^i) \odot M\|_2$

Algorithm 1 *PL* + Boundary Attack

Given original image $\mathbf{x}^{org}$, image in the target class $\mathbf{x}^{target}$, hard-label black-box classifier $f(\mathbf{x}) : \mathbb{R}^N \rightarrow \{0, 1, ..., C\}$

Generate $\mathbf{M} \in \mathbb{R}^N$ according to (4.2). Initial step size γ and β. Let $\mathbf{x}^1 = \mathbf{x}^{target}$

for $i = 1 : N_0$ **do**

  Generate random noise $\boldsymbol{\eta} \in \mathbb{R}^N$ and project it such that $\langle \boldsymbol{\eta}, \mathbf{x}^{org} - \mathbf{x}^i \rangle = 0$

  i) Perform orthogonal step:

$$\mathbf{x}_o^{i+1} = \mathbf{x}^{org} + \frac{1}{\sqrt{1+\gamma^2}}\left(\gamma \frac{\|(\mathbf{x}^{org} - \mathbf{x}^i) \odot \mathbf{M}\|_2}{\|\boldsymbol{\eta} \odot \mathbf{M}\|_2}\boldsymbol{\eta} - (\mathbf{x}^{org} - \mathbf{x}^i)\right) \tag{4.8}$$

  ii) Perform step towards original image:

$$\mathbf{x}^{i+1} = \mathbf{x}_o^{i+1} + \beta \mathbf{M} \odot (\mathbf{x}^{org} - \mathbf{x}_o^{i+1}) \tag{4.9}$$

  if $\mathbf{x}^{i+1}$ is not adversarial **then**

    $\mathbf{x}^{i+1} = \mathbf{x}^i$

  Increase γ and β if the attack success rate is too high. Otherwise, decrease them.

return $\mathbf{x}^{i+1}$

CHAPTER 5

Experimental results

5.1 Quantitative evaluation

Experiment setup We first randomly generate 50 image pairs from ImageNet test dataset ([DDS09]). Then we use Sign-OPT in [CSC19] and Boundary Attack in [BRB17] to optimize the loss function. Experiments are conducted on three different network architectures: Inception ([SVI16]), ResNet-50, and ResNet-101 ([HZR16]). We mainly focus on the targeted black-box attack setting, where the initial samples are the images that are correctly classified as the targeted class by the classifiers. Besides, μ_1, μ_2, and μ_3 in (Eqn. 4.5) are set to 1, 0.3, and 0.5 respectively. r_0 is 1/10 of the width of input images and s_0 equals to 0.4.

In the experiments, each input image is normalized to $[-0.5, 0.5]$. PL and L_2-norm are calculated as we mentioned in (4.6) and (4.7). Also, we mainly use *median distortion* as the metric. *median distortion* for x queries is the median adversarial perturbation across all examples under a specific metric, i.e., L_2-norm.

We can have the following observations: i) The total perturbation across PL and L_2-norm are comparable. However, by using PL to guide noise according to image features, we can assign noise to pixels based on the spatial distribution of image features. ii) PL is compatible with different network architectures and different optimization methods. iii) Compared to the original attack metric, PL can achieve better visual quality by suppressing unpleasant artifacts, such as ghosting effects.

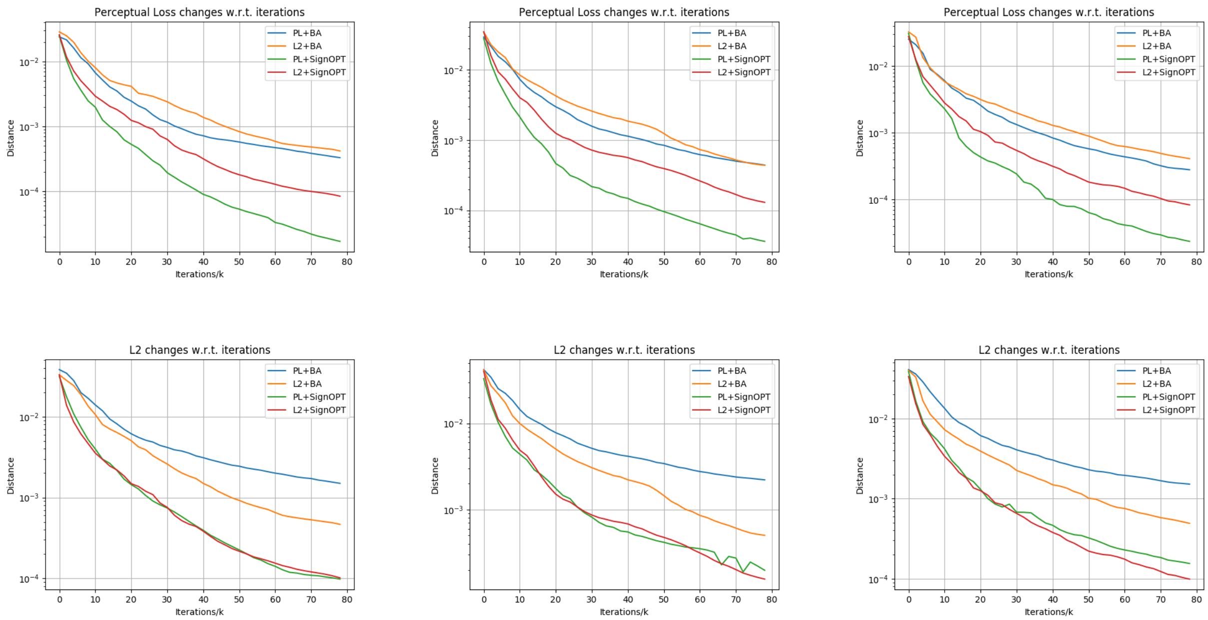


Figure 5.1: Performance v.s. iteration. The first row: PL; Second row: L_2. From left to right, experiments are conducted on Inception-v3, ResNet-50, and ResNet-101 network architectures on ImageNet.

5.2 Human perception evaluation

To demonstrate the effectiveness of our methods, we further conducted human perception evaluation among 21 volunteers. We prepared 20 attack images and generated the corresponding adversarial images using L_2 and PL metric by querying the models 20k times. We present a subject with the original image and adversarial images generated by two different metrics together. Subjects are asked to evaluate which adversarial image looks closer to the original one. The information of attack methods is hidden from the subjects, and the questions can only be answered by looking at the image quality.

Preference	PL	L_2	Cannot determine
# (ratio)	219 (52.1%)	71 (16.9%)	130 (31.0%)

Table 5.2: Human evaluation results

Attack	Inception-v3			ResNet-50			ResNet-101		
	queries	L_2	PL	queries	L_2	PL	queries	L_2	PL
L_2+BA	10k	1.0e−2	8.0e−3	10k	1.0e−2	8.4e−3	10k	7.2e−3	5.8e−3
	20k	5.0e−3	4.1e−3	20k	5.0e−3	4.3e−3	20k	3.9e−3	3.2e−3
	40k	1.5e−3	1.4e−3	40k	2.2e−3	1.9e−3	40k	1.5e−3	1.3e−3
PL+BA	10k	1.4e−2	6.7e−3	10k	1.4e−2	7.3e−3	10k	1.3e−2	6.0e−3
	20k	6.1e−3	2.5e−3	20k	7.8e−3	3.0e−3	20k	6.1e−3	2.6e−3
	40k	3.1e−3	7.1e−4	40k	4.1e−3	1.1e−3	40k	3.0e−3	8.4e−4
L_2+SignOPT	10k	3.5e−3	2.9e−3	10k	4.9e−3	4.0e−3	10k	3.4e−3	2.8e−3
	20k	1.5e−3	1.2e−3	20k	1.5e−3	1.2e−3	20k	1.2e−3	1.0e−3
	40k	3.8e−4	3.2e−4	40k	6.8e−4	5.6e−4	40k	3.8e−4	3.1e−4
PL+SignOPT	10k	4.0e−3	2.0e−3	10k	4.4e−3	2.1e−3	10k	4.2e−3	2.3e−3
	20k	1.4e−3	5.3e−4	20k	1.8e−3	4.6e−4	20k	1.3e−3	4.3e−4
	40k	4.0e−4	9.0e−5	40k	5.5e−4	1.5e−4	40k	4.7e−4	1.0e−4

Table 5.1: Algorithm performance comparison. Column: objective function + attack method. Row: different evaluating metric. Using the same optimization method, our results are better in terms of PL metric and even comparable in L_2 metric.

As is shown in the table 5.2, among 420 responses, there are 52% answers indicate the adversarial examples generated using PL are closer to the original image, which is significantly larger than 16.9% using L_2. Besides, 31% of the responses indicate that they cannot distinguish a better adversarial image. Our interpretation is that: i) Some images have too few high variant features such as edges and texture so that our metric essentially regresses to L_2. ii) Due to the input size of classifiers, images are restricted to relatively low resolution, making details difficult to be identified. Nevertheless, those responses do indicate that our methods are not degrading the visual quality.

5.3 Conclusion

In this work, we propose a new metric, *perceptual loss*, for adversarial attack based on low-level image features for better perceptual quality. The metric relies on our novel low-level

Figure 5.2: Visualization of experimental results. First row: PL (ours); Second row: L_2-distance. From left to right: adversarial image, a zoomed patch, the corresponding noise, zoomed noise patch. Our method can effectively reduce the ghosting artifacts within the red box.

feature classifier and is compatible with different optimization methods. With the total distortion amount (L_2) comparable, our method can smartly change noise distribution to improve human perceptual quality, which is also verified by our human perception evaluation.

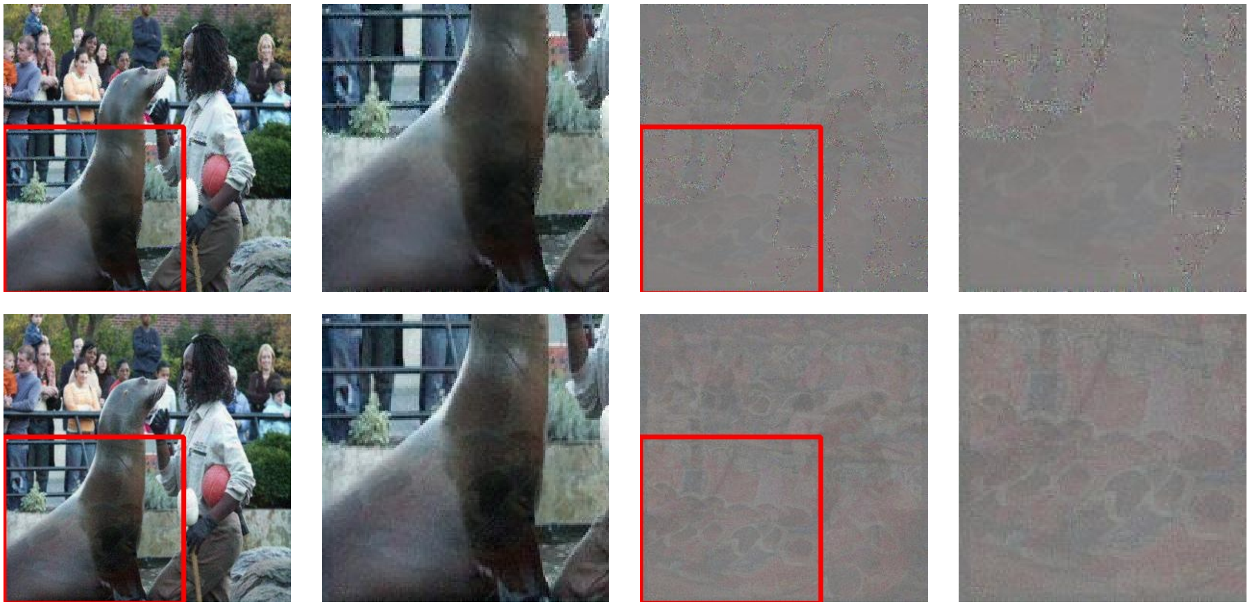


Figure 5.3: Visualization of experimental results. First row: PL (ours); Second row: L_2-distance. From left to right: adversarial image, a zoomed patch, the corresponding noise, zoomed noise patch. Our method can effectively reduce the ghosting artifacts that appear on the back of the sea lion.

Figure 5.4: Visualization of experimental results. First row: PL+Sign-OPT; Second row: L_2+Sign-OPT. From left to right: adversarial images at 5k, 10k, 15k, 20k. Notice the strong watermark in the background using L_2 metric.

APPENDIX A

Optimized using Sign-OPT

In this chapter, we discuss how we utilized the optimization method in Sign-OPT ([CSC19]) to minimize the *Perceptual Loss (PL)*. Note that the function $g(\boldsymbol{\theta})$ is formulated in [CSC19] and [CLC18], which essentially describe how good a perturbation $\boldsymbol{\theta}$ is.

Algorithm 2 *PL* + Sign-OPT

Given original image $\mathbf{x}^{org}$, image $\mathbf{x}^{target}$ in the target class t, hard-label black-box classifier $f(\mathbf{x}) : \mathbb{R}^N \to \{0, 1, ..., C\}$

Define function $g(\boldsymbol{\theta}) \stackrel{\text{def}}{:=} min._{\lambda>0}$ s.t. $f(\mathbf{x}^{org} + \lambda \frac{\boldsymbol{\theta}}{||\boldsymbol{\theta} \odot \boldsymbol{M}||}) = t$

Generate $\mathbf{M} \in \mathbb{R}^N$ according to the procedures described in the paper. Let $\theta_{\mathbf{0}} = \mathbf{x}^{target} - \mathbf{x}^{org}$

for $i = 0 : N_0$ **do**

  Generate random noise $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots, \boldsymbol{\eta}_Q \in \mathbb{R}^N$ from Gaussian or Uniform distribution

  for $q = 1 : Q$ **do**

    $\boldsymbol{\eta}_q \leftarrow (\boldsymbol{\eta}_q \odot \frac{1}{\mathbf{M}})^2$

  Compute $\nabla \hat{g}(\boldsymbol{\theta}_i) = \frac{1}{Q} \sum_{q=1}^{Q} \text{sign}(g(\boldsymbol{\theta}_i + \epsilon \boldsymbol{\eta}_q) - g(\boldsymbol{\theta}_i)) \cdot \boldsymbol{\eta}_q$

  Choose an appropriate step size γ using line search

  Update $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \gamma \nabla \hat{g}(\boldsymbol{\theta}_i)$

return $\mathbf{x}^{org} + \boldsymbol{\theta}_{i+1}$

APPENDIX B

Adversarial image visualization

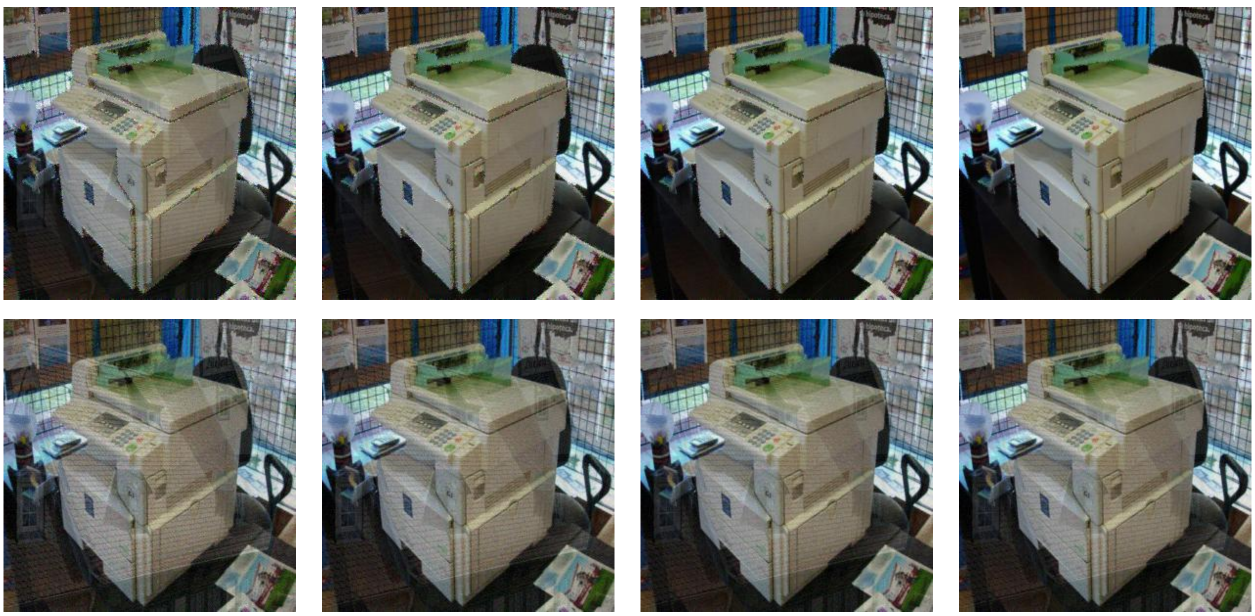


Figure B.1: From left to right: adversarial images at 5k, 10k, 15k, 20k. First row: PL+Sign-OPT; Second row: L_2+Sign-OPT

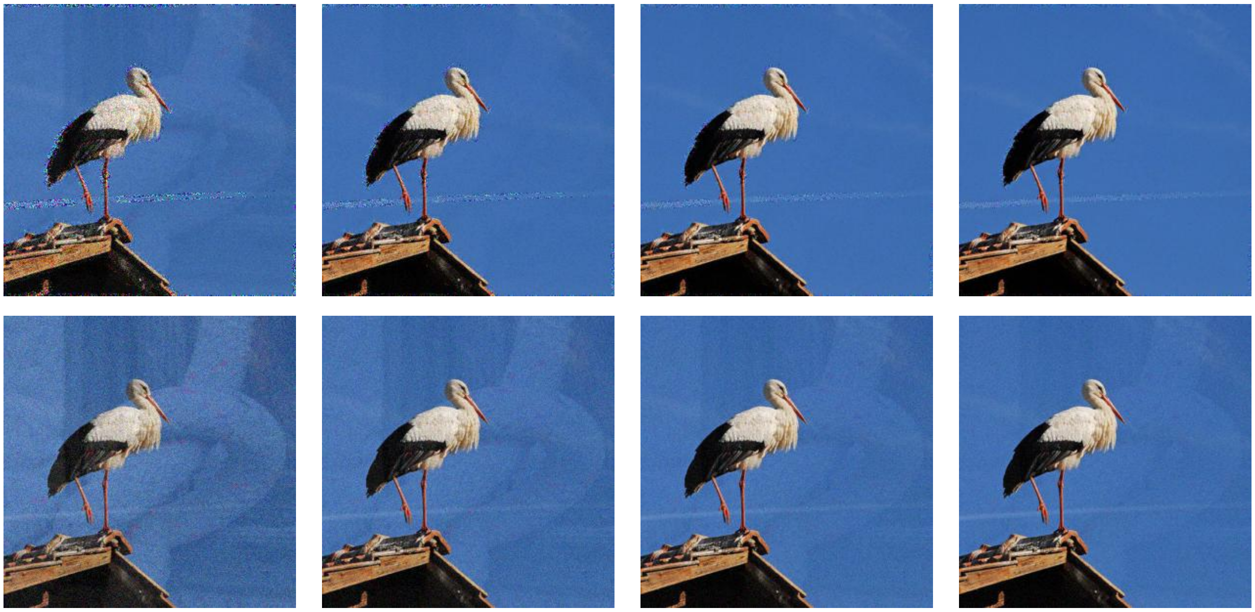


Figure B.2: From left to right: adversarial images at 5k, 10k, 15k, 20k. First row: PL+Sign-OPT; Second row: L_2+Sign-OPT

Figure B.3: From left to right: adversarial images at 5k, 10k, 15k, 20k. First row: PL +Sign-OPT; Second row: L_2+Sign-OPT

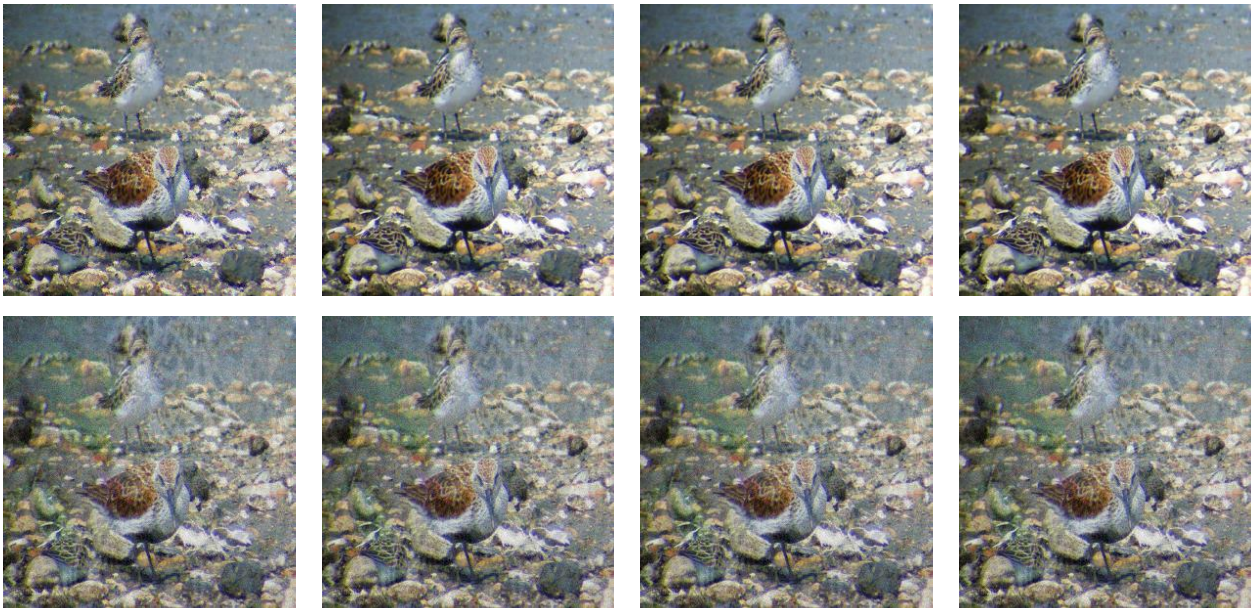


Figure B.4: From left to right: adversarial images at 5k, 10k, 15k, 20k. First row: PL+Sign-OPT; Second row: L_2+Sign-OPT

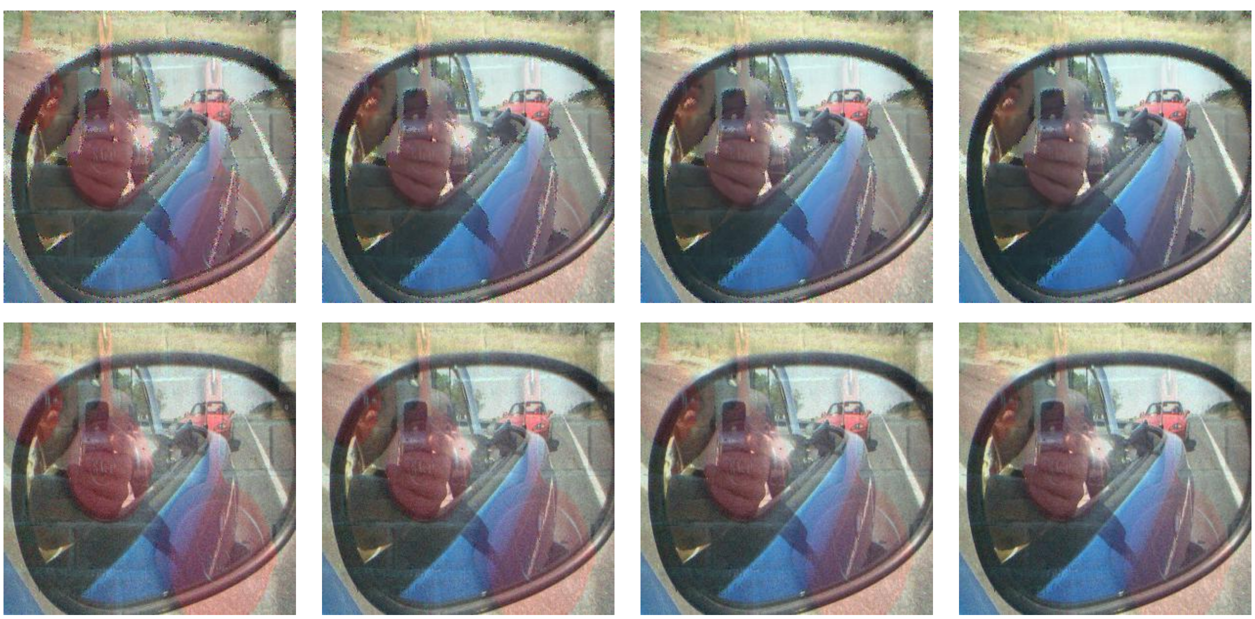


Figure B.5: From left to right: adversarial images at 5k, 10k, 15k, 20k. First row: PL+Sign-OPT; Second row: L_2+Sign-OPT

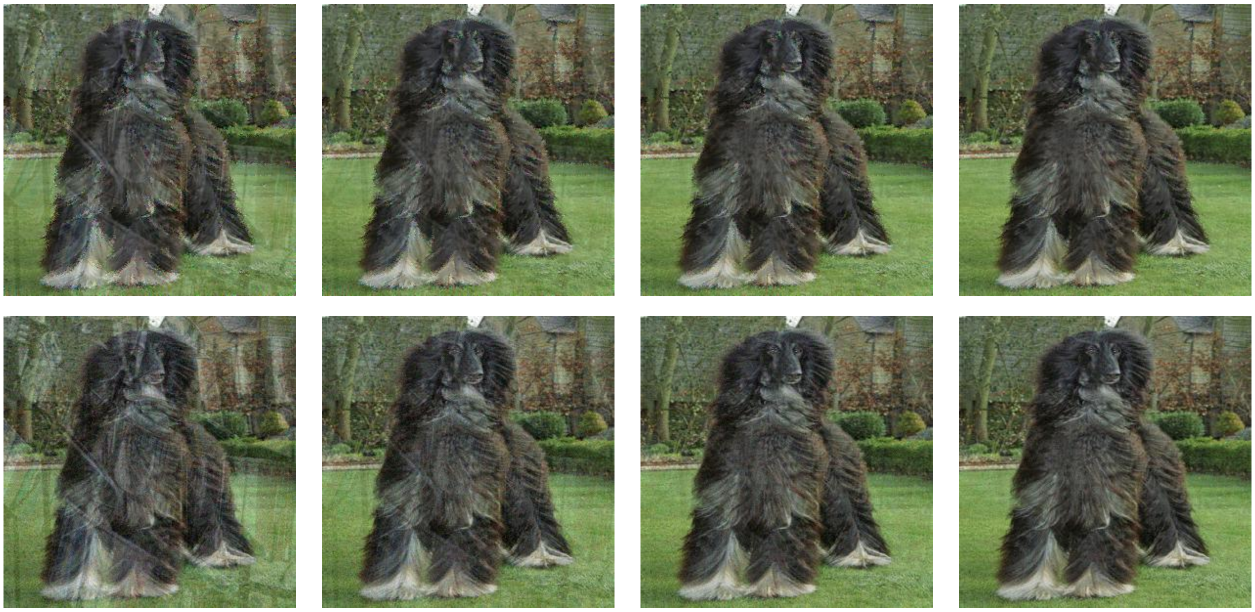


Figure B.6: From left to right: adversarial images at 5k, 10k, 15k, 20k. First row: PL+Sign-OPT; Second row: L_2+Sign-OPT

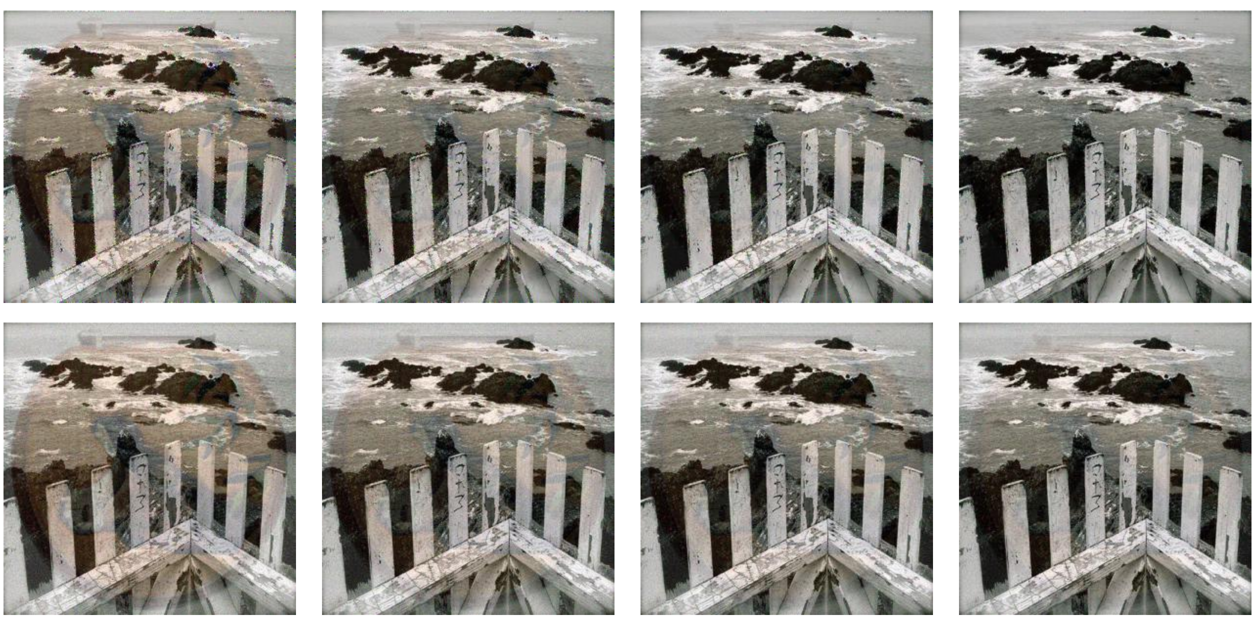


Figure B.7: From left to right: adversarial images at 5k, 10k, 15k, 20k. First row: L_2+Sign-OPT; Second row: $PercLoss$+Sign-OPT

Figure B.8: From left to right: adversarial images at 5k, 10k, 15k, 20k. First row: L_2+Sign-OPT; Second row: $PercLoss$+Sign-OPT

REFERENCES

[ASC19] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B Srivastava. "Genattack: Practical black-box attacks with gradient-free optimization." In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1111–1119, 2019.

[BRB17] Wieland Brendel, Jonas Rauber, and Matthias Bethge. "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models." *arXiv preprint arXiv:1712.04248*, 2017.

[BSD09] Peter J. Bex, Samuel G. Solomon, and Steven C. Dakin. "Contrast sensitivity in natural scenes depends on edge as well as spatial frequency structure." *Journal of Vision*, **9**(10):1–1, 09 2009.

[Can86] John Canny. "A computational approach to edge detection." *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[CH19] Francesco Croce and Matthias Hein. "Sparse and imperceivable adversarial attacks." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4724–4732, 2019.

[CLC18] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. "Query-efficient hard-label black-box attack: An optimization-based approach." *arXiv preprint arXiv:1807.04457*, 2018.

[CSC19] Minhao Cheng, Simranjit Singh, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. "Sign-OPT: A Query-Efficient Hard-label Adversarial Attack." *arXiv preprint arXiv:1909.10773*, 2019.

[CW17] Nicholas Carlini and David Wagner. "Towards evaluating the robustness of neural networks." In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

[CZS17] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models." In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26, 2017.

[DD90] Russell L DeValois and Karen K DeValois. *Spatial vision.* Oxford university press, 1990.

[DDS09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[DHL15] James Davis, Yi-Hsuan Hsieh, and Hung-Chi Lee. "Humans perceive flicker artifacts at 500 Hz." *Scientific reports*, **5**:7861, 2015.

[ETS17] Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. "A rotation and a translation suffice: Fooling cnns with simple transformations." *arXiv preprint arXiv:1712.02779*, **1**(2):3, 2017.

[FA91] William T. Freeman and Edward H Adelson. "The design and use of steerable filters." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (9):891–906, 1991.

[GMP17] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. "On the (statistical) detection of adversarial examples." *arXiv preprint arXiv:1702.06280*, 2017.

[GMP19] Diego Gragnaniello, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. "Perceptual Quality-preserving Black-Box Attack against Deep Learning Image Classifiers." *arXiv preprint arXiv:1902.07776*, 2019.

[GRC17] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. "Countering adversarial images using input transformations." *arXiv preprint arXiv:1711.00117*, 2017.

[GSS14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572*, 2014.

[Hai92] Richard F Haines. *The effects of video compression on acceptability of images for monitoring life sciences experiments*, volume 3239. National Aeronautics and Space Administration, Office of Management . . . , 1992.

[HLN18] Graham Hudson, Alain Léger, Birger Niss, István Sebestyén, and Jørgen Vaaben. "JPEG-1 standard 25 years: past, present, and future reasons for a success." *Journal of Electronic Imaging*, **27**(4):040901, 2018.

[HP18] Hossein Hosseini and Radha Poovendran. "Semantic adversarial examples." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1614–1619, 2018.

[HZR16] 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*, pp. 770–778, 2016.

[IEA18] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. "Black-box adversarial attacks with limited queries and information." *arXiv preprint arXiv:1804.08598*, 2018.

[LLW18] Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. "Towards imperceptible and robust adversarial example attacks against neural networks." In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[MC17] Dongyu Meng and Hao Chen. "Magnet: a two-pronged defense against adversarial examples." In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 135–147, 2017.

[MGF17] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. "On detecting adversarial perturbations." *arXiv preprint arXiv:1702.04267*, 2017.

[PIM08] L.J. Ping, B. Igor, N.L. M, P.S. S, and Y. Simon. *Information Computing And Automation (In 3 Volumes) - Proceedings Of The International Conference*. World Scientific Publishing Company, 2008.

[RBM19] Mohammad Saeed Rad, Behzad Bozorgtabar, Urs-Viktor Marti, Max Basler, Hazim Kemal Ekenel, and Jean-Philippe Thiran. "SROBB: Targeted Perceptual Loss for Single Image Super-Resolution." *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019.

[SBR18] Mahmood Sharif, Lujo Bauer, and Michael K Reiter. "On the suitability of lp-norms for creating and preventing adversarial examples." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1605–1613, 2018.

[SVI16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[SZM19] Ayon Sen, Xiaojin Zhu, Liam Marshall, and Robert Nowak. "Should Adversarial Attacks Use Pixel p-Norm?" *arXiv preprint arXiv:1906.02439*, 2019.

[SZS13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks." *arXiv preprint arXiv:1312.6199*, 2013.

[TKP17] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. "Ensemble adversarial training: Attacks and defenses." *arXiv preprint arXiv:1705.07204*, 2017.

[WSL19] Jinjian Wu, Guangming Shi, and Weisi Lin. "Survey of visual just noticeable difference estimation." *Frontiers of Computer Science*, **13**(1):4–15, 2019.

[ZAF19] Hanwei Zhang, Yannis Avrithis, Teddy Furon, and Laurent Amsaleg. "Smooth Adversarial Examples." *arXiv preprint arXiv:1903.11862*, 2019.

[ZB20] T Zhao and T Blu. "The Fourier-Argand Representation: An Optimal Basis of Steerable Patterns." *IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society*, 2020.

[ZDS17] Zhengli Zhao, Dheeru Dua, and Sameer Singh. "Generating natural adversarial examples." *arXiv preprint arXiv:1710.11342*, 2017.

[ZLL19] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. "Towards Large yet Imperceptible Adversarial Image Perturbations with Perceptual Color Distance." *arXiv preprint arXiv:1911.02466*, 2019.