Adversarial Robustness against Perceptual Attacks

Suraj Singireddy*[1], Sumit Kumar Jha†[2], and Rickard Ewetz‡[3]

[1]Computer Science Department, University of Texas at San Antonio, TX 78249
[2]Computer Science Department, University of Texas at San Antonio, TX 78249
[3]Electrical and Computer Engineering Department, University of Central Florida, FL 32826

Abstract

Neural networks are vulnerable to adversarial examples, which are carefully crafted inputs designed to cause misclassification. Particularly in image classification, recent research has focused on providing guarantees against imperceptible perturbations, often defined using the L_p norm family or other formulations. This paper introduces a novel perturbation space based on the Structural Similarity Index Measure (SSIM), a perception-based metric. We demonstrate that existing defenses fail to protect against imperceptible perturbations in this space. Furthermore, we propose a defense method leveraging adversarial training, which significantly improves performance against both L_p-based and SSIM-based attacks.

1 Introduction

Neural networks have demonstrated remarkable performance in various image classification tasks. However, they remain susceptible to adversarial examples, which are subtly modified natural images that lead to misclassification. The presence of adversarial examples suggests that these models do not interpret the semantic content of images in the same way humans do, thereby creating a security risk for decision-making systems based on neural networks.

Previous research [5, 7, 9] has focused on developing more robust classifiers by providing guarantees against adversaries constrained to an L_p threat model. However, recent findings [6, 14] have shown that it is possible to create imperceptible adversarial perturbations with large L_p errors using more advanced heuristic measures to quantify perceptual differences. These perturbations can fool defended networks, indicating that the L_p norm may not adequately represent the space of imperceptible perturbations.

Numerous image quality metrics [4, 12, 16, 17] have been developed to measure distortion in image processing applications, and they show strong correlation with human perception across various image quality databases [8]. In this work, we utilize existing image quality metrics to establish more meaningful measures of adversarial robustness. Our contributions are as follows:

- We propose a new formulation of image classifier robustness based on a class of adversarial perturbations defined by the SSIM image quality metric, as an alternative to L_p norms. We also develop variants of existing state-of-the-art attacks that search for adversarial examples within this class.
- We demonstrate that current defenses are unable to accurately classify images generated using our methods, and we present a defense approach that enhances robustness against a range of perceptual attacks without compromising accuracy on benign images.

2 Motivation

The majority of adversarial robustness research primarily focuses on adversaries constrained by L_p-bounded attacks.

*suraj.singireddy@utsa.edu
†sumit.jha@utsa.edu
‡rickard.ewetz@eecs.ucf.edu

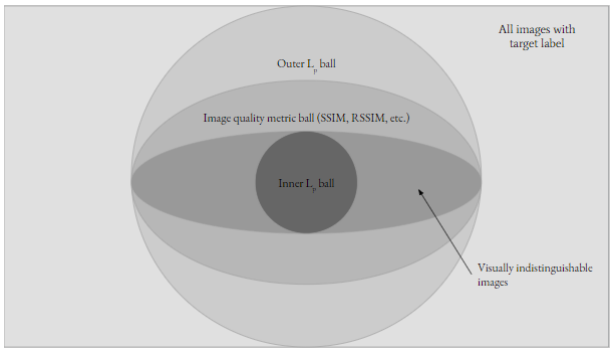

Figure 1: Illustration of adversarial spaces around an image. The most common threat model constrains adversaries to the inner L_p ball, which does not capture the space of imperceptible perturbations. Increasing the size of the L_p ball to capture all imperceptible perturbations necessitates an infeasibly large radius. We propose utilizing image quality metrics that better approximate the space of imperceptible perturbations, ensuring the smallest region encompassing all possible attacks is defendable.

However, this threat model does not accurately encompass the full range of possible perturbations an adversary can employ [2]. Figure 1 illustrates various relevant adversarial spaces, including the inner L_p ball, which is the largest L_p ball that only contains visually indistinguishable images; the set of images visually indistinguishable from the original image, representing the ideal threat model; the outer L_p ball, the smallest L_p ball that encompasses all visually indistinguishable images; and the proposed perturbation space. The inner L_p ball fails to account for all potential adversarial perturbations, rendering it an insufficient threat model despite its prevalence in adversarial robustness research.

Certified defenses can only guarantee protection against attacks within a small L_p radius [5, 10], but it is feasible to construct visually similar images with large L_p errors. Figure 2 demonstrates that shifting an image by one pixel to the right can yield the desired effect. Observe that L_p-based defenses cannot feasibly be improved to accurately classify such perturbations; the shifted image has an $L_\infty = 0.506$, and any classifier capable of resisting all adversarial perturbations with an $L_\infty > 0.5$ must output a single label across the entire input space.

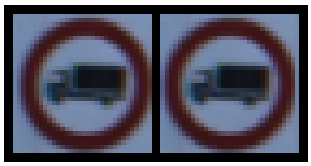

Figure 2: An example of a potential adversarial attack. The adversarial image (right) was created by shifting the original image left by 1 pixel. Its L_2 and L_∞ errors are 4.549 and 0.506, respectively, but visually, it is very similar to the original. The SSIM error is 0.116.

Figure 3: Clean images (top row) and SSIM adversarial examples (bottom row) generated by attacking an undefended network. Mean L_2 error, L_∞ error, and SSIM error of SSIM adversarial examples are 0.629, 0.136, and 0.003 respectively.

A well-defined mathematical threat model that accurately reflects the range of perturbations a real-world adversary can produce is essential. To achieve this, we employ the SSIM metric, which is designed to quantify the impact of distortion effects on human visual perception. As depicted in Figure 2, translating an image by one pixel can generate an image with low visual distortion and moderate SSIM error. Although the SSIM ball of radius 0.116 is sizable, it remains significantly smaller than the corresponding L_p balls necessary to defend images against translation attacks. Metrics that closely approximate human perception demand smaller radii to encompass the space of imperceptible perturbations.

The toy example of translating an image by one pixel theoretically demonstrates the existence of adversarial examples that appear unperturbed yet possess high L_p errors. We provide practical evidence for such attacks using the SSIM metric. Figure 4 presents images generated by an SSIM-bounded adversary, which exhibit high L_∞ errors

Figure 4: SSIM adversarial examples (top row) and images generated using the PGD-L_∞ attack with $\epsilon = 0.136$ (bottom row). Both have a mean L_∞ of 0.136, but the average SSIM error is 0.003 for SSIM attacked images and 0.320 for PGD-L_∞ attacked images. As a result, SSIM images appear unperturbed, while PGD-L_∞ attacked images are highly distorted.

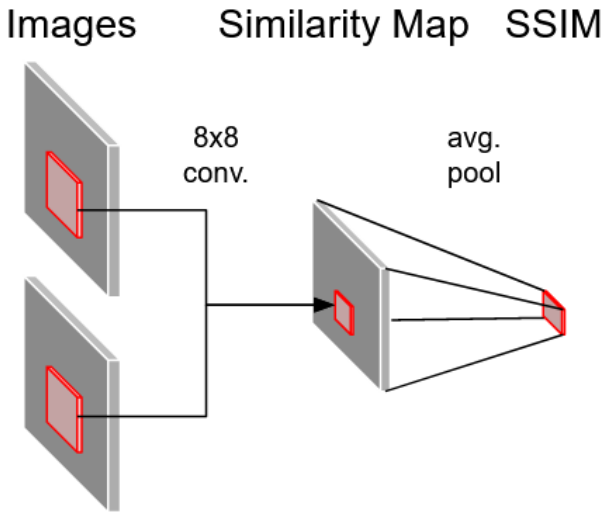

Figure 5: Visualization of SSIM computation. Using the formula in Equation 1, a local SSIM value is computed for each 8x8 window; the overall SSIM value is the mean of local SSIM values for each window.

but appear visually unperturbed due to their low SSIM errors. We compare this adversary to the PGD attack with $\epsilon = 0.14$. Although the mean L_∞ errors of the two attacks are equal, the PGD attack fails to find images with minimal visual distortion. An L_∞-based defense that resists the SSIM attack must also defend against all images within the L_∞ ball of radius 0.14, which includes numerous images with significant visual distortion. Employing image quality metrics like SSIM enables us to capture only those images that appear visually similar to the original; we contend that this smaller set of images is theoretically easier to defend.

3 Background

3.1 Structural Similarity Index Measure (SSIM)

The Structural Similarity Index Measure (SSIM) is an image quality metric proposed by Wang et al. [16] that quantifies the perceived difference between a reference image and a distorted image. SSIM is designed to evaluate the change in structural information, meaning that if an image is perturbed in a way that retains its information content, the perturbed image will have a high SSIM value compared to the original. This metric has demonstrated strong correlation with human subjective assessments of image quality.

$$\mathrm{SSIM}(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \times \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \quad (1)$$

As illustrated in Figure 5, SSIM is computed on greyscale images by applying the formula in Equation 1 to corresponding spatial patches (e.g., $w \times w$ windows). Let $x_{i,j,w}$ represent the $w \times w$ window with its top-left corner at (i, j); the local SSIM for images x and y can be calculated as $\mathrm{SSIM}(x_{i,j,w}, y_{i,j,w})$. A similarity value is computed for each spatial patch, producing a local similarity map across the image. The mean of these similarity values serves as the overall SSIM value. An SSIM of 1 indicates identical patches, while an SSIM of 0 suggests no structural similarity between patches. SSIM computation can be extended to RGB images by converting the RGB channels to greyscale.

Wang et al. [16] recommend computing SSIM using 11x11 windows weighted by a circular-symmetric Gaussian weighting function. This weighting function assigns higher weight to central pixels and lower weight to those near the image border. However, in our experiments, we found that adversarial attacks minimizing this metric gen-

erate visually distorted images. Since an adversary can significantly manipulate border pixels with minimal impact on the Gaussian-weighted SSIM, the metric performs poorly at assessing perceptual distortion of images in adversarial contexts. We thus adopt the sliding window approach proposed by Wang et al. [15], which assigns equal weight to all pixels. Although this method can cause blocking artifacts [16], it remains effective in generating imperceptible adversarial perturbations in practice.

4 SSIM Adversarial Attacks

The SSIM metric can be utilized to create perceptual adversarial attacks. Generally, an adversary aims to find the minimal perturbation δ such that $x' = x + \delta$ is an adversarial example for a given classifier F; meaning $F(x') \neq F(x)$ (for an untargeted attack) or $F(x') = c$ (for a targeted attack with class $c \neq F(x)$). The notion of a minimal δ is defined concerning a distance measure $\mathcal{D}$; in this case, $\mathcal{D} = 1 - \text{SSIM}$. This nonconvex optimization problem is generally NP-hard; however, several approximate algorithms have been developed for computing minimal adversarial perturbations. We enumerate a few such algorithms and describe their modifications to generate SSIM-bounded perturbations.

4.1 PGD-SSIM

Projected gradient descent (PGD) [7] is an adversarial attack algorithm that iteratively computes distance-constrained, loss-maximizing perturbations using the following update rule:

$$x^{t+1} = \Pi_{x'|\mathcal{D}(x',x)<\epsilon}(x^t + \alpha \nabla_x L_\theta(x^t, y)) \quad (2)$$

where L is the neural network loss function, y is the ground-truth label, α is the step size, and Π is a function that projects its input onto the valid set of perturbations, defined by a distance metric $\mathcal{D}$ and a maximum perturbation magnitude ϵ. The SSIM metric does not satisfy the properties of a distance metric; however, the slight modification $\mathcal{D} = \sqrt{1 - \text{SSIM}}$ constitutes an approximate distance metric [1] and can, therefore, be used to normalize the step size.

4.2 CW-SSIM

Carlini and Wagner [3] propose the following method for generating adversarial examples given the pre-softmax activations Z of a classifier, a base image x, its label y, and confidence κ:

$$\text{minimize} \quad \mathcal{D}(x, x+\delta) + c \cdot f(x+\delta) \text{ s.t.} \quad x + \delta \in [0,1]^{n^2} \quad (3)$$

where $\mathcal{D}$ is a distance metric and f is defined as

$$f(x) = \max(\max Z(x)i, i \neq y - Z(x)y, -\kappa) \quad (4)$$

The constant c is a hyperparameter and can be optimized using one-dimensional optimization techniques such as binary search to find an adversarial image of minimum distance. Directly applying this method with $\mathcal{D} = 1 -$ SSIM is sufficient to produce adversarial examples.

Unlike PGD, which only guarantees an upper bound on the distortion of the generated adversarial example (and in practice, nearly always finds adversarial examples on the boundary of the valid set), the CW attack can find minimally distorted adversarial examples. However, CW is significantly slower in comparison; thus, we opt to use the PGD attack in our proposed defense.

5 Adversarial Defenses

We now focus on developing classifiers that are robust against the previously described attacks. Formally, we aim to solve the following saddle-point formulation of adversarial robustness:

$$\min_\theta \mathbb{E}(x, y) \in \mathcal{S}[\max \delta : \text{SSIM}(x, x+\delta) \leq \epsilon L_\theta(x+\delta, y)] \quad (5)$$

Although this problem is intractable in general (the inner maximization problem is non-convex and the outer minimization problem is non-concave), approximate solutions can be found using first-order methods. It is common to use an adversarial attack algorithm (such as FGSM or PGD) to solve the inner maximization problem and a first-order optimizer (such as SGD or Adam) to solve the outer minimization problem; this technique is more commonly known as adversarial training. Our proposed defense introduces attacked images generated using the PGD-SSIM attack into the training procedure.

(a) (b) (c)

Figure 6: Clean images (top row) and adversarial examples generated by various attack methods (PGD-L_2, PGD-SSIM, CW-SSIM, and Shadow Attack). Images generated by attacking an undefended network (a) and networks trained adversarially against the PGD-L_2 attack (b) and the PGD-SSIM attack (c).

	Undefended	AT-PGD-L_2	AT-PGD-SSIM
Benign	**92.58%**	75.68%	78.71%
PGD-L_2	17.29%	50.49%	**53.12%**
PGD-SSIM	9.57%	45.61%	**52.83%**
Shadow Attack	8.59%	20.90%	**27.44%**
CW-SSIM	0.003	0.009	**0.010**

Table 1: Accuracy of defended networks against various adversarial attacks. The CW-SSIM attack generates adversarial examples of minimal SSIM error; since the attack always succeeds, we instead report the average SSIM error of the generated examples.

6 Experimental Results

We conducted experiments using our proposed adversarial training method on the German Traffic Sign Recognition Benchmark (GTSRB) dataset. For these experiments, we employed the ResNet50 network architecture, training each network for 100 epochs. Our experiments involved training against two types of adversaries: the traditional L_2-bounded adversary (represented by the PGD-L_2 attack with $\epsilon = 0.5$) and a perceptually bounded adversary (represented by the PGD-SSIM attack with $\epsilon = 0.02$).

To evaluate the networks, we subjected them to various attacks, including the Shadow Attack [6]. Table 1 shows the test set accuracy results. The SSIM-trained network outperformed the L_2-trained network for all attacks while maintaining higher benign accuracy. Our findings indicate a tradeoff between adversarial robustness and accuracy [11, 13], which can be modulated by the strength of the adversary (in this case, by adjusting ϵ). Therefore, our results suggest that SSIM training is strictly superior to L_2 training. Moreover, SSIM-based attacks were more successful against the L_2-defended network than the PGD-L_2 attack, demonstrating that using L_2-based attacks can lead to overestimations of robustness.

7 Conclusions

In this paper, we proposed a novel adversarial training method that focuses on improving robustness against perceptually bounded adversaries. By leveraging structural similarity (SSIM) as a perceptual metric, we were able

to create more visually consistent adversarial examples. Our experiments on the German Traffic Sign Recognition Benchmark (GTSRB) dataset demonstrated that SSIM-trained networks outperformed their L_2-trained counterparts in terms of both adversarial robustness and benign accuracy.

Our results indicate a tradeoff between adversarial robustness and accuracy, which can be modulated by the strength of the adversary, as shown by adjusting the value of ϵ. Our findings suggest that SSIM-based adversarial training is strictly superior to L_2-based training, as it provides better robustness against various attacks, including SSIM-based ones. Furthermore, our study showed that using L_2-based attacks for evaluating robustness can lead to overestimations, highlighting the importance of considering perceptual metrics when designing defenses against adversarial attacks.

As future work, we plan to explore the use of other perceptual metrics for adversarial training and to investigate the impact of these methods on other datasets and network architectures. Additionally, we aim to develop more efficient and effective optimization techniques for generating perceptually consistent adversarial examples. Ultimately, our goal is to design deep learning models that are robust against a wide range of adversaries, while maintaining high performance on benign data.

References

[1] Dominique Brunet, Edward R Vrscay, and Zhou Wang. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, 21(4):1488–1499, 2011.

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

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (S&P)*, pages 39–57. IEEE, 2017.

[4] Damon M Chandler and Sheila S Hemami. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE transactions on image processing*, 16(9):2284–2298, 2007.

[5] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[6] Amin Ghiasi, Ali Shafahi, and Tom Goldstein. Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates. *arXiv preprint arXiv:2003.08937*, 2020.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[8] Marius Pedersen and Jon Yngve Hardeberg. Full-reference image quality metrics: Classification and evaluation. *Foundations and Trends® in Computer Graphics and Vision*, 7(1):1–80, 2012.

[9] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

[10] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples. *arXiv preprint arXiv:1811.01057*, 2018.

[11] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

[12] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

[13] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[14] Yajie Wang, Shangbo Wu, Wenyi Jiang, Shengang Hao, Yu-an Tan, and Quanxin Zhang. Demiguise attack: Crafting invisible semantic adversarial perturbations with perceptual similarity. *arXiv preprint arXiv:2107.01396*, 2021.

[15] Zhou Wang and Alan C Bovik. A universal image quality index. *IEEE signal processing letters*, 9(3):81–84, 2002.

[16] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[17] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.