R-LPIPS: An Adversarially Robust Perceptual Similarity Metric

Sara Ghazanfari [1] **Siddharth Garg** [1] **Prashanth Krishnamurthy** [1] **Farshad Khorrami** [1] **Alexandre Araujo** [1]

Abstract

Similarity metrics have played a significant role in computer vision to capture the underlying semantics of images. In recent years, advanced similarity metrics, such as the Learned Perceptual Image Patch Similarity (LPIPS), have emerged. These metrics leverage deep features extracted from trained neural networks and have demonstrated a remarkable ability to closely align with human perception when evaluating relative image similarity. However, it is now well-known that neural networks are susceptible to adversarial examples, i.e., small perturbations invisible to humans crafted to deliberately mislead the model. Consequently, the LPIPS metric is also sensitive to such adversarial examples. This susceptibility introduces significant security concerns, especially considering the widespread adoption of LPIPS in large-scale applications. In this paper, we propose the Robust Learned Perceptual Image Patch Similarity (R-LPIPS) metric, a new metric that leverages adversarially trained deep features. Through a comprehensive set of experiments, we demonstrate the superiority of R-LPIPS compared to the classical LPIPS metric. The code is available at `https://github.com/SaraGhazanfari/R-LPIPS`.

1. Introduction

The ability to compare data points is fundamental in many areas of machine learning. For many years, The ℓ_p distance metric, for instance, is a well-established mathematical tool for measuring differences between data points. However, in the context of computer vision, these metrics primarily focus on pixel-wise differences and fail to capture the semantic information of the images. This limitation becomes especially evident in high-dimensional settings, where two high-definition images depicting the same scene, *i.e.*, sharing the same underlying informational content, are far apart in terms of ℓ_p distance metrics.

Perceptual metrics (Wang et al., 2003; 2004; Hore & Ziou, 2010; Zhang et al., 2011; Mantiuk et al., 2011; Zhang et al., 2018) have been adopted for their ability to closely align with human perception when assessing relative image similarity. These metrics successfully capture the underlying semantics of images, providing a more accurate reflection of human judgment. To compute the "distance" between images, these metrics operate on the features of the images instead of the raw images in the image space. For example, the Learned Perceptual Image Patch Similarity (Zhang et al., 2018) (LPIPS) metric takes the Euclidean distance over the deep features (latent space) of a trained neural network. This new semantic measure has been shown to outperform all previous metrics by large margins due to the capabilities of a neural network to learn *good features*.

Notwithstanding their remarkable success of neural networks in a range of tasks, neural networks are also known to be sensitive to adversarial perturbations (Goodfellow et al., 2014; Madry et al., 2017), *i.e.*, small perturbations invisible to humans crafted to deliberately mislead the model. Given that the LPIPS metric is based on the feature of a trained neural network, it should not come as a surprise that this metric is also sensitive to adversarial perturbations (Kettunen et al., 2019), *i.e.*, invisible perturbations to an image which considerably modify the LPIPS value. This raises significant security concerns as similarity metrics are already in wide use, for instance, in detecting cases of online copyright infringement and digital forensics.

In this paper, we propose a thorough analysis of the LPIPS metric and empirically show that this metric, which is based on the learned feature of a trained network, is sensitive to adversarial attacks. Then, we introduce the Robust Learned Perceptual Image Patch Similarity (R-LPIPS) which leverages adversarially trained deep features and shows that is robust to adversarial perturbations. Our contributions can be summarized as follows.

1. We show that the LPIPS metric is sensitive to adversarial perturbation by showing that there exist small ℓ_∞ perturbations such that the LPIPS between a reference image and the perturbed image is large.

[1] Department of Electrical and Computer Engineering, New York University, NY, USA. Correspondence to: Sara Ghazanfari <sg7457@nyu.edu>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2. We propose the use of Adversarial Training (Madry et al., 2017) to build a new Robust Learned Perceptual Image Patch Similarity (R-LPIPS) that leverages adversarially trained deep features.
3. Based on an adversarial evaluation, we demonstrate the robustness of R-LPIPS to adversarial examples compared to the LPIPS metric.
4. Finally, based on the work of Laidlaw et al. (2021), we showed that the perceptual defense achieved over LPIPS metrics could easily be broken by stronger attacks developed based on R-LPIPS.

2. Related Work

In this section, we provide a comprehensive review of related works on perceptual and similarity metrics. Additionally, we propose a concise review of adversarial attacks and adversarial robustness. Finally, we offer a brief summary of research that explores the intersection of robustness and similarity metrics.

Similarity Metrics. The ℓ_2 Euclidean distance, a classic per-pixel measure, assumes pixel-wise independence and is often used for regression problems. However, it is insufficient for structured outputs like images due to its inability to effectively capture perceptual changes like blurring. Peak Signal-to-Noise Ratio (PSNR) measures the quality degradation in a reconstructed or compressed image/video by calculating the ratio of peak signal power to the mean squared error. Despite its wide usage, it does not correlate well with perceived image quality. The Structural Similarity Index (SSIM), proposed by (Wang et al., 2004), is a perceptual metric that quantifies the structural similarity between two images or video frames. It takes into account luminance, contrast, and structural similarities and has been shown to correlate well with human visual perception. SSIM calculates local measures of similarity by comparing small image patches and then computes the average similarity across the entire image. The Feature Similarity Index for Image Quality Assessment (FSIM), proposed by (Zhang et al., 2011), is a perceptual metric that evaluates image quality by quantifying the similarity between two images based on their features. FSIM uses phase congruency for feature significance and image gradient magnitude for feature similarity, ensuring consistency across varying lighting conditions. These two metrics are often considered to be a better indicator of perceived image quality compared to PSNR, as they correlate better with human visual perception.

More recently, the Learned Perceptual Image Patch Similarity (LPIPS) proposed by (Zhang et al., 2018) was developed and aimed at providing a more accurate measure of the perceptual similarity between two images. Instead of comparing raw pixel data, LPIPS uses deep learning to calculate the perceptual difference between images. Specifically, it uses a deep convolutional neural network, pretrained on an image classification task, to extract features from the images. Then, it computes the distance between these feature vectors to calculate the perceptual similarity. Any neural network architecture can be used for the LPIPS metric, (Zhang et al., 2018) experimented with several well-known architectures such as SqueezeNet (Iandola et al., 2016), AlexNet (Krizhevsky et al., 2017), and VGG (Simonyan & Zisserman, 2014) and showed that the AlexNet architecture offers the best performance.

To evaluate the quality of this metric with respect to human perception compared to other perceptual metrics, Zhang et al. (2018) introduced the Berkeley-Adobe Perceptual Patch Similarity (BAPPS) dataset. The BAPPS dataset is a large-scale, highly diverse dataset of perceptual judgments used to evaluate perceptual similarity metrics. It contains pairs of images along with human judgments of their perceptual similarity, which serves as ground truth data.

Adversarial Examples. Since the discovery of adversarial examples (Szegedy et al., 2013), significant research has been devoted to developing attacks (Goodfellow et al., 2014; Kurakin et al., 2018; Carlini & Wagner, 2017; Croce & Hein, 2020; 2021) and defenses (Goodfellow et al., 2014; Madry et al., 2017; Pinot et al., 2019; Araujo et al., 2020; 2021; Meunier et al., 2022; Araujo et al., 2023), resulting in an ongoing battle between the two. Most of these defenses relied on smoothing the local neighborhood around each point, resulting in very small gradients that the attacks were based on. However, it has become apparent that many of the proposed empirical defenses could be circumvented with stronger attacks (Athalye et al., 2018). In the context of a classification task, one of the best attacks, called Projected Gradient Descent (PGD) (Madry et al., 2017) consists in maximizing the cross-entropy loss with respect to a perturbation added to the input and then projecting the perturbation to a specific ℓ_p ball. This attack also led to one of the strongest empirical defenses (Athalye et al., 2018) called adversarial training (AT) which trains neural networks with adversarial examples crafted with PGD attack.

Adversarial Robustness & Similarity Metrics. Perceptual similarity metrics based on deep features inherit both the emergent properties (good features) and the sensitivity to adversarial perturbation. To the best of our knowledge, the robustness of LPIPS has only been investigated in the work proposed by (Kettunen et al., 2019). They introduced a self-ensembled metric (E-LPIPS), which operates in the space of natural images. However, this approach may have limitations, as ensembling models have been shown to be ineffective in defending against adversarial examples (Atha-

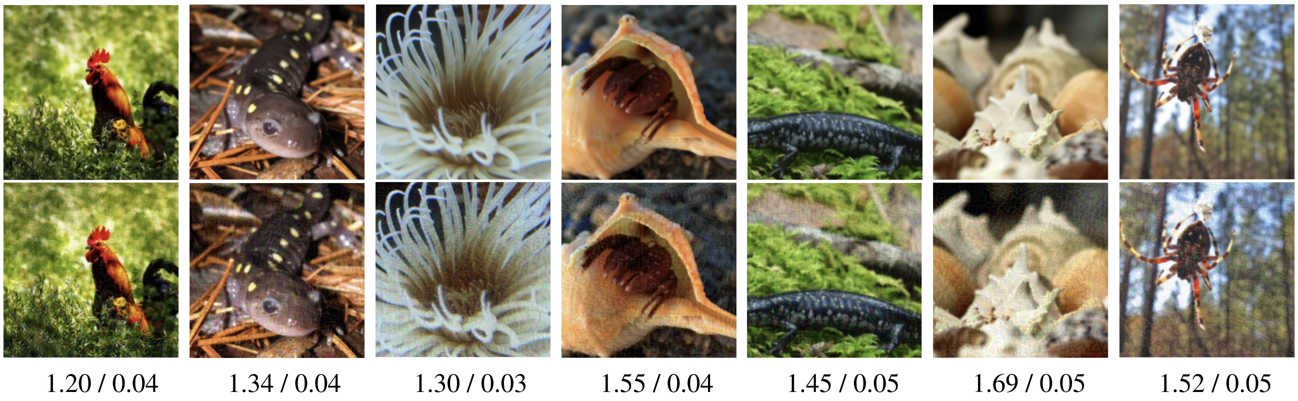


Figure 1. Adversarial examples generated using PGD with $\|\delta\|_\infty \leq 0.05$ on ImageNet-100 validation set. Original and perturbed images are shown in the first and second rows, respectively. The LPIPS/R-LPIPS values for these images are mentioned below each image. In contrast with LPIPS values that are quite large, the R-LPIPS are very small and correctly reflect the small difference between images.

lye et al., 2018). Another line of work has proposed to use the LPIPS metric to craft a perceptual attack (Laidlaw et al., 2021). They build upon PGD and introduced a new attack called Perceptual Projected Gradient Descent (PPGD) which consists in projecting with the LPIPS metric instead of an ℓ_p norm. Furthermore, they combined an AT scheme with PPGD, called PAT, and demonstrated strong defenses against adversarial attacks that generalize to unforeseen threat models.

3. Robust Perceptual Similarity Metric

In this section, we build upon LPIPS and adversarial training and introduce R-LPIPS a new *robust* perceptual similarity metric. Moreover, based on this new robust metric we propose two new strong perceptual attacks.

3.1. Adversarially Trained Perceptual Similarity Metric

The LPIPS metric (Kettunen et al., 2019) is defined as the ℓ_2 norm of deep features of a trained convolutional neural network. More formally, for inputs $x, x_0 \in \mathcal{X}$, the LPIPS metric is defined as follows:

$$d(x, x_0) = \sum_j \frac{1}{W_j H_j} \sum_{h,w} \left\|\phi^j(x) - \phi^j(x_0)\right\|_2^2 \quad (1)$$

where $\phi^j(\cdot)$ is defined as:

$$\phi^j(x) = w_j \odot o^j_{hw}(x) \quad (2)$$

and $o^j(x)$ and $o^j(x_0)$ are the internal activations of a trained convolutional neural network scaled channel-wise by vector w_j. Then, the ℓ_2 norm of the weighted activations is normalized by the width and height of filters.

In order to build the LPIPS metric, Zhang et al. (2018) used the features of the AlexNet classification model trained on the ImageNet dataset (Deng et al., 2009). Then, they "tune" the metric by learning the weights w_j on the BAPPS dataset. More formally, the loss used to "tune" the metric is defined as follows:

$$l_{\text{ce}}\left[g_\theta\left(d_w(x, x_0), d_w(x, x_1)\right), h\right]$$

where l_{ce} is the cross-entropy loss, x_0, x_1 are distortions of the reference images x from the BAPPS dataset, $h \in (0, 1)$ is a perceptual score, g_θ is a small network parameterized by θ, trained to map distances to h score and d is the distance defined in Equation (1) and it is parameterized by w.

To adversarially train the LPIPS metric, we leverage the adversarial training scheme introduced by Madry et al. (2017) and introduce an adversarial perturbation δ at each step of the training on x_0:

$$\min_{\theta,w} \max_{\delta:\|\delta\|_p \leq \varepsilon} l_{\text{ce}}\left[g_\theta\left(d_w(x, x_0 + \delta), d_w(x, x_1)\right), h\right] \quad (3)$$

The new weights w trained with adversarial training become the building block of R-LPIPS following the same construct as LPIPS in Equation (2).

3.2. New Attacks based on R-LPIPS

Recently, LPIPS has been employed instead of ℓ_p norms to produce perceptual adversarial attacks. The general constrained optimization scheme to craft an adversarial example with respect to the LPIPS metric is defined as follows:

$$\max_{\tilde{x}} \quad l_m\left[f(\tilde{x}), y\right] \quad \text{s.t.} \quad \|\phi(x) - \phi(\tilde{x})\|_2 \leq \varepsilon$$

where $l_{\text{m}} = \max_{i \neq y}(f(\tilde{x})_y - f(\tilde{x})_i)$ is the margin loss used by Carlini & Wagner (2017), $f(\cdot)$ is the classifier, and $\phi(\cdot)$ is the network that generates feature vectors. However, this constrained optimization scheme is not trivial. Therefore, Laidlaw et al. (2021) relax the problem and proposed two

Table 1. Results of Naturally and Adversarially Trained LPIPS with respect to adversarial attacks generated using ℓ_2 and ℓ_∞ norms against the BAPPS dataset. First, we can observe from table (a) that LPIPS is not robust to adversarial attacks crafted with ℓ_∞-PGD or ℓ_2-PGD. We note that the ℓ_∞ attacks are causing the largest drops in 2AFC score compared to ℓ_2-PGD attacks. Tables (b) and (c) show the 2AFC score under attack for R-LPIPS trained with ℓ_∞ and ℓ_2. The natural 2AFC score of R-LPIPS remains mostly the same as LPIPS while the 2AFC score under attack is considerably improved.

		Natural 2AFC	**ℓ_∞-PGD (ϵ =8/255)**			**ℓ_2-PGD (ϵ =1.0)**		
			x_0	x_1	x_0/x_1	x_0	x_1	x_0/x_1
(a)	**LPIPS**							
	Traditional	74.58	64.22	64.36	63.47	71.17	72.92	69.48
	CNN-based	83.52	70.02	68.37	68.92	80.06	79.35	78.53
	Superres	71.36	58.92	58.46	59.74	65.07	65.54	63.53
	Deblur	60.92	53.68	51.70	53.92	58.15	57.45	56.55
	Color	65.53	58.72	51.74	54.76	61.97	57.84	60.07
	Frameinterp	63.01	53.99	52.60	51.47	58.89	58.05	55.01
(b)	**R-LPIPS with ℓ_∞ AT**							
	Traditional	70.94	66.29	66.30	63.89	69.58	70.42	70.34
	CNN-based	83.04	75.17	74.80	73.74	81.28	81.68	80.77
	Superres	71.77	63.28	61.42	61.34	68.06	67.18	66.19
	Deblur	60.83	54.85	53.58	57.46	59.53	58.57	58.74
	Color	65.55	57.98	55.56	59.50	62.32	63.69	59.80
	Frameinterp	63.27	56.63	53.82	58.95	61.38	56.05	58.46
(c)	**R-LPIPS with ℓ_2 AT**							
	Traditional	73.19	67.07	65.11	65.63	71.17	71.35	72.14
	CNN	83.40	73.46	72.28	71.98	81.56	81.70	80.27
	Superres	71.70	60.42	59.89	58.21	67.17	68.10	65.44
	Deblur	61.19	54.24	53.05	52.20	58.69	57.58	58.58
	Color	65.71	59.21	53.83	57.69	63.44	61.38	60.40
	Frameinterp	63.53	53.67	53.56	58.23	61.57	57.58	55.38

perceptual attack methods, Perceptual Projected Gradient Descent (PPGD) and Lagrangian Perceptual Attack (LPA) based on the LPIPS metric to craft adversarial perturbations with better perceptual properties.

Perceptual Projected Gradient Descent (PPGA) tries to find the optimal δ by using first-order Taylor's approximation and rewriting the optimization formula as:

$$\max_{\delta} \quad l\left[f(x), y\right] + \nabla l\left[f(x), y)\right]^\top \delta \quad \text{s.t.} \quad \|J\delta\|_2 \leq \eta$$

where J is the Jacobian matrix of $\phi(\cdot)$ at x, δ is the perturbation size applied to x, and η is the step size. The second method, Lagrangian Perceptual Attack (LPA), uses a Lagrange multiplier to add the constraint to the optimization formula and perform the optimization:

$$\max_{\tilde{x}} \quad l\left[f(\tilde{x}), y\right] - \lambda \max\left(0, \|\phi(\tilde{x}) - \phi(x)\|_2 - \varepsilon\right)$$

In this work, we build upon the work of Laidlaw et al. (2021) and propose the R-PPGD and R-LPA attack scheme which consist of the same optimization but our R-LPIPS is replaced with the classical LPIPS metric.

4. Experiments

In this section, we present a comprehensive set of experiments to demonstrate the superiority of R-LPIPS compared to the classical LPIPS metric. More precisely, we aim at answering the following questions:

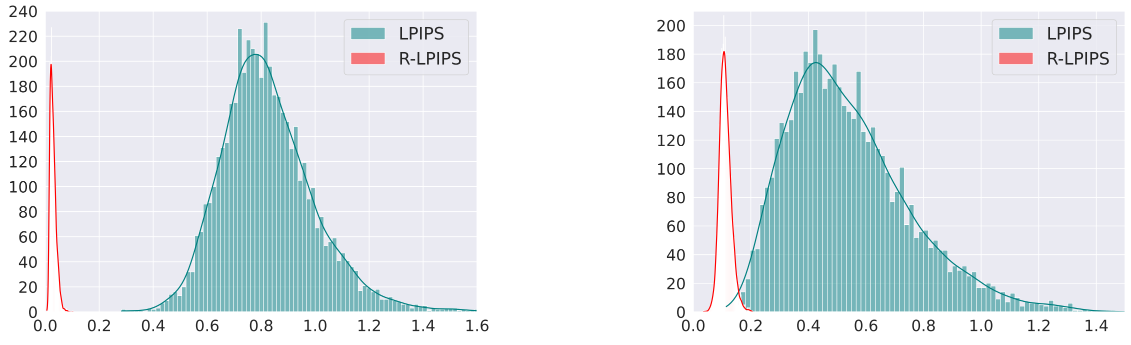
(Q1) Is LPIPS vulnerable to adversarial examples?

(Q2) How robust is R-LPIPS compared to LPIPS?

(Q3) Can R-LPIPS leads to stronger attacks?

4.1. Vulnerabilities of LPIPS (Q1)

To illustrate the lack of robustness of the LPIPS metric, we present two sets of results. First, we present the result of ℓ_∞-PGD and ℓ_2-PGD against the LPIPS metric in Table 1a with $\varepsilon = 8/255$ and $\varepsilon = 1$ respectively on x_0, x_1 independently, and x_0/x_1 together. To evaluate the performance of LPIPS over clean and adversary data, we use the 2FAC score which was employed in Zhang et al. (2018). It can be observed that the 2AFC score under attack on different distortions is significantly lower, up to 15.15% lower for ℓ_∞-PGD and



(a) OPT with LPIPS

(b) OPT with R-LPIPS

Figure 2. The histogram of LPIPS and R-LPIPS distances between the clean and OPT adversarial examples for ImageNet-100 validation set. Figure (a) shows the LPIPS and R-LPIPS distribution of adversarial examples generated using the OPT optimization scheme and the LPIPS metric with $\|\delta\|_\infty \leq 0.05$. Although LPIPS is fooled and assigns large values to the semantically identical images, the R-LPIPS shows complete robustness and considers the perturbation to be small. Figure (b) shows the same setup except R-LPIPS is used instead of LPIPS during the optimization of adversarial attack. Although R-LPIPS values in (b) are greater than the values in (a), they are still far from the threshold (0.5) and are quite smaller than the values of LPIPS.

for 8% lower ℓ_2-PGD, compared to the natural 2AFC score of the LPIPS metric over different distortions.

Second, we propose a new optimization scheme, called OPT, to demonstrate that there exist adversarial examples with small ℓ_∞ perturbations such that the LPIPS metric is large. More formally, by defining $\phi(\cdot)$ as the model that generates the feature vectors, we define the optimization formula for the attack as follows:

$$\max_{\delta:\|\delta\|_\infty \leq \varepsilon} l_{\text{MSE}}\left[\phi(x+\delta), \phi(x)\right]$$

where l_{MSE} is the mean squared error loss and we choose $\varepsilon = 0.05$ for the optimization. We perform this attack on the validation set of ImageNet-100 and present the distribution of the values of LPIPS in Figure 2a in blue. Based on the result presented by Laidlaw et al. (2021), two images with an LPIPS value greater than 0.5 are observable to humans. The histogram in Figure 2a demonstrates that nearly all images of the ImageNet-100 validation set have an LPIPS value over 0.5 while having a difference of 0.05 in ℓ_∞ which is considered very small and nearly imperceptible to humans. Figure 1 illustrates this difference. The top row shows clean reference images while the bottom row shows OPT adversarial images and the left value below shows the LPIPS value.

4.2. Robust LPIPS (R-LPIPS) (Q2)

After observing the vulnerabilities of LPIPS, we propose a new robust perceptual similarity metric called R-LPIPS. In order to develop R-LPIPS, we leverage the training scheme of LPIPS and adversarial training. The setup is explained in detail in Section 3.1. First, we conducted an adversarial training with ℓ_∞ and ℓ_2 norms and evaluated both R-LPIPS against the BAPPS dataset. Table 1 presents results for natural and under attack images with ℓ_∞-PGD and ℓ_2-PGD against adversarial training conducted with ℓ_∞ (Table 1b) and ℓ_2 (Table 1c) norms.

Table 2. Accuracy of Perceptual Adversarial Training (PAT) variants on CIFAR-10 against PPGD/LPA and R-PPGD/R-LPA attacks with the constraint of 0.5 for perturbation size measured by LPIPS and R-LPIPS. Although PAT variants show relative robustness to PPGD/LPA attacks, they are completely vulnerable to attacks generated by R-PPGD and R-LPA.

	PPGA	**LPA**	**R-PPGA**	**R-LPA**
PAT-self	13.1	2.1	3.1	0.0
PAT-AlexNet	26.6	9.8	4.3	0.2

The first interesting result to observe from Table 1 is that the natural 2AFC score is preserved across all data distortions, except for a slight decrease observed for the traditional distortion. We can even observe slight improvements in the natural 2AFC score for some distortions, as the model shows a better generalization. To evaluate robustness of R-LPIPS, we compute a perturbation with PGD attack with ℓ_∞ and ℓ_2 norms with $\varepsilon = 8/255$ and $\varepsilon = 1$, respectively on x_0, x_1 independently, and x_0/x_1 together. We observe a consistent increase in robustness of R-LPIPS compared to the original LPIPS metric trained without AT. We also note that R-LPIPS with ℓ_∞-AT seems to provide better results. In the following, we refer to ℓ_∞-AT LPIPS as the R-LPIPS.

To further demonstrate the robustness of R-LPIPS with re-

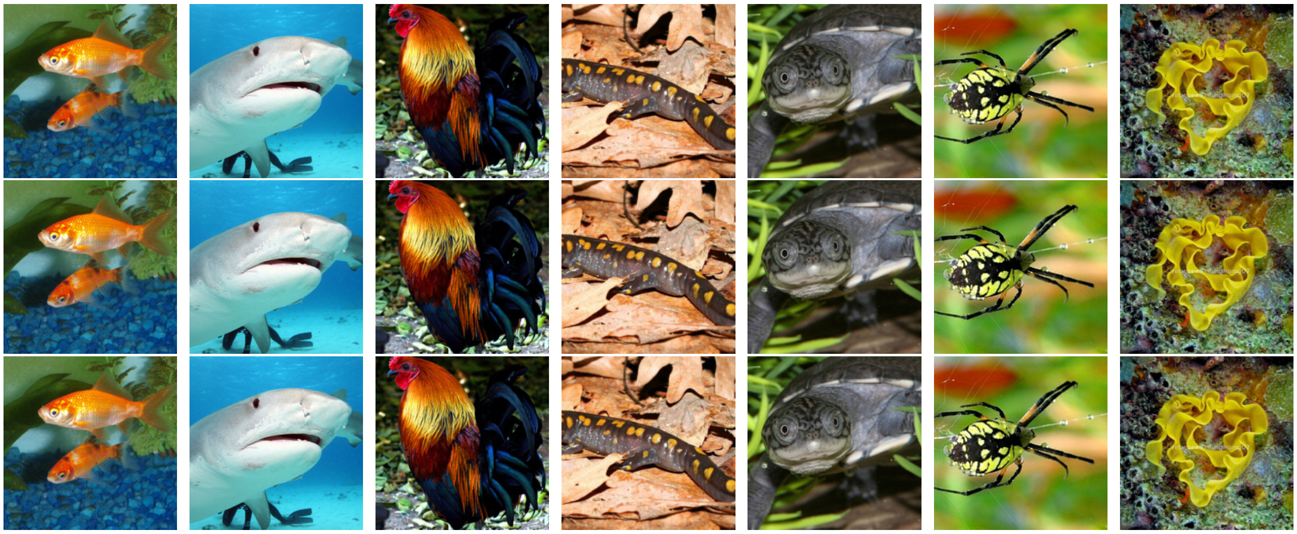


Figure 3. The adversarial images generated by the R-PPGA/R-LPA attacks. Original images are shown in the first row, and the adversarial images generated by R-PPGA and R-LPA (which are bounded by 0.5) are added to the second and third rows. Similar to adversarial images generated by PPGA/LPA (Laidlaw et al., 2021), the perturbations between images in this figure are invisible to human eyes.

spect to LPIPS, we computed the R-LPIPS metric on the OPT adversarial examples. The distribution of the values of R-LPIPS is shown in Figure 2a. One can observe that the two distributions are entirely separated and the values of R-LPIPS are very small when the ℓ_∞ perturbation is small. Figure 2b provides the same experiments but with OPT with R-LPIPS instead of LPIPS. The adversarial examples are therefore stronger but the distinction between LPIPS and R-LPIPS is still significant. To better illustrate, Figure 1 provides a set of images with the comparison between the LPIPS value and the R-LPIPS value.

4.3. Perceptual Adversarial Attack with R-LPIPS (Q3)

Perceptual Adversarial Training (PAT) was also proposed by Laidlaw et al. (2021) to train the model using perceptual attacks and come up with a perceptually robust model. In this section, we combined the PPGA and LPA attacks with our robust perceptual distance metric and developed attacks named R-PPGA and R-LPA. To compare the strength of attacks generated by LPIPS and R-LPIPS, we reproduced the accuracy of PAT to attacks generated by PPGA and LPA on CIFAR-10 and performed an experiment to compute the accuracy of PAT to R-PPGA and R-LPA attacks. Our results (Table 2) shows that PAT has relative robustness to attacks constrained with LPIPS, and is highly vulnerable to attacks bounded by R-LPIPS. The significant drop in the accuracy of PAT when exposed to R-PPGA and R-LPA attacks motivated us to visualize the adversarial data for attacks generated based on R-LPIPS; the results are shown in Figure 3. The first row consists of the original images, and the second and third rows are the adversarial images generated by R-PPGA and R-LPA attacks, respectively.

5. Conclusion & Future Work

Conclusion. In this paper, we showed that the LPIPS metric is vulnerable to adversarial examples and proposed R-LPIPS, a perceptual similarity metric that has been trained adversarially. During the process of adversarial training, the w_l weights are optimized while leaving the backbone weights of the model (AlexNet architecture) unchanged. Our findings reveal that R-LPIPS exhibits superior generalization and robustness across various data distortions when subjected to ℓ_∞-PGD and ℓ_2-PGD attacks. Additionally, we have investigated strong perceptual attacks using R-LPIPS, namely R-PPGA and R-LPA, and demonstrated their superiority over the previously established state-of-the-art attacks.

Future work. First, the R-LPIPS metric, which is an adversarially trained version of LPIPS achieved through ℓ_∞ AT on x_0, could be further explored by applying AT to x_1 or x_0 and x_1. Assessing the robustness of these different versions would offer valuable insights. Second, R-LPIPS can be used as a defense mechanism in similar settings as the PAT training scheme and LPIPS. It would be interesting to explore the development of R-PAT, which has the potential to be a more universal perceptual adversarial defense. Evaluating its performance under attack would provide valuable insights and potentially demonstrate superior results. Finally, by using adversarial training to develop R-LPIPS, the new metric inherits its drawback which is the lack of theoretical guarantees. An interesting future direction would be to devise guarantees to a perceptual metric.

Acknowledgments

This work was supported in part by the Army Research Office under grant number W911NF-21-1-0155 and by the New York University Abu Dhabi (NYUAD) Center for Artificial Intelligence and Robotics, funded by Tamkeen under the NYUAD Research Institute Award CG010.

References

Araujo, A., Meunier, L., Pinot, R., and Negrevergne, B. Advocating for multiple defense strategies against adversarial examples. In *ECML PKDD 2020 Workshops*, 2020.

Araujo, A., Negrevergne, B., Chevaleyre, Y., and Atif, J. On lipschitz regularization of convolutional layers using toeplitz matrix theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Araujo, A., Havens, A. J., Delattre, B., Allauzen, A., and Hu, B. A unified algebraic perspective on lipschitz neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, 2018.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 2017.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, 2020.

Croce, F. and Hein, M. Mind the box: l_1-apgd for sparse adversarial attacks on image classifiers. In *International Conference on Machine Learning*, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 2009.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Hore, A. and Ziou, D. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, 2010.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Kettunen, M., Härkönen, E., and Lehtinen, J. E-lpips: robust perceptual image similarity via random transformation ensembles. *arXiv preprint arXiv:1906.03973*, 2019.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 2017.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial examples in the physical world. In *Artificial intelligence safety and security*. 2018.

Laidlaw, C., Singla, S., and Feizi, S. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations*, 2021.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 2011.

Meunier, L., Delattre, B. J., Araujo, A., and Allauzen, A. A dynamical system perspective for lipschitz neural networks. In *International Conference on Machine Learning*. PMLR, 2022.

Pinot, R., Meunier, L., Araujo, A., Kashima, H., Yger, F., Gouy-Pailler, C., and Atif, J. Theoretical evidence for adversarial robustness through randomization. *Advances in neural information processing systems*, 2019.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2003.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 2004.

Zhang, L., Zhang, L., Mou, X., and Zhang, D. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 2011.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.