On the Suitability of L_p-norms for Creating and Preventing Adversarial Examples

Mahmood Sharif[†] Lujo Bauer[†] Michael K. Reiter[‡]
[†]Carnegie Mellon University
[‡]University of North Carolina at Chapel Hill
{mahmoods,lbauer}@cmu.edu, reiter@cs.unc.edu

Abstract

Much research has been devoted to better understanding adversarial examples, which are specially crafted inputs to machine-learning models that are perceptually similar to benign inputs, but are classified differently (i.e., misclassified). Both algorithms that create adversarial examples and strategies for defending against adversarial examples typically use L_p-norms to measure the perceptual similarity between an adversarial input and its benign original. Prior work has already shown, however, that two images need not be close to each other as measured by an L_p-norm to be perceptually similar. In this work, we show that nearness according to an L_p-norm is not just unnecessary for perceptual similarity, but is also insufficient. Specifically, focusing on datasets (CIFAR10 and MNIST), L_p-norms, and thresholds used in prior work, we show through online user studies that "adversarial examples" that are closer to their benign counterparts than required by commonly used L_p-norm thresholds can nevertheless be perceptually distinct to humans from the corresponding benign examples. Namely, the perceptual distance between two images that are "near" each other according to an L_p-norm can be high enough that participants frequently classify the two images as representing different objects or digits. Combined with prior work, we thus demonstrate that nearness of inputs as measured by L_p-norms is neither necessary nor sufficient for perceptual similarity, which has implications for both creating and defending against adversarial examples. We propose and discuss alternative similarity metrics to stimulate future research in the area.

1. Introduction

Machine learning is quickly becoming a key aspect of many technologies that impact us on a daily basis, from automotive driving aids to city planning, from smartphone cameras to cancer diagnosis. As such, the research community has invested substantial effort in understanding *adversarial examples*, which are inputs to machine-learning systems that are perceptually similar to benign inputs, but that are misclassified, i.e., classified differently than the benign inputs from which they are derived (e.g., [1, 29]). An attacker who creates an adversarial example can cause an object-recognition algorithm to incorrectly identify an object (e.g., as a worm instead of as an panda) [7], a street-sign recognition algorithm to fail to recognize a stop sign [5], or a face-recognition system to fail to identify a person [27]. Because of the potential impact on safety and security, better understanding the susceptibility of machine-learning algorithms to adversarial examples, and devising defenses, has been a high priority.

A key property of adversarial examples that makes them dangerous is that human observers do not recognize them as adversarial. If a human recognizes an input (e.g., a person wearing a disguise at an airport) as adversarial, then any potential harm may often be prevented by traditional methods (e.g., physically detaining the attacker). Hence, most research on creating adversarial examples (e.g., [2, 23]) or defending against them (e.g., [17, 12]) focuses on adversarial examples that are *imperceptible*, i.e., a human would consider them perceptually similar to benign images.

The degree to which an adversarial example is imperceptible from its benign original is usually measured using L_p-norms, e.g., L_0 (e.g., [23]), L_2 (e.g., [29]), or L_∞ (e.g., [7]). Informally, for images, L_0 measures the number of pixels that are different between two images, L_2 measures the Euclidean distance between two images, and L_∞ measures the largest difference between corresponding pixels in two images. These measures of imperceptibility are critical for creating adversarial examples and defending against them. On the one hand, algorithms for creating adversarial examples seek to enhance their imperceptibility by producing inputs that both cause misclassification and whose distance from their corresponding benign originals has small L_p-norm. On the other hand, defense mechanisms assume that if the difference between two inputs is below a specific

L_p-norm threshold then the two objects belong to the same class (e.g., [17]).

Hence, the completeness and soundness of attacks and defenses commonly rely on the assumption that some L_p-norm is a reasonable measure for perceptual similarity, i.e., that if the L_p-norm of the difference between two objects is below a threshold then the difference between those two objects will be imperceptible to a human, and vice versa. Recent work has shown that one direction of this assumption does not hold: objects that are indistinguishable to humans (e.g., as a result of slight rotation or translation) can nevertheless be very different as measured by the L_p-norm of their difference [4, 10, 34].

In this paper we further examine the use of L_p-norms as a measure of perceptual similarity. In particular, we examine whether pairs of objects whose difference is small according to an L_p-norm are indeed similar to humans. Focusing on datasets, L_p-norms, and thresholds used in prior work, we show that small differences between images according to an L_p-norm do not imply perceptual indistinguishably. Specifically, using the CIFAR10 [13] and MNIST [15] datasets, we show via online user studies that images whose distance—as measured by the L_p-norm of their difference—is below thresholds used in prior work can nevertheless be perceptibly very different to humans. The perceptual distance between two images can in fact lead people to classify two images differently. For example, we find that by perturbing about 4% of pixels in digit images to achieve small L_0 distance from benign images (an amount comparable to prior work [23]), humans become likely to classify the resulting images correctly only 3% of the time.

Combined with previous work, our results show that nearness between two images according to an L_p-norm is neither necessary nor sufficient for those images to be perceptually similar. This has implications for both attacks and defenses against adversarial inputs. For attacks, it suggests that even though a candidate attack image may be within a small L_p distance from a benign image, this does not ensure that a human would find the two images perceptually similar, or even that a human would classify those two images consistently (e.g., as the same person or object). For defenses, it implies defense strategies that attempt to train machine-learning models to correctly classify what ought to be an adversarial example may be attempting to solve an extremely difficult problem, and may result in ill-trained machine-learning models.

To stimulate future research on developing better similarity metrics for comparing adversarial examples with their benign counterparts, we propose and discuss several alternatives to L_p-norms. In doing so, we hope to improve attacks against machine-learning algorithms, and, in return, defenses against them.

Next, we review prior work and provide background (Sec. 2). We then discuss the necessity and sufficiency of conditions for perceptual similarity, and show evidence that L_p-norms lead to conditions that are neither necessary nor sufficient (Sec. 3–4). Finally, we discuss alternatives to L_p-norms and conclude (Sec. 5–6).

2. Background and Related Work

In concurrent research efforts, Szegedy et al. and Biggio et al. showed that specifically crafted small perturbations of benign inputs can lead machine-learning models to misclassify them [1, 29]. The perturbed inputs are referred to as *adversarial examples* [29]. Given a machine-learning model, a sample $\hat{x}$ is considered as an adversarial example if it is *similar* to a benign sample x (drawn from the data distribution), such that x is correctly classified and $\hat{x}$ is classified differently than x. Formally, for a classification function F, a class c_x of x, a distance metric D, and a threshold ϵ, $\hat{x}$ is considered to be an adversarial example if:

$$F(x) = c_x \wedge F(\hat{x}) \neq c_x \wedge D(x, \hat{x}) \leq \epsilon \qquad (1)$$

The leftmost condition ($F(x) = c_x$) checks that x is correctly classified, the middle condition ($F(\hat{x}) \neq c_x$) ensures that $\hat{x}$ is incorrectly classified, and the rightmost condition ($D(x, \hat{x}) \leq \epsilon$) ensures that x and $\hat{x}$ are similar (i.e., their distance is small) [1].

Interestingly, the concept of similarity is ambiguous. For example, two images may be considered similar because both contain the color blue, or because they are indistinguishable (e.g., when performing ABX tests [20]). We believe that prior work on adversarial examples implicitly assumes that similarity refers to *perceptual or visual similarity*, as stated by Engstrom et al. [4]. As Goodfellow et al. explain, adversarial examples are "particularly disappointing because a popular approach in computer vision is to use convolutional network features as a space where Euclidean distance approximates perceptual distance" [7]. In other words, adversarial examples are particularly of interest because they counter our expectation that neural networks specifically, and machine-learning models in general, represent perceptually similar inputs with features that are similar (i.e., close) in the Euclidean space.

A common approach in prior work has been to use L_p-distance metrics (as D) in attacks that craft adversarial examples and defenses against them (e.g., [2]). For non-negative values of p, the L_p distance between the two d-dimensional inputs x and $\hat{x}$ is defined as [25]:

$$||x - \hat{x}||_p = \Big(\sum_{i=1}^{d} |x_i - \hat{x}_i|^p \Big)^{\frac{1}{p}}$$

The main L_p distances used in the literature are L_0, L_2, and L_∞. Attacks using L_0 attempt to minimize the number of pixels perturbed [2, 23]; those using L_2 attempt to

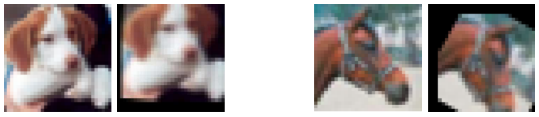
minimize the Euclidean distance between x and $\hat{x}$ [2, 29]; and attacks using L_∞ attempt to minimize the maximum change applied to any coordinate in x [2, 7].

To defend against adversarial examples, prior work has focused on either training more robust deep neural networks (DNNs) that are not susceptible to small perturbations [7, 11, 14, 17, 24, 29, 12], or developing techniques to detect adversarial examples [6, 8, 18, 19, 35]. *Adversarial training* is a particular defense that has achieved a relatively high success [7, 11, 14, 17, 29]. In this defense, adversarial examples with bounded L_p distance (usually, $p = \infty$) are generated in each iteration of training DNNs, and the DNN is trained to correctly classify those examples. Insufficiency and lack of necessity in L_p-distance metrics have direct implication on adversarial training, as we discuss in Sec. 3.

Despite the goal for adversarial examples to be perceptually similar to benign samples, little prior work on adversarial examples has explicitly explored or accounted for human perception. The theoretical work of Wang et al. is an exception, as they treated the human as an oracle to seek for the conditions under which DNNs would be robust [31]. In contrast, we take an experimental approach to show that L_p distances may be inappropriate for defining adversarial examples. Our findings are in line with research in psychology, which has found that distance metrics in geometric spaces may not always match humans' assessment of similarity (e.g., [30]). Concurrently to our work, Elsayed et al. showed that adversarial examples can mislead humans as well as DNNs [3]. While they considered images of higher dimensionality than we consider in this work (see Sec. 4), they allowed perturbations of higher norm than commonly found in practice.

Some work proposed generating adversarial examples via techniques other than minimizing L_p-norms. In three different concurrent efforts, researchers proposed to use minimal geometric transformations to generate adversarial examples [4, 10, 34]. As we detail in the next section, geometric transformations evidence that conditions on L_p-norms are unnecessary for ensuring similarity. In other work, researchers proposed to achieve imperceptible adversarial examples by maximizing the Structural Similarity (SSIM) with respect to benign images [25]. SSIM is a measure of perceived quality of images that has been shown to align with human assessment [33]. It is a differentiable metric with values in the interval [-1,1] (where a values closer to 1 indicate higher similarity). By maximizing SSIM, the researchers hoped to increase the similarity between the adversarial examples and their benign counterparts. In Sec. 5 we provide a preliminary analysis of SSIM as a perceptual similarity metric for adversarial examples.

In certain cases, perceptual similarity to a reference image is not a goal for an attack (e.g., images that seem incomprehensible or benign to humans, but are classified as street signs by machine-vision algorithms [16, 21, 28]). While such attacks are important to defend against, this paper focuses on studying the notion of perceptual similarity that is relied upon in the majority of the literature on adversarial examples.



(a) (b)

Figure 1: Translations and rotations can fool DNNs. (a) A dog image (right) resulting from transforming a benign image (left) is classified as a cat. (b) A horse image (right) resulting from transforming a benign image (left) is classified as a truck. Images from Engstrom et al. [4].

3. Necessity and Sufficiency of Conditions for Perceptual Similarity

To effectively find adversarial examples and defend against them, the parameter choice in Eqn. 1 should help us capture the set of all interesting adversarial examples. In particular, the selection of D and ϵ should capture the samples that are perceptually similar to benign samples. Ideally, we should be able to define *necessary* and *sufficient* conditions for perceptual similarity via D and ϵ.

3.1. Necessity

The condition $C := D(x, \hat{x}) \leq \epsilon$ is a necessary condition for perceptual similarity if:

$$\hat{x} \text{ is perceptually similar to } x \Rightarrow C$$

Finding necessary conditions for perceptual similarity is important for the development of better attacks that find adversarial examples, as well as better defenses. If the condition C used in hopes of ensuring perceptual similarity is unnecessary (i.e., there exist examples that are perceptually similar to benign samples, but do not satisfy C), then the search space of attacks may be too constrained and some stealthy adversarial examples may not be found. For defenses, and especially for adversarial training (because the DNNs are specifically trained to prevent adversarial inputs that satisfy C), unnecessary conditions for perceptual similarity may lead us to fail at defending against adversarial examples that do not satisfy C.

Attacks that craft adversarial examples via applying slight geometric transformations (e.g., translations and rotations) to benign samples [4, 10, 34] evidence that when using L_p-distance metrics as the measure of distance, D, we may wind up with unnecessary conditions for perceptual similarity. Such geometric transformations result in

small perceptual differences with respect to benign samples, yet they result in large L_p distances. Fig. 1 shows two adversarial examples resulting from geometric transformation of 32×32 images. While the adversarial examples are similar to the benign samples, their L_p distance is large: $L_0 \geq$3,010 (maximum possible is 3,072), $L_2 \geq$15.83 (maximum possible is 55.43), and $L_\infty \geq$0.87 (maximum possible is 1).[1] These distances are much larger that what has been used in prior work (e.g., [2]). Indeed, because small L_p is not a necessary condition for perceptual similarity, state-of-the-art defenses meant to defend against L_p-bounded adversaries fail at defending against adversarial examples resulting from geometric transformations [4].

3.2. Sufficiency

The condition $C := D(x, \hat{x}) \leq \epsilon$ is a sufficient condition for perceptual similarity if:

$$C \Rightarrow \hat{x} \text{ is perceptually similar to } x$$

Alternately, C is insufficient, if it is possible to demonstrate that a sample $\hat{x}$ is close to x under D, while in fact $\hat{x}$ is not perceptually similar to x. The sufficiency of C for perceptual similarity is also important for both attacks and defenses. In the case of attacks, an adversary using insufficient conditions for perceptual similarity may craft misclassified samples that she may deem as perceptually similar to benign samples under C, when they are not truly so. For defenses, the defender may be attempting to solve an extremely difficult problem by requiring that a machine-learning model would classify inputs in a certain way, while even humans may be misled by such inputs due to their lack of perceptual similarity to benign inputs. In the case of adversarial training, if C is insufficient, we may even train DNNs to classify inputs differently than how humans would (thus, potentially poisoning the DNN).

In the next section we show that the L_p-norms commonly used in prior work may be insufficient for ensuring perceptual similarity. We emphasize that our findings should *not* be interpreted as stating that the adversarial examples reported on by prior work are not imperceptible. Instead, our findings highlight that commonly used L_p-norms and associated thresholds in principle permit algorithms for crafting adversarial examples to craft samples that are not perceptually similar to benign ones, leading to the undesirable outcomes described above. Specific instances of algorithms that use those L_p-norms and thresholds could still produce imperceptible adversarial examples because the creation of these examples is constrained by factors other than the chosen L_p-norms and thresholds.

[1] We use RGB pixel values in the range [0,1].

4. Experiment Design and Results

To show that a small L_p distance ($p \in \{0, 2, \infty\}$) from benign samples is insufficient for ensuring perceptual similarity to these samples, we conducted three online user studies (one for each p). The goal of each study was to show that, for small values of ϵ, it is possible to find samples that are close to benign samples in L_p, but are not perceptually similar to those samples to a human. In what follows, we present the high-level experimental design that is common among the three studies. Next, we report on our study participants. Then, we provide the specific design details for each study and the results.

4.1. Experiment Design

Due to the many ways in which two images can be similar, it is unclear whether one can learn useful input from users by directly asking them about the level of similarity between image pairs. Therefore, we rely on indirect reasoning to determine whether an image $\hat{x}$ is perceptually similar to x. In particular, we make the following observation: if we ask mutually exclusive sets of users about the contents of $\hat{x}$ and x, and they disagree, then we learn that x and $\hat{x}$ are *definitely* dissimilar; otherwise, we learn that they are *likely* similar. Our observation is motivated by Papernot et al.'s approach to determine that their attack is imperceptible to humans [23].

Motivated by the above-mentioned observation, we followed a between-subject design for each study, assigning each participant to one of three conditions: C_B, C_{AI}, and C_{AP}. In all the conditions, participants were shown images and were asked to select the label (i.e., category) that best describes the image out of ten possible labels (e.g., the digit shown in the image). The conditions differed in the nature of images shown to the participants. Participants in C_B ("benign" condition) were shown unaltered images from standard datasets. Participants in C_{AI} ("adversarial and imperceptible" condition) were shown adversarial examples of images in C_B that fool state-of-the-art DNNs. The images in C_{AI} have small L_p distances to images in C_B, and *were not* designed to mislead humans. Participants in C_{AP} ("adversarial and perceptible" condition) were shown variants of the images in C_B that are close in L_p distance to their counterparts in C_B, but were designed to mislead both humans and DNNs. To lower the mental burden on participants, each participant was asked only 25 image-categorization questions. Because the datasets we used contain thumbnail images (see below), we presented the images to participants in three different sizes: original size, resized $\times 2$, and resized $\times 4$. Additionally to categorizing images, we asked participants in all conditions about their level of confidence in their answers on a 5-point Likert scale (one denotes low confidence and five denotes high confidence). The protocol was approved by Carnegie Mellon's review board.

Conceptually, the responses of participants in C_B help us estimate humans' accuracy on benign images (i.e., their likelihood to pick the labels consistent with the ground truth). By comparing the accuracy of users in C_{AI} to C_B, we learn whether the attack and the threshold on L_p distance that we pick result in imperceptible attacks. We hypothesize that images in C_{AI} are likely to be categorized correctly by users. Hence, the attack truly crafts adversarial examples, and poses risk to the integrity of DNNs at the chosen threshold. In contrast, by comparing C_{AP} with C_B, we hope to show that, for the same threshold, it is possible to find instances that mislead humans and DNNs. Namely, we hypothesize that the accuracy of users on C_{AP} is significantly lower than their accuracy on C_B. If our hypothesis is validated, we learn that small L_p distance does not ensure perceptual similarity.

Datasets and DNNs. In the studies, we used images from the MNIST [15] and CIFAR10 [13] datasets. MNIST is a dataset of 28×28 pixel images of digits, while CIFAR10 is a dataset of 32×32 pixel images that contain ten object categories: airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks. Both MNIST and CIFAR10 are widely used for developing both attacks on DNNs and defenses against them (e.g., [7, 17, 23]).

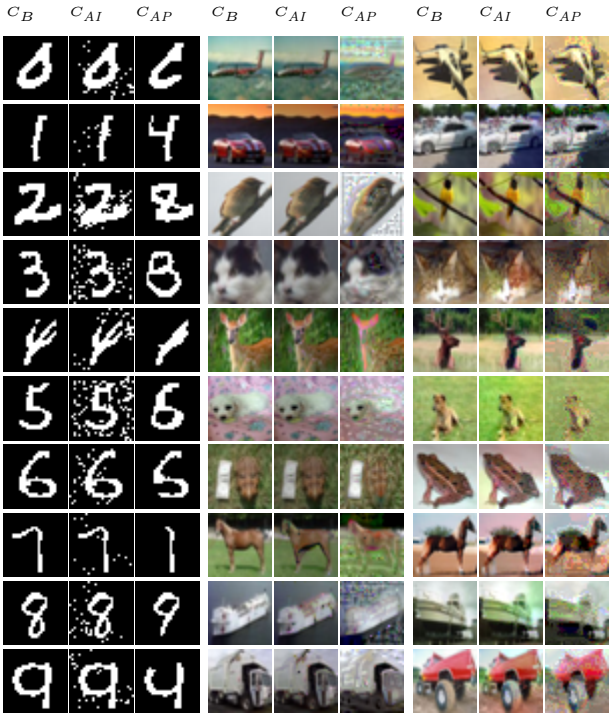
In conditions C_{AI} and C_{AP}, we created attacks against two DNNs published by Madry et al. [17]—one for MNIST, and another for CIFAR10. The MNIST DNN is highly accurate, achieving 98.8% accuracy on the MNIST test set. The CIFAR10 DNN also achieves a relatively high accuracy—87.3% on CIFAR10's test set. More notably, both DNNs are two of the most resilient models to adversarial examples (specifically, ones with bounded L_∞ distance from benign samples) known to date.

4.2. Participants

We recruited a total of 399 participants from the United States through the Prolific crowdsourcing platform. Their ages ranged from 18 to 78, with a mean of 32.85 and standard deviation of 11.54. The demographics of our participants were slightly skewed toward males: 59% reported to be males, 39% reported to be females, and 1% preferred not to specify their gender. 34% of the participants were students, and 84% were employed at least partially. Participants took an average of roughly six minutes to complete the study and were compensated $1.

4.3. Insufficiency of L_0

Study Details. We used the MNIST dataset and DNN to test whether it is possible to perturb images only slightly on L_0 while simultaneously misleading humans and DNNs. C_B was assigned 75 randomly selected images from the test set of MNIST. All 75 images were correctly classified by the DNN. We used the Jacobian Saliency Map Approach (JSMA) [23], as implemented in Cleverhans [22], to craft adversarial examples for C_{AI}. We limited the amount of change applied to an image to at most 15% of the pixels, and found that JSMA was able to find successful attacks for 68 out of 75 images. For successful attacks, JSMA perturbed 4.90% (2.91% standard deviation) of pixels on average—this is comparable to the result of Papernot et al. [23]. Images in C_{AP} were crafted manually. Using a photo-editing software[2] while simultaneously receiving feedback from the DNN, a member of our team edited the 75 images from C_B while attempting to minimize the number of pixels changed such that the resulting image would be misclassified by both humans and the DNN. The average L_0 distance of images in C_{AP} from their counterparts in C_B is 4.48% (2.45% standard deviation). Examples of the images in the three conditions are shown in Fig. 2a. We note that because creating the images for C_{AP} involves time-consuming manual effort, we limited each condition to at most 75 images.

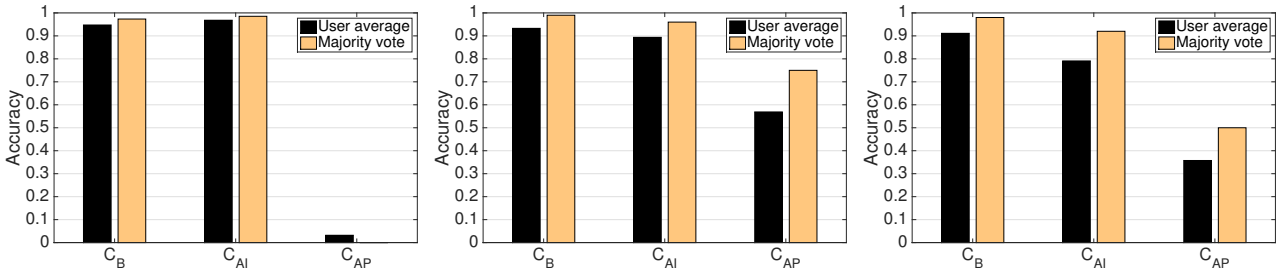


(a) L_0 (b) L_2 (c) L_∞

Figure 2: Sample images from the three conditions we had for each L_p-norm. Each row shows three variants of the same image.

A total of 201 participants were assigned to the L_0 study: 73 were assigned to C_B, 59 to C_{AI}, and 69 to C_{AP}.

[2] GIMP (`https://www.gimp.org/`)



(a) L_0 (b) L_2 (c) L_∞

Figure 3: User performance for the three L_p-norms that we studied. For each condition (C_B, C_{AI}, and C_{AP}), we report the users' average accuracy, and the accuracy when labeling each image by the majority vote (over the labels provided by the participants). Accuracy is the fraction of labels that are consistent with the ground truth.

Experiment results. We computed the users' accuracy (i.e., how often their responses agreed with the ground truth), and the accuracy when classifying each image according to the majority vote over all labels provided by the users. Fig. 3a shows the results. We found that users' average accuracy was high for the unaltered images of C_B (95%) and adversarial images of C_{AI} (97%), but low for the images of C_{AP} (3%). In fact, if we classify images via majority votes, none of the images of C_{AP} would be classified correctly. The difference between users' average accuracy in C_B and C_{AI} is not statistically significant according to a t-test ($p = 0.34$). In contrast, the difference between users' average accuracy in C_{AP} and other conditions is statistically significant ($p < 0.01$). Users in all the conditions were confident in their responses—the average confidence levels ranged from 4.19 (C_{AP}) to 4.47 (C_B).

The results support our hypotheses. While it is possible to find adversarial examples with small L_0 distance from benign samples, it is possible, for the same distance, to find samples that are not imperceptible to humans. In fact, humans may be highly confident that those samples belong to other classes.

4.4. Insufficiency of L_2

Study Details. We used the CIFAR10 dataset and DNN to test L_2 for insufficiency. We randomly picked 100 images from the test set that were correctly classified by the DNN for C_B. For each image in C_B, we created (what we hoped would be) an imperceptible adversarial example for C_{AI}. Images in C_{AI} have a fixed L_2 distance of 6 from their counterparts in C_B. Because we did not find evidence in the literature for an upper bound on L_2 distance that is is still imperceptible to humans, we chose a distance of 6 empirically—our results (below) support this choice. To create the adversarial examples, we used an iterative gradient descent approach, in the vein of prior work [2], albeit with two notable differences. First, we used an algorithm by Wang et al. [33] to initialize the attack to an image that has high SSIM to the benign image, but lies at a fixed L_2 distance from it. The rationale behind this was to increase the perceptual similarity between the adversarial image and the benign image. Second, we ensured that the L_2-norm of the perturbation is fixed by normalizing it to 6 after each iteration of the attack. The images of C_{AP} we generated via a similar approach to those of C_{AI}. The only difference is that we initialized the attack with an image that has low SSIM with respect to its counterpart benign image using Wang et al.'s algorithm. Fig. 2b shows a sample of the images that we used in the L_2 study.

In total, we had 99 participants assigned to the L_2 study: 25 were assigned to C_B, 38 to C_{AI}, and 36 to C_{AP}.

Experiment results. Users' average accuracy and the accuracy of the majority vote are shown in Fig. 3b. On the benign images of C_B, users had an average accuracy of 93%. Their average accuracy on the images of C_{AI} was 89%, not significantly lower ($p \approx 1$). Moreover, we found that users in C_B and C_{AI} were almost equally confident about their choices (averages of 4.31 and 4.37). We thus concluded that 6 is a reasonable bound for L_2 attacks. In stark contrast, we found that C_{AP} users' average accuracy dropped to 57% and confidence to 2.97 ($p < 0.01$ for both). In other words, users' likelihood to make mistakes increased by 36%, on average, and their confidence in their decisions decreased remarkably.

The results support our hypotheses, as a significant fraction of attack samples with bounded L_2 can be perceptually different than their corresponding benign samples.

4.5. Insufficiency of L_∞

Study Details. Similarly to the L_2 study, we used the CIFAR10 dataset and DNN also for the L_∞ study. We again picked 100 random images from the test set for C_B. For C_{AI}, we generated adversarial examples with L_∞ distance of 0.1 from benign examples, as done by Goodfellow et al. [7]. We generated the adversarial examples using the Projected Gradient Descent algorithm [17], with a simple

tweak to enhance imperceptibility: after each gradient descent iteration, we increased SSIM with respect to benign images using Wang et al.'s algorithm [33]. Examples in C_{AP} were generated using a similar algorithm, but we decreased the SSIM with respect to benign images instead of increasing it.

In total, we had 99 participants assigned to the study: 36 were assigned to C_B, 31 to C_{AI}, and 32 to C_{AP}.

Experiment results. The accuracy results are summarized in Fig. 3c. On benign examples, users' average accuracy was 91%. Their average confidence score was 4.45. The attacks in C_{AI} were not completely imperceptible: users' average accuracy decreased to 79% and confidence score to 3.98 (both significant with $p < 0.01$). However, attacks in C_{AP} were significantly less similar to benign images: users' average accuracy and confidence score were 36% and 3.04 ($p < 0.01$).

Our hypotheses hold also for L_∞—a significant fraction of attacks with relatively small L_∞ can be perceptually different to humans than benign images.

5. Discussion

The results of the user studies confirm our hypotheses—defining similarity using L_0, L_2, and L_∞-norms can be insufficient for ensuring perceptual similarity in some cases. Here, we discuss some of the limitations of this work, and discuss some alternatives for L_p-norms.

5.1. Limitations

A couple of limitations should be considered when interpreting our results. First, we demonstrated our results on two DNNs and for two datasets, and so they may not apply for every DNN and image-recognition task. The DNNs that we considered are among the most resilient models to adversarial examples to date. Consequently, we believe that the attacks we generated against them for C_{AI} and C_{AP} would succeed against other DNNs. Depending on the chosen thresholds, our results may or may not directly apply to specific combinations of norms and image-recognition tasks that we did not consider in this work. While studying more combinations may be useful, we believe that our findings are impactful in their current form, as the combinations of norms, thresholds, and datasets we considered are commonly used in practice (e.g., [2, 7, 23, 29]). We note that it may be possible to achieve sufficient conditions for perceptual similarity by using lower thresholds than in our experiments. Stated differently, it may be impossible to fool humans using lower thresholds. However, decreasing thresholds may also prevent algorithms for crafting attacks from finding successful adversarial examples (namely, the ones that are part of C_{AI} in the experimental conditions).

Second, we estimated similarity using a proxy: whether participants' categorizations of perturbed images were consistent with their categorizations of their benign counterparts. However, similarity has different facets that may or may not be interesting, depending on the threat model being studied. For instance, in some cases we may want to estimate whether certain attacks are inconspicuous or not (e.g., to learn whether TSA agents would detect disguised individuals attempting to mislead surveillance systems). In such cases, we want to measure whether adversarial images are "close enough" to benign images to the extent that a human observer cannot reliably tell between adversarial and non-adversarial images. We believe that future work should develop a better understanding of this and other notions of similarity and how to assess them as a means to help us improve current attacks and defenses.

5.2. Alternatives to L_p-norms

We next list several alternative distance metrics for assessing similarity and provide a preliminary assessment of their suitability as a replacement for L_p-norms.

Our results show that by using the same threshold for all samples, one may generate adversarial images that mislead humans—thus, they are not imperceptible. As a solution, one may consider setting a different threshold for every sample to ensure that the attacks' output would be imperceptible. In this case, a principled automated method would be needed to set the sample-dependent thresholds in order to create attacks at scale. Without such a method, human feedback may be needed in the process for every sample.

Other similarity-assessment measures, such as SSIM [33] and the "minimal" transform needed to align two images, that have been used in prior work (e.g., [10, 26]) may be considered to replace L_p-norms. Additionally, one may consider image-similarity metrics that have not been previously used in the adversarial examples literature, such as Perceptual Image Diff [36] and the Universal Quality Index [32]. Such metrics should be treated with care, as they may lead to unnecessary and insufficient conditions for perceptual similarity. For example, SSIM is sensitive to small geometric transformation (e.g., the SSIM between the images in Fig. 1b is 0.36 out of 1, which is relatively low [33]). So, using SSIM to define similarity may lead to unnecessary conditions. Moreover, as demonstrated in Fig. 4, SSIM may be high even when two images are not similar. Thus, SSIM may lead to insufficient conditions for similarity.

The recent work of Jang et al. suggests three metrics to evaluate the similarity between adversarial examples and benign inputs [9]. The metrics evaluate adversarial perturbations' quality by how much they corrupt the Fourier transform, their effect on edges, or their effect on the images' gradients. This work appears to take a step in the right direction. However, further validation of the proposed metrics is needed. For instance, we speculate that slight ge-

(a) (b) (c)

Figure 4: SSIM can be high between two images containing different objects or subjects. Despite showing different animals or subjects (the bottom image in (b) was created by swapping the faces of actress Judi Dench and actor Daniel Craig), the SSIM value between the images in (a) and (b) is high—0.89 between the top images, and 0.95 between the bottom images. Images in (c) were created by adding uniform noise to images in (a). The SSIM value between (a) and (c) is 0.77 for the images at the top, and 0.87 for the images at the bottom. (Sources of images in (a) and (b): `https://goo.gl/Mxo9mK`, `https://goo.gl/GEd6Bs`, `https://goo.gl/mvvFZ1`, and `https://goo.gl/vwuK9t`.)

ometric transformations to a benign image might affect the gradients in the image dramatically. Thus, metrics based on gradients alone may be unnecessary for defining conditions for perceptual similarity.

Lastly, one may consider using several metrics simultaneously to define similarity (e.g., by allowing geometric transformations and perturbations with small L_p-norm to craft adversarial examples [4]). While this may be a promising direction, metrics should be combined with special care. As the conjunction of one or more unnecessary conditions leads to an unnecessary condition, and the disjunction of one or more insufficient conditions leads to an insufficient condition, simply conjoining or disjoining metrics may not solve the (in)sufficiency and (un)necessity problems of prior definitions of similarity.

Finding better measures for assessing perceptual similarity remains an open problem. Better similarity measures could help improve both algorithms for finding adversarial examples and, more importantly, defenses against them. In the absence of such measures, we recommend that future research rely not only on known metrics for perceptual similarity assessment, but also on human-subject studies.

6. Conclusion

In this work, we aimed to develop a better understanding of the suitability of L_p-norms for measuring the similarity between adversarial examples and benign images. Specifically, we complemented our knowledge that conditions on L_p distances used for defining similarity are unnecessary in some cases—i.e., they may not capture all imperceptible attacks—and showed that they can also be insufficient—i.e., they may lead one to conclude that an adversarial instance is similar to a benign instance when it is not so. Hence, L_p-distance metrics may be unsuitable for assessing similarity when crafting adversarial examples and defending against them. We pointed out possible alternatives for L_p-norms to assess similarity, though they seem to have limitations, too. Thus, there is a need for further research to improve the assessment of similarity when developing attacks and defenses for adversarial examples.

References

[1] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Proc. ECML PKDD*, 2013.

[2] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *Proc. IEEE S&P*, 2017.

[3] G. F. Elsayed, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, and J. Sohl-Dickstein. Adversarial examples that fool both human and computer vision. *arXiv preprint arXiv:1802.08195*, 2018.

[4] L. Engstrom, D. Tsipras, L. Schmidt, and A. Madry. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

[5] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2017.

[6] R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[8] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

[9] U. Jang, X. Wu, and S. Jha. Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In *Proc. ACSAC*, 2017.

[10] C. Kanbak, S.-M. Moosavi-Dezfooli, and P. Frossard. Geometric robustness of deep networks: analysis and improvement. *arXiv preprint arXiv:1711.09115*, 2017.

[11] A. Kantchelian, J. Tygar, and A. D. Joseph. Evasion and hardening of tree ensemble classifiers. In *Proc. ICML*, 2016.

[12] J. Z. Kolter and E. Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851*, 2017.

[13] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

[14] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. In *ICLR*, 2017.

[15] Y. LeCun, C. Cortes, and C. J. Burges. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/.

[16] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang. Trojaning attack on neural networks. In *Proc. NDSS*, 2018.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. PADL Workshop*, 2017.

[18] D. Meng and H. Chen. MagNet: a two-pronged defense against adversarial examples. In *Proc. CCS*, 2017.

[19] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. In *ICLR*, 2017.

[20] W. Munson and M. B. Gardner. Standardizing auditory tests. *The Journal of the Acoustical Society of America*, 22(5):675–675, 1950.

[21] A. M. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proc. CVPR*, 2015.

[22] N. Papernot, N. Carlini, I. Goodfellow, R. Feinman, F. Faghri, A. Matyasko, K. Hambardzumyan, Y.-L. Juang, A. Kurakin, R. Sheatsley, A. Garg, and Y.-C. Lin. Cleverhans v2.0.0: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768*, 2017.

[23] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Proc. IEEE Euro S&P*, 2016.

[24] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Proc. IEEE S&P*, 2016.

[25] F. Riesz. Untersuchungen über systeme integrierbarer funktionen. *Mathematische Annalen*, 69(4):449–497, 1910.

[26] A. Rozsa, E. M. Rudd, and T. E. Boult. Adversarial diversity and hard positive generation. In *Proc. CVPRW*, 2016.

[27] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proc. CCS*, 2016.

[28] C. Sitawarin, A. N. Bhagoji, A. Mosenia, M. Chiang, and P. Mittal. Darts: Deceiving autonomous cars with toxic signs. *arXiv preprint arXiv:1802.06430*, 2018.

[29] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[30] A. Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[31] B. Wang, J. Gao, and Y. Qi. A theoretical framework for robustness of (deep) classifiers under adversarial noise. In *Proc. ICLR Workshop*, 2017.

[32] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, 2002.

[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[34] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song. Spatially transformed adversarial examples. In *Proc. ICLR 2018*, 2018.

[35] W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proc. NDSS*, 2018.

[36] H. Yee, S. Pattanaik, and D. P. Greenberg. Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Transactions on Graphics (TOG)*, 20(1):39–65, 2001.