ROBUSTNESS FROM PERCEPTION

Saeed Mahloujifar[1], **Chong Xiang**[1], **Vikash Sehwag**[1], **Sihui Dai**[1], **Prateek Mittal**[1]
[1]Princeton University

ABSTRACT

One of the main motivations behind the study of adversarial examples is to understand the security implications of "imperceptible" changes in the inputs that can change the prediction of machine learning models. One important criticism of the studies of adversarial robustness is that they mostly use ℓ_p distance metrics as a proxy for human perception. It is not clear if ℓ_p bounded perturbations are always imperceptible, or if all imperceptible perturbations have a small ℓ_p norm. Hence, it is often argued that we first have to formally and exactly model human perception. In this paper, we prove that modeling human perception is at least as hard as finding a robust model. We show that an "ideal" perception formulation would immediately imply robustness in the information-theoretic sense.

1 INTRODUCTION

Research on adversarial examples studies the effect of imperceptible perturbations on classification accuracy Goodfellow et al. (2014); Kurakin et al. (2016); Gu & Rigazio (2014); Qin et al. (2019). Specifically, we consider adversaries who add a small noise ϵ to an image x to get another image $x' = x+\epsilon$ with two conditions. First, we need x' to be close to x and perceived the same by a human. In other words, the adversary needs ϵ to be an "imperceptible" change. On the other hand, we require to change the prediction of a specific classifier h on the perturbed input, namely, $h(x) \neq h(x')$. To account for the imperceptibility of the noise, researchers often consider ℓ_p bounded perturbations as a proxy for human perception Madry et al. (2017); Shafahi et al. (2018); Cullina et al. (2018); Cohen et al. (2019); Salman et al. (2019). One common criticism here is that ℓ_p norms are only toy examples and do not capture human perception Gilmer et al. (2018); Laidlaw et al. (2020). In other words, even if we get robustness against ℓ_p bounded adversaries, we would have still need to consider adversaries that add noises beyond ℓ_p perturbation and are still imperceptible. The question then arises:

Do we need to first model human perception in a precise way, and then incorporate the perception model in the design of classifiers that are robust to imperceptible adversarial examples?

In this work, we argue that formulating human perception is a harder task than robustness at least in the information-theoretic sense. In particular, we show that if one can formulate "imperceptibility" as an algorithm then we can construct a robust inference algorithm that achieves strong robustness guarantees against imperceptible adversarial examples. In the rest of the paper, we first model "imperceptibility" as a relationship with certain properties. Then we show that given such a relationship, one can achieve strong robustness guarantees using a randomized inference algorithm. Although our result is information-theoretic, In the end, we provide a simple toy example for such a relationship and the robustness we achieve from it.

2 PERCEPTUAL EQUIVALENCE RELATION

We model imperceptibility as a relation between pair of images x and x'. In particular, we consider a relation P where $P(x, x') = 1$ means that x and x' are perceived identically in human eyes. On the other hand if $P(x, x') = 0$ then x and x' have visible differences, and humans would not consider them identically. Note that here the relationship P is different from the concept that the learner tries to learn. In particular, it is possible that $P(x, x') = 0$ while the true labels for x and x' are the same. For instance, considering the case of learning a classifier for classifying cats and dogs, for any two distinct cat images in the dataset we would probably have $p(x, x') = 0$ although the labels are equal

for these two photos. This is because humans would perceive different cat photos differently based on many other aspects that exist in the image. For example, changing the color of the cat, or the background can completely change the human perception.

Why relation instead of metric? As stated above, we model perception as a relationship rather than a distance metric. In particular, two images x and x' either look the same or do not. On the other hand, a metric would define a real value for a distance between two images. We believe that an ideal perception formulation should not be based on real value distances between different images as humans do not assign numbers to the difference between identically looking images.

Additionally we believe the ideal perception relationship P should have several properties:

1. Reflective: For any image x we have $P(x, x) = 1$. This means that two images that are exactly identical are equivalent in human eyes too.
2. Symmetric: For all pairs of images x, x' such that $P(x, x') = 1$ we have $P(x', x) = 1$. This is expected because we the order of two images is irrelevant in the perception relationship.
3. Transitive: For any three images x, x', x'' such that $P(x, x') = 1$ and $P(x', x'') = 1$, then we have $P(x, x'') = 1$. Again, recall that perceptual relation is supposed to capture a relationship between pairs of images (x, x') in a way that $P(x, x') = 1$ if and only if x and x' have no difference in human eyes. Using this logic, we can justify the transitivity for perceptual relation. Namely, if x and x' create the same mental picture in human brain, and x' and x'' also create the same mental picture, then x and x'' should be also considered equal in human eyes. Note that for distance based metrics (which are different than our approach), this property does not hold. For instance if the perceptual relation is modeled based on a threshold τ on ℓ_p distance of two images, then this property does not hold because x and x'' can potentially have distance 2τ. Indeed, we believe this shows one of the shortcomings with ℓ_p metrics in modeling human perception.

We know that any relationship with the three properties above makes an equivalence relation in the image space. For such a relation P, we use $[x]_P$ to denote the equivalence class of x, that is, $[x]_P = \{y \in S; P(x, y) = 1\}$, the subset of all images x' that are perceived identical to x. These equivalence classes will cluster the space into disjoint clusters. In the next section, we will see how this clustering can immediately give us robustness against adversarial examples that are imperceptible with respect to P.

3 Perceptual Robustness

Before stating our main theorem about robustness, we define robustness under an arbitrary perceptual relation that satisfies the conditions mentioned in Section 2.

Definition 1 (Perceptual robustness)**.** *The robust accuracy under perception relation P, for a distribution μ and a classifier h and a concept function c is defined as:*

$$\mathsf{AdvRisk}_P(h, \mu, c) = \mathop{\mathbf{E}}_{x \leftarrow \mu}[\max_{x' \in [x]_P} \Pr[h(x') \neq c(x)].$$

As we mentioned in Section 2, the perceptual relation P is distinct from the concept function in the sense that there could exist x and x' such that $P(x, x') = 0$ while $c(x) = c(x')$. However, for the tasks that humans can perform, such as classifying cats and dogs, we would have consistency between P and c meaning that if $P(x, x') = 1$ then we should have $c(x) = c(x')$. This is because if two images are identical to a human, then the labels assigned to them should be equal, independent from the task. Below we formalize this notion.

Definition 2 (Consistent concept functions)**.** *A concept function c is consistent with relation P if*

$$\forall x, x' \text{ s.t } P(x, x') = 1; c(x) = c(x').$$

Now, we define a transformation that uses the equivalent classes $[x]_P$ for the perceptual relation and transforms the distribution of images to a new distribution.

Definition 3. *For an equivalence relation P and a distribution μ we define a transformed distribution μ_P that is distributed based on the following probability distribution function:*

$$\forall x \in \text{Supp}(\mu) : \mu_P(x) = \int_{\text{Supp}(\mu)} P(x, y) d\mu(y)$$

Now we have the following Theorem about robustness against perception relations that have all the properties of an equivalence relation. The proof is provided in the supplemental material.

Theorem 4. *Let P be a perception relation that has all properties of an equivalence relation. Also, let c be a concept function consistent with P and μ a distribution of instances. If there is classifier h_p that has high benign accuracy on μ_p, namely $\mathsf{Risk}(h_p, \mu_p, c) \leq \epsilon$, Then there is a classifier h for distribution μ with high robust accuracy. That is,*

$$\mathsf{AdvRisk}_P(h, \mu, c) \leq \epsilon.$$

3.1 Algorithmic Aspects of Perceptual Robustness

Theorem 6 shows how one can use the perceptual relationship to achieve robustness in three steps. The first step transforms the distribution of training examples to another distribution. The second step uses this transformed distribution to do standard training. The third step use the model obtained in the second step to infer the correct label for a given instance in a robust and accurate way. the distribution transformation algorithm, the training algorithm, and the inference algorithm are described Algorithms 1 and 2 and 3 respectively. The first algorithm is the uniform sampling step which is used for transforming the distribution.

Algorithm 1 Uniform Sampling

Input

x	An instance sampled from distribution D
P	A circuit capturing the perceptual relation

Output

$\tilde{x}$	A uniformly sampled instance x such that $P(x, \tilde{x}) = 1$

1: Create a circuit P_x that is same as P but with x hard-coded as one of the inputs.
2: Uniformly sample an instance $\tilde{x}$ such that $P_x(\tilde{x}) = 1$.

Given the sampling algorithm, the remaining steps are easy. We need to implement the training and inference by first transforming the training and test sets using our transformation.

Algorithm 2 Robust Training

Input

S	A dataset S
P	A circuit describing the perceptual relation
L	A learning algorithm

Output

h	A classifier

1: For all for $(x_i, y_i) \in S$ call the Uniform Sampling algorithm on x_i based on P to get x'_i and construct a dataset $S' = \{(x'_1, y_1), \ldots, (x'_n, y_n)\}$.
2: run L on S' to get a classifier h and output h.

Theorem 6 proves that the inference algorithm 3 will be robust to imperceptible changes. However, one might still question if we can implement the sampling algorithm. In particular, it might be computationally hard to uniformly sample another image that is imperceptible from the original image. Note that one can use rejection sampling scheme where many samples are generated uniformly at random from the the whole space until an imperceptible image is found. This technique will not be effective in cases where the space is high dimensional. Can we propose any better strategy for sampling imperceptible images? To answer this question, we have the following Theorem which is directly followed by the work of Bellare et al. (2000). The proof sketch for this Theorem is presented in Appendix A.

Theorem 5. *Given an NP oracle (i.e. a SAT solver), the Uniform Sampling Algorithm 1 can be implemented in probabilistic polynomial time.*

Algorithm 3 Robust Inference

Input

x	query point
P	A circuit describing the perceptual relation
h	A classifier

Output

y	The predicted label

1: Call Uniform Sampling algorithm on x using P to get x'
2: Call h on x' to get y and output y.

Table 1: Caption

Setting	Vanilla	Adversarial	Robust
Accuracy	99.64%	1.94%	99.51%

Relation to Randomized Smoothing We also note that the proposed framework of sampling from the neighboring imperceptible images and running the classifier on them is similar to the ideas used in randomized smoothing Cohen et al. (2019); Salman et al. (2019); Li et al. (2018). Randomized smoothing usually deals with robustness to perturbations of bounded l_p norm. l_p distances do not satisfy the transitivity requirement and our theorems do not apply to them. However, the idea of randomized smoothing is closely related to the idea used here for achieving robustness. Namely, randomized smoothing has a sampling procedure where given an input instance x, we first sample from around x according to a symmetric distribution. Based on the tail analysis of these distributions, one can argue that if two given instances x and x' are close in l_p distance then the sampling distribution applied on them would sample from distributions that are close and hence one can achieve provable robustness. On the other hand, in Algorithm 1, we uniformly sample from the exact equivalence class which means the imperceptible images are mapped to exactly identical distributions and hence we have achieved stronger robustness guarantees.

4 Toy Candidates for Perceptual Relationship

The conditions of perceptual relationship described in Definition 2 are not satisfied for ℓ_p distances that are common in the study of adversarial robustness. However, there are several examples that satisfy all these examples. For example, consider an arbitrary feature extractor F that maps the input instances into a set of discrete features. One possible way of defining a perceptual metric is use F with exact matching, namely $P(x, x') = 1$ if and only if $F(x) = F(x')$. This generic framework covers many types of perturbation. For example, if F down-samples an input image to a less detailed image, and then the corresponding candidate for the perceptual metric will only output $p(x, x') = 1$ if the down-sampled version of x and x' are exactly equal. Or if F is a feature extractor that removes the rotation noise, then for corresponding perceptual metric would output $P(x, x') = 1$ if x and x' are rotated versions of each other. In the rest of the section, to better demonstrate the effectiveness of our algorithm, we use our algorithm on a simple version of patch attacks with a fixed patch location.

The patch attack is an attack that happens on image classifiers where the input image might be perturbed in one local area in the image. We focus on the case that there is one patch at one fixed location. This threat model satisfied all three properties in Section 2. We provide a simple evaluation on ImageNette, which is a 10-class subset of ImageNet dataset. All images are resized and cropped into 224×224 pixels. We use ResNet-50 as our base classifier and we consider a 40×40 patch at the left upper corner of the image.

Effectiveness of patch attacks. Originally, without any defense, one single patch at the image corner can degrade the model accuracy from 99.64% to 1.94%.

robustness. We observe that if we have a perfect perceptual metric, we can achieve a robust accuracy (99.51%) that is close to the vanilla model (99.64%)

REFERENCES

Mihir Bellare, Oded Goldreich, and Erez Petrank. Uniform generation of np-witnesses using an np-oracle. *Information and Computation*, 163(2):510–526, 2000.

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

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of evasion adversaries. *arXiv preprint arXiv:1806.01471*, 2018.

Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

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

Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *arXiv preprint arXiv:2006.12655*, 2020.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. *arXiv preprint arXiv:1809.03113*, 2018.

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.

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pp. 5231–5240. PMLR, 2019.

Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sebastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. *arXiv preprint arXiv:1906.04584*, 2019.

Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.

A APPENDIX

Theorem 6. *Let P be an perception relation that has all properties of an equivalence relation. Also let c be a concept function consistent with P and μ a distribution of instances. if there is classifier h_p that has high benign accuracy on μ_p, namely $\mathsf{Risk}(h_p, \mu_p, c) \leq \epsilon$. Then there is a classifier h for distribution μ with high robust accuracy. That is,*

$$\mathsf{AdvRisk}_P(h, \mu, c) \leq \epsilon.$$

Proof. We first show how to construct the classifier h using h_P. On a given instance x, h will first sample a point x_p uniformly at random from $[X]_P$. Then h will query h_p on x_p to get y and outputs y.

We now show that this classifier will have high robust accuracy.

$$\begin{aligned}
\mathsf{AdvRisk}(h, \mu, c) &= \mathop{\mathbf{E}}_{x \leftarrow \mu}[\max_{x' \in [x]_p} \Pr[h(x') \neq c(x)]] \\
&= \mathop{\mathbf{E}}_{x \leftarrow \mu}[\max_{x' \in [x]_p} \Pr_{x_p \leftarrow [x']_P}[h_p(x_p) \neq c(x)]] \\
&= \mathop{\mathbf{E}}_{x \leftarrow \mu}[\max_{x' \in [x]_p} \Pr_{x_p \leftarrow [x]_P}[h_p(x_p) \neq c(x)]] \\
&= \mathop{\mathbf{E}}_{x \leftarrow \mu}[\Pr_{x_p \leftarrow [x]_P}[h_p(x_p) \neq c(x)]] \\
&= \mathop{\mathbf{E}}_{x \leftarrow \mu}[\Pr_{x_p \leftarrow [x]_P}[h_p(x_p) \neq c(x_p)]] \\
&= \mathop{\mathbf{E}}_{x_p \leftarrow \mu_p}[h_p(x_p) \neq c(x_p)] \\
&= \mathsf{Risk}(h_p, \mu_p, c) = \epsilon.
\end{aligned}$$

This finishes the proof. □

Theorem 7. *Given an NP oracle (i.e. a SAT solver), the Uniform Sampling Algorithm 1 can be implemented in probabilistic polynomial time.*

Proof. Based on the work of (Bellare et al., 2000) we know that given an NP-oracle, one can sample a valid proof uniformly at random among all the valid proofs that prove the presence of an instance x in any given language $\Pi \in NP$. Here, we want to sample images that are imperceptible from x based on a relation P. As described in algorithm 1, we need to uniformly sample x' such that $P(x, x') = 1$. If we look at P as the verifier for an NP language, x as the instance and x' as the proof. This verification algorithm will give us a language that includes all valid images as each input x has a proof x. Sampling a proof for an instance x in this language would exactly translate to sampling an imperceptible image x', uniformly at random. Therefore, the result of Bellare et al. (2000) shows us that we can sample imperceptible images uniformly at random in probabilistic polynomial time, given an NP oracle. □