

ACHIEVING ADVERSARIAL ROBUSTNESS REQUIRES AN ACTIVE TEACHER*

Chao Ma¹⁾ and Lexing Ying

Department of Mathematics, Stanford University, California, USA

Email: chaoma@stanford.edu, lexing@stanford.edu

Abstract

A new understanding of adversarial examples and adversarial robustness is proposed by decoupling the data generator and the label generator (which we call the teacher). In our framework, adversarial robustness is a conditional concept—the student model is not absolutely robust, but robust with respect to the teacher. Based on the new understanding, we claim that adversarial examples exist because the student cannot obtain sufficient information of the teacher from the training data. Various ways of achieving robustness is compared. Theoretical and numerical evidence shows that to efficiently attain robustness, a teacher that actively provides its information to the student may be necessary.

Mathematics subject classification: 68T07, 68T99.

Key words: Adversarial robustness, Decoupled supervised learning, Active teacher.

1. Introduction

The existence of adversarial examples restricts the application of deep learning in many fields with high demand on the robustness and security, such as autonomous driving and health care. Hence, improving adversarial robustness of deep neural networks has experienced extensive study, both theoretically and practically [1, 15]. Originally, adversarial examples are found to be perturbed images whose perturbations are imperceptible to humans but cause huge error to the neural networks [2, 37]. In most existing works, however, adversarial robustness is defined as robustness with respect to perturbations measured by the l_p distance (e.g. [12, 37]). Specifically, a model $f_\theta(\cdot)$ is considered to be robust if the adversarial loss

$$L_{\text{adv}}(f_\theta) = \mathbb{E}_{(\mathbf{x}, y)} \max_{\|\delta\|_p \leq \epsilon} l(f_\theta(\mathbf{x} + \delta), y) \quad (1.1)$$

is small, where ϵ is a pre-defined value and l is some loss function [25]. This simplification helps analysis and implementation. In spite of this, the robustness with small l_p perturbations is very different from the robustness with respect to human-imperceptible perturbations [32]. A human-imperceptible perturbation may not have small l_p norm [5, 46], and a perturbation with small l_p norm may also not necessarily be imperceptible to humans [35]. In Figure 1.1, inspired by optical illusions, we show an example of difference between some l_p distances and human perception. This difference makes current “adversarially robust” models easily broken by newly-designed attacks. Besides l_p distances, other measures, such as Wasserstein distance [43] and structural similarity (SSIM) [41], are also shown to be different from human perception [32].

* Received March 23, 2021 / Revised version received May 5, 2021 / Accepted May 7, 2021 /

Published online October 12, 2021 /

¹⁾ Corresponding author

In this paper, we propose a conditional explanation of adversarial robustness, which highlights the role of human labeler in defining the adversarial examples. Specifically, we decouple the data generator with the labeler, and make two definitions: the teacher is an object or a mechanism that assigns true labels to data points, and the student is a machine learning model used to learn from the data and labels. Within our framework, adversarial robustness is not a universal concept defined unconditionally for any learning problem (like l_p robustness), but rather a relative concept conditioned on a certain teacher. The teacher is usually human, but can also be other objects such as physical processes or neural networks. A student model is said to be (strongly) adversarially robust with respect to a teacher if it can correctly classify any data the teacher can classify with certainty. This is possible because in our framework the teacher has an “uncertain set”, and it does not assign labels to data within this set. Hence a robust student model does not need to have the same decision boundary as the teacher. A weaker version of adversarial robustness is also defined by considering the data produced by an “attack”, instead of all the data that the teacher can classify. This weak definition of adversarial robustness can cover the l_p robustness, but in a more proper way. We show that our definitions of adversarial robustness are not equivalent with the l_p robustness by simple illustrative examples— l_p robust classifier may not be adversarially robust, vice versa.

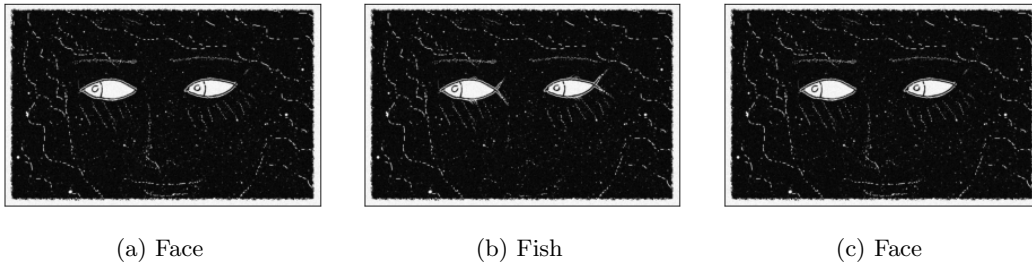


Fig. 1.1. Difference between human perception and l_2 distances illustrated by an optical illusion. **(a)** The image looks like a face; **(b)** The image looks like two fishes; **(c)** The image in (a) adding a noise. The l_0 , l_1 and l_2 distances between (a) and (b) are 15037, 2534.44 and 43.02, respectively. The l_0 , l_1 and l_2 distances between (a) and (c) are 812311, 63413.43 and 89.17, respectively. Though the images in (a) and (c) are perceptually the same, their l_p distances are greater than the distances between (a) and (b), which are perceptually different. (The original image is taken from <https://pixabay.com/illustrations/fairy-tale-fish-portrait-1077859/>)

Based on this new understanding, we point out two reasons that cause adversarial examples: (1) Some features the student uses to make classification are imperceptible to the teacher. (2) The training data do not provide sufficient information of the classification mechanism of the teacher, e.g. which feature the teacher uses to make classification. Combining the two reasons above, we argue that the adversarial examples are caused by insufficient (out-of-distribution) information of the teacher provided by the training data. Without necessary information, the student model cannot select the robust solution among many solutions that perform well on the original data distribution. Therefore, to achieve adversarial robustness, or at least alleviate adversarial vulnerability, more teacher information should be provided to the student model. This can be achieved in two ways:

1. **An active student:** The student model asks information from the teacher, and the teacher passively answers the student’s questions, and does not provide extra information.

2. **An active teacher:** The teacher directly provides information to the student about how it makes classification.

We show theoretically that the first way is not always efficient. Specifically, we prove that in some cases an active student cannot get enough information to achieve robustness in a reasonable time from a passive teacher. Hence, we conclude that an active teacher is required to achieve real adversarial robustness. By simple illustrative examples we show how an active teacher helps the student to learn a robust model, and better robustness can be achieved when more information is provided by the teacher.

Our contributions are summarized as follows:

- We propose a new conditional framework of understanding adversarial robustness. In this framework, the teacher is decoupled from the distribution that generates the data, and robustness is defined as a relative concept of a student model with respect to the teacher.
- Based on the new understanding of adversarial robustness, we demonstrate that achieving robustness requires additional teacher information except the original training data.
- Using both theoretical and empirical approaches, we show that an active teacher helps attaining robustness, while a passive teacher with an active student may not be as efficient.

2. Related Work

Adversarial examples were first introduced in [37]. The work identified data points that are very close to another point (imperceptible to human) but lead to totally different predictions of the model. Several attack methods were then proposed based on the idea of finding the direction in the input space in which the model's output changes fastest [12, 21, 22, 27, 29]. Due to the significance of the security of machine learning models, defenses for adversarial attacks also received extensive study ([3, 14, 30, 42], etc). Adversarial training [12, 21, 25, 38] is a class of methods that can effectively defense against certain attacks. It trains a robust model by including adversarial examples into the training set. Large volume of works arise during an arm race between attacks and defenses. Interested readers can refer to [1] or [15] for a thorough review of the attack and defense methods in different application fields.

On the other side of practical methods, theoretical understanding of adversarial examples also drew attention. Explanations of adversarial vulnerability of machine learning models were provided from different perspectives, including linearity [12], decision boundary geometry [8, 26], low flexibility of the networks [7], non-robust features [17], etc. In particular, [17] proposed that adversarial examples exist because the model learns non-robust features. This viewpoint can be put into our framework: non-robust features, though with good generalization performance, are not used by the teacher, the student cannot reject these features since the training data do not provide enough teacher information.

Mathematical analysis were also conducted, e.g. to show the inevitable existence of adversarial examples [6, 33], the trade-off between adversarial robustness and clean data accuracy [39], the trade-off between robustness and classifier complexity [28], and the provable robustness of highly over-parameterized models [45].

Due to its benefits on analysis and implementation, the l_p distances are used to quantify robustness in most works mentioned above, especially the cases of $p = 0, 2, \infty$. However, l_p

distance is obviously different from human perception. In [5, 46], data pairs that are imperceptible to human but have large l_p distances are identified. On the other side, [35] found image pairs that are close measured by the l_p norm but look very different for humans. Attempts are made to find metrics that align better with human perception, such as the Wasserstein distance [43, 44], SSIM [13] and other perceptibility metrics [18, 23]. However, human experiments and statistical tests in [32] show significant difference between human perception and these metrics.

Among all the theoretical explanations of adversarial examples, the understanding provided in [40] is most relevant to our work. Like what we do in this paper, the authors of [40] also decouple the data generator and the label generator (which they call the oracle), and compare topological properties of the oracle and the student model. They claim that adversarial examples are caused by the difference of the two (pseudo)metric spaces corresponding to the student and the oracle. Our work is different from theirs in at least two ways: (1) After decoupling the data generator and the teacher, we directly compare the decision regions and decision boundaries of the teacher and the student, instead of considering metric spaces. The metric spaces help mathematical analysis, but are hard to verify and identify in practice. (2) Based on the decoupled understanding of adversarial examples, we further explore and compare possible ways to achieve adversarial robustness, and suggest that an active teacher is required to efficiently align student decision regions with those of the teacher in order to achieve adversarial robustness.

3. A Conditional Framework of Adversarial Robustness

3.1. Decoupling data generator and teacher in supervised learning

In this section we introduce a conditional framework to understand adversarial examples and adversarial robustness. We start from a decoupled understanding of supervised learning problems. Traditional formulation of supervised learning problems consists of two parts: a joint distribution of data and label (\mathbf{x}, y) , and a student model which learns the relation between \mathbf{x} and y using the training data sampled from the distribution. Compared with the traditional ones, our formulation of supervised learning decouples the process of generating \mathbf{x} and y , and consists of three components: the data generator, the teacher, and the student.

- **The data generator** is a distribution μ from which data points \mathbf{x} are sampled, to form training and testing data sets.
- **The teacher** is a mechanism to assign labels to the data points. It takes data \mathbf{x} as input and outputs a label y associated with the data. The teacher can be a deterministic function or a stochastic mechanism. For practical machine learning problems the teacher is usually human. We use T to denote the teacher.
- **The student** is a machine learning model trained using a set of data and labels generated by the data generator and the teacher, to learn the labeling rules of the teacher. The student takes data points as inputs and the predicted labels for the input data as outputs. We use S to denote the student.

Fig. 3.1 shows the learning procedure of our machine learning model: the data generator generates data, the teacher assigns labels to the data, forming a dataset, and finally the student is trained using the dataset.

In our formulation we decouple the data generator and the teacher, so we can study the teacher alone. The decoupled perspective highlights that the teacher can work out of the data distribution μ , and we do not have access to all the information of the teacher by just sampling data from μ . As we are going to clarify in the next section, this is the essential reason for the existence of adversarial examples. Finally, note that the traditional formulation can also be included into our framework, by considering the data generator to be the marginal distribution (of \mathbf{x}) and the teacher to be the conditional distribution (of y conditioned on \mathbf{x}).

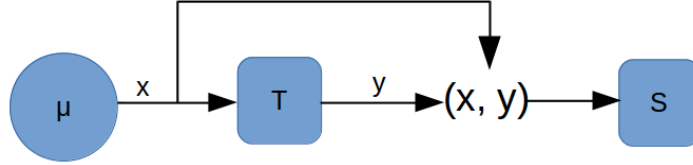


Fig. 3.1. The learning procedure of the machine learning model considered in this paper.

3.2. The conditional definition of adversarial examples

By decoupling the teacher from the data generator, we can now examine adversarial examples and define adversarial robustness in a conditional way. Specifically, adversarial robustness is a property of a student model conditioned on a certain teacher. It involves both the student and the teacher.

We first illustrate the ideas by a simple example. Assume $\mathbf{x} \in X \subset \mathbb{R}^d$. Consider a binary classification problem with two classes A and B . Since sometimes a classifier cannot assign a label with high confidence for any \mathbf{x} in X , we assume that the teacher can output three values: A , B , and U . Here A and B mean the input data belongs to classes A and B , respectively, and U means the teacher is uncertain with the input data. This kind of classifiers are also studied as “selective classifier” in previous works [4, 9]. Let $\Omega_A \subset \mathbb{R}^d$ be the set in which the teacher outputs A :

$$\Omega_A := \{\mathbf{x} \in X : T(\mathbf{x}) = A\}.$$

Ω_B and Ω_U are similarly defined. We require $\Omega_A \cup \Omega_B \cup \Omega_U = X$.

Remark 3.1. The existence of class U is reasonable given that even for humans it is very common to be uncertain with some hard-to-classify images. We can understand the model as a classification problem with three classes but we are only interested in two of them. In traditional understanding of supervised learning the class U is not highlighted because the data distribution is coupled with the teacher and naturally concentrates in $\Omega_A \cup \Omega_B$. But to address adversarial robustness the uncertain class U becomes important because we have to consider adversarially generated unnatural data distributions.

Remark 3.2. The most interesting teachers are humans, which is the case for most CV and NLP problems. However, it can also be objects such as machine learning models, e.g. in the case of knowledge distillation [16]. For a simple example, assume we have a neural network $N : \mathbb{R}^d \rightarrow [0, 1]$, which predicts the probability that the input belongs to class A . Then the

teacher can be defined as

$$T(\mathbf{x}) = \begin{cases} A & \text{if } N(\mathbf{x}) > 0.99, \\ B & \text{if } N(\mathbf{x}) < 0.01, \\ U & \text{otherwise,} \end{cases}$$

i.e. the classes A and B are assigned only when the neural network has high confidence. Hence, adversarial robustness can be considered with respect to general teachers, as in the examples below.

Now we can give a formal description of adversarial examples within our framework. Usually an adversarial example is defined as a data point wrongly classified by a machine learning model, which is very close to another correctly classified data point, and the difference between the two data points are imperceptible to humans. In our framework, we let humans be the teacher and the machine learning model be the student. We highlight the fact that the student gives different prediction from the teacher, then the above definition of the adversarial examples can be rephrased as follow:

An adversarial example is a data point \mathbf{x} that satisfies $T(\mathbf{x}) = A$ or B and $T(\mathbf{x}) \neq S(\mathbf{x})$.

Later examples will show that, as long as adversarial examples described above exist, there will naturally be adversarial examples perceptually close to a correctly classified data point.

In the above statement, an adversarial example can be understood as a data point that the teacher can classify with high confidence, but the student gives different label from the teacher. This kind of data exists because the student is trained by data sampled from μ , but μ cannot provide full information of the teacher, e.g. the support of μ cannot fully cover Ω_A and Ω_B . As an illustrative example, (See the left panel of Figure 3.2) let $\mathbf{x} = (x_1, x_2) \in [-1, 1]^2$ and the teacher is induced by a linear model:

$$T(\mathbf{x}) = \begin{cases} A & \text{if } x_1 \geq 0.5, \\ B & \text{if } x_1 \leq -0.5, \\ U & \text{otherwise.} \end{cases}$$

On the other side, assume that μ is a uniform distribution on $[0.5, 1] \times [0.5, 1] \cup [-1, -0.5] \times [-1, -0.5]$. Then if the student makes max margin classification, the decision boundary will be close to $x_1 + x_2 = 0$. Adversarial examples appear in the second and fourth quadrant (as show by the grey areas in the figure).

With the above definition of adversarial examples, we can state the following definition of adversarial robustness:

A student model S is adversarially robust with respect to a teacher T , if $T(\mathbf{x}) = S(\mathbf{x})$ for all $\mathbf{x} \in \Omega_A \cup \Omega_B$.

We call this definition Strong Adversarial Robustness, because it requires the student to generalize on any distribution in $\Omega_A \cup \Omega_B$, i.e. it should give correct classification on any data point that the teacher can classify with high confidence. It is clear that in this situation no perturbation imperceptible to the teacher can lead to a change of classification of the student. Note that strong adversarial robustness does not require the student to be the same as the teacher, due to the existence of U .

The definition of strong adversarial robustness can be extended to multi-class classification problems. Assume there are $K \geq 2$ classes denoted by C_1, C_2, \dots, C_K , and let Ω_{C_k} be the regions where T outputs C_k , for $k = 1, 2, \dots, K$. Then, we have the following definition for strong adversarial robustness:

Definition 3.1 (Strong adversarial robustness). Let μ , T , S be the data distribution, teacher, and student, respectively. Then, S is strongly adversarially robust with respect to T if

$$S(\mathbf{x}) = T(\mathbf{x}), \quad \forall \mathbf{x} \in \bigcup_{k=1}^K \Omega_{C_k}.$$

Besides strong adversarial robustness, we can also define a weaker version of adversarial robustness. In this case, we consider an attack \mathcal{A} which takes the original data distribution and the student model as inputs and a family of adversarial data distributions as output. We say the student S is adversarially robust with respect to the teacher T and the attack \mathcal{A} if S generalizes as well as T on all the distributions generated by the attack \mathcal{A} . A mathematical definition is given as follows.

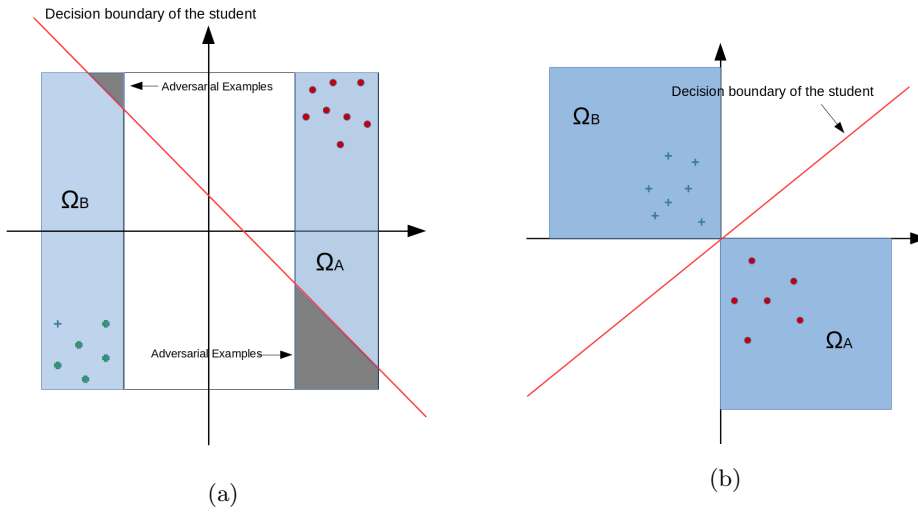


Fig. 3.2. (Left) The simple illustration of adversarial examples. l_2 robust model may not be adversarially robust. (Right) Strongly adversarially robust model may not be l_2 robust.

Definition 3.2 (Adversarial robustness with respect to an attack). Let μ , T , S be the data distribution, teacher, and student, respectively. Let \mathcal{A} be the attack, and

$$\mathcal{P} = \mathcal{A}(\mu, S),$$

where \mathcal{P} is a family of adversarial distributions given by \mathcal{A} with input μ and S . Then, the student S is $(\mathcal{A}, \mu, \epsilon)$ -adversarially robust with respect to T , if

$$\inf_{\nu \in \mathcal{P}} \mathbb{P}_{\mathbf{x} \sim \nu} \left(T(\mathbf{x}) = S(\mathbf{x}) | \mathbf{x} \in \bigcup_{k=1}^K \Omega_{C_k} \right) > 1 - \epsilon. \quad (3.1)$$

By the definition above, the student is adversarially robust if it can generalize well over the distributions generated by a specific attack, on the regions where the teacher performs with certainty. The attack can take many forms. For example, the attack with small l_p perturbations produces all the distributions whose support is within a small distance δ of the support of μ :

$$\mathcal{P} = \{ \nu : \forall \mathbf{x} \in \text{supp}(\nu), \exists \mathbf{x}' \in \text{supp}(\mu), \text{ s.t. } \|\mathbf{x} - \mathbf{x}'\|_p \leq \delta \}.$$

Note that the weak adversarial robustness with above attack is not exactly equivalent with the commonly studied l_p robustness. Because in our definition we only require the student to classify correctly in the region where the teacher can make confident classification, instead of giving the same classification within the l_p ball with radius δ centered at any data point \mathbf{x} . (See the conditional probability in (3.1)) As a result, our definition of adversarial robustness does not conflict with the clean data accuracy (the accuracy on μ)—the student can be robust at the same time of having good accuracy on μ . This is a more proper definition of l_p robustness.

As a second example, the attack can also be all the distributions whose Radon-Nikodym derivative with respect to μ is close to 1:

$$\mathcal{P} = \left\{ \nu : \frac{d\nu}{d\mu} \in [c, \frac{1}{c}] \right\},$$

for some constant $c > 0$. As a third example, it can also depend on the student S , such as the fast gradient method:

$$\mathcal{P} = \left\{ \nu = \Gamma \# \mu : \Gamma(\mathbf{x}) = \mathbf{x} + \delta \frac{\mathbf{g}}{\|\mathbf{g}\|}, \mathbf{g} = \frac{\partial S(\mathbf{x})}{\partial \mathbf{x}} \right\}.$$

Finally, strong adversarial robustness can be viewed as robustness with an attack that produces all the probability distributions on X .

3.3. Relation with l_p robustness

As we mentioned above, l_p robustness is appropriately covered by Definition 3.2. In this section, we focus on traditional l_2 robustness and compare it with our definition of strong robustness. Using simple illustrative examples, we show that l_2 robust students may not be strongly robust, and strongly robust students may not be l_2 robust, either.

The example in the left panel of Fig. 3.2 shows a student that is l_2 robust but not strongly adversarially robust with respect to the teacher. In the example, the teacher conducts classification with only x_1 , and does not use the feature x_2 . Hence, data points with the same x_1 but different x_2 are imperceptible to the teacher. However, the student gathers teacher information only from the training data, hence it is reasonable for it to make max margin classification. Using the max margin decision boundary, the student is l_2 robust even when the perturbation is large, but adversarial examples exist. For example, for a data point in the grey area on the upper-left part of the figure, the student will make wrong classification, while for the teacher this data looks similar to the ones on the bottom-left side because they have the same x_1 .

In the right panel of Fig. 3.2, we show an example that adversarial robustness does not imply l_2 robustness. In this example, the two classes lie in the second and the fourth quadrants, respectively. And there is no margin between the two classes. The student with decision boundary shown by the red line is strongly adversarially robust with respect to the teacher, because the decision boundary passes through the origin. However, since there is no margin between Ω_A and Ω_B , the student is not l_2 robust with any $\epsilon > 0$, because for any ϵ we can always find a sample in $\Omega_A \cup \Omega_B$ whose l_2 distance from the decision boundary is smaller than ϵ . In real problems, such “zero margin” situation is quite common. The teacher’s decision might have sudden jumps from one class to another in a small region, for example, when the teacher decides the sign of a number, or compares the sizes of two objects. Humans are usually good at these tasks.

4. Adversarial Robustness Requires Active Teacher

By the new understanding of adversarial examples, we tentatively conclude that adversarial examples exist because the student does not have sufficient information of the teacher. To achieve adversarial robustness, additional teacher information should be incorporated into the student. This can be achieved in two ways:

1. **A passive teacher and an active student:** In this approach, the teacher provides information to the student only when the student asks for information from the teacher. For instance, in addition to the training data, the student generates extra data and asks the teacher to classify these data. Then, the new data and labels are included to the training set to train an updated student model. (This is like an interactive way of adversarial training).
2. **An active teacher:** The teacher directly tells the student information on how it makes classification, such as the features used, invariances, sparsity, or the structure of the model, etc. Then, the student tries to encode the information into its learning procedure, e.g. taking specially designed network structure and learning algorithm.

In this section, we show that an active teacher is preferred, and may even be necessary, for the student to be adversarially robust.

In the setting of a passive teacher, we theoretically prove that a simple query-based active student cannot efficiently learn robust models. On the other hand, in the setting of an active teacher, we show by numerical examples how can the teacher “teach” the student to be robust.

4.1. A passive teacher and an active student

An active student can acquire teacher information in many different ways. In this section, we consider one of the most natural ways to ask for teacher information—feature query. Specifically, every time the student provides the teacher with a feature, and the teacher returns the correlation of the feature with the labels (the correlation is computed in a data distribution generated by the attack, hence it helps achieving adversarial robustness and cannot be approximated with μ). In this way, the student asks the teacher “to what extent do you use this feature to make classification”, and the teacher answers the question with a score. Then, the student updates itself according to the teacher’s answer. Intuitively, the student can learn a robust classifier if it identifies all the features used by the teacher to make classification. However, since there are numerous possibilities when choosing the features to query, it can be hard to find the right ones. In this section, we borrow the theories of hardness of learning to show that in some cases it is impossible to efficiently learn a robust student with feature querying, even though we have a very weak attack which only produces one single adversarial distribution.

Mathematically, we put our “feature query” setting into the statistical query framework [19]. Let $X = \{0, 1\}^d$, \mathcal{D} be some probability distribution on X . Let $T : X \rightarrow \{0, 1\}$ be the teacher. Then, a statistical query $\text{STAT}(T, \mathcal{D})$ takes a function $\chi : X \times \{0, 1\} \rightarrow \{0, 1\}$ and returns $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \chi(\mathbf{x}, T(\mathbf{x}))$ with some tolerance α , i.e. the returned value lies in $[\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \chi(\mathbf{x}, T(\mathbf{x})) - \alpha, \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \chi(\mathbf{x}, T(\mathbf{x})) + \alpha]$. Obviously, the correlation of a feature $h : X \rightarrow \{0, 1\}$ with the labels, $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} h(\mathbf{x})T(\mathbf{x})$, is a statistical query with $\chi(\mathbf{x}, T(\mathbf{x})) = h(\mathbf{x})T(\mathbf{x})$. Statistical queries are powerful because it can return the correlation of any feature with the teacher’s output with high accuracy, of course including those features used by the teacher.

It is proven in [19] that parity functions are not efficiently learnable from statistical queries. A parity function with respect to a subset S of $\{1, 2, \dots, d\}$ is defined as

$$f_S(\mathbf{x}) = \left(\sum_{i \in S} \mathbf{x}_i \right) \bmod 2, \quad (4.1)$$

i.e. f_S takes 1 if an odd number of entries in $\{\mathbf{x}_i, i \in S\}$ is 1, and takes 0 otherwise.

Theorem 4.1 (Theorem 5 of [19]). *Let \mathcal{F}_d be all parity functions over $\{0, 1\}^d$ and \mathcal{D} be the uniform distribution on $\{0, 1\}^d$. Then, for any fixed accuracy ε , there does not exist polynomials $p(d)$ and $q(d)$, and an algorithm using statistical queries with tolerance $\alpha \geq 1/q(d)$, such that for any $f \in \mathcal{F}_d$ the algorithm can return a hypothesis h within $p(d)$ statistical queries that satisfies*

$$\mathbb{P}_{\mathbf{x} \sim \mathcal{D}}(h(\mathbf{x}) = f(\mathbf{x})) > 1 - \varepsilon. \quad (4.2)$$

Based on the theorem above, we can show that an active student using feature queries cannot always learn adversarially robust classifiers efficiently. Still consider $X = \{0, 1\}^d$. Now, let μ be the uniform distribution on two points $(0, 0, \dots, 0)$ and $(1, 0, \dots, 0)$, \mathcal{A} be an attack, and ν be the uniform distribution on X , which is generated by the attack \mathcal{A} . That is to say, the output of this weak attack contains only one distribution, and even does not depend on the student. Finding a robust classifier requires the student to generalize on ν . Consider the set of teachers \mathcal{T} to be all parity functions over X with the first coordinate included, i.e.

$$\mathcal{T} = \{f_S(\mathbf{x}) : S \subset [d], 1 \in S\}. \quad (4.3)$$

Then, since μ only supports on two points, teachers in \mathcal{T} can be learned by a simple linear regression on μ . However, by Theorem 4.1, they cannot be learned efficiently on ν using feature queries. Hence, the student cannot learn adversarially robust classifiers with respect to the teachers in \mathcal{T} , if the attack gives the distribution ν . To summarize, we have the following theorem.

Theorem 4.2. *Let \mathcal{T} , μ , ν , \mathcal{A} be defined above. Let T be a teacher from \mathcal{T} and S be a student which has access to the data pairs $(\mathbf{x}, T(\mathbf{x}))$ where \mathbf{x} is sampled from μ . Besides, the student can get feature queries $\mathbb{E}_{\mathbf{x} \sim \nu} h(\mathbf{x}) T(\mathbf{x})$ for any feature $h : X \rightarrow \{0, 1\}$, with a tolerance α that satisfies $\alpha \geq 1/q(d)$ for some polynomial $q(d)$. Then, for any fixed $\varepsilon > 0$, there does not exist a polynomial $p(d)$ such that for any $T \in \mathcal{T}$ the student can learn an $(\mathcal{A}, \mu, \varepsilon)$ -Adversarially robust classifier with respect to T within $p(d)$ feature queries.*

Proof. For any $d \geq 2$, let $X_0 = \{(0, x_2, x_3, \dots, x_d) : x_i \in \{0, 1\}, i = 2, 3, \dots, d\}$. Assume that the conclusion of Theorem 4.2 does not hold. Then, there exists an algorithm that for any teacher $T \in \mathcal{T}$ it can learn a student model h that satisfies

$$\mathbb{P}_{\mathbf{x} \sim \mathcal{D}}(h(\mathbf{x}) = T(\mathbf{x})) > 1 - \epsilon \quad (4.4)$$

with at most $p(d)$ feature queries and a tolerance $\alpha \geq 1/q(d)$. Here, ϵ is a constant and $p(\cdot), q(\cdot)$ are two polynomials, which may depend on ϵ . Equation (4.4) implies

$$\mathbb{P}_{\mathbf{x} \sim \mathcal{D}}(h(\mathbf{x}) \neq T(\mathbf{x})) \leq \epsilon,$$

which can be rewritten as

$$\frac{1}{2^d} \sum_{\mathbf{x} \in \{0, 1\}^d} \mathbf{1}_{h(\mathbf{x}) \neq T(\mathbf{x})} \leq \epsilon.$$

Therefore,

$$\sum_{\mathbf{x} \in X_0} \mathbf{1}_{h(\mathbf{x}) \neq T(\mathbf{x})} \leq \sum_{\mathbf{x} \in \{0,1\}^d} \mathbf{1}_{h(\mathbf{x}) \neq T(\mathbf{x})} \leq 2^d \epsilon,$$

which directly gives

$$\frac{1}{2^{d-1}} \sum_{\mathbf{x} \in X_0} \mathbf{1}_{h(\mathbf{x}) \neq T(\mathbf{x})} \leq 2\epsilon,$$

and hence

$$\mathbb{P}_{\mathbf{x} \sim \text{unif}(X_0)}(h(\mathbf{x}) = T(\mathbf{x})) > 1 - 2\epsilon. \quad (4.5)$$

By the definition, \mathcal{T} conditioned on X_0 contains all the parity functions of x_2, \dots, x_3 . Hence, Equation (4.5) is contradictory with Theorem 4.1. This completes the proof. \square

The hardness of learnability has recently been studied in the setting of neural networks [10, 11, 24]. Therefore, it is possible to extend Theorem 4.2 to more general teachers, e.g. neural network models.

4.2. An active teacher

On the other hand, if the teacher actively provides information to the student, then it is possible to efficiently learn robust classifiers. For the same problem in Theorem 4.2, if the student knows from the teacher that it is a parity function, then the student can check whether the teacher considers the i -th coordinate by querying two data points $(0, 0, \dots, 0)$ and $(0, \dots, 0, 1, 0, \dots, 0)$ where the 1 in the second data point appears in the i -th coordinate. In this way, the student can learn the teacher within $d + 1$ data queries. Hence, we have the following theorem:

Theorem 4.3. *Let \mathcal{T} , μ , ν , \mathcal{A} be defined the same as in Theorem 4.2. The teacher T comes from \mathcal{T} . Let S be a student that can make data query from the teacher, i.e. get $T(\mathbf{x})$ from the teacher for any \mathbf{x} . Then, if the student knows the teacher is a parity function, it can learn a strongly adversarially robust classifier within $d + 1$ data queries.*

Proof. Let $\mathbf{0} = (0, 0, \dots, 0)$, and $\mathbf{e}_i = (0, \dots, 0, 1, 0, \dots, 0)$ for $i = 1, 2, \dots, d$, where the 1 appears on the i -th coordinate. If the student knows that the teacher T comes from parity functions, it can query $T(\mathbf{0})$ and $\{T(\mathbf{e}_i)\}_{i=1}^d$. Then for any $i = 1, 2, \dots, d$, $T(\mathbf{0}) = T(\mathbf{e}_i)$ implies $i \notin S$, while $T(\mathbf{0}) \neq T(\mathbf{e}_i)$ implies $i \in S$. \square

Therefore, information directly from an active teacher may help the student find a robust classifier more efficiently. In the following we support this claim by two numerical examples.

Example 4.1. Consider a binary classification problem. Let $\mathbf{x} \in \mathbb{R}^{100}$ be the input data, and \mathbf{x}_i be the i -th element of \mathbf{x} . Assume $\mathbf{x}_i \in [0, 1]$. To assign the label, the teacher only compares the first element x_1 and the last element x_{100} . The teacher assigns label $y = -1$ if $x_1 > x_{100}$, and $y = 1$ if $x_1 \leq x_{100}$. Obviously a strongly adversarially robust classifier for this problem cannot be l_2 robust, because there is no margin between the two classes. For each class, we uniformly sample 1000 training data. Linear regression (without bias) is used as the student model. Let the linear regression model be

$$\hat{y} = \alpha^T \mathbf{x}, \quad (4.6)$$

and let α_i be the coefficients corresponding to \mathbf{x}_i . Then the strongly adversarially robust model satisfies $\alpha_1 < 0$, $\alpha_{100} = -\alpha_1$, and $\alpha_i = 0$ for other i .

If the student does not have any additional information besides the training data, a plain linear regression is conducted with 100 variables. A dense vector will be produced, and it will be easy to find adversarial examples by changing \mathbf{x}_i 's other than \mathbf{x}_1 and \mathbf{x}_{100} according to the sign of the corresponding coefficients. Specifically, for some \mathbf{x} correctly classified by the student, assume $\mathbf{x}_1 > \mathbf{x}_{100}$ without loss of generality, we can construct $\hat{\mathbf{x}}$ by

$$\hat{\mathbf{x}}_1 = \mathbf{x}_1, \hat{\mathbf{x}}_{100} = \mathbf{x}_{100}, \hat{\mathbf{x}}_i = \mathbf{x}_i + \epsilon \text{sign}(\alpha_i) \text{ for } i = 2, 3, \dots, 99.$$

Then, the prediction of $\hat{\mathbf{x}}$ can be flipped as long as $\epsilon > \alpha^T \mathbf{x} / \sum_{i=2}^{99} |\alpha_i|$, while the difference between $\hat{\mathbf{x}}$ and \mathbf{x} is always imperceptible to the teacher. Fig. 4.1 shows the coefficients and some adversarial examples in the form of 10×10 images. On the other hand, if the student is provided with additional information directly from the teacher beyond the training data, better adversarial robustness may be achieved. For example, if the student is told that the teacher only considers \mathbf{x}_1 and \mathbf{x}_{100} , then the student can choose to use a sparse model

$$\hat{y} = \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_{100}. \quad (4.7)$$

Training this sparse model with the same set of training data, we obtain the model

$$\hat{y} = -2.01 \mathbf{x}_1 + 1.98 \mathbf{x}_{100}, \quad (4.8)$$

which is much more robust than the plain linear regression model, because perturbing pixels other than \mathbf{x}_1 and \mathbf{x}_{100} can no longer change the prediction of the model. However, adversarial examples still exist for those $\mathbf{x}_1, \mathbf{x}_{100}$ that satisfies $-\mathbf{x}_1 + \mathbf{x}_{100} > 0$ but $-2.01 \mathbf{x}_1 + 1.98 \mathbf{x}_{100} < 0$, e.g. $\mathbf{x}_1 = 1$, $\mathbf{x}_{100} = 1.01$. If we incorporate further information, e.g. the teacher is a linear model that takes integer coefficients, then we can round the coefficients in (4.8) and get a model with strong adversarial robustness.

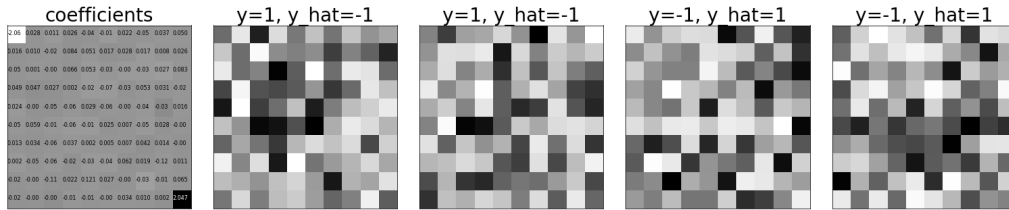


Fig. 4.1. (Left) the coefficients of the student model corresponding to every entries of \mathbf{x} . (Others) some adversarial examples of the student model.

Example 4.2. This example shows that when multiple features can be picked to make generalizable classification, additional teacher information can help the student find the features that lead to a robust model. In this problem, the data are images of a disk or a square with random size and location, and the student model is asked to classify between disks and squares. Except the shapes, we add textures in the squares as a confounding feature. Examples of the data are shown on the first row of Fig. 4.2. In the data distribution which generates the training and testing data (μ), squares always have textures while disks always do not. Therefore, both features—shape and texture—can be used to build a generalizable classifier. However, for the teacher (human) shape and texture have different meanings and the teacher expect the student

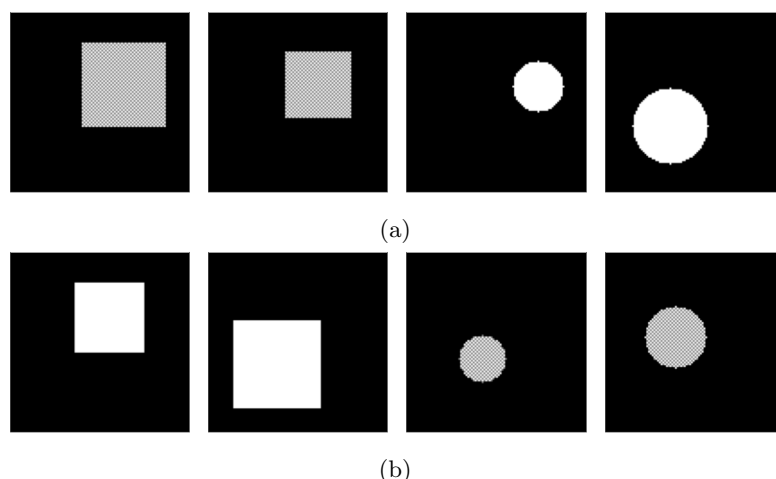


Fig. 4.2. (a) Examples of the training and testing data. The location and size of the squares and disks are chosen randomly. The squares have textures while the disks do not. (b) Examples of the data generated by the attack from adversarial distribution ν . Textures appear in the disks instead of the squares..

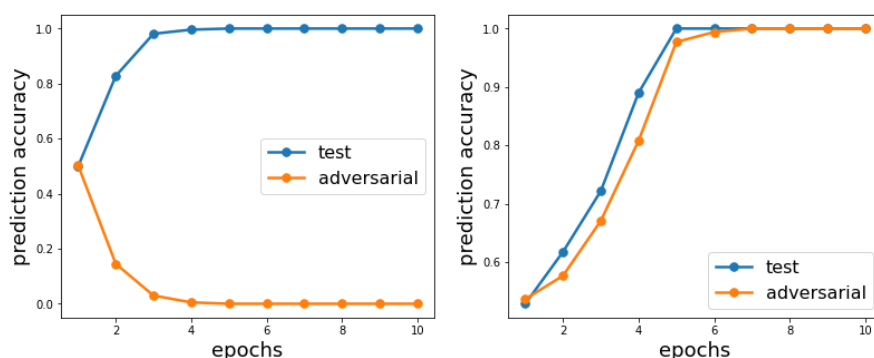


Fig. 4.3. The prediction accuracy on test samples and adversarial samples. (Left) The training data are directly used to train the neural network. (Right) A max pooling of kernel size 3 and stride 1 is conducted before the data is fed into the convolution layers.

to use shape for classification. Hence, as an adversarial data distribution (ν), we generate disks with texture and squares without texture, as shown on the second row of Fig. 4.2.

A convolutional neural network is utilized to learn the problem on a training set including 1000 images, with 500 squares and 500 disks. Experiment details are provided in the appendix. 1000 test samples are randomly generated from μ , and another 1000 adversarial examples are generated from ν . The left panel of Fig. 4.3 shows the accuracy on the test samples and adversarial samples during the training process, when no teacher information except the training data is provided. It shows clearly that the student learns to make classification using textures, hence as the test accuracy goes to 1 the adversarial accuracy goes to 0. On the other hand, if the teacher tells the student that the classification should be made depending on the shape, then the student can conduct a low-pass filtering to the images before feeding them into the neural network, to filter out the texture. For this problem, specifically, we use a max pooling with kernel size 3 and stride 1 to act as the filtering. The results are shown on the right panel of Fig. 4.3. In this case the adversarial accuracy is nearly as good as the test accuracy.

5. Discussion

In this paper we make three points about the cause of and the solution to adversarial examples. First, the teacher and the data generator should be considered separately, and adversarial robustness is a relevant concept between the student and the teacher. Second, adversarial examples are caused by the insufficiency of information provided by the training data about the teacher. Third, to solve the insufficiency of teacher information, we suggest that an active teacher is more preferred than an active student with a passive teacher. In the case where the teacher is human, our study suggests that human labelers should provide more information besides the labels and the model should be designed to incorporate the additional information. This is similar to the case when people are learning. For example, when human teachers teach image recognition to human students, they usually describe features about the objects. The description of features certainly contains more information than just labels.

Moreover, in complicated learning problems the features are often hierarchical. In deep learning, one often prefers end-to-end training and relies on the models to automatically learn the hierarchical structure of the features. Our study, however, demonstrates that including information of the feature hierarchy may help the student model be robust. Similar methodology has been studied in a different context. In [34], it is shown that decomposition learning can be efficient when end-to-end learning is impossible.

Strictly speaking, any model or algorithm encodes certain prior information and hence exhibits certain “implicit bias”. The model performs well when its implicit bias coincides with the prior of the teacher. This is especially crucial in the over-parameterized regime where there are many solutions which perfectly fit the training data but only a small fraction of them generalize well. In the case of adversarial robustness, however, we require another level of implicit bias: the solutions picked by the model not only have to generalize well on the test data provided by the data generator, but also need to generalize to regions that are not sufficiently represented by the data generator. This is also a topic studied by out-of-distribution generalization [20, 36] and distribution shift [31]. However, existing models cannot provide satisfactory implicit bias to learn human-like classifiers. They are either too simple (like the linearity of linear regression and the sparsity of LASSO), or hard to interpret (like deep neural networks). It is very important to design models that can directly and explicitly incorporate interpretable information provided by humans. We leave this as a major direction of future work.

Finally, other than achieving robustness, a model whose prior knowledge is better aligned with that of humans may also help in few-shot learning, meta-learning, and model interpretation. It is an inevitable step to achieve higher levels of artificial intelligence than today’s deep learning.

A. Experiment Details

In this section we provide the experimental details of the second example in Section 4.2. The experiments are run on a 2020 Macbook Pro 13’ with 16GB RAM, and the neural networks are implemented and trained by Pytorch.

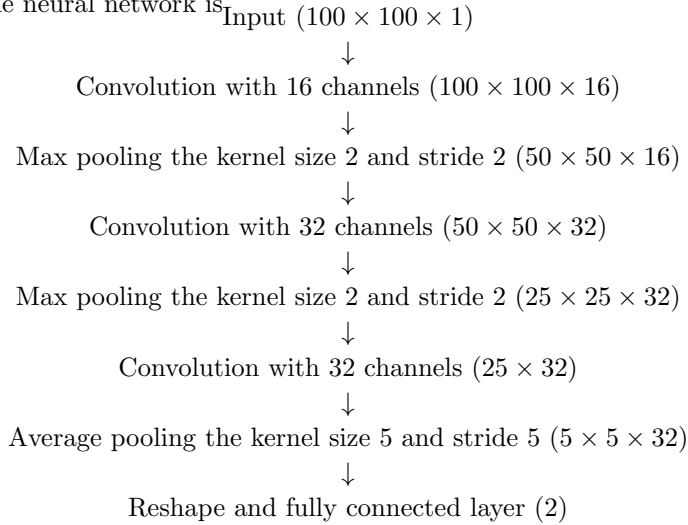
A.1. Data

The data are images with 100×100 pixels. The half side length of the squares and the radius of the disks are uniformly sampled from integers within $[12, 25]$, which roughly corresponds to $[a/8, a/4]$, where a is the side length of the images. The centers of the shapes are then uniformly sampled from pixels so that the whole shape is within the image. For example, for a square with half side length 20, the center is a pixel (x, y) with x and y uniformly sampled from integers within $[20, 78]$. Then, the area of the square is $[x - 20, x + 20] \times [y - 20, y + 20]$. The images are in gray scale, with each pixel taking values in $[0, 1]$. The background pixels take values 0 while the pixels in the shape are 1. The texture is added to the shape by changing the value from 1 to 0.5 for the pixels (x, y) with $x + y$ being even and leaving the value at other pixels unchanged.

For training, we sample 500 images of squares with texture and 500 images of disks without texture, forming a data set consisting of 1000 images. For testing, we sample 1000 new image each time testing is conducted. The images still consist of squares with textures and disks without textures. The probability of squares is 0.5. When measuring adversarial performance, each time we sample 1000 images of squares without texture and disks with texture. The probabilities of squares and disks are still 0.5.

A.2. Model

We use a multi-layer convolutional neural network (CNN) as the student model. The neural network has 3 convolution layers and 2 max pooling layers in the middle of convolution layers. An average pooling and a fully connected layer follow the convolution layers. Specifically, the architecture of the neural network is



The cross entropy loss is used as the loss function. Adam is taken as the optimizer, with learning rate 0.001 and default momentum factors (0.9, 0.999). The network is trained by 10 epochs and the batch size is 50.

References

- [1] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, **6** (2018), 14410–14430.

- [2] 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 *Joint European conference on machine learning and knowledge discovery in databases*, Springer, (2013), 387–402.
- [3] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations*, 2018.
- [4] Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, **11**:5 (2010).
- [5] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, (2019), 1802–1811.
- [6] A. Fawzi, H. Fawzi, and O. Fawzi. Adversarial vulnerability for any classifier. In *Advances in neural information processing systems*, (2018), 1178–1187.
- [7] A. Fawzi, O. Fawzi, and P. Frossard. Analysis of classifiers robustness to adversarial perturbations. *Machine Learning*, **107**:3 (2018), 481–508.
- [8] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, (2016), 1632–1640.
- [9] Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *Advances in neural information processing systems*, (2017), 4878–4887.
- [10] S. Goel, A. Gollakota, Z. Jin, S. Karmalkar, and A. Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. *arXiv preprint arXiv:2006.12011*, 2020.
- [11] S. Goel, A. Gollakota, and A. Klivans. Statistical-query lower bounds via functional gradients. *arXiv preprint arXiv:2006.15812*, 2020.
- [12] I.J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [13] D. Gragnaniello, F. Marra, G. Poggi, and L. Verdoliva. Perceptual quality-preserving black-box attack against deep learning image classifiers. *arXiv preprint arXiv:1902.07776*, 2019.
- [14] C. Guo, M. Rana, M. Cisse, and L. Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.
- [15] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang and A.K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, **17**:2 (2020), 151–178.
- [16] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [17] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, (2019), 125–136.
- [18] M. Jordan, N. Manoj, S. Goel, and A.G. Dimakis. Quantifying perceptual distortion of adversarial examples. *arXiv preprint arXiv:1902.08265*, 2019.
- [19] M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, **45**:6 (1998), 983–1006.
- [20] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.
- [21] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.
- [22] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.
- [23] C. Laidlaw, S. Singla, and S. Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *arXiv preprint arXiv:2006.12655*, 2020.
- [24] R. Livni, S. Shalev-Shwartz, and O. Shamir. On the computational efficiency of training neural networks. In *Advances in neural information processing systems*, (2014), 855–863.

- [25] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [26] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, and S. Soatto. Analysis of universal adversarial perturbations. *arXiv preprint arXiv:1705.09554*, 2017.
- [27] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), 2574–2582.
- [28] P. Nakkiran. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.
- [29] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, IEEE, (2016), 372–387.
- [30] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, IEEE, (2016), 582–597.
- [31] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N.D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.
- [32] A. Sen, X.J. Zhu, L. Marshall, and R. Nowak. Should adversarial attacks use pixel p-norm? *arXiv preprint arXiv:1906.02439*, 2019.
- [33] A. Shafahi, W.R. Huang, C. Studer, S. Feizi, and T. Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.
- [34] S. Shalev-Shwartz, O. Shamir, and S. Shammah. Failures of gradient-based deep learning. *arXiv preprint arXiv:1703.07950*, 2017.
- [35] M. Sharif, L. Bauer, and M.K. Reiter. On the suitability of lp-norms for creating and preventing adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (2018), 1605–1613.
- [36] Y. Sun, X.L. Wang, Z. Liu, J. Miller, A.A. Efros, and M. Hardt. Test-time training for out-of-distribution generalization. *arXiv preprint arXiv:1909.13231*, 2019.
- [37] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [38] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.
- [39] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.
- [40] B.L. Wang, J. Gao, and Y.J. Qi. A theoretical framework for robustness of (deep) classifiers against adversarial examples. *arXiv preprint arXiv:1612.00334*, 2016.
- [41] 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 (2004), 600–612.
- [42] E. Wong and Z. Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, PMLR, (2018), 5286–5295.
- [43] E. Wong, F.R. Schmidt, and J.Z. Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. *arXiv preprint arXiv:1902.07906*, 2019.
- [44] K.W. Wu, A.H. Wang, and Y.L. Yu. Stronger and faster wasserstein adversarial attacks. *arXiv preprint arXiv:2008.02883*, 2020.
- [45] Y. Zhang, O. Plevrakis, S.S. Du, X.G. Li, Z. Song, and S. Arora. Over-parameterized adversarial training: An analysis overcoming the curse of dimensionality. *arXiv preprint arXiv:2002.06668*, 2020.
- [46] Z.Y. Zhao, Z.R. Liu, and M. Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 1039–1048.