Improving Adversarial Robustness: An Information-Theoretic Perspective

Anonymous Authors[1]

Abstract

Deep neural networks (DNNs) are known to be vulnerable to adversarial examples. Many explanations for this intriguing phenomenon have emerged in recent years, attempting to attribute the vulnerability to different properties of DNNs (e.g., linearity, overfitting and large local variations). In this paper, we propose to understand adversarial examples through the lens of information theory. We prove that under mild assumptions, not employing minimal sufficient statistics as feature representation is a necessary condition for the existence of adversarial examples. This result provides theoretical support for the recent observation that information bottleneck (IB), a framework to approximate the minimal sufficient statistics, can achieve higher robustness than other learning objectives. We establish a bound on the adversarial robustness of IB, which provides insights into learning robust models. Besides, we present a framework to rigorously characterize the factors that contribute to the transferability of adversarial examples across different models trained using IB. We perform extensive experiments to validate the insights from our theoretical analysis.

1. Introduction

DNNs are expressive models that have achieved state-of-the-art performance across various domains, such as speech recognition and computer vision. However, they are known to lack robustness to *adversarial examples*, which look almost identical to legitimate ones for a human but can mislead DNNs to make incorrect predictions. Adversarial examples are transferable from one model to another trained for the same or even different learning tasks. Transferability presents an obstacle to secure deployment of DNNs for it enables simple black-box attacks against DNNs: an adversary can craft adversarial examples for a local model while using them to attack a target model that it does not have access to. All these properties of adversarial examples have triggered people's concern about the trustworthiness of DNNs.

Many recent works have been focused on the theoretical understanding of adversarial examples and gleaning the insights for designing more robust DNNs. Intuitively, DNNs exhibit large local variations so that small alterations of the input can flip the labeling decision Szegedy et al. (2013); Nayebi & Ganguli (2017). Other ideas to characterize adversarial examples include inspecting the data density Feinman et al. (2017), the uncertainty to Gaussian data corruptions Ford et al. (2019), the geometry of decision boundaries relative to data distribution Tanay & Griffin (2016), and the dimensionality of the local data manifold Ma et al. (2018), etc.

In this paper, we characterize adversarial examples from an information-theoretic perspective, which is complementary to existing interpretations. Indeed, many works have emerged to leverage information theory for explaining adversarial examples Galloway et al. (2018); Alemi et al. (2016). These works all reach the same conclusion that proper regularization of feature complexity can help enhance the robustness of a model. However, they either only provided empirical evidence or posed a conjecture by regarding DNNs as a communication channel. We build upon these early works and present a rigorous discussion about the connection between feature complexity and model robustness.

Our contributions can be summarized as follows: (1) We prove that under mild assumptions, not using minimal sufficient statistics for feature representation is a necessary condition for the existence of adversarial examples. This result provides theoretical backing for the recent empirical observation that information bottleneck (IB), a framework to approximate the minimal sufficient statistic, exhibits higher robustness than other learning objectives such as cross-entropy. (2) We establish a theoretical bound on the robustness of IB. Our bound justifies the observation that IB can achieve higher robustness with a smoother stochastic encoder and more regularization on feature complexity. Our bound also directly implies a new method to further improve the robustness of the vanilla IB Alemi et al. (2016), which is to combine IB with adversarial training. (3) We

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

study the transferability of adversarial examples under the IB framework. We theoretically show that two low-risk classifiers are susceptible to transferability-based black-box attacks. Our result also uncovers various interesting factors that affect the success rate of such attacks. (4) We perform extensive experiments to validate the insights gained from the theoretical analysis and demonstrate the effectiveness of adversarial training when combined with IB on MNIST and CIFAR against various attacks.

2. Related Work

Adversarial examples are carefully crafted data points that are indistinguishable from benign for human perception systems but can cause machine learning models to make arbitrary prediction errors. Existing attacks mainly rely on the gradient information to efficiently search for adversarial examples in the high-dimensional space. For instance, the fast gradient sign method (FGSM) has been proposed to add perturbations along the gradient directions Goodfellow et al. (2015). Other examples include optimization algorithms that search for minimal perturbation Carlini & Wagner (2017); Liu et al. (2017). Several defenses attempt to improve the model robustness by obfuscating gradients so that no useful gradient information can be exploited Guo et al. (2017); Dhillon et al. (2018); Song et al. (2017). However, it has been shown that these models based on obfuscated gradients are still vulnerable to more sophisticated gradient estimation methods Athalye et al. (2018).

Prior work on adversarial examples attributed the vulnerability of deep networks to high local variations Szegedy et al. (2013); Goodfellow et al. (2015). A number of defenses inspired by this explanation were later proposed, which regularize the variations of the model for small input changes Gu & Rigazio (2014); Cisse et al. (2017). The vulnerability of a model is often characterized by the minimal perturbation that is sufficient to change the prediction Moosavi-Dezfooli et al. (2016). Such geometric viewpoint has been leveraged to explain the empirical observation that classifiers tend to be more robust to random perturbations than adversarial perturbations Fawzi et al. (2016).

A handful of recent work started to study adversarial examples from an information-theoretic perspective. Galloway et al. identified a neural network with a noisy communication channel and ascribed sensitivity to adversarial examples to the excess complexity of feature representations. Our paper is inspired by this work but further provides rigorous proofs that uncover the connection between adversarial examples and feature representation complexity. We also present a series of strategies to robustify the model based on the IB principle. The early work of IB and its use case in machine learning can be traced back to Tishby et al. (2000); Tishby & Zaslavsky (2015); Shwartz-Ziv & Tishby (2017). Recently, IB has demonstrated its effectiveness on various tasks in machine learning, including reinforcement learning-based control Goyal et al. (2018); Peng et al. (2018) and modeling data generation distirbutions Jeon et al. (2019). Alemi et al. presented a variational method for efficient training with IB and experimentally showed that it improves the model robustness against adversarial attacks. Our work establishes an adversarial robustness bound for IB, which provides actionable insights for improving the model robustness.

3. A Necessary Condition for the Existence of Adversarial Examples

Let $f : \mathcal{X} \to \mathcal{Y}$ be a neural network used for classification, where $\mathcal{X}$ and $\mathcal{Y}$ are the domains for input data signals and labels, respectively. f can be regarded as the composition of an encoder T and a decoder g, i.e., $f = T \circ g$. For any $x \in \mathcal{X}$, $T(x)$ produces a feature representation which is subsequently used by g to predict the label for x.

We consider the following definition of adversarial examples. This definition is based on the notion of an oracle which can be understood as a human decision-maker or the unknown input-output function considered in learning theory Shalev-Shwartz & Ben-David (2014).

Definition 1 (Adversarial Examples Jacobsen et al. (2018))**.** *A ϵ-bounded adversarial example x' of x for a neural network f fulfills: (1) $f(x') \neq o(x')$ where $o(\cdot)$ is the oracle; (2) x' is created by an attack algorithm $\mathcal{A}$ which maps x to x'; (3) $\|x - x'\| \leq \epsilon$, where $\|\cdot\|$ is a norm on $\mathcal{X}$ and $\epsilon > 0$.*

Empirical evidence has shown that model robustness can be improved if low-complexity features are employed via proper operations such as principal component analysis Bhagoji et al. (2018) or feature selection Bao et al. (2018). In this section, we explicate the relationship between adversarial examples and the amount of information within the feature representation.

Suppose the input source X and its corresponding target Y are drawn from some joint distribution $p(x, y)$. The feature representation $T(X)$ can be treated as a statistic of X. For notational convenience, the random variable $T(X)$ will be occasionally abbreviated to T. T is subject to the Markovian relation $T - X - Y$. The complexity of a feature representation can thus be formalized using its mutual information with the data source, i.e., $I(T; X)$. Moreover, the most succinct feature representations, formally defined as minimal sufficient statistics, are the ones that maximally compress the information about Y in the data X.

Definition 2 (Minimal sufficient statistic)**.** *Suppose that X is a sample from a distribution indexed by ground truth Y. A function $T(X)$ is said to be a minimal sufficient statistic* *for Y if*

$$T(X) \in \mathit{arg}\min_{S} I(X; S(X)) \tag{1}$$

$$\textit{s.t. } I(Y; S(X)) = \max_{T'} I(Y; T'(X)) \tag{2}$$

i.e., it is a statistic that has smallest mutual information with X while having largest mutual information with Y.

Let $\hat{Y} = g(T)$ be the output of neural network and define the probability of prediction error as $P_e = P[\hat{Y} \neq Y]$. Indeed, P_e is closely related to the conditional entropy of Y given T, i.e., $H(Y|T)$. From Fano's inequality Cover & Thomas (2012), we know that the lower bound on P_e is determined by $H(Y|T)$:

$$P_e \geq \frac{H(Y|T) - 1}{\log |\mathcal{Y}|} \tag{3}$$

and $P_e = 0$ implies $H(Y|T) = 0$. Moreover, in practice, $H(Y|T)$ is often approximated via a variational method Kolchinsky et al. (2017); Alemi et al. (2016); Belghazi et al. (2018), in which case $H(Y|T)$ is equal to the cross-entropy loss. As a result, lower P_e implies lower $H(Y|T)$ in practical implementations. Based on the above discussions, we can consider $H(Y|T)$ a proxy for the error probability.

Assumption 1 (Monotonic entropy-error assumption)**.** *For two statistics T and T' of X, let the probability of error associated with T and T' be $P_e = P[g(T) \neq Y]$ and $P'_e = P[g(T') \neq Y]$. We assume that if $P_e \leq P'_e$, then $H(Y|T) \leq H(Y|T')$.*

The following theorem shows that under the monotonic entropy-error assumption, not using minimal sufficient statistics is a necessary condition for the existence of adversarial examples.

Theorem 1. *Suppose that Assumption 1 holds and there exist adversarial examples for the neural network $f(\cdot) = g(T(\cdot))$. Then, $T(X)$ is not a minimal sufficient statistic.*

The proof idea of Theorem 1 is to construct another statistic that achieves lower mutual information with X but higher with Y. Due to the space limitation, we omit detailed proof to the supplementary material.

We now present a toy example to show that using minimal sufficient statistics will eradicate adversarial examples. Consider that data is generated from the following distribution: $Y \sim \text{Bernoulli}(0.5)$ and $p(X|Y = i) = \mathcal{N}(\mu_i, \Sigma)$. It is known that under this data distribution, the Bayes optimal predictor is a linear model, which has the form $o(x) = 1$ if $\theta^T x > c$ and $o(x) = 0$ otherwise, where $\theta = \Sigma^{-1}(\mu_1 - \mu_0)$ and $c = \frac{1}{2}(-\mu_0^T \Sigma^{-1} \mu_0 + \mu_1^T \Sigma^{-1} \mu_1)$. We can regard $o(x)$ as the oracle. Moreover, it can be proved that in this case, the log likelihood ratio $T(X) = \theta^T x$ is a minimal sufficient statistic. In this simple example, the minimal sufficient statistic equates to the oracle classifier and thus have no adversarial examples.

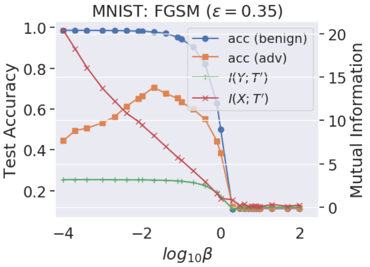

Figure 1. The gap between benign training accuracy and adversarial testing accuracy decreases with $I(X;T)$. The model is trained with the Variational IB Alemi et al. (2016) on the `MNIST` dataset.

Although the theorem implies that learning a minimal sufficient statistic from data can help boost adversarial robustness, it is shown in Pitman-Koopman-Darmois therem Koopman (1936) that sufficient statistic whose dimension does not depend on the sample size exists only for an exponential family. This limits interest in solving exact minimal sufficient statistics for complex learning tasks. We will discuss alternative more practical strategies to approach the goal.

4. Improving Adversarial Robustness via IB

In this section, we will introduce the IB framework and show that it allows us to naturally extend the concept of minimal sufficient statistic to any joint distribution of X and Y. Within the framework, one solves for the feature representation T that maximizes mutual information with Y, corresponding to the sufficiency for Y, while minimizing the mutual information with X, corresponding to the minimality of the statistic. More specifically, suppose that the encoder is parameterized by θ, the following Lagrangian formulation of IB is often used for learning θ:

$$\max_{\theta} I(Y; T) - \beta I(X; T) \tag{4}$$

where β is a constant that controls the tradeoff between learning performance and minimality of the extracted feature. In practice, the value of β is tuned via a line search Alemi et al. (2016).

Figure 1 illustrates how β affects $I(X;T)$, $I(Y;T)$, testing accuracy on benign and adversarial examples for a model trained with the IB learning objective (4). We can see that $I(X;T)$ decreases monotonically with β. The adversarial robustness (i.e., the testing accuracy on adversarial examples) increases as $I(X;T)$ becomes lower until the point where $I(Y;T)$ begins to shrink. This shows that "sufficiency" and "minimality" are both needed for better robustness and partially justifies Theorem 1. However, when focused on the gap between benign training accuracy and adversarial testing accuracy, we find that this gap continually narrows as $I(X;T)$ decreases.

Next, we will present a bound on this gap that can theoretically explain the above observation. The bound can also provide insights into other factors that affect the adversarial robustness of a model trained with IB.

4.1. Adversarial Robustness Bound

Due to the prevalence of using stochastic encoders when learning with IB, we now extend the encoder definition to incorporate stochasticity. Let the encoder T be defined via $p(t|x)$ for $t \in \mathcal{T}$ and $x \in \mathcal{X}$ and $p(t|x)$ is parameterized by θ. This definition subsumes the deterministic encoder previously discussed, which corresponds to $p(t|x)$ assigning probability mass one to a single value in $\mathcal{T}$ and zero probability elsewhere.

Let X be the random variables corresponding to benign data. We define the adversarial data distribution induced by the attack algorithm A as $X'(A)$. For notational convenience, we suppress the dependence on A in the remainder of the paper. The distributions of X and X' are denoted by $p(x)$ and $q(x)$, which have the support $\mathcal{X}_b$ and $\mathcal{X}_a$, respectively. Note that $q(x) = \sum_{x' \in \mathcal{X}_b, A(x')=x} p(x')$. Let T and T' be the random variables representing features generated by the benign data and adversarial data. The distributions of T and T' have support $\mathcal{T}_b$ and $\mathcal{T}_a$, respectively.

Theorem 2. *Suppose that $p(t|x)$ is a L-lipschitz function of x for any given t. Then,*

$$\begin{aligned}&|I(Y;T) - I(Y;T')| \le |\mathcal{T}_b \cup \mathcal{T}_a|\psi(L\epsilon)\\ &+ \max\{C_1\sqrt{|\mathcal{T}_b|}(I(X;T))^{\frac{1}{2}} + C_2|\mathcal{T}_b|^{\frac{3}{4}}(I(X;T))^{\frac{1}{4}},\\ &C_3\sqrt{|\mathcal{T}_a|}(I(X';T'))^{\frac{1}{2}} + C_4|\mathcal{T}_a|^{\frac{3}{4}}(I(X';T'))^{\frac{1}{4}}\} \qquad (5)\end{aligned}$$

where C_1, C_2, C_3 and C_4 are some constants depending only on $p(x)$ and the attack algorithm. ϵ is the bound on the magnitude of the perturbation exerted by the attack algorithm. $\psi(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ is a monotonically increasing function defined as

$$\psi(x) = \begin{cases} 0 & x = 0 \\ x\log(1/x) & 0 < x < 1/e \\ 1/e & x > 1/e \end{cases} \qquad (6)$$

The proof of the theorem is inspired by Shamir et al. (2010). Note that the theorem holds for any fixed $p(t|x)$, not just ones which optimize (4). In particular, the theorem holds for any $p(t|x)$ found by an IB algorithm, even if $p(t|x)$ is not a global optimizer.

$I(Y;T)$ represents the utility of T in predicting Y. Since $I(Y;T) = H(Y) - H(Y|T)$, where $H(Y)$ does not depend on the encoder T that we would like to learn and $H(Y|T)$, as aforementioned, is often approximated by cross-entropy loss in practice, we can treat $I(Y;T)$ as the accuracy of the model on benign data. Similarly, $I(Y;T')$ can be interpreted as the model performance when tested on adversarial data. Thus, the left-hand side of the bound signifies the vulnerability of a model to adversarial examples when solely trained on benign data. We term the performance gap on the left-hand of the bound *Oblivious Vulnerability*.

The bound in (5) indicates that the oblivious vulnerability is controlled by $I(X;T)$ and $I(X';T')$. Note that when using IB to train on benign data, $I(X;T)$ is explicitly minimized by the learning objective. This sheds light on the reason for the empirical observation that IB can achieve moderate robustness against adversarial examples even without other more sophisticated defenses Alemi et al. (2016). A viable way of controlling $I(X';T')$ is to perform adversarial training—adding adversarial examples into the training dataset—so that $I(X';T')$ can be incorporated into the learning objective and minimized. The theorem suggests that combining IB with adversarial training should lead to better robustness than using the vanilla IB. However, this simple, actionable insight has not been explored in the prior work. We will verify this idea implied by our theorem in Section 6.

Moreover, the Lipschitz constant of the encoder also plays a part in the bound. Theorem 2 suggests that a smoother encoder will lead to lower oblivious vulnerability. Intuitively, with a smooth encoder, the feature distribution of adversarial examples will change very little from that of benign data; thus, the decoder will achieve similar performance on both adversarial and benign data. This consequence of our theorem will also be validated in Section 6.

Note that we treat $I(Y;T)$ in the bound as the training performance on benign data. Our bound has not yet taken into account the error of estimating $I(Y;T)$ due to the availability of only finite training samples. However, estimation errors can be incorporated straightforwardly by combining (5) with the results in Shamir et al. (2010) which provides an upper bound for $|\hat{I}(Y;T) - I(Y;T)|$, where $\hat{I}(Y;T)$ is the estimate of $I(Y;T)$ with finite samples. Thus, it is easy to see that $|\hat{I}(Y;T) - I(Y;T')|$ can be bounded by a sum of our current bound and the one of $|\hat{I}(Y;T) - I(Y;T)|$. It is shown in Shamir et al. (2010) that $|\hat{I}(Y;T) - I(Y;T')|$ is also controlled by $I(X;T)$; therefore, the bound that incorporates estimation errors will reveal the same insight as our current one.

4.2. Lipschitz Constant of Stochastic Encoders

Existing works have developed various techniques to compute the Lipschitz constant for a deterministic neural network Szegedy et al. (2013); Cisse et al. (2017). Specifically, neural networks are the composition of functions at different layers, and the overall Lipschitz constant can be bounded by the Lipschitz constant of each layer. However, how to obtain the Lipschitz constant for a stochastic encoder is not yet studied in the previous work.

We investigate the problem of computing the Lipschitz constant for a stochastic encoder as it is widely used when training with IB. We consider a typical stochastic encoder

$$p^*(t|x) = \mathcal{N}(t|f^\mu, f^\sigma) \tag{7}$$

where f^μ and f^σ are encoding networks whose outputs are both K-dimensional vectors. Let $f^\mu_k(x)$ denotes the kth coordinate of $f^\mu(x)$. $f^\sigma_k(x)$ is similarly defined. f^μ encodes the mean and f^σ_e encodes a diagonal covariance matrix. The following theorem shows that for any t, $p^*(t|x)$ defined above is a Lipschitz continuous function of x with a Lipschitz constant dependent on the Lipschitz constants of the mean and covariance encoders.

Theorem 3. *Suppose that f^μ_e and f^σ_e are both Lipschitz continuous functions with Lipschitz constants L^μ and L^σ, respectively. Let $\sigma_{min} = \min_x \min_{k=1,\dots,K} f^\sigma_k(x)$. Suppose that σ_{min} is a positive number. Then,*

$$\begin{aligned}&|p^*(t|x) - p^*(t|x')| \\ &\le \frac{K}{(\sqrt{2\pi})^{K-1}\sigma^2_{min}}(2\frac{L^\sigma}{e^{3/2}} + \frac{L^\mu}{e^{1/2}})\|x - x'\|\end{aligned} \tag{8}$$

This theorem shows that we can control the Lipschitz constant of this stochastic encoder via controlling its mean and variance network, which are deterministic functions. Their Lipschitz constants can be regularized using existing techniques Oberman & Calder (2018); Qian & Wegman (2018); Finlay et al. (2019).

5. Transferability Analysis for IB

One of the most intriguing properties of adversarial examples is their transferability across different models. In this section, we theoretically study the transferability of adversarial perturbations between any two low-risk classifiers that share the same stochastic encoder. This setting is practical considering the prevalence of pre-training and also a first-step to prove transferability for more general settings.

We adopt a decision-theoretic framework with $0-1$ binary class on $\mathcal{Y} \times \mathcal{Y}$. Let $l(y, y') = 0$ if $y = y'$ and $l(y, y') = 1$ otherwise, where y and y' denote true and predicted labels, respectively. A classifier is a mapping from $\mathcal{X}$ to $\mathcal{Y}$ and its risk is defined as $\mathcal{R}[f] = \mathbb{E}_{(X,Y)}[l(f(X), Y)]$.

Moreover, an attack algorithm $\mathcal{A}$ is said to be δ-effective on f_1 if $P[f(X) \neq f(\mathcal{A}(X))] \geq 1 - \delta$. In a nutshell, the goal of the transferability analysis is to prove that under some conditions for two classifiers f_1 and f_2, if $\mathcal{A}$ is δ-effective on f_1, then it will also be δ-effective on f_2.

Intuitively, for two low-risk classifiers f_1 and f_2, their classification boundaries are close on the benign feature distribution. If the encoder is smooth, then the small perturbation added into the benign data will lead to a small change in its feature distribution. As a result, the decision boundaries of the two classifiers are still close on the adversarial feature distribution. This is formally characterized in Lemma 1.

Lemma 1. *Suppose that $p(t|x)$ is a L-Lipschitz function of x and the magnitude of adversarial perturbations are bounded by ϵ. Then, we have for all $t \in \mathcal{T}$, $|p(t) - q(t)| \leq L\epsilon$, where $p(t)$ and $q(t)$ are benign and adversarial feature distribution, respectively.*

Therefore, when there is an adversarial example that can effectively attack classifier f_1, which means the adversarial instance will change the predicted label to be different from the ground truth, it will also be able to change the prediction for another low-risk classifier f_2 with high probability.

Theorem 4. *Suppose that $p(t|x)$ is a Lipschitz continuous function with the Lipschitz constant L. Suppose that f_1 and f_2 are both low risks classifiers with $P[f(x) \neq y] \leq \tau$, $f \in \{f_1, f_2\}$, and f_1 and f_2 share the same encoder $p(t|x)$, i.e., $f_1 = g_1 \circ T$ and $f_2 = g_2 \circ T$. If $A(\cdot)$ is δ-effective attack strategy on f_1, then $A(\cdot)$ is δ'-effective on f_2, where $\delta' = \delta + 4\tau + |\mathcal{T}|L\epsilon$.*

This theorem indicates that the loss of attack effectiveness when transferring from one classifier to another is dependent on the performance of two classifiers on benign data, the smoothness of the encoder, as well as the adversarial perturbation magnitude. Note that L decreases implicitly with the domain size $|\mathcal{T}|$ of the feature space with the rate $\mathcal{O}(1/|\mathcal{T}|)$. This can be illustrated by the following example: consider that the stochastic encoder is $p^*(t|x)$ in (7) with the domain constrained on the values $\{t_i\}_{i=1}^{|\mathcal{T}|}$. Then, the encoder can be expressed as $p(t|x) = p^*(t|x)/\sum_{i=1}^{|\mathcal{T}|} p^*(t_i|x)$ for $t \in \{t_i\}_{i=1}^{|\mathcal{T}|}$. Denote the Lipschitz constant of $p^*(t|x)$ by L^*, which is a finite constant. Since $p(t|x) \leq p^*(t|x)/(|\mathcal{T}|p_{\min})$, where $p_{\min} = \min_{t \in \{t_i\}_{i=1}^{|\mathcal{T}|}} p^*(t|x)$, the Lipschitz constant L of $p(t|x)$ is bounded by $\mathcal{O}(L^*/|\mathcal{T}|)$. Although the domain size $|\mathcal{T}|$ of the feature space also appears in the bound, it will be canceled out with the implicit $1/|\mathcal{T}|$ factor in L. Therefore, the bound in Theorem 4 will not affect much by specific quantization methods employed in the feature space.

6. Experimental Results

In this section, we present empirical results to justify our theoretical analysis, including the adversarial robustness bound and the transferability results of IB. We also demonstrate the effectiveness of the IB framework for both normal and adversarial training as well as the connection between adversarial training and learning succinct feature representations.

Experimental Setup The evaluations are conducted on two popular datasets, namely, `MNIST` LeCun & Cortes (2010) and `CIFAR-10` Krizhevsky & Hinton (2009), which are widely used in previous literature Papernot et al. (2016); Alemi et al. (2016); Madry et al. (2017); Carlini & Wagner (2017). It is worth noting that we use pristine data (*i.e.*, no data augmentation) in the experiments in order to clearly explore the effectiveness of IB and adversarial training Alemi et al. (2016). For the same purpose, other regularization techniques (e.g., BatchNormalization Ioffe & Szegedy (2015) or Dropout Srivastava et al. (2014)) are not used in the evaluations. We make use of five popular attack algorithms to generate adversarial examples, including the Fast Gradient Sign Method (FGSM) Goodfellow et al. (2015), the Basic Iterative Method (BIM) Kurakin et al. (2016a), the Projected Gradient Descent (PDG) Madry et al. (2017), DeepFool Moosavi-Dezfooli et al. (2016), and the Carlini and Wagner (CW) method Carlini & Wagner (2017).

We adopt the variational method in Alemi et al. (2016) to approximate the IB objective so as to perform efficient training. Our experiments also leverage similar model architectures to Alemi et al. (2016). Specifically, standard multi-layer perceptrons (MLPs) and CNNs-MLPs are designed for `MNIST` and `CIFAR` training, respectively. To build a stronger baseline and improve the robustness of networks, we expand each dense layer in Alemi et al. (2016) to several duplicate layers—replacing $1024-1024-2K$ with $\{512\}_{\times N}-\{256\}_{\times N}-2K'$, where K and K' denote the mean/variance network dimensions of the stochastic encoder, N denote the number of duplicated layers. For CNNs-MLPs architecture, we use two-layer CNN filters followed by the expanded MLP layers. These changes significantly improve the robustness of the vanilla model in Alemi et al. (2016) (e.g., more than 40% accuracy improvement on adversarial examples generated by FGSM with the perturbation magnitude $\epsilon = 0.3$). Except for the validations of Theorem 4 which require varied encoder architectures to derive different Lipschitz constants, all the other experiments adopt the aforementioned architectures with $N = 3$ by default. More details about model architectures and the settings of various attack algorithms are available in the supplementary material.

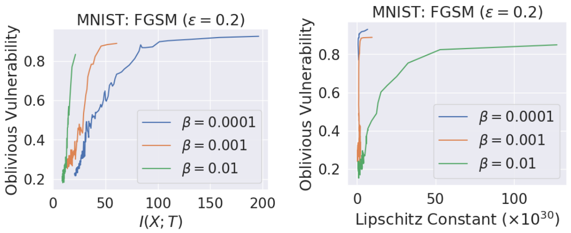

Figure 2. Validation of Theorem 2. Oblivious vulnerability is defined as LSH of (5) in the theorem, i.e., the train-test accuracy gap when training is performed on benign data and testing is on adversarial examples.

Validation of the Adversarial Robustness Bound. We first empirically validate the adversarial robustness bound in Theorem 2. Our theorem shows that the upper bound on the oblivious vulnerability (i.e., the train-test accuracy gap when training is performed on benign data and testing is on adversarial examples) increases with the Lipschitz constant of the stochastic encoder L and the feature complexity $I(X;T)$. Similar to Alemi et al. (2016), we implement the stochastic encoder as follows: we take the first half of the output of an MLP to be the mean and apply a softplus transform to the remaining half to get the standard deviation Alemi et al. (2016); the stochastic encoder is then specified by a Gaussian distribution defined by the mean and standard deviation network. In this case, it is easy to verify that the Lipschitz constant of the entire MLP is an upper bound on the Lipschitz constants of both mean and standard deviation networks. Moreover, by Theorem 3, we know that the Lipschitz constant of the stochastic encoder is controlled by the Lipschitz constants of mean and standard deviation networks. Thus, we compute the Lipschitz constant for the entire MLP network as a proxy for the smoothness of the stochastic encoder. The result is illustrated in Figure 2, where the oblivious vulnerability is computed by subtracting the testing accuracy of the learned model from the training accuracy in each epoch and $I(X;T)$ is the mutual information value in each epoch. Figure 2 shows that lower Lipschitz constant and $I(X;T)$ leads to lower oblivious vulnerability, which conforms with the theorem.

Controlling $I(X';T')$ via Adversarial Training. Another implication of the bound in Theorem 2 is that we can reduce the oblivious vulnerability via controlling $I(X';T')$, i.e., the complexity of features generated adversarial examples, because this term appears in the right-hand side of (5). As discussed before, a simple way to regularize $I(X';T')$ is to combine adversarial training with IB. We test this idea on the `MNIST` dataset with adversarial examples generated by different attacks. The result is shown in Table 1. We can see that applying adversarial training to IB can successfully control $I(X';T')$ and achieves much higher robustness than the vanilla IB.

Transferability. We validate the transferability result for IB presented in Theorem 4, which states that for a fixed black-box attack and two target models that share the same encoder, the attack success rate drop $\delta' - \delta$ due to the trans-

Table 1. Comparison of adversarial robustness and $I(X';T')$ between vanilla IB and IB combined with adversarial training.

Setting			FGSM	BIM	PGD
			$(\epsilon = 0.3)$	$(\epsilon = 0.3)$	$(\epsilon = 0.3)$
$I(X';T')$	$\beta = 1e^{-1}$	IB	90.56	17.80	17.78
		IB + adv.	**5.22**	**5.05**	**5.04**
	$\beta = 1e^{-2}$	IB	62.39	22.24	21.78
		IB + adv.	**10.73**	**9.23**	**9.11**
	$\beta = 1e^{-3}$	IB	32.70	21.91	23.66
		IB + adv.	**25.88**	**19.26**	**18.42**
Test acc. on adv. examples	$\beta = 1e^{-1}$	IB	70.96	72.84	74.92
		IB + adv.	**91.43**	**87.34**	**86.75**
	$\beta = 1e^{-2}$	IB	68.86	66.46	66.67
		IB + adv.	**91.86**	**90.08**	**89.78**
	$\beta = 1e^{-3}$	IB	59.87	59.88	57.97
		IB + adv.	**91.54**	**89.59**	**89.88**

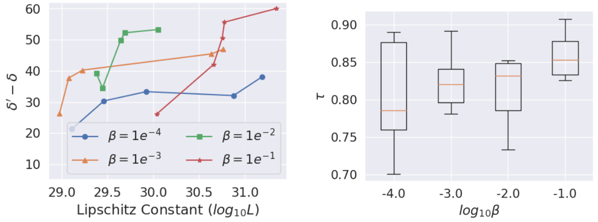

Figure 3. Validation of Theorem 4. $\delta' - \delta$ indicates the attack success rate drop due to the transfer from one model to another. τ is the maximum of the testing accuracy of a pair of models.

fer (hereafter referred to as the transferability gap) increases with the error τ of the local and remote model on benign data as well as the Lipschitz constant of the shared encoder. In this experiment, we train multiple pairs of neural networks; the models in each pair shares the same stochastic encoder while different pairs use different encoder architectures and therefore have different Lipschitz constants L. It is noted that different encoder architectures will also lead to different testing errors on benign data, i.e., τ. It is, therefore, difficult to perfectly control one variable and inspect the effect of another on the transferability gap. We perform statistical tests on the linear correlation between $\delta' - \delta$, L and τ. The Pearson correlation coefficient between L and $\delta' - \delta$ is 0.57 with the p-value 0.0015 and the one between τ and $\delta' - \delta$ is 0.51 with the p-value 0.0054. Hence, both L and τ have positive linear correlations with the transferability gap, which conforms with our analysis. Figure 3 illustrates the trend of the transferability gap when the Lipschitz constant changes. Since τ correlates with β in our experiment (shown in the right penal of Figure 3), thus we separate trend lines for different β to better control the effect of τ. Although τ exhibits variations even for a fixed β, the left panel of Figure 3 shows that the transferability gap is more sensitive to the encoder Lipschitz constant and increases almost monotonically with L.

Comparing IB with Cross-Entropy. We compare the IB learning objective with CE. We train neural networks using each objective on benign data and then test their performance on both benign data and adversarial examples generated by different attacks. The result on `MNIST` is presented in the first two rows of Table 2. Although both objectives achieve similar generalization on benign data, IB can lead to significantly higher robustness than CE. We also tested the two learning objectives when combined with adversarial training. We generate the training adversarial examples using FGSM with the perturbation magnitude $\epsilon = 0.3$. The performance of the resulting models is tested on benign data, adversarial data generated from the same attack with different magnitudes as well as different attacks. The last two rows of Table 2 shows the result, which again demonstrates the superiority of learning with IB to CE. We can see that CE with adversarial training can generalize well to the testing adversarial examples generated by the same attack and magnitude as the training ones; however, adversarial training on CE cannot confer robustness to adversarial examples generated from a different attack. Particularly, it can only achieve testing 2% accuracy on adversarial examples generated by DeepFool, which is a striking difference from the 97.94% accuracy achieved when the training and testing data adversarial examples come from the same source. This phenomenon is also observed in Kurakin et al. (2016b). By contrast, adversarial training when combined with IB achieves higher robustness to the attacks unexpected during training. Similar results can be observed on `CIFAR` (see Table 3). However, the performance gap between CE and IB on `CIFAR` is smaller than that on `MNIST`. This indicates the difficulty of using IB to control information flow in complex learning tasks.

Adversarial training: an implicit regularizer for feature complexity. We perform adversarial training with different ratios of adversarial examples added into the training dataset. As shown in Figure 4, this gives rise to models with different levels of robustness measured in terms of testing accuracy on an adversarial testing set. At the same time, we measure the mutual information between data inputs and feature representation, which is equal to the entropy of features for a deterministic neural network. We used an all-CNNs network and estimated the entropy of features extracted by the last convolutional layer using Maximum Likelihood and JVHM estimators Jiao et al. (2014). As illustrated by Figure 4, the estimated entropy (blue dots) decreases as the robustness (red dots) increases for different attacks and datasets. In other words, the complexity of the feature representation is lower when the models become more robust. Thus, it seems that adding adversarial examples into the training set serves as an implicit regularizer for feature complexity. How to theoretically prove this interesting observation is thus far an open question.

Table 2. Comparison of IB and CE with/without adversarial training on MNIST. The adversarial examples used for adversarial training are generated by FGSM with $\epsilon = 0.3$. $\beta = 0.001$.

Setting	Benign	FGSM			DeepFool	CW (L_2)	BIM	PGD
	-	$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 0.5$	$m = 10^2$	$c = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.2$
IB	**98.64**	**85.88**	**62.94**	**40.63**	**34.53**	**53.30**	**58.41**	**51.79**
CE	98.63	63.39	1.38	1.04	1.79	18.33	2.18	2.12
IB + adv.	**98.53**	**92.13**	**90.43**	**55.27**	**43.08**	**49.69**	**58.97**	**56.65**
CE + adv.	97.94	78.95	89.73	50.33	2.54	16.81	21.82	21.36

Table 3. Comparison of IB and CE with/without adversarial training on CIRAR. The adversarial examples used for adversarial training are generated by FGSM with $\epsilon = 0.3$. $\beta = 0.001$.

Setting	Benign	FGSM			DeepFool	CW (L_2)	BIM	PGD
	-	$\epsilon = 0.1$	$\epsilon = 0.3$	$\epsilon = 0.5$	$m = 10^2$	$c = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.2$
IB	**67.06**	**34.92**	**20.15**	**14.73**	**23.47**	**21.37**	**19.06**	**20.04**
CE	65.39	17.33	12.23	10.35	18.5	11.63	18.67	18.19
IB + adv.	**63.14**	**44.48**	**58.64**	**56.28**	**26.64**	**13.20**	**21.52**	**22.86**
CE + adv.	61.67	29.4	57.04	48.44	21.79	11.65	19.69	19.78

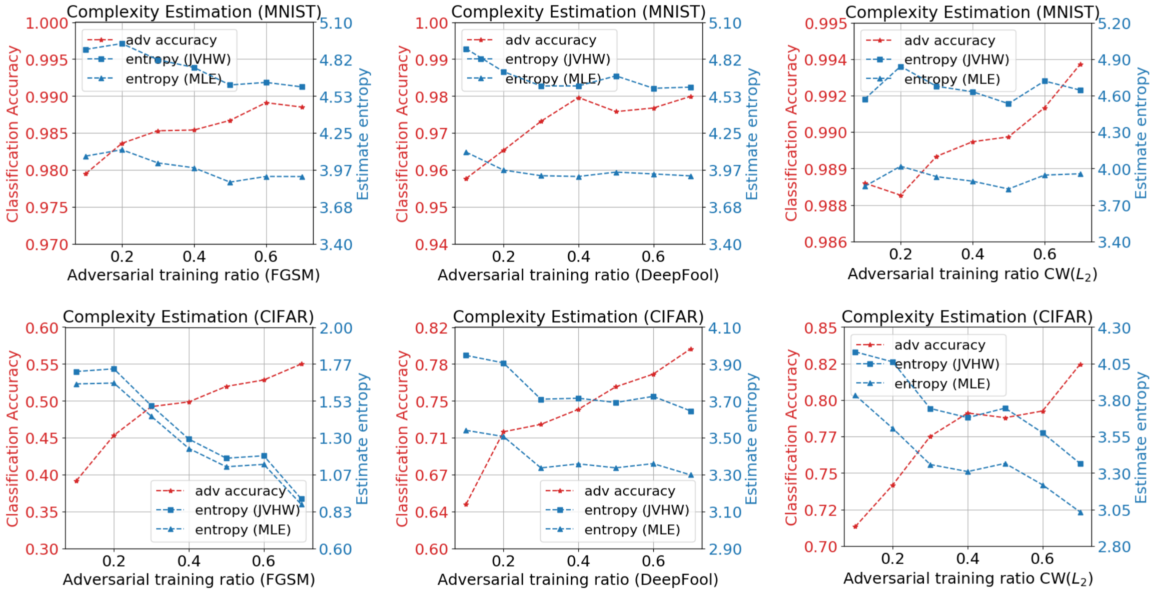

Figure 4. Estimated entropy of feature representation for models with different robustness, which are generated by adversarial training with different ratios. The three columns correspond to the results on three adversarial attacks (FGSM, DeepFool, CW).

7. Conclusion

This paper presents an information-theoretic view of adversarial examples. We prove that under mild assumptions, failing to learn the minimal sufficient statistics as feature representation is a necessary condition for the existence of adversarial examples. This result justifies the widely used IB principle for learning in adversarial environments. We prove an adversarial robustness bound for learning with IB, which attributes the robustness of IB to the continuity of the stochastic encoder as well as the feature complexity. We also prove a bound of transferability between any two low-risk models based on IB which share the same stochastic representation encoder. We hope our theoretic analysis and results can provide more insights into adversarial examples and further facilitate the design of robust machine learning models.

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Bao, Z., Muñoz-González, L., and Lupu, E. C. Mitigation of adversarial attacks through embedded feature selection. *arXiv preprint arXiv:1808.05705*, 2018.

Belghazi, I., Rajeswar, S., Baratin, A., Hjelm, R. D., and Courville, A. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Bhagoji, A. N., Cullina, D., Sitawarin, C., and Mittal, P. Enhancing robustness of machine learning systems via data transformations. In *Information Sciences and Systems (CISS), 2018 52nd Annual Conference on*, pp. 1–5. IEEE, 2018.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 39–57. IEEE, 2017.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. *arXiv preprint arXiv:1704.08847*, 2017.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

Dhillon, G. S., Azizzadenesheli, K., Lipton, Z. C., Bernstein, J., Kossaifi, J., Khanna, A., and Anandkumar, A. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

Fawzi, A., Moosavi-Dezfooli, S.-M., and Frossard, P. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, pp. 1632–1640, 2016.

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

Finlay, C., Oberman, A. M., and Abbasi, B. Improved robustness to adversarial examples using lipschitz regularization of the loss, 2019. URL https://openreview.net/forum?id=HkxAisC9FQ.

Ford, N., Gilmer, J., and Cubuk, E. D. Adversarial examples are a natural consequence of test error in noise, 2019. URL https://openreview.net/forum?id=S1xoy3CcYX.

Galloway, A., Golubeva, A., and Taylor, G. W. A rate-distortion theory of adversarial examples. 2018.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Goyal, A., Islam, R., Strouse, D., Ahmed, Z., Larochelle, H., Botvinick, M., Levine, S., and Bengio, Y. Transfer and exploration via the information bottleneck. 2018.

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

Guo, C., Rana, M., Cisse, M., and van der Maaten, L. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 448–456. JMLR.org, 2015.

Jacobsen, J.-H., Behrmann, J., Zemel, R., and Bethge, M. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*, 2018.

Jeon, I., Lee, W., and Kim, G. IB-GAN: Disentangled representation learning with information bottleneck GAN, 2019. URL https://openreview.net/forum?id=ryljV2A5KX.

Jiao, J., Venkat, K., and Weissman, T. Order-optimal estimation of functionals of discrete distributions. *CoRR*, abs/1406.6956, 2014. URL http://arxiv.org/abs/1406.6956.

Kolchinsky, A., Tracey, B. D., and Wolpert, D. H. Nonlinear information bottleneck. *arXiv preprint arXiv:1705.02436*, 2017.

Koopman, B. O. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409, 1936.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016a.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial machine learning at scale. *CoRR*, abs/1611.01236, 2016b.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. In *ICLR*, 2017.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

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

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

Nayebi, A. and Ganguli, S. Biologically inspired protection of deep networks from adversarial attacks. *arXiv preprint arXiv:1703.09202*, 2017.

Oberman, A. M. and Calder, J. Lipschitz regularized deep neural networks generalize. 2018.

Papernot, N., McDaniel, P. D., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. The limitations of deep learning in adversarial settings. In *EuroS&P*, pp. 372–387. IEEE, 2016.

Peng, X. B., Kanazawa, A., Toyer, S., Abbeel, P., and Levine, S. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *CoRR*, abs/1810.00821, 2018.

Qian, H. and Wegman, M. N. L2-nonexpansive neural networks. *CoRR*, abs/1802.07896, 2018.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shamir, O., Sabato, S., and Tishby, N. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Song, Y., Kim, T., Nowozin, S., Ermon, S., and Kushman, N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

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

Tanay, T. and Griffin, L. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *Information Theory Workshop (ITW), 2015 IEEE*, pp. 1–5. IEEE, 2015.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.