Learning Adversarially Robust Representations via Worst-Case Mutual Information Maximization

Sicheng Zhu [1 *] **Xiao Zhang** [1 *] **David Evans** [1]

Abstract

Training machine learning models that are robust against adversarial inputs poses seemingly insurmountable challenges. To better understand adversarial robustness, we consider the underlying problem of learning robust representations. We develop a notion of *representation vulnerability* that captures the maximum change of mutual information between the input and output distributions, under the worst-case input perturbation. Then, we prove a theorem that establishes a lower bound on the minimum adversarial risk that can be achieved for any downstream classifier based on its representation vulnerability. We propose an unsupervised learning method for obtaining intrinsically robust representations by maximizing the worst-case mutual information between the input and output distributions. Experiments on downstream classification tasks support the robustness of the representations found using unsupervised learning with our training principle.

1. Introduction

Machine learning has made remarkable breakthroughs in many fields, including computer vision (He et al., 2016) and natural language processing (Devlin et al., 2019), especially when evaluated on classification accuracy on a given dataset. However, adversarial vulnerability (Szegedy et al., 2014; Engstrom et al., 2017), remains a serious problem that impedes the deployment of the state-of-the-art machine learning models in safety-critical applications, such as autonomous driving (Eykholt et al., 2018) and face recognition (Sharif et al., 2016). Despite extensive efforts to improve model robustness, state-of-the-art adversarially robust training methods (Mądry et al., 2018; Zhang et al., 2019) still fail to produce robust models, even for simple classification tasks on CIFAR-10 (Krizhevsky et al., 2009).

In addition to many ineffective empirical attempts for achieving model robustness, recent studies have identified intrinsic difficulties for learning in the presence of adversarial examples. For instance, a line of works (Gilmer et al., 2018; Fawzi et al., 2018; Mahloujifar et al., 2019; Shafahi et al., 2018) proved that adversarial vulnerability is inevitable if the underlying input distribution is concentrated. Schmidt et al. (2018) showed that for certain learning problems, adversarially robust generalization requires more sample complexity compared with standard one, whereas Bubeck et al. (2019) constructed a specific task on which adversarially robust learning is computationally intractable.

Motivated by the apparent empirical and theoretical difficulties of robust learning with adversarial examples, we focus on the underlying problem of learning adversarially robust representations (Garg et al., 2018; Pensia et al., 2020). Given an input space $\mathcal{X} \subseteq \mathbb{R}^d$ and a feature space $\mathcal{Z} \subseteq \mathbb{R}^n$, any function $g : \mathcal{X} \to \mathcal{Z}$ is called a representation with respect to $(\mathcal{X}, \mathcal{Z})$. Adversarially robust representations denote the set of functions from $\mathcal{X}$ to $\mathcal{Z}$ that are less sensitive to adversarial perturbations with respect to some metric Δ defined on $\mathcal{X}$. Note that one can always get an overall classification model by learning a downstream classifier given a representation, thus learning representations that are robust can be viewed as an intermediate step for the ultimate goal of finding adversarially robust models. In this sense, learning adversarially robust representations may help us better understand adversarial examples, and perhaps more importantly, bypass some of the aforementioned intrinsic difficulties for achieving model robustness.

In this paper, we give a general definition for robust representations based on mutual information, then study its implications on model robustness for a downstream classification task. Finally, we propose empirical methods for estimating and inducing representation robustness.

Contributions. Motivated by the empirical success of standard representation learning using the mutual information maximization principle (Bell & Sejnowski, 1995; Hjelm et al., 2018), we first give a formal definition on *representa-*

[*]Equal contribution [1]Department of Computer Science, University of Virginia. Correspondence to: Sicheng Zhu <sz6hw@virginia.edu>, Xiao Zhang <xz7bc@virginia.edu>, David Evans <evans@virginia.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

tion vulnerability as the maximum change of mutual information between the representation's input and output against adversarial input perturbations bound in an ∞-Wasserstein ball (Section 3). Under a Gaussian mixture model, we established theoretical connections between the robustness of a given representation and the adversarial gap of the best classifier that can be based on it (Section 3.1). In addition, based on the standard mutual information and the representation vulnerability, we proved a fundamental lower bound on the minimum adversarial risk that can be achieved for any downstream classifiers built upon a representation with given representation vulnerability (Section 3.2).

To further study the implication of robust representations, we first propose a heuristic algorithm to empirically estimate the vulnerability of a given representation (Section 4), and then by adding a regularization term on representation vulnerability in the objective of mutual information maximization principle, provide an unsupervised way for training meaningful and robust representations (Section 5). We observe a direct correlation between model and representation robustness in experiments on benchmark image datasets MINST and CIFAR-10 (Section 6.1). Experiments on downstream classification tasks and saliency maps further show the effectiveness of our proposed training method in obtaining more robust representations (Section 6.2).

Related Work. With similar motivations, several different definitions of robust features have been proposed in literature. The pioneering work of Garg et al. (2018) considered a feature to be robust if it is insensitive to input perturbations in terms of the output values. However, their definition of feature robustness is not invariant to scale changes. Based on the linear correlation between feature outputs and true labels, Ilyas et al. (2019) proposed a definition of robust features to understand adversarial examples, whereas Eykholt et al. (2019) proposed to study robust features whose outputs will not change with respect to small input perturbations. However, these two definitions either require the additional label information or restrict the feature space to be discrete, thus are not general. The most closely related work to ours is Pensia et al. (2020), which considered Fisher information of the output distribution as the indicator of feature robustness and proposed a robust information bottleneck method for extracting robust features. Compared with Pensia et al. (2020), our definition is defined for the worst-case input distribution perturbation, whereas Fisher information can only capture feature's sensitivity near the input distribution. In addition, our proposed training method for robust representations is better in the sense that it is unsupervised.

Notation. We use small boldface letters such as $\boldsymbol{x}$ to denote vectors and capital letters such as X to denote random variables. Let $(\mathcal{X}, \Delta)$ be a metric space, where $\Delta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is some distance metric. Let $\mathcal{P}(\mathcal{X})$ denote the set of all probability measures on $\mathcal{X}$ and $\delta_{\boldsymbol{x}}$ be the dirac measure at $\boldsymbol{x} \in \mathcal{X}$. Let $\mathcal{B}(\boldsymbol{x}, \epsilon, \Delta) = \{\boldsymbol{x}' \in \mathcal{X} : \Delta(\boldsymbol{x}', \boldsymbol{x}) \leq \epsilon\}$ be the ball around $\boldsymbol{x}$ with radius ϵ. When Δ is free of context, we simply write $\mathcal{B}(\boldsymbol{x}, \epsilon) = \mathcal{B}(\boldsymbol{x}, \epsilon, \Delta)$. Denote by $\operatorname{sgn}(\cdot)$ the sign function such that $\operatorname{sgn}(x) = 1$ if $x \geq 0$; $\operatorname{sgn}(x) = -1$ otherwise. Given $f : \mathcal{X} \to \mathcal{Y}$ and $g : \mathcal{Y} \to \mathcal{Z}$, define $g \circ f$ as their composition such that for any $\boldsymbol{x} \in \mathcal{X}$, $(g \circ f)(\boldsymbol{x}) = g(f(\boldsymbol{x}))$. We use $[m]$ to denote $\{1, 2, \ldots, m\}$ and $|\mathcal{A}|$ to denote the cardinality of a finite set $\mathcal{A}$. For any $\boldsymbol{x} \in \mathbb{R}^d$, the ℓ_p-norm of $\boldsymbol{x}$ is defined as $\|\boldsymbol{x}\|_p = (\sum_{i \in [d]} x_i^p)^{1/p}$ for any $p \geq 1$. For any $\boldsymbol{\theta} \in \mathbb{R}^d$ and positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, denote by $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ the d-dimensional Gaussian distribution with mean vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}$.

2. Preliminaries

This section introduces the main ideas we build upon: mutual information, Wasserstein distance and adversarial risk.

Mutual information. Mutual information is an entropy-based measure of the mutual dependence between variables:

Definition 2.1. Let (X, Z) be a pair of random variables with values over the space $\mathcal{X} \times \mathcal{Z}$. The *mutual information* of (X, Z) is defined as:

$$\mathrm{I}(X; Z) = \int_{\mathcal{Z}} \int_{\mathcal{X}} p_{XZ}(\boldsymbol{x}, \boldsymbol{z}) \log \left(\frac{p_{XZ}(\boldsymbol{x}, \boldsymbol{z})}{p_X(\boldsymbol{x}) p_Z(\boldsymbol{z})} \right) d\boldsymbol{x} d\boldsymbol{z},$$

where p_{XZ} is the joint probability density function of (X, Z), and p_X, p_Z are the marginal probability density functions of X and Z, respectively.

Intuitively, $\mathrm{I}(X; Z)$ tells us how well one can predict Z from X (and X from Z, since it is symmetrical). By definition, $\mathrm{I}(X; Z) = 0$ if X and Z are independent; when X and Z are identical, $\mathrm{I}(X; X)$ equals to the entropy $\mathrm{H}(X)$.

Wasserstein distance. Wasserstein distance is a distance function defined between two probability distributions on a given metric space:

Definition 2.2. Let $(\mathcal{X}, \Delta)$ be a metric space with bounded support. Given two probability measures μ and ν on $(\mathcal{X}, \Delta)$, the *p-th Wasserstein distance*, for any $p \geq 1$, is defined as:

$$\mathrm{W}_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} \Delta(\boldsymbol{x}, \boldsymbol{x}')^p \, d\gamma(x, x') \right)^{1/p},$$

where $\Gamma(\mu, \nu)$ is the collection of all probability measures on $\mathcal{X} \times \mathcal{X}$ with μ and ν being the marginals of the first and second factor, respectively. The *p-th Wasserstein ball* with respect to μ and radius $\epsilon \geq 0$ is defined as:

$$\mathcal{B}_{\mathrm{W}_p}(\mu, \epsilon) = \{\mu' \in \mathcal{P}(\mathcal{X}) : \mathrm{W}_p(\mu', \mu) \leq \epsilon\}.$$

The ∞-Wasserstein distance is defined as the limit of p-th Wasserstein distance, $\mathrm{W}_\infty(\mu, \nu) = \lim_{p\to\infty} \mathrm{W}_p(\mu, \nu)$.

Adversarial risk. Adversarial risk captures the vulnerability of a given classification model to input perturbations:

Definition 2.3. Let $(\mathcal{X}, \Delta)$ be the input metric space and $\mathcal{Y}$ be the set of labels. Let μ_{XY} be the underlying distribution of the input and label pairs. For any classifier $f : \mathcal{X} \to \mathcal{Y}$, the *adversarial risk* of f with respect to $\epsilon \geq 0$ is defined as:

$$\mathrm{AdvRisk}_\epsilon(f) = \Pr_{(\boldsymbol{x},y)\sim\mu_{XY}} \big[\exists\, \boldsymbol{x}' \in \mathcal{B}(\boldsymbol{x}, \epsilon) \text{ s.t. } f(\boldsymbol{x}') \neq y\big].$$

Adversarial risk with $\epsilon = 0$ is equivalent to standard risk, namely $\mathrm{AdvRisk}_0(f) = \mathrm{Risk}(f) = \Pr_{(\boldsymbol{x},y)\sim\mu}[f(\boldsymbol{x}) \neq y]$. For any classifier $f : \mathcal{X} \to \mathcal{Y}$, we define the *adversarial gap* of f with respect to ϵ as:

$$\mathrm{AG}_\epsilon(f) = \mathrm{AdvRisk}_\epsilon(f) - \mathrm{Risk}(f).$$

3. Adversarially Robust Representations

In this section, we first propose a definition of representation vulnerability, and then prove several theorems that bound achievable model robustness based on representation vulnerability. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space and $\mathcal{Z} \subseteq \mathbb{R}^n$ be some feature space. In this work, we define a representation to be a function g that maps any input $\boldsymbol{x}$ in $\mathcal{X}$ to some vector $g(\boldsymbol{x}) \in \mathcal{Z}$. A classifier, $f = h \circ g$, maps an input to a label in a label space $\mathcal{Y}$, and is a composition of a downstream classifier, $h : \mathcal{Z} \to \mathcal{Y}$, with a representation, $g : \mathcal{X} \to \mathcal{Z}$. As is done in previous works (Garg et al., 2018; Ilyas et al., 2019), we define a feature as a function from $\mathcal{X}$ to $\mathbb{R}$, so can think of a representation as an array of features.

Inspired by the empirical success of standard representation learning using the mutual information maximization principle (Hjelm et al., 2018), we propose the following definition of *representation vulnerability*, which captures the robustness of a given representation against input distribution perturbations in terms of mutual information between its input and output.

Definition 3.1. Let $(\mathcal{X}, \mu_X, \Delta)$ be a metric probability space of inputs and $\mathcal{Z}$ be some feature space. Given a representation $g : \mathcal{X} \to \mathcal{Z}$ and $\epsilon \geq 0$, the *representation vulnerability* of g with respect to perturbations bounded in an ∞-Wasserstein ball with radius ϵ is defined as:

$$\mathrm{RV}_\epsilon(g) = \sup_{\mu_{X'}\in\mathcal{B}_{\mathrm{W}_\infty}(\mu_X,\epsilon)} \big[\mathrm{I}(X; g(X)) - \mathrm{I}(X'; g(X'))\big],$$

where X and X' denote random variables that follow μ_X and $\mu_{X'}$, respectively.

Representation vulnerability is always non-negative, and higher values indicate that the representation is less robust to adversarial input distribution perturbations. More formally, given parameters $\epsilon \geq 0$ and $\tau \geq 0$, a representation g is called (ϵ, τ)-*robust* if $\mathrm{RV}_\epsilon(g) \leq \tau$.

Notably, using the ∞-Wasserstein distance does not restrict the choice of the metric function Δ of the input space. This metric Δ corresponds to the perturbation metric for defining adversarial examples. Thus, based on our definition of representation vulnerability, our following theoretical results and empirical methods work with any adversarial perturbation, including any ℓ_p-norm based attack.

Compared with existing definitions of robust features (Garg et al., 2018; Ilyas et al., 2019; Eykholt et al., 2019), our definition is more general and enjoys several desirable properties. As it does not impose any constraint on the feature space, it is invariant to scale change[1] and it does not require the knowledge of the labels. The most similar definition to ours is from Pensia et al. (2020), who propose to use statistical Fisher information as the evaluation criteria for feature robustness. However, Fisher information can only capture the average sensitivity of the log conditional density to small changes on the input distribution (when $\epsilon \to 0$), whereas our definition is defined with respect to the worst-case input distribution perturbations in an ∞-Wasserstein ball, which is more aligned with the adversarial setting. As will be shown next, our representation robustness notion has a clear connection with the potential model robustness of any classifier that can be built upon a representation.

3.1. Gaussian Mixture

We first study the implications of representation vulnerability under a simple Gaussian mixture model. We consider $\mathcal{X} \subseteq \mathbb{R}^d$ as the input space and $\mathcal{Y} = \{-1, 1\}$ as the space of binary labels. Assume μ_{XY} is the underlying joint probability distribution defined over $\mathcal{X} \times \mathcal{Y}$, where all the examples $(\boldsymbol{x}, y) \sim \mu_{XY}$ are generated according to

$$y \sim \mathrm{Unif}\{-1, +1\}, \quad \boldsymbol{x} \sim \mathcal{N}(y \cdot \boldsymbol{\theta}^*, \boldsymbol{\Sigma}^*), \tag{3.1}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}^* \in \mathbb{R}^{d\times d}$ are given parameters. The following theorem, proven in Appendix A.1, connects the vulnerability of a given representation with the adversarial gap of the best classifier based on the representation.

Theorem 3.2. Let $(\mathcal{X}, \|\cdot\|_p)$ be the input metric space and $\mathcal{Y} = \{-1, 1\}$ be the label space. Assume the underlying data are generated according to (3.1). Consider the feature space $\mathcal{Z} = \{-1, 1\}$ and the set of representations,

$$\mathcal{G}_{\text{bin}} = \{g : \boldsymbol{x} \mapsto \mathrm{sgn}(\boldsymbol{w}^\top \boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X} \;\big|\; \|\boldsymbol{w}\|_2 = 1\}.$$

Let $\mathcal{H} = \{h : \mathcal{Z} \to \mathcal{Y}\}$ be the set of non-trivial downstream

[1] Scale-invariance is desirable for representation robustness. Otherwise, one can always divide the function by some large constant to improve its robustness, e.g., Garg et al. (2018).

classifiers.[2] Given $\epsilon \geq 0$, for any $g \in \mathcal{G}_{\text{bin}}$, we have

$$\int_{\frac{1}{2}-\mathrm{AG}_\epsilon(f^*)}^{\frac{1}{2}} \mathrm{H}_2'(\theta)d\theta \leq \mathrm{RV}_\epsilon(g) \leq \int_{\frac{1}{2}-\frac{1}{2}\mathrm{AG}_\epsilon(f^*)}^{\frac{1}{2}} \mathrm{H}_2'(\theta)d\theta,$$

where $f^* = \operatorname{argmin}_{h\in\mathcal{H}} \mathrm{AdvRisk}_\epsilon(h \circ g)$ is the optimal classifier based on g, $\mathrm{H}_2(\theta) = -\theta\log\theta - (1-\theta)\log(1-\theta)$ is the binary entropy function and H_2' denotes its derivative.

For this theoretical model for a simple case, Theorem 3.2 reveals the strong connection between representation vulnerability and the adversarial gap achieved by the optimal downstream classifier based on the representation. Note that the binary entropy function $\mathrm{H}_2(\theta)$ is monotonically increasing over $(0, 1/2)$, thus the first inequality suggests that low representation vulnerability guarantees a small adversarial gap if we train the downstream classifier properly. On the other hand, the second inequality implies that adversarial robustness cannot be achieved for any downstream classifier, if the vulnerability of the representation it uses is too high. As discussed in Section 6.1, the connection between representation vulnerability and adversarial gap is also found to hold empirically for image classification benchmarks.

3.2. General Case

This section presents our main theoretical results regarding robust representations. First, we present the following lemma, proven in Appendix A.2, that characterizes the connection between adversarial risk and input distribution perturbations bounded in an ∞-Wasserstein ball.

Lemma 3.3. Let $(\mathcal{X}, \Delta)$ be the input metric space and $\mathcal{Y}$ be the set of labels. Assume all the examples are generated from a joint probability distribution $(X, Y) \sim \mu_{XY}$. Let μ_X be the marginal distribution of X. Then, for any classifier $f : \mathcal{X} \to \mathcal{Y}$ and $\epsilon > 0$, we have

$$\mathrm{AdvRisk}_\epsilon(f) = \sup_{\mu_{X'} \in \mathcal{B}_{\mathrm{W}_\infty}(\mu_X, \epsilon)} \Pr\left[f(X') \neq Y\right],$$

where X' denotes the random variable that follows $\mu_{X'}$.

The next theorem, proven in Appendix A.3, gives a lower bound for the adversarial risk for any downstream classifier, using the worst-case mutual information between the representation's input and output distributions.

Theorem 3.4. Let $(\mathcal{X}, \Delta)$ be the input metric space, $\mathcal{Y}$ be the set of labels and μ_{XY} be the underlying joint probability distribution. Assume the marginal distribution of labels μ_Y is a uniform distribution over $\mathcal{Y}$. Consider the feature space $\mathcal{Z}$ and the set of downstream classifiers $\mathcal{H} = \{h : \mathcal{Z} \to \mathcal{Y}\}$. Given $\epsilon \geq 0$, for any $g : \mathcal{X} \to \mathcal{Z}$, we have

$$\inf_{h\in\mathcal{H}} \mathrm{AdvRisk}_\epsilon(h \circ g) \geq 1 - \frac{\mathrm{I}(X; Z) - \mathrm{RV}_\epsilon(g) + \log 2}{\log|\mathcal{Y}|},$$

where X is the random variable that follows the marginal distribution of inputs μ_X and $Z = g(X)$.

Theorem 3.4 suggests that adversarial robustness cannot be achieved if the available representation is highly vulnerable or the standard mutual information between X and $g(X)$ is low. Note that $\mathrm{I}(X; g(X)) - \mathrm{RV}_\epsilon(g) = \inf\{\mathrm{I}(X'; g(X')) : X' \sim \mu_{X'} \in \mathcal{B}_{\mathrm{W}_\infty}(\mu_X, \epsilon)\}$, which corresponds to the worst-case mutual information between input and output of g. Therefore, if we assume robust classification as the downstream task for representation learning, then the representation having high worst-case mutual information is a necessary condition for achieving adversarial robustness for the overall classifier.

In addition, we remark that Theorem 3.4 can be extended to general p-th Wasserstein distances, if the downstream classifiers are evaluated based on robustness under distributional shift[3], instead of adversarial risk. To be more specific, if using W_p metric to define representation vulnerability, we can then establish an upper bound on the maximum distributional robustness with respect to the considered W_p metric for any downstream classifier based on similar proof techniques of Theorem 3.4.

4. Measuring Representation Vulnerability

This section presents an empirical method for estimating the vulnerability of a given representation using i.i.d. samples. Recall from Definition 3.1, for any $g : \mathcal{X} \to \mathcal{Z}$, the representation vulnerability of g with respect to the input metric probability space $(\mathcal{X}, \mu_X, \Delta)$ and $\epsilon \geq 0$ is defined as:

$$\mathrm{RV}_\epsilon(g) = \underbrace{\mathrm{I}(X; g(X))}_{J_1} - \underbrace{\inf_{\mu_{X'} \in \mathcal{B}_{\mathrm{W}_\infty}(\mu_X, \epsilon)} \mathrm{I}(X'; g(X'))}_{J_2}. \tag{4.1}$$

To measure representation vulnerability, we need to compute both terms J_1 and J_2. However, the main challenge is that we do not have the knowledge of the underlying probability distribution μ_X for real-world problem tasks. Instead, we only have access to a finite set of data points sampled from the distribution. Therefore, it is natural to consider sample-based estimator for J_1 and J_2 for practical use.

The first term J_1 is essentially the mutual information between X and $Z = g(X)$. A variety of methods have been proposed for estimating mutual information (Moon et al.,

[2] To be more specific, we do not consider the case where h is a constant function. Under our problem setting, there are two elements in $\mathcal{H}$, namely $h_1(z) = z$, $h_2(z) = -z$, for any $z \in \mathcal{Z}$.

[3] See Sinha et al. (2018) for a rigorous definition of distributional robustness.

1995; Darbellay & Vajda, 1999; Suzuki et al., 2008; Kandasamy et al., 2015; Moon et al., 2017). The most effective estimator is the mutual information neural estimator (MINE) (Belghazi et al., 2018), based on the dual representation of KL-divergence (Donsker & Varadhan, 1983):

$$\widehat{\mathrm{I}}_m(X;Z) = \sup_{\theta\in\Theta} \mathbb{E}_{\widehat{\mu}^{(m)}_{XZ}}[T_\theta] - \log\big(\mathbb{E}_{\widehat{\mu}^{(m)}_X\otimes\widehat{\mu}^{(m)}_Z}[\exp(T_\theta)]\big),$$

where $T_\theta : \mathcal{X}\times\mathcal{Z}\to\mathbb{R}$ is the function parameterized by a deep neural network with parameters $\theta\in\Theta$, and $\widehat{\mu}^{(m)}_{XZ}$, $\widehat{\mu}^{(m)}_X$ and $\widehat{\mu}^{(m)}_Z$ denote the empirical distributions[4] of random variables (X,Z), X and Z respectively, based on m samples. In addition, Belghazi et al. (2018) empirically demonstrates the superiority of the proposed estimator in terms of estimation accuracy and efficiency, and prove that it is strongly consistent: for all $\varepsilon>0$, there exists $M\in\mathbb{Z}$ such that for any $m\geq M$, $|\widehat{\mathrm{I}}_m(X;Z)-\mathrm{I}(X;Z)|\leq\varepsilon$ almost surely. Given the established effectiveness of this method, we implement MINE to estimate $\mathrm{I}(X;g(X))$ as the first step.

Compared with J_1, the second term J_2 is much more difficult to estimate, as it involves finding the worst-case perturbations on μ_X in a ∞-Wasserstein ball in terms of mutual information. As with the estimation of J_1, we only have a finite set of instances sampled from μ_X. On the other hand, due to the non-linearity and the lack of duality theory with respect to the ∞-Wasserstein distance (Champion et al., 2008), it is inherently difficult to directly solve an ∞-Wasserstein constrained optimization problem, even if we work with the empirical distribution of μ_X. To deal with the first challenge, we replace μ_X with its empirical measure $\widehat{\mu}^{(m)}_X$ based on i.i.d. samples. Then, to avoid the need to search through the whole ∞-Wasserstein ball, we restrict the search space of $\mu_{X'}$ to be the following set of empirical distributions:

$$\mathcal{A}(\mathcal{S},\epsilon) = \left\{\frac{1}{m}\sum_{i=1}^m \delta_{\boldsymbol{x}'_i} : \boldsymbol{x}'_i\in\mathcal{B}(\boldsymbol{x}_i,\epsilon)\ \forall i\in[m]\right\}, \quad (4.2)$$

where $\mathcal{S}=\{\boldsymbol{x}_i : i\in[m]\}$ denotes the given set of m data points sampled from μ_X. Note that the considered set $\mathcal{A}(\mathcal{S},\epsilon)\subseteq\mathcal{B}_{\mathrm{W}_\infty}(\widehat{\mu}^{(m)}_X,\epsilon)$, since each perturbed point $\boldsymbol{x}'_i$ is at most ϵ-away from $\boldsymbol{x}_i$. Finally, making use of the dual formulation of KL-divergence that is used in MINE, we propose the following empirical optimization problem for estimating J_2:

$$\min_{\mu_{X'}} \widehat{\mathrm{I}}_m\big(X';g(X')\big) \ \text{ s.t. } \ \mu_{X'}\in\mathcal{A}(\mathcal{S},\epsilon), \quad (4.3)$$

where we simply set the empirical distribution $\widehat{\mu}^{(m)}_{X'}$ to be the same as $\mu_{X'}$. In addition, we propose a heuristic alternating minimization algorithm to solve (4.3) (see Appendix B for the pseudocode and a complexity analysis of the proposed algorithm). More specifically, our algorithm alternatively performs gradient ascent on θ for the inner maximization problem of estimating $\widehat{\mathrm{I}}_m(X';g(X'))$ given μ'_X, and searches for the set of worst-case perturbations on $\{\boldsymbol{x}'_i : i\in[m]\}$ given θ based on projected gradient descent.

[4]Given a set of m samples $\{\boldsymbol{x}_i\}_{i\in[m]}$ from a distribution μ, we let $\widehat{\mu}^{(m)}=\frac{1}{m}\sum_{i\in[m]}\delta_{\boldsymbol{x}_i}$ be the empirical measure of μ.

5. Learning Robust Representations

In this section, we present our method for learning adversarially robust representations. First, we introduce the mutual information maximization principle for representation learning (Linsker, 1989; Bell & Sejnowski, 1995). Mathematically, given an input probability distribution μ_X and a set of representations $\mathcal{G}=\{g:\mathcal{X}\to\mathcal{Z}\}$, the maximization principle proposes to solve this optimization problem:

$$\max_{g\in\mathcal{G}} \ \mathrm{I}\big(X;g(X)\big). \quad (5.1)$$

Although this principle has been shown to be successful for learning good representations under the standard setting (Hjelm et al., 2018), it becomes ineffective when considering adversarial perturbations (see Table 1 for an illustration). Motivated by the theoretical connections between feature sensitivity and adversarial risk for downstream robust classification shown in Section 3, we stimulate robust representations by adding a regularization term based on representation vulnerability:

$$\max_{g\in\mathcal{G}} \quad \mathrm{I}(X;g(X)) - \beta\cdot\mathrm{RV}_\epsilon(g), \quad (5.2)$$

where $\beta\geq 0$ is the trade-off parameter between $\mathrm{I}(X;g(X))$ and $\mathrm{RV}_\epsilon(g)$. When $\beta=0$, (5.2) is same as the objective for learning standard representations (5.1). Increasing the value of β will produce representations with lower vulnerability, but may undesirably affect the standard mutual information $\mathrm{I}(g(X);X)$ if β is too large. In particular, we set $\beta=1$ in the following discussions, which allows us to simplify (5.2) to obtain the following optimization problem:

$$\max_{g\in\mathcal{G}} \min_{\mu_{X'}\in\mathcal{B}_{\mathrm{W}_\infty}(\mu_X,\epsilon)} \mathrm{I}\big(X';g(X')\big). \quad (5.3)$$

The proposed training principle (5.3) aims to maximize the mutual information between the representation's input and output under the worst-case input distribution perturbation bounded in a ∞-Wasserstein ball. We remark that optimization problem (5.3) aligns well with the results of Theorem 3.4, which shows the importance of the learned feature representation achieving high worst-case mutual information for a downstream robust classification task.

As with estimating the feature sensitivity in Section 4, we do not have access to the underlying μ_X. However, the inner

minimization problem is exactly the same as estimating the worst-case mutual information J_2 in (4.1), thus we can simply adapt the proposed empirical estimator (4.3) to solve (5.3). To be more specific, we reparameterize g using a neural network with parameter $\psi \in \Psi$ and use the following min-max optimization problem:

$$\max_{\psi \in \Psi} \min_{\mu_{X'} \in \mathcal{A}(S,\epsilon)} \widehat{\mathrm{I}}_m\big(X'; g_\psi(X')\big). \tag{5.4}$$

Based on the proposed algorithm for the inner minimization problem, (5.4) can be efficiently solved using a standard optimizer, such as stochastic gradient descent.

6. Experiments

This section reports on experiments to study the implications of robust representations on benchmark image datasets. Instead of focusing directly improving model robustness, our experiments focus on understanding the proposed definition of robust representations as well as its implications. Based on the proposed estimator in Section 4, Section 6.1 summarizes experiments to empirically test the relationship between representation vulnerability and model robustness, by extracting internal representations from the state-of-the-art pre-trained standard and robust classification models. In addition, we empirically evaluate the general lower bound on adversarial risk presented in Theorem 3.4. In Section 6.2, we evaluate the proposed training principle for learning robust representations on image datasets, and test its performance with comparisons to the state-of-the-art standard representation learning method in a downstream robust classification framework. We also visualize saliency maps as an intuitive criteria for evaluating representation robustness.

We conduct experiments on MNIST (LeCun & Cortes, 2010), Fashion-MNIST (Xiao et al., 2017), SVHN (Netzer et al., 2011), and CIFAR-10 (Krizhevsky et al., 2009), considering typical ℓ_∞-norm bounded adversarial perturbations for each dataset ($\epsilon = 0.3$ for MNIST, 0.1 for Fashion-MNIST, $4/255$ for SVHN, and $8/255$ for CIFAR-10). We use the PGD attack (Mądry et al., 2018) for both generating adversarial distributions in the estimation of worst-case mutual information and evaluating model robustness. To implement our proposed estimator (4.3), we adopt the *encode-and-dot-product* model architecture in Hjelm et al. (2018) and adjust it to adapt to different forms of representations. We leverage implementations from Engstrom et al. (2019a) and Hjelm et al. (2018) in our implementation[5]. Implementation details are provided in Appendix D.1.

6.1. Representation Robustness

To evaluate our proposed definition on representation vulnerability and its implications for downstream classification models, we conduct experiments on image benchmarks using various classifiers, including VGG (Simonyan & Zisserman, 2015), ResNet (He et al., 2016), DenseNet (Huang et al., 2017) and the simple convolutional neural network in Hjelm et al. (2018) denoted as Baseline-H.

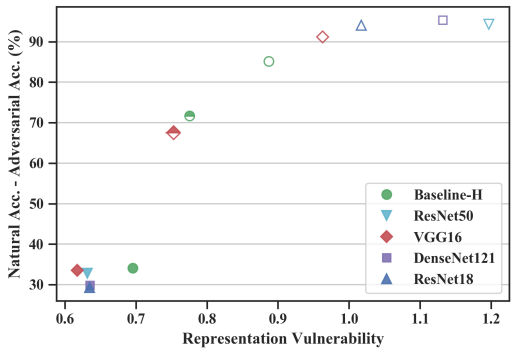

Figure 1. Correlations between the representation vulnerability and the CIFAR-10 model's natural-adversarial accuracy gap. Filled points indicate robust models (trained with $\epsilon = 8/255$), half-filled are models adversarially trained with $\epsilon = 2/255$, and unfilled points are standard models.

Correlation with model robustness. Theorem 3.2 establishes a direct correlation between our representation vulnerability definition and achievable model robustness for the synthetic Gaussian-mixture case, but we are not able to theoretically establish that correlation for arbitrary distributions. Here, we empirically test this correlation on image benchmark datasets. Figure 1 summarizes the results of these experiments for CIFAR-10, where we set the logit layer as the considered representation space. The adversarial gap decreases with decreasing representation vulnerability in an approximately consistent relationship. Models with low logit layer representation vulnerability tend to have low natural-adversarial accuracy gap, which is consistent with the intuition behind our definition and with the theoretical result on the synthetic Gaussian-mixture case. This suggests the correlation between representation vulnerability and model robustness may hold for general case.

Adversarial risk lower bound. Theorem 3.4 provides a lower bound on the adversarial risk that can be achieved by any downstream classifier as a function of representation vulnerability. To evaluate the tightness of this bound, we estimate the normal-case and worst-case mutual information $\mathrm{I}(X; g(X))$ of layer representation g for different models, and empirically evaluate the adversarial risk of the models. Figure 2 shows the results, where we again set the logit layer as the feature space for a more direct comparison. The lower bound of adversarial risk is calculated according to Theorem 3.4 and is converted to the upper bound of ad-

[5] https://github.com/schzhu/learning-adversarially-robust-representations

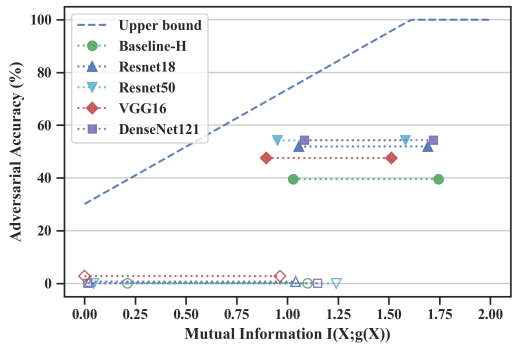

Figure 2. Normal and worst case mutual information for logit-layer representations. Each pair of points shows the result of a specific model—the left point indicates the worst case mutual information and the right for the normal mutual information. Filled points are robust models; hollow points are standard models.

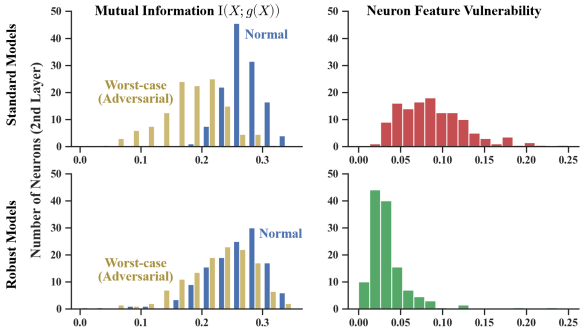

Figure 3. Distribution of mutual information $I(X; g(X))$ and feature vulnerability in the second convolutional layer of Baseline-H. The upper plots are for standard models, and the lower plots are for robust models. The total number of neurons is 128.

versarial accuracy for reference. In particular, for standard models, both the estimated worst-case mutual information and the adversarial accuracy are close to zero, whereas the computed upper bounds on adversarial accuracy are around 30%. We empirically observed around 50% adversarial accuracy for robust models, whereas the bounds computed using the estimated worst-case mutual information and Theorem 3.4 are about 75%. This shows that Theorem 3.4 gives a reasonably tight bound for a model's adversarial accuracy with respect to the logit-layer representation robustness.

Figure 2 also indicates that even the robust models produced by adversarial training have representations that are not sufficiently robust to enable robust downstream classifications. For example, robust DenseNet121 in our evaluations has the highest logit layer worst-case mutual information of 1.08, yet the corresponding adversarial accuracy is upper bounded by 77.0% which is unsatisfactory for CIFAR-10. Such information theoretic limitation also justifies our training principle of worst-case mutual information maximization, since on the other hand the adversarial accuracy upper bound calculated by normal-case mutual information does not constitute a limitation for most robust models in our experiments (as in Figure 1, most robust models achieve adversarial accuracy close to 100%).

Internal feature robustness. We further investigate the implications of our proposed definition from the level of individual features. Specifically for neural networks, we consider the function from the input to each individual neuron within a layer as a feature. The motivations for considering feature robustness comes from the fact that mutual information in terms of the whole representation is controlled by the sum of all the features' mutual information (see Appendix C for a rigorous argument) and robust features are potentially easier to train (Garg et al., 2018). As an illustration, we evaluate the robustness of all the convolutional kernels in the second layer of the Baseline-H model. Each neuron evaluated here is a composite convolutional kernel (all kernels in the first layer connected to a second layer kernel) with image input size 10×10. Figure 3 shows the results that are averaged over two independently trained models for each type. This result reveals the apparent difference in feature robustness between a standard model and the adversarially-trained robust model, even in lower layers. Although in this case the result does not prohibit a robust downstream model for lower layers neurons, for neurons in higher layers the difference becomes more distinct and the vulnerability of neurons can thus be the bottleneck of achieving high model robustness. The different feature robustness according to our definition also coincide with the saliency maps of features (see Figure 5 in Appendix D.2), where the saliency maps of robust features are apparently more interpretable compared to those of standard features.

6.2. Learning Robust Representations

Our worst-case mutual information maximization training principle provides an unsupervised way to learn adversarially robust representations. Since there are no established ways to measure the robustness of a representation, empirically testing the robustness of representations learned by our training principle poses a dilemma. To avoid circular reasoning, we evaluate the learned representations by running a series of downstream adversarial classification tasks and comparing the performance of the best models we are able to find for each representation. In addition, recent work shows that the interpretability of saliency map has certain connections with robustness (Etmann et al., 2019; Ilyas et al., 2019), thus we study the saliency map as an alternative criteria for evaluating robust representations.

The unsupervised representation learning approach based on mutual information maximization principle in Hjelm

Representation (g)	Classifier (h)	MLP h		Linear h	
		Natural	Adversarial	Natural	Adversarial
Hjelm et al. (2018)	Standard	58.77 ± 0.22	0.22 ± 0.08	47.01 ± 0.53	0.15 ± 0.03
Hjelm et al. (2018)	Robust	29.75 ± 1.49	15.08 ± 0.63	22.79 ± 1.42	10.28 ± 0.52
Ours	Standard	$\mathbf{62.54 \pm 0.12}$	14.06 ± 0.69	$\mathbf{50.29 \pm 0.58}$	10.98 ± 0.49
Ours	Standard (E.S.)	51.59 ± 3.34	27.53 ± 0.81	48.55 ± 0.63	13.52 ± 0.16
Ours	Robust	52.34 ± 0.17	$\mathbf{31.52 \pm 0.31}$	43.55 ± 0.10	$\mathbf{25.15 \pm 0.10}$
Fully-Supervised Standard		86.33 ± 0.17	0.07 ± 0.02	86.36 ± 0.13	0.02 ± 0.01
Fully-Supervised Robust		70.71 ± 0.58	40.50 ± 0.27	72.44 ± 0.59	39.98 ± 0.16

Table 1. Comparisons of different representation learning methods on CIFAR-10 in downstream classification settings. *E.S.* denotes early stopping under the criterion of the best adversarial accuracy. We present mean accuracy and the standard deviation over 4 repeated trials.

et al. (2018) achieves the state-of-the-art results in many downstream tasks, including standard classification. We further adopt their encoder architecture in our implementation, and extend their evaluation settings to adversarially robust classification. Specifically, we truncate the front part of Baseline-H with a 64-dimensional latent layer output as the representation g and train it by the worst-case mutual information maximization principle using only unlabeled data (removing the labels from the normal training data). We test two architectures (two-layer multilayer perceptron and linear classifier) for implementing the downstream classifier h and train it using labeled data after the encoder g has been trained using unlabeled data. Appendix D.1 provides additional details on the experimental setup.

Downstream classification tasks. Comparison results on CIFAR-10 are demonstrated in Table 1 (see Appendix D.2 for a similar results for MNIST, Fashion-MNIST, and SVHN). The fully-supervised models are trained for reference, from which we can see the simple model architecture we use achieves a decent natural accuracy of 86.3%; the adversarially-trained robust model reduces accuracy to around 70% with adversarial accuracy of 40.5%. The baseline, with g and h both trained normally, resembles the setting in Hjelm et al. (2018) and achieves a natural accuracy of 58.8%. For representations learned using worst-case mutual information maximization, the composition with standard two-layer multilayer perceptron (MLP) h achieves a non-trivial (compared to the 0.2% for the standard representation) adversarial accuracy of 14.1%. When h is further trained using adversarial training, the robust accuracy increases to 31.5% which is comparable to the result of the robust fully-supervised model. As an ablation, the robust h based on standard g achieves an adversarial accuracy of 15.1%, yet the natural accuracy severely drops below 30%, indicating that a robust classifier cannot be found using the vulnerable representation. The case where h is a simple linear classifier shows similar results. These comparisons show that the representation learned using worst-case mutual information maximization can make the downstream classification more robust over the baseline and approaches the robustness of fully-supervised adversarial training. This provides evidence that our training principle produces adversarially robust representations.

Another interesting implication given by results in Table 1 is that robustly learned representations may also have better natural accuracy (62.5%) over the standard representation (58.8%) in downstream classification tasks on CIFAR-10. This matches our experiments in Figure 2 where logit layer representations in robust models conveys more normal-case mutual information (up to 1.75) than those in standard models (up to 1.25). However, this is not the case on MNIST dataset as in Table 3. We conjecture that this is because the information conveyed by robust representations has better generalizations, and the generalization is more of a problem on CIFAR-10 than on MNIST (Schmidt et al., 2018).

Saliency maps. A saliency map is commonly defined as the gradient of a model's loss with respect to the model's input (Etmann et al., 2019). For a classification model, it intuitively illustrates what the model looks for in changing its classification decision for a given sample. Recent work (Etmann et al., 2019; Ilyas et al., 2019) indicates, at least in some synthetic settings, that the more alignment the saliency map has with the input image, the more adversarially robust the model is. As an additional test of representation robustness, we calculate the saliency maps of standard and robust representations g by the mutual information maximization loss with respect to the input. Figure 4 shows that the saliency maps of the robust representation appear to be much less noisy and more interpretable in terms of the alignment with original images. Intuitively, this shows that robust representations capture relatively higher level visual concepts instead of pixel-level statistical clues (Engstrom et al., 2019b). The more interpretable saliency maps of representation learned by our training principle further support its effectiveness in learning adversarially robust representation.

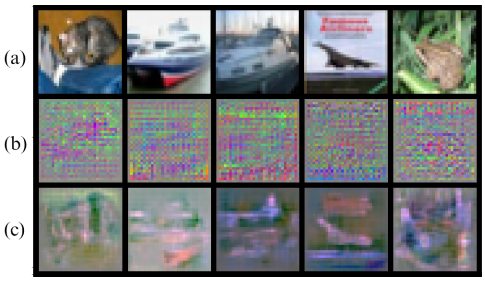

Figure 4. Visualization of saliency maps of different models on CIFAR-10: (a) original images (b) representations learned using Hjelm et al. (2018) (c) representations learned using our method.

7. Conclusion

We proposed a novel definition of representation robustness based on the worst-case mutual information, and showed both theoretical and empirical connections between our definition and model robustness for a downstream classification task. In addition, by developing estimation and training methods for representation robustness, we demonstrated the connection and the usefulness of the proposed method on benchmark datasets. Our results are not enough to produce strongly robust models, but they provide a new approach for understanding and measuring achievable adversarial robustness at the level of representations.

Acknowledgements

This work was partially funded by an award from the National Science Foundation SaTC program (Center for Trustworthy Machine Learning, #1804603) and additional support from Amazon, Baidu, and Intel.

References

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Hjelm, R. D., and Courville, A. C. Mutual information neural estimation. In *International Conference on Learning Representations*, 2018.

Bell, A. J. and Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, 2019.

Champion, T., De Pascale, L., and Juutinen, P. The ∞-wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM Journal on Mathematical Analysis*, 40(1):1–20, 2008.

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

Darbellay, G. A. and Vajda, I. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Donsker, M. D. and Varadhan, S. S. Asymptotic evaluation of certain markov process expectations for large time, IV. *Communications on Pure and Applied Mathematics*, 36 (2):183–212, 1983.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., and Mądry, A. A rotation and a translation suffice: Fooling CNNs with simple transformations. *arXiv:1712.02779*, 2017.

Engstrom, L., Ilyas, A., Santurkar, S., and Tsipras, D. Robustness (python library), 2019a. URL https://github.com/MadryLab/robustness.

Engstrom, L., Ilyas, A., Santurkar, S., Tsipras, D., Tran, B., and Mądry, A. Adversarial robustness as a prior for learned representations. *arXiv:1906.00945*, 2019b.

Etmann, C., Lunz, S., Maass, P., and Schoenlieb, C. On the connection between adversarial robustness and saliency map interpretability. In *International Conference on Machine Learning*, 2019.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning visual classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Eykholt, K., Gupta, S., Prakash, A., and Zheng, H. Robust classification using robust feature augmentation. *arXiv:1905.10904*, 2019.

Fawzi, A., Fawzi, H., and Fawzi, O. Adversarial vulnerability for any classifier. In *Advances in Neural Information Processing Systems*, 2018.

Garg, S., Sharan, V., Zhang, B., and Valiant, G. A spectral view of adversarially robust features. In *Advances in Neural Information Processing Systems*, 2018.

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. Adversarial spheres. *arXiv:1801.02774*, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2018.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Mądry, A. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, 2019.

Kandasamy, K., Krishnamurthy, A., Poczos, B., Wasserman, L., et al. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems*, 2015.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

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

Linsker, R. How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation*, 1(3):402–411, 1989.

Mahloujifar, S., Diochnos, D. I., and Mahmoody, M. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *AAAI Conference on Artificial Intelligence*, 2019.

Mądry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Moon, K. R., Sricharan, K., and Hero, A. O. Ensemble estimation of mutual information. In *IEEE International Symposium on Information Theory*. IEEE, 2017.

Moon, Y.-I., Rajagopalan, B., and Lall, U. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, 1995.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Pensia, A., Jog, V., and Loh, P.-L. Extracting robust and accurate features via a robust information bottleneck. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Scarlett, J. and Cevher, V. An introductory guide to Fano's inequality with applications in statistical estimation. *arXiv:1901.00555*, 2019.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Mądry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, 2018.

Shafahi, A., Huang, W. R., Studer, C., Feizi, S., and Goldstein, T. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2018.

Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *ACM Conference on Computer and Communications Security*, 2016.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Suzuki, T., Sugiyama, M., Sese, J., and Kanamori, T. Approximating mutual information by maximum likelihood density ratio estimation. In *New challenges for feature selection in data mining and knowledge discovery*, 2008.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Learning Representations*, 2019.