Adversarial Training Reduces Information and Improves Transferability

Matteo Terzi[1], Alessandro Achille[2], Marco Maggipinto[1], Gian Antonio Susto[1]

[1]Department of Information Engineering, University of Padova
[2]AWS
{terzimat,marco.maggipinto}@dei.unipd.it, gianantonio.susto@unipd.it, achille@cs.ucla.edu

Abstract

Recent results show that features of adversarially trained networks for classification, in addition to being robust, enable desirable properties such as invertibility. The latter property may seem counter-intuitive as it is widely accepted by the community that classification models should only capture the minimal information (features) required for the task. Motivated by this discrepancy, we investigate the dual relationship between Adversarial Training and Information Theory. We show that the Adversarial Training can improve linear transferability to new tasks, from which arises a new trade-off between transferability of representations and accuracy on the source task. We validate our results employing robust networks trained on CIFAR-10, CIFAR-100 and ImageNet on several datasets. Moreover, we show that Adversarial Training reduces Fisher information of representations about the input and of the weights about the task, and we provide a theoretical argument which explains the invertibility of deterministic networks without violating the principle of minimality. Finally, we leverage our theoretical insights to remarkably improve the quality of reconstructed images through inversion.

Introduction

In the last 10 years, Deep Neural Networks (DNNs) dramatically improved the performance in any computer vision task. However, the impressive accuracy comes at the cost of poor robustness to small perturbations, called *adversarial perturbations*, that lead the models to predict, with high confidence, a wrong class (Goodfellow, Shlens, and Szegedy 2014; Terzi, Susto, and Chaudhari 2020). This undesirable behaviour led to a flourishing of research works ensuring robustness against them. State-of-the-art approaches for robustness are provided by Adversarial Training (AT) (Madry et al. 2017) and its variants (Zhang et al. 2019). The rationale of these approaches is to find worst-case examples and feed them to the model during training or constraining the output to not change significantly under small perturbations. However, robustness is achieved at the expense of a decrease in accuracy: the more a model is robust, the lower its accuracy will be (Tsipras et al. 2019). This is a classic "waterbed effect" between precision and robustness ubiquitous in optimal control and many other fields. Interestingly, robustness is not the only desiderata of adversarially trained models: their representations are semantically meaningful and they can be used for other Computer Vision (CV) tasks, such as generation and (semantic) interpolation of images. More importantly, AT enables invertibility, that is the ability to reconstruct input images from their representations (Ilyas et al. 2019) by solving a simple optimization problem. This is true also for *out-of-distribution* images meaning that robust networks do not *destroy* information about the input. Hence, how can we explain that, while robust networks preserve information, they lack in generalization power?

In this context, obtaining good representations for a task has been the subject of *representation learning* where the most widely accepted theory is Information Bottleneck (IB) (Tishby, Pereira, and Bialek 2000; Alemi et al. 2016; Achille and Soatto 2018b) which calls for reducing information in the activations, arguing it is necessary for generalization. More formally, let x be an input random variable and y be a target random variable, a good representation z of the input should be maximally expressive about for y while being as concise as possible about x. The solution of the optimal trade-off can be found by optimizing the Information Lagrangian:

$$\min_z -I(z,y) + \beta I(z,x)$$

where β controls how much information about x is conveyed by z. Both AT and IB at their core aim at finding good representations: the first calls for representations that are *robust* to input perturbations while the latter finds *minimal* representations sufficient for the task. How are these two methods related? Do they share some properties? More precisely, does the invertibility property create a contradiction on IB theory? In fact, if generalization requires discarding information in the data that is not necessary for the task, it should not be possible to reconstruct the input images.

Throughout this paper we will (i) investigate the research questions stated above, with particular focus on the connection between IB and AT and as a consequence of our analysis, (ii) we will reveal new interesting properties of robust models.

Contributions and Related Works

A fundamental result of IB is that, in order to generalize well on a task, z has to be sufficient and minimal, that is, it should contain only the information necessary to predict y, which in our case is a target class. Apparently, this is in contradiction with the evidence that robust DNNs are invertible maintaining almost all the information about the input x even if is *not* necessary for the task. However, what matters for generalization is not the information in the activations, but information in the weights (PAC-Bayes bounds) (Achille and Soatto 2019). Reducing information in the weights, yields to reduction in the *effective* information in the activations at test time. Differently from IB theory, (Achille and Soatto 2019) claims that the network does not need to destroy information in the data that is not needed for the task: it simply needs to make it inaccessible to the classifier, but otherwise can leave it lingering in the weights. That is the case for ordinary learning. As for AT, robustness is obtained at cost of lower accuracy on natural images (Madry et al. 2017; Tsipras et al. 2019), suggesting that only the robust features are extracted by the model (Ilyas et al. 2019): How can be this conciliated with invertibility of robust models? This paper shows that, while AT *preserves* information about the data that is irrelevant for the task in the weights (to the point where the resulting model is invertible), the information that is *effectively* used by the classifier does not contain all the details about the input x. In other words, the network is not effectively invertible: what really matters is the accessible information stored in the weights. In order to visualize this fact, we will introduce *effective images*, that are images that represent what the classifier "sees". Inverting learned representations is not new, and it was solved in (Mahendran and Vedaldi 2015; Yosinski et al. 2015; Ulyanov, Vedaldi, and Lempitsky 2018; Kingma and Welling 2013); however, these methods either inject external information through priors or explicitly impose the task of reconstruction contrary to robust models.

The main contribution of this work can be summarized as follows. If representations contain all the information about the input x, then adversarially trained models should be better at transfering features to different tasks, where aspects of the data that were irrelevant to the task it was (pre)-trained on were neither destroyed nor ignored, but preserved. To test this hypothesis, we perform linear classification (finetune the last layer) for different tasks. We show that AT improves linear transferability of deep learning models across diverse tasks which are *sufficiently different* from the source task/dataset. Specifically, the farther two tasks are (as measured by a task distance), the higher the performance improvement that can be achieved by training a linear classifier using an adversarially-trained model (feature, or backbone) compared to an ordinarily trained model. Related to this, in (Shafahi et al. 2019) the transferability of robustness to new tasks is studied experimentally; differently, in the present work we study the linear transferability of natural accuracy. Moreover, we also analytically show that, confirming empirical evidence (Ilyas et al. 2019), once we extract robust features from a backbone model, *all* the models using these features have to be robust.

We will also show that adversarial regularization is a lower-bound of the regularizer in the Information Lagrangian, so AT in general results in a loss of accuracy for the task at hand. The benefit is increased transferability, thus showing a classical tradeoff of robustness (and its consequent transferability) and accuracy on the task for which it is trained. This is a classic "waterbed effect" between precision and robustness ubuiquitous in optimal control. Regarding the connection with IB, we show analytically that AT reduces the effective information in the activations about the input, as defined by (Achille and Soatto 2019). Moreover, we show empirically that adversarial training also reduces information in the weights and its consequences.

Finally, we show that injecting effective noise once during the inversion process dramatically improves reconstruction of images in term of convergence and quality of fit.

In order to facilitate the reading, the manuscript is organized as follows. The section "Preliminaries and Notation" provides the necessary notation,"AT Reduces Information" presents all the theoretical building blocks by showing the connection between AT and IB. Based on the previous results, the section "Does Invertibility Contradict IB?" shows why there is no contradiction between minimality of representations and invertibility of robust models, and "Transferability-accuracy Trade-Off" shows that robust features can transfer better to new tasks.

Preliminaries and Notation

We introduce here the notation used in this paper. We denote a dataset of N samples with $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ where $x \in X$ is an input, and $y \in Y$ is the target class in the finite set $Y = \{1, \dots, K\}$. More in general, we refer to y as a random variable defining the "task". In this paper we focus on classification problems using cross-entropy $L_{\mathcal{D}}(w) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(x, y; w)]$ on the training set $\mathcal{D}$ as objective where $\ell(x, y; w) = -\log p_w(y|x)$ and $p_w(y|x)$ is encoded by a DNN. The loss $L_{\mathcal{D}}(w)$ is usually minimized using *stochastic gradient descent* (SGD) (Bottou, Curtis, and Nocedal 2018), which updates the weights w with a noisy estimate of the gradient computed from a mini-batch of samples. Thus, weights update can be expressed by a stochastic diffusion process with non-isotropic noise (Li, Tai et al. 2017). In order to measure the (asymmetric) dissimilarity between distributions we use the Kullbach-Liebler divergence between $p(x)$ and $q(x)$ given by $\mathrm{KL}(\, p(x) \,\|\, q(x) \,) = \mathbb{E}_{x\sim p(x)} \big[\log(p(x)/q(x)) \big]$. It is well-known that the second order approximation of the KL-divergence is $\mathbb{E}_x \mathrm{KL}(\, p_w(y|x) \,\|\, p_{w+\delta w}(y|x) \,) = \delta w^t F \delta w + o(\|\delta w\|^2)$ where F is the *Fisher Information Matrix* (FIM), defined by $F = \mathbb{E}_{x,y\sim p(x)p_w(y|x)}[\nabla \log p_w(y|x) \nabla \log p_w(y|x)^t] = \mathbb{E}_{x\sim p(x)p_w(y|x)}[-\nabla_w^2 \log p_w(y|x)]$. The FIM gives a local measure of how much a perturbation δw on parameters w, will change $p_w(y|x)$ with respect to KL divergence (Martens 2014). Finally, let x and z be two random variables. The *Shannon mutual information* is defined as $I(x; z) = \mathbb{E}_{x\sim p(x)}[\mathrm{KL}(\, p(z|x) \,\|\, p(z) \,)]$. Throughout this paper, we indicate the *representations* before the linear layer as $z = f_w(x)$, where $f_w(x)$ is called *feature extractor*.

Adversarial Training

AT aims at solving the following min-max problem:

$$\begin{cases} \min_w \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(x^\star, y; w)] \\ \delta^\star = \operatorname{argmax}_{\|\delta\|_2 < \varepsilon} \ell(x+\delta, y; w) \\ x^\star = x + \delta^\star \end{cases} \tag{1}$$

In the following we denote $\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(x^\star, y; w)]$ with $L^\star_{\mathcal{D}}(w)$. We remark that by $\mathbb{E}_{(x,y)\sim\mathcal{D}}$ we mean the empirical expectation over N elements of the dataset. Intuitively, the objective of AT is to ensure stability to small perturbations on the input. With cross-entropy loss this amounts to require that $\mathrm{KL}(\, p_w(x+\delta) \,\|\, p_w(x)\,) \le \gamma$, with γ small. Depending on ε, we can write Equation (1) as:

$$\begin{aligned} &\min_w \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(x, y; w)] + \\ &\beta \max_{\|\delta\|_2 \le \varepsilon} \mathrm{KL}(\, p_w(y|x+\delta) \,\|\, p_w(y|x)\,) \end{aligned} \tag{2}$$

which is the formulation introduced in (Zhang et al. 2019) when using cross-entropy loss. We define the (weak) inversion of features as:

Definition 0.1 (Inversion)**.** *Let $\bar{z} = f_w(x)$ be the final representation (before linear classifier) of an image x, and let f_w be the robust feature extractor. The reconstructed image (inversion) is the solution of the following problem:*

$$\hat{x}(w; z) = f_w^{-1}(\bar{z}) = \operatorname*{argmin}_{x'} \|\bar{z} - f_w(x')\|_2 \tag{3}$$

where the initial condition of x' is white noise $x'(0) \sim N(0.5, \sigma)$, where σ is the noise scale.

AT Reduces Information

In this section, we analytically show why a robust network, even if it is invertible at test time, is effectively not invertible as a consequence of noise injected by SGD. We first define the Fisher $F_{z|x}$ of representations w.r.t. inputs.

Definition 0.2. *The FIM $F_{z|x}$ of representations w.r.t the input distribution is defined as:*

$$\begin{aligned} F_{z|x} &= \mathbb{E}_{x\sim p(x)} \mathbb{E}_{z\sim p_w(z|x)} \nabla_x \log p_w(z|x) \nabla_x \log p_w(z|x)^t \\ &= \mathbb{E}_{x\sim p(x)} S(z|x) \end{aligned} \tag{4}$$

where $S(z|x)$ is the sensitivity matrix of the model at a fixed input location x.

In the next proposition we relate AT to Equation (4), showing that, requiring the stability of f_w w.r.t. x is equivalent to regularize the FIM $F_{z|x}$.

Proposition 0.3. *Let $\delta \in X$ be a small perturbation such that $\|\delta\|_2 = \varepsilon$.*[1] *Then,*

$$\max_{\|\delta\|_2} \mathrm{KL}(\, p_w(z|x+\delta) \,\|\, p_w(z|x)\,) \approx \frac{\varepsilon^2}{2} v^t_{\lambda_1} S(z|x) v_{\lambda_1} \tag{5}$$

where v_{λ_1} is the (unit-norm) eigen-vector corresponding to the first principal eigenvalue λ_1.

Hence, AT is equivalent to regularize the Fisher of representation z with respect to inputs x. By applying white Gaussian noise instead of adversarial noise, Equation (5) would become $\mathrm{KL}(\, p_w(z|x+\delta) \,\|\, p_w(z|x)\,) \approx \frac{\varepsilon^2}{2n} \operatorname{tr} S(z|x)$, where n is the input dimension. It is easy to see that $\frac{\operatorname{tr} S(z|x)}{n} \le v^t_{\lambda_1} S(z|x) v_{\lambda_1}$, meaning that Gaussian Noise Regularization (GNR) is upper bounded by AT: the inefficiency of GNR increases as the input dimension increases, causing that many directions preserve high curvature. (Tsipras et al. 2019) showed that AT, for a linear classification problem with hinge loss, is equivalent to penalize the ℓ_2-norm of weights. The next example shows that when using cross-entropy loss, penalizing the Fisher $F_{z|x}$ yields a similar result.

Example 0.4 (Binary classification)**.** Assume a binary classification problem where $y \in \{-1, 1\}$ Let $p(y=1|x) = 1 - p(y=-1|x) = \mathtt{sigmoid}(w^t x)$. Then we have:

$$F_{z|x} = cww^t \ , \ \operatorname{tr}(F_{z|x}) = c\,\|w\|_2^2 \, , \ c = \mathbb{E}_x[p(-1|x)p(1|x)]$$

The previous example may suggest that with ℓ_2-perturbations AT may reduce the l_2-norm of the weights. We trained robust models with different ε (with the same seed) to verify this claim: as reported in Figure 5, we discovered that it is true only for $\varepsilon > 1$, pointing out that there may exist two different regimes.

What we are interested in is the relation between the Shannon Mutual Information $I(z, x)$ and the Fisher Information in the activations $F_{z|x}$. However, in adversarial training there is nothing that is stochastic but SGD. For this reason, (Achille and Soatto 2019) introduced *effective information*. The idea under this definition is that, even though the network is deterministic at the end of training, what matters is the noise that SGD injects to the classifier. Thus, the effective information is a measure of the information that the network effectively uses in order to classify. Before continuing, we need to quantify this noise applied to weights.

Definition 0.5 (Information in the Weights)**.** *The complexity of the task $\mathcal{D}$ at level β, using the posterior $Q(w|\mathcal{D})$ and the prior $P(w)$, is*

$$\begin{aligned} C_\beta(\mathcal{D}; P, Q) = \mathbb{E}_{w\sim Q(w|\mathcal{D})}[L_{\mathcal{D}}(p_w(y|x))] + \\ \beta \underbrace{\mathrm{KL}(\, Q(w|\mathcal{D}) \,\|\, P(w)\,)}_{\text{Information in the Weights}}, \end{aligned} \tag{6}$$

where $\mathbb{E}_{w\sim Q(w|\mathcal{D})}[L_{\mathcal{D}}(p_w(y|x))]$ is the (expected) reconstruction error of the label under the "noisy" weight distribution $Q(w|\mathcal{D})$; $\mathrm{KL}(\, Q(w|\mathcal{D}) \,\|\, P(w)\,)$ measures the entropy of $Q(w|\mathcal{D})$ relative to the prior $P(w)$. If $Q^*(w|\mathcal{D})$ minimizes Equation (6) for a given β, we call $\mathrm{KL}(\, Q^*(w|\mathcal{D}) \,\|\, P(w)\,)$ the Information in the Weights *for the task $\mathcal{D}$ at level β.*

Given the prior $P(w) \sim N(0, \lambda^2 I)$, the solution of the optimal trade-off is given by the distribution $Q(w|\mathcal{D}) \sim N(w^\star, \Sigma^\star)$ such that $\Sigma^\star = \frac{\beta}{2}\left(F_w + \frac{\beta}{2\lambda^2} I\right)^{-1}$ with $F_w \approx$

[1] We would like to note that the practical implementation only requires $\|\delta\|_2 \le \varepsilon$. However, in practice, it is possible to see that for small ε, the norm of δ is almost always ε.

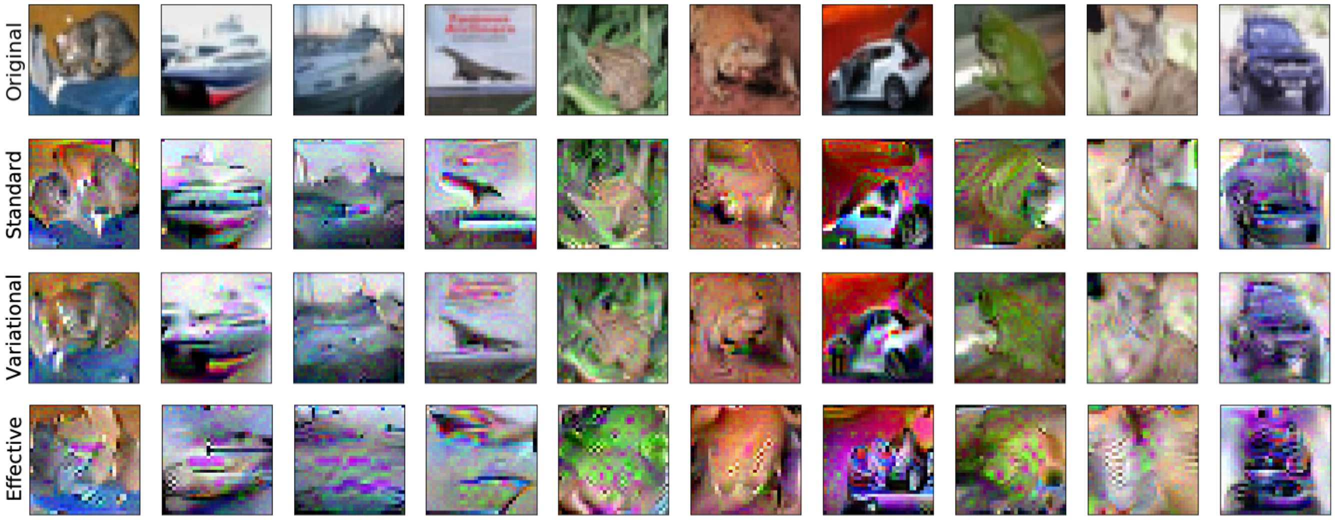


Figure 1: Inversion using the standard and variational ResNet-50 model. (Top row) Original images. (Second row) Images reconstructed optimizing Equation (3). (Third row) Images reconstructed by adding noise only once. (Bottom row) Effective images obtained optimizing Equation (10).

$\nabla_w^2 L_{\mathcal{D}}(w)$. The previous definition tells us that if we perturb uninformative weights, the loss is only slightly perturbed. This means that information in the activations that is not preserved by such perturbations is not used by the classifier.

Definition 0.6. *(Effective Information in the Activations (Achille and Soatto 2019)). Let w be the weights, and let $n \sim N(0, \Sigma_w^*)$, with $\Sigma_w^* = \beta F^{-1}(w)$ be the optimal Gaussian noise minimizing Equation* (6) *at level β for a prior $N(0, \lambda^2 I)$. We call* effective information *(at noise level β) the amount of information about x that is not destroyed by the added noise:*

$$I_{\mathit{eff},\beta}(x; z) = I(x; z_n), \tag{7}$$

where $z_n = f_{w+n}(x)$ are the activations computed by the perturbed weights $w + n \sim N(w, \Sigma_w^*)$.

By Prop. 4.2(i) in (Achille and Soatto 2019) we have that the relation between $F_{z|x}$ and effective information is given by:

$$I_{\text{eff},\beta}(x; z) \approx H(x) - \mathbb{E}_x\Big[\frac{1}{2}\log\Big(\frac{(2\pi e)^k}{|F_{z|x}|}\Big)\Big], \tag{8}$$

where $H(x)$ is the entropy of input distribution. Equation (8) shows that AT compresses data similarly to IB. With AT, the noise is injected in the input x and *not* only in the weights. In order to reduce the *effective* information that the representations have about the input (relative to the task), it is sufficient to decrease $|F_{z|x}|$, that is, increasing ε. In the Supplementary Material, we show how details about x are discarded varying ε.

AT reduces the information in the weights We showed that AT reduces effective information about x in the *activation*. However, (Achille and Soatto 2019) showed that to have guarantees about generalization and invariance to nuisances at test time one has to control the trade off between sufficiency for the task and information the weights have about the dataset. A natural question to ask is whether reducing information in the activations implies reducing information in the weights, that is the mutual information $\beta I(w; D)$ between the weights and the dataset. The connection between weights and activation is given by the following formula (Achille and Soatto 2019):

$$F_{z|x} = \frac{1}{\beta}\nabla_x f_w \cdot J_f F_w J_f^t \, \nabla_x f_w \tag{9}$$

where $\nabla_x f_w(x)$ is the Jacobian of the representation given the input, and $J_f(x)$ is the Jacobian of the representation with respect to the weights. Decreasing the Fisher Information that the weights contain about the training set decreases the effective information between inputs and activations. However, the vice-versa may not be true in general. In fact, it is sufficient that $\|\nabla_x f_w\|$ decreases. Indeed, this fact was used in several works to enhance model robustness (Virmaux and Scaman 2018; Fazlyab et al. 2019). However, as we show in Figure 4, AT reduces information in the features as the embedding defined by $|F_w^{-1}|$, that is, the log-variance of parameters is increased when increasing the ε applied on training. Experiments are done with a ResNet-18 on CIFAR-10. Interestingly, this provides the evidence that it is possible to achieve robustness without reducing $\|\nabla_x f_w\|$.

Does Invertibility Contradict IB?

Robust representations are (almost) invertible, even for out-of-distribution data (Engstrom et al. 2019). Figure 1 shows examples of inversions using Equation (3). However, past literature claims that a classifier should store only information useful for the task. This is even more surprising as robust features should discard useful details more than standard models. This fact empirically proves that it is not necessary to remove information about the input to generalize well (Behrmann et al. 2018). Moreover, when f is an invertible map, the Shannon information $I(x,z)$ is infinite. So, how can invertibility and minimality of representations be conciliated? Where is the excess of information which explains the gap? As shown in the section "AT Reduces Information", the main problem of standard IB, is that it requires to operate in the activations during training and there is no guarantee that information is also reduced at test time, which is not as AT shows. The crucial point shown in (Achille and Soatto 2019) and in the previous sections, is that it is still possible to maintain information about input at test time while making the information inaccessible for the classifier. Moreover, an important result in this paper, is that it is possible to visualize the images that are effectively "seen" by the classifier in computing the prediction. By leveraging Definition 0.6, we define the *effective image*.

Definition 0.7 (Effective image)**.** *Let $\bar{z} = f_w(x)$, and let f_w be the model trained with $\|\delta\|_2 \leq \varepsilon$. We define effective image $x_{\mathit{eff},\varepsilon}$ at level ε, the solution of the following problem:*

$$x_{\mathit{eff},\varepsilon}(x;z) = \operatorname*{argmin}_{x'} \|f_{w+n}(x) - f_w(x')\|_2 \qquad (10)$$

where $n \sim N(w, \Sigma^\star)$ and $\Sigma^\star = \beta F^{-1}(w)$.

The idea under effective images is to simulate the training conditions by artificially injecting the noise that approximates SGD. In this manner we can visualize how AT controls the conveyed information. In Figure 1 we show some examples. Interestingly, robust features are not always good features: in fact, due to the poor diversity of the dataset (CIFAR-10), the feature *color green* is highly correlated with class *frog*.

Adding effective noise (once) improves inversion The quality of inversion depends on the capability of gradient flow to reach the target representation $\hat{z}$. Starting from regions that are distant from training and test points f_w may be less smooth. Intuitively, especially during the first phase of optimization, it can be beneficial to inject noise to escape from local minima. Surprisingly, we discover that by injecting effective noise once, reconstruction is much faster and the quality of images improves dramatically. At the beginning of optimization, we perturb weights with $\bar{n} \sim N(0, \Sigma^\star)$ and solve the inversion with $f_{w+\bar{n}}$. By visually comparing row 2 and 3 of Figure 1, it is easy to see that injecting noise as described above, improves the quality of reconstruction. In support of this, in Figure 2 we numerically assess the quality of representations using the loss $L_{inv}(x,z)$. The variational model, besides improving quality of fit, also allows fast convergence: convergence is achieved after roughly 200 iterations while the deterministic model converges after 8k iterations ($\sim 40\times$).

Transferability-accuracy Trade-Off

The insights from the previous sections motivate the following argument: if in robust models information is still there, is it possible that features not useful for the original task y_1 are useful for other tasks? In a sense, z is a well-organized semantic compression of x such that it approximately allows to linearly solve the new task $y_2|z$. How well the task y_2 is solved depends on how z is organized. In fact, even though z is optimal for y_1 and for reconstructing x, it still could be not optimal for y_2. This intuition suggests that having robust features z is more beneficial than having a standard model when the distance $d(y_2, y_1)$ between tasks y_2 and y_1 is such that features from the source models are not easily adaptable to the new task. Thus, there may exist a trade-off between accuracy on a given task and stability to distributions changes: "locally", standard models work better as feature extractor, but globally this may not be true. In order to test our hypothesis, we (i) analyze the structure of representations extracted from adversarially-trained models, (ii) provide a theoretical motivation and (iii) experimentally confirm the theory by showing the emergence of a trade-off in transferability.

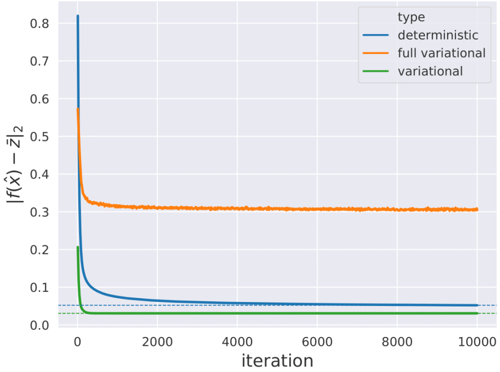


Figure 2: Comparison of $L_{inv}(x,z)$ of (orange) Effective images, (green) variational and (blue) deterministic models.

Recently, (Frosst, Papernot, and Hinton 2019) showed that more entangled features, that is more class-independent, allow for better generalization and robustness. In order to understand the effect of AT, in Figure 6 we show the t-SNE (Maaten and Hinton 2008) embedding of final representations for different values of ε: as ε increases, the entanglement increases at the expenses of less discriminative features. Thus, robust models capture more high-level features instead of the ones useful only for the task at hand.

Effective Transferable Information

Interestingly, Fisher Information theory presented in the section "AT Reduces Information" can be applied even to provide an theoretical intuition about transferability of robust models.

Since AT reduces $F_{w|D}$, it reduces the information that the network has about the dataset $\mathcal{D}$. In fact:

$$I(w;\mathcal{D}) \approx H(w) - \mathbb{E}_{\mathcal{D}}\Big[\frac{1}{2}\log\Big(\frac{(2\pi e)^k}{|F_{w|\mathcal{D}}|}\Big)\Big], \qquad (11)$$

where $F_{w|\mathcal{D}} \approx \nabla_{\mathcal{D}} w^t F_w \nabla_{\mathcal{D}} w$. From the previous proposition we can see that there are two ways of reducing the information $I(w;\mathcal{D})$. The first is reducing $|F_w|$ and the other is making the weights w more stable with respect to perturbation of the datasets. For example, the latter can be accomplished by choosing a suitable optimization algorithm or a particular architecture. Reducing the Fisher $F_{w|D}$, implies that the representations vary less when perturbing the

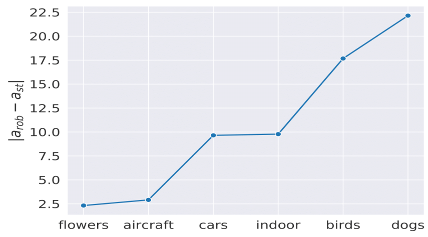


Figure 3: Accuracy gap between the robust and standard model as the distance from the source task increases.

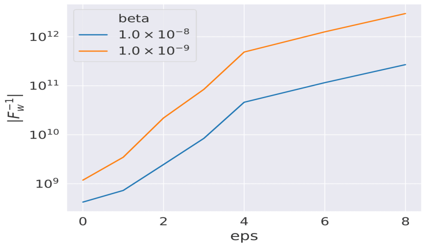


Figure 4: Flatness of Fisher information as measured by the norm of embedding (log-variance).

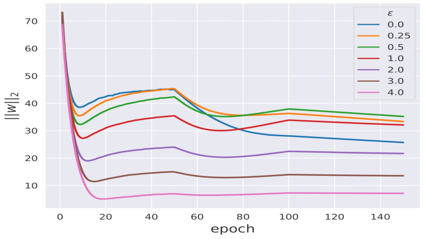


Figure 5: Norm of weights for different ε. Robust and standard training differ in the dynamics of $\|w\|_2$.

	C-10				**C-100**			
	C-100	F-MNIST	MNIST	SVHN	C-10	F-MNIST	MNIST	SVHN
Rob	**44.92**	**76.89**	**88.11**	**58.34**	74.47	**84.85**	**94.96**	**70.61**
St	35.76	67.15	64.17	36.6	**80.18**	76.10	79.46	55.6

Table 1: Transfer accuracy [%] starting from CIFAR-10 (left) and CIFAR-100 (right).

dataset with $\delta\mathcal{D}$. This explains that fact that AT is more robust to distribution shifts. We would like to remark again that there are two ways for transferring better: one is to reduce $\|\nabla_{\mathcal{D}} w\|_2$ and the other one is reducing $|F_w|$.

Transferability Experiments

We employ CIFAR-10 (Krizhevsky, Nair, and Hinton 2009), CIFAR-100 (Krizhevsky, Nair, and Hinton 2009) and ImageNet (Deng et al. 2009) as source datasets. All the experiments are obtained with ResNet-50 and $\varepsilon = 1$ for CIFAR and $\varepsilon = 3$ for ImageNet as described in (Ilyas et al. 2019) and in the Appendix. In Table 1 we show performance of fine-tuning for the networks pretrained on CIFAR-10 and CIFAR-100 transferring to CIFAR-10, CIFAR-100 F-MNIST (Xiao, Rasul, and Vollgraf 2017), MNIST (LeCun and Cortes 2010) and SVHN (Netzer et al. 2011). Details of target datasets are given in Appendix. Results confirm our hypothesis: when a task is ”visually” distant from the source dataset, the robust model performs better. For example, CIFAR-10 images are remarkably different from the SVHN or MNIST ones. Moreover, as we should expect, the accuracy gap (and thus the distance) is *not* symmetric: while CIFAR-100 is a good proxy for CIFAR-10, the opposite is not true. In fact, when fine-tuning on a more complex dataset, from a robust model is possible to leverage features that the standard model would discard. According to (Cui et al. 2018), we employ Earth Mover’s Distance (EMD) as a proxy of dataset distance, and we extract the *order* between datasets. As we show in Figure 7, the distance correlates well with the accuracy gap between robust and standard *across* all the tasks. Table 2 shows similar results using models pretrained on ImageNet. The robust model provides better performance in all the benchmarks being them quite different from the original tasks. We also report experiments on more difficult datasets namely Aircraft (Maji et al. 2013), Birds (Wah et al. 2011), Cars (Krause et al. 2013), Dogs (Khosla et al. 2011)[2], Flowers (Nilsback and Zisserman 2008), Indoor (Sharif Razavian et al. 2014) that would have not been suitable for transfering from simpler tasks like CIFAR-10 and CIFAR-100. Not surprisingly the robust model shows lower accuracy compared to the standard one since images are very similar to those contained in the ImageNet dataset. For examples, Dogs images are selected from ImageNet. Also with ImageNet, as shown by Figure 3, the difference in accuracy between the two model is correlated with distance. We can see that the furthest the task the higher the difference in accuracy in favor of the robust model. For the sake of space, we report similar results for other source and target datasets in the Appendix. Finally, in table 3 we analyze the impact of using a bigger architecture. It is noticeable that with the more complex network (ResNet50) the gap is reduced in cases where the standard model is better and it is increased in cases where the robust one is better.

Robustness of fine-tuned models Are the fine-tuned models still robust? As already experimentally shown by (Ilyas et al. 2019; Shafahi et al. 2019), an advantage of using $f_w(\cdot)$ as a feature extraction is that then the new model $A_2 f_w(\cdot) + b_2$ is robust for the new task. Indeed, it is sufficient to show that the Fisher $F_{y|x}$ is bounded from above by $F_{z|x}$, that is, the linear classifier can only reduce information.

Lemma 0.8. *Let $z = f_w$ be the feature extractor, $y = Az + b$, with $A \in \mathbb{R}^{k \times p}$, where $k < p$. Let $F_{z|x}$ be the Fisher of its activations about the input. Then, it holds:* $\operatorname{tr} F_{y|x} \leq \operatorname{tr} F_{z|x}$.

Conclusions

Existing works about robust models (Madry et al. 2017; Ilyas et al. 2019; Tsipras et al. 2019) showed that there exists

[2] The Stanford Dogs has been built using images and annotations from ImageNet.

IMG	C-10	C-100	F-MNIST	MNIST	SVHN	Aircraft	Birds	Cars	Dogs	Flowers	Indoor
Rob	**93.78**	**77.94**	**90.09**	**98.03**	**76.90**	33.81	35.91	40.47	66.25	93.15	63.06
St	84.72	64.48	86.38	93.91	50.46	**36.72**	**53.58**	**50.12**	**88.39**	**95.48**	**72.84**

Table 2: Transfer accuracy [%] of a ResNet50 pretrained on ImageNet.

		ResNet50				**ResNet18**			
	C-100	C-10	F-MNIST	MNIST	SVHN	C-10	F-MNIST	MNIST	SVHN
0	**Rob**	74.47	**84.85**	**94.96**	**70.61**	68.89	**83.40**	**94.61**	**61.08**
	St	**80.18**	76.10	79.46	55.60	**76.50**	76.30	77.98	49.32
1	**Rob**	85.67	**89.22**	**98.33**	**91.34**	82.40	**87.59**	97.82	**89.68**
	St	**87.80**	88.65	97.75	91.12	**85.11**	86.11	**97.84**	88.62
2	**Rob**	94.82	**92.58**	**99.24**	**96.63**	94.59	**92.48**	**99.30**	**96.39**
	St	**95.20**	91.78	99.22	96.60	**95.10**	92.03	99.15	96.29

Table 3: Performance comparison using different architectures transfering from CIFAR-100.

a trade-off between robustness of representations and accuracy for the task. This paper extends this property showing the parameters of robust models are the solution of a trade-off between usability of features for other tasks and accuracy for the source task. By leveraging results in (Achille and Soatto 2019, 2018a), we show that AT has a compression effect similarly to IB, and we explain how a network can be invertible and lose accuracy for the task at the same time. Moreover, we show that AT also reduces information in the weights, extending the notion of effective information from perturbations of the weights, to perturbations of the input.

We also show that effective noise can be also useful to improve reconstruction of images both in terms of convergence and quality of reconstruction.

Finally, we provide an analytic argument which explains why robust models can be better at transferring features to other tasks. As a corollary of our analysis, to train a generic feature extractor for several tasks, it is best to train adversarially, unless one already knows the specific task for which the features are going to be used.

Acknowledgments

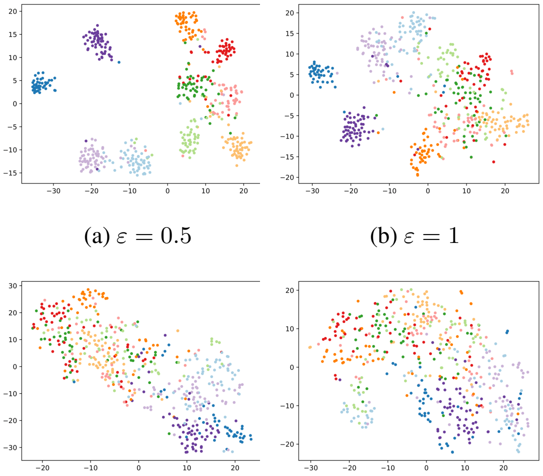
We would like to thank Stefano Soatto for providing valuable feedback on the manuscript.

References

Achille, A.; and Soatto, S. 2018a. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research* 19(1): 1947–1980.

Achille, A.; and Soatto, S. 2018b. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence* 40(12): 2897–2905.

Achille, A.; and Soatto, S. 2019. Where is the information in a deep neural network? *arXiv preprint arXiv:1905.12213* .



(a) $\varepsilon = 0.5$

(b) $\varepsilon = 1$

(c) $\varepsilon = 2$

(d) $\varepsilon = 3$

Figure 6: t-SNE of features extracted from a batch of 512 images with a robust ResNet-18 model trained on CIFAR-10 for different values of ε. As ε increases, features become less discriminative.

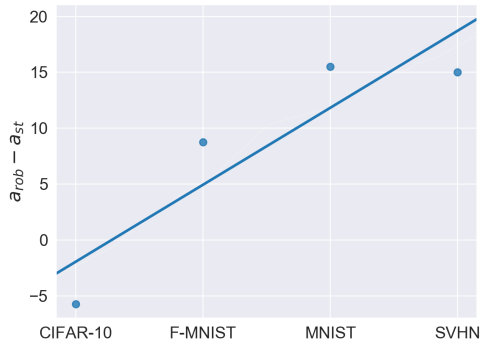


Figure 7: Accuracy gap between the robust and standard model transfering from CIFAR-100.

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

Behrmann, J.; Grathwohl, W.; Chen, R. T.; Duvenaud, D.; and Jacobsen, J.-H. 2018. Invertible residual networks. *arXiv preprint arXiv:1811.00995* .

Bottou, L.; Curtis, F. E.; and Nocedal, J. 2018. Optimization methods for large-scale machine learning. *Siam Review* 60(2): 223–311.

Cui, Y.; Song, Y.; Sun, C.; Howard, A.; and Belongie, S. 2018. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4109–4118.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 248–255. Ieee.

Engstrom, L.; Ilyas, A.; Santurkar, S.; Tsipras, D.; Tran, B.; and Madry, A. 2019. Learning perceptually-aligned representations via adversarial robustness. *arXiv preprint arXiv:1906.00945* .

Fazlyab, M.; Robey, A.; Hassani, H.; Morari, M.; and Pappas, G. 2019. Efficient and accurate estimation of lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, 11423–11434.

Frosst, N.; Papernot, N.; and Hinton, G. 2019. Analyzing and improving representations with the soft nearest neighbor loss. *arXiv preprint arXiv:1902.01889* .

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv:1412.6572* .

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, 125–136.

Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Fei-Fei, L. 2011. Novel Dataset for Fine-Grained Image Categorization. In *First Workshop on Fine-Grained Visual Categorization, CVPR.*

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*. Sydney, Australia.

Krizhevsky, A.; Nair, V.; and Hinton, G. 2009. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html* 6.

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

Li, Q.; Tai, C.; et al. 2017. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2101–2110. JMLR. org.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov): 2579–2605.

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

Mahendran, A.; and Vedaldi, A. 2015. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5188–5196.

Maji, S.; Kannala, J.; Rahtu, E.; Blaschko, M.; and Vedaldi, A. 2013. Fine-Grained Visual Classification of Aircraft. Technical report.

Martens, J. 2014. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193* .

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning .

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE.

Shafahi, A.; Saadatpanah, P.; Zhu, C.; Ghiasi, A.; Studer, C.; Jacobs, D.; and Goldstein, T. 2019. Adversarially robust transfer learning. *arXiv preprint arXiv:1905.08232* .

Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *CVPR workshops*, 806–813.

Terzi, M.; Susto, G. A.; and Chaudhari, P. 2020. Directional adversarial training for cost sensitive deep learning classification applications. *Engineering Applications of Artificial Intelligence* 91: 103550.

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

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=SyxAb30cY7.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9446–9454.

Virmaux, A.; and Scaman, K. 2018. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, 3835–3844.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

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

Yosinski, J.; Clune, J.; Nguyen, A.; Fuchs, T.; and Lipson, H. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* .

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573* .