

# The Robust Manifold Defense: Adversarial Training using Generative Models

Andrew Ilyas  
ailyas@mit.edu  
MIT EECS

Ajil Jalal  
ajiljalal@utexas.edu  
UT Austin

Eirini Asteri  
eirini@utexas.edu  
UT Austin

Constantinos Daskalakis  
costis@mit.edu  
MIT EECS

Alexandros G. Dimakis  
dimakis@austin.utexas.edu  
UT Austin

December 27, 2017

*Problems worthy of attack,  
prove their worth by fighting back.*

## Abstract

Deep neural networks are demonstrating excellent performance on several classical vision problems. However, these networks are vulnerable to *adversarial examples*, minutely modified images that induce arbitrary attacker-chosen output from the network. We propose a mechanism to protect against these adversarial inputs based on a generative model of the data. We introduce a pre-processing step that projects on the range of a generative model using gradient descent before feeding an input into a classifier. We show that this step provides the classifier with robustness against first-order, substitute model, and combined adversarial attacks. Using a min-max formulation, we show that there may exist adversarial examples even in the range of the generator, natural-looking images extremely close to the decision boundary for which the classifier has unjustified high confidence. We show that adversarial training on the generative manifold can be used to make a classifier that is robust to these attacks.

Finally, we show how our method can be applied even without a pre-trained generative model using a recent method called the deep image prior. We evaluate our method on MNIST, CelebA and Imagenet and show robustness against the current state of the art attacks.

## 1 Introduction

Deep neural network (DNN) classifiers are currently demonstrating excellent performance for various computer vision tasks. These models work well for benign inputs but recent work has shown it is possible to make very small changes to an input image and drastically fool state-of-the-art models [45, 18]. These *adversarial examples* are barely perceivable to humans, can be targeted to create desired labels even with black-box access to the classifiers and can be made robust as real objects in the physical world [36, 26, 5].

This phenomenon is receiving a tremendous amount of recent attention (e.g. [31, 25, 48, 19, 22, 46] and references therein) for two good reasons: First a classifier that can be easily fooled by non-perceivable noise poses a security threat in any real deployment. Second, it illustrates that even our best models can be *making correct predictions for the wrong reasons*. This relates to interpretability and trust [39, 29, 14] in modern complex models which is an important emerging topic.

Typical methods of attack involve modifying pixel values while keeping a small  $\ell_2$  or  $\ell_\infty$  distance from the original image. Very recent work however has shown that small rotations [15] or spatial transformations [4] can also fool classifiers. We would like to propose an extended definition of adversarial examples that captures all these important aspects, building on legal theory and the *reasonable person test* (see e.g. [35]): *A pair of inputs  $x, x'$  is an adversarial example for a classifier, if a reasonable person would say they are of the same class but the classifier produces significantly different outputs.* This definition is useful: if someone has defaced a stop sign so that a reasonable person could confuse it for a different sign, nobody can accuse a classifier for making the same mistake. On the contrary, attacks like the robust physical perturbations of traffic signs shown in [16] would never make a reasonable person think this is not a stop sign.

Many attempts have been made to defend DNNs against adversarial examples. We survey the literature in the subsequent section, but the overall message is that defending all possible methods of attack, as previously defined, remains challenging. Our intuition is that adversarial examples exist because an original natural image  $x$  is perturbed into  $x_{adv}$ , a point that is *far from the manifold of natural images*. Our classifier has never been trained on objects far from natural images so it can behave in unexpected ways. Furthermore, the natural image manifold is low-dimensional but the noisy objects that can be reached with even small perturbations, is very high dimensional and hence much harder to learn.

In this paper we make the critical assumption that we have a generative model for the data we are working on. This generative model can be either explicit (i.e. produce likelihoods) or an implicit model like a Generative Adversarial Network [17]. Several methods train neural networks to project an image on the manifold [33, 42] but these are end-to-end differentiable and hence easy to attack [9]. We use the compressed sensing inversion method [7] instead: Given an input image  $x$  and a classifier  $C$ , do not feed the image directly as an input to the classifier, but rather treat it as noisy measurements of another true image  $x_{true}$  in the range of a (pre-trained) generator  $G(z)$ . We solve a minimization problem to find a  $z^*$  such that  $G(z^*)$  is close to the input image, and feed  $G(z^*)$  to the classifier. This minimization is solved by gradient descent which makes it a non-differentiable method of projecting on the manifold. Since thousands of gradient steps are required, it is not easy to “unfold” this operation and attack a differentiable substitute model, as we show in our evaluation section.

We formulate this method (called *Invert and Classify (INC)*) and show that it is able to resist first-order and black-box attacks. We then explore its robustness even further by formulating a min-max optimization problem where the adversary has much more power: the process tries to simply find any two points in the latent code  $z, z'$  that produce images that are close, i.e.  $\|G(z) - G(z')\|$  is small, but the classifier produces very different outputs i.e.  $\|C(G(z)) - C(G(z'))\|$  is large. By Lagrangifying the constraints and using a first-order method, we are able to solve this problem and find pairs of adversarial points on the manifold with unjustified drastically different classifier confidence. This shows that natural problematic points exist, an idea also supported by the recent work in [2], which shows it for an artificially constructed classifier over spheres. We thus seek to

robustify the system further.

We show how this Min-Max attack can be used to robustify INC by using adversarial training on these examples on the manifold. We show that our proposed INC classifier is robust to various types of attacks including end-to-end substitution models. The accuracy of the classifier drops compared to clean-image performance but the inversion operation seems to provide effective protection.

Our last innovation deals with robust classification without a pre-trained generative model. This is relevant for several rich datasets like ImageNet where it is hard to train an accurate generative model. To address this problem, we rely on Deep Image Prior (DIP) [47]: An untrained convolutional neural network for which the latent code is kept fixed in some random value, but the weights are trained to match a desired output image. Ulyanov et al. [47] showed how this can be used for denoising, inpainting, and super-resolution without any pre-training on a dataset.

We define the Deep Image Prior INC method that uses such untrained generators and can still be used to create robust classifiers for Imagenet. We show that the deep image prior INC protection maintains the accuracy of (top-1) ResNet152 for ImageNet at 30% – 50% under BIM attacks for  $\varepsilon = 0.01 - 0.10$ . The price of this robustness is that the accuracy on clean images drops from 71% to 35% – 50%, for top-1 classification in 1000 classes.

## 1.1 Contributions

Concretely, our contributions are summarized as follows:

- We formulate and present the “Invert-and-Classify” (INC) algorithm, which protects a classifier  $C$  by projecting its inputs onto the range of a given generator  $G$  which effectively serves as a prior  $P(x)$  for the classification. We demonstrate that the algorithm induces robustness across a wide variety of attacks, including first-order methods, substitute models, and enhanced attacks combining the two.
- By formulating a min-max optimization problem that can be viewed as an overpowered attack on  $C$ , we demonstrate that there may in fact exist problematic pairs  $(z, z')$  in the domain of a generator that interact with the hard decision boundaries of the classifier such that  $G(z)$  and  $G(z')$  are close but their classifications are far. Through adversarial training we soften the classifier’s decision boundaries and demonstrate robustness to the same min-max optimization attack.
- We propose a possible modification of the INC algorithm for settings in which good generative models are unavailable (e.g. for Imagenet), where we instead use the *structural prior* given by an untrained generator, as introduced in [47]. We show that this Deep Image Prior defense can actually defend against adversarial attacks for the ImageNet dataset.

## 2 A Min-Max Formulation

### 2.1 Step 1: Defending using GANs

Given a classifier  $C_\theta$  parametrized by a vector of parameters  $\theta$ , we want to defend it by filtering its input through a generator that samples natural inputs. This would be a pre-trained generative model that is assumed to produce natural inputs from all different categories that we are classifying.

More precisely, for some hyperparameter  $\eta$  and given an input  $x$ , we perform the following procedure that we call *Invert and Classify (INC)*:

1. Perform gradient descent in  $z$  space to minimize  $\|G(z) - x\|_2$ . Let  $z^*$  be the point returned by gradient descent. Ideally,  $z^* = \arg \min_z \|G(z) - x\|_2$ .
2. If the “projection”  $G(z^*)$  of  $x$  is far from  $x$ , i.e. if  $\|G(z^*) - x\|_2 \geq \eta$ , we reject the input  $x$  as “unnatural,” since it lies far from the range-of- $G$  manifold.
3. Otherwise, we apply our classifier on the projected input, outputting a class according to the distribution  $C_\theta(G(z^*))$ .

## 2.2 Step 2: An Overpowered Attack

Given some input  $x$ , one way to attack INC is to search for some  $x'$  that is close to  $x$  and also close to the manifold, so that the classification of their projections  $G(z), G(z')$  is significantly different.

If such an attack exists, then (by triangle inequality) there must exist  $z$  and  $z'$  such that  $G(z)$  and  $G(z')$  are close, yet  $C_\theta(G(z))$  and  $C_\theta(G(z'))$  are far. The following optimization problem captures the furthest  $C_\theta(G(z))$  and  $C_\theta(G(z'))$  can be subject to some constraint on the distance of  $G(z)$  and  $G(z')$ . This provides an upper bound on the magnitude of the INC attack:

$$\sup_{z, z'} \|C_\theta(G(z)) - C_\theta(G(z'))\|_2^2, \quad (1)$$

$$\|G(z') - G(z)\|_2^2 \leq \eta^2. \quad (2)$$

This optimization problem upper-bounds the size of the attack to INC. In fact, it also captures the loss that may arise from a potential imperfect optimization in the first step of INC, where the input is projected into the range of  $G$ . Namely, the value of the objective function for a solution  $z$  and  $z'$  to the above optimization problem also captures the maximum loss in accuracy under a scenario where  $x = G(z)$  and the projection of  $x$  onto the range of  $G$  identified in the first step of INC via gradient descent is  $G(z')$ . Of course, the projection  $G(z')$  of  $x$  into the range of  $G$  is not chosen adversarially, so the above optimization problem captures an “overpowered adversary,” serving as an upper bound on the loss from both adversarial attacks and suboptimality in step one of INC.

We can Lagrangify the constraints in the above formulation to obtain the following min-max formulation:

$$\inf_{\lambda \geq 0} \sup_{z, z'} \|C_\theta(G(z)) - C_\theta(G(z'))\|_2^2 + \lambda \cdot (\|G(z) - G(z')\|_2^2 - \eta^2). \quad (3)$$

As usual, the Lagrange multipliers make sure that if any constraints are violated, then the inf player can make the objective  $-\infty$ , by setting the multiplier  $\lambda$  to  $-\infty$ . Hence, the sup player must respect these constraints, and the inf player will have to set these multipliers to 0, so ultimately the right objective will be maximized by the sup player.

We found experimentally that we can identify good solutions to (3) using gradient descent, as we describe in Section 4.3. We show attacks for a gender classifier trained on CelebA [30] and protected with the INC framework using a BEGAN [6] generator. Our experiments show that it is possible to obtain pairs of images  $(x, x') = (G(z), G(z'))$  that lie on the range of  $G$  and are close in  $\ell_2$  (or  $\ell_\infty$  distance), yet the gender classifier produces drastically different outputs on these images.

Another interesting experimental finding is that the pairs of images that were identified through our attack framework appeared to change gender-relevant features to confuse the classifier, often producing images whose class was ambiguous even for human observers. We show some of these pairs and how they were classified in Section 4.3. The main limitation of the INC-protected classifier is that the predictions have very high confidence (close to 0/1 classification probabilities) even for ambiguous images, and that small changes on an image cause abrupt changes in the classification probabilities.

Finally, we note that while the classifier outputs drastically different distributions on the pairs of images  $(G(z), G(z'))$  that we identified, the “noise” introduced by the suboptimality of gradient descent in the first step of INC—the (non-convex) projection step—seems to make the process more robust. In particular, when the images  $G(z)$  and  $G(z')$  were given as inputs to the INC protected classifier, the outputs were not drastically different. This is because the latent codes recovered by gradient descent on these inputs were not exactly  $z, z'$  and this projection noise seemed to protect INC from the low-probability set of adversarial examples.

This is clearly an interesting empirical finding for the robustness of INC, but we would not hope to base its security on the noise introduced by gradient descent in the projection step. So in the next section we use these adversarial pairs of inputs to robustify our classifier using adversarial training.

### 2.3 Step 3: Robustifying INC through Adversarial Training

Anticipating the attack outlined in Section 2.2, we can take our defense approach outlined in Section 2.1 one step further, retraining the parameters of the classifier to minimize the damage from the attack. This results in the following min-max formulation:

$$\inf_{\theta} \mu \left( \sup_{z, z'} \|C_{\theta}(G(z)) - C_{\theta}(G(z'))\|_2^2 \right) + (1 - \mu) \left( \frac{1}{N} \sum_{i=1}^N f(y^{(i)}, C_{\theta}(x^{(i)})) \right), \quad (4)$$

$$\text{s.t. } \|G(z) - G(z')\|_2^2 \leq \eta^2. \quad (5)$$

where the mixing weight  $\mu \in (0, 1)$  is some hyperparameter, and the second term of the objective function above evaluates the performance of the classifier on the input set  $\{x^{(i)}, y^{(i)}\}_{i=1}^N$ , where  $f$  is a traditional classification loss function ( $f$  can actually be the same as that used to train the classifier initially, e.g. the cross-entropy loss).

We can again fold the constraints into a Min-Max formulation as follows:

$$\begin{aligned} \inf_{\theta, \lambda \leq 0} \sup_{z, z'} & \left( \mu \|C_{\theta}(G(z)) - C_{\theta}(G(z'))\|_2^2 + (1 - \mu) \frac{1}{N} \sum_{i=1}^N f(y^{(i)}, C_{\theta}(x^{(i)})) \right. \\ & \left. + \lambda \cdot (\|G(z) - G(z')\|_2^2 - \eta^2) \right). \end{aligned} \quad (6)$$

We tried this retraining approach and report our findings in Section 4.3.

### 2.4 Step 4: What to do if no GAN is available

The approach described in Sections 2.1–2.3 can be used to robustify any classifier if we have a good generative model for the inputs of interest. We propose a way to extend our approach to settings

---

**Algorithm 1:** Invert-and-classify

---

**Input** : Input image  $x$ , Generator  $G$ , Classifier  $C$

**Output:** Classifier prediction  $C(G(z^*))$

**begin**

$z_0 \sim \mathcal{N}(0, 1)$

**for**  $t \leftarrow 1$  **to**  $T$  **do**

$z_t \leftarrow z_{t-1} - \eta_t (\nabla_z \|x - G(z)\|_2^2)|_{z=z_{t-1}}$

$z^* \leftarrow z_T$

**return**  $C(G(z^*))$

---

where no pretrained generative model is available. We illustrate our approach for classification of image data.

We propose to use a Deep Image Prior (DIP) [47], i.e. an untrained generative model  $G_\phi(z)$  with a convolutional neural network topology. The surprising result is that training *over the weights*  $\phi$  to approximate a given input image  $x$  effectively projects  $x$  onto the manifold of natural images. We leverage this idea to defend classifiers on Imagenet without relying on a pre-trained generative model. Our method is identical to INC with the only difference being that the projection step optimizes over weights  $\phi$ .

Specifically, given an input image  $x \in \mathbb{R}^n$ , we search for  $\hat{x}$ , such that  $\hat{x}$  is a natural image, and  $\|x - \hat{x}\|_2$  is small. The Deep Image Prior (DIP) method tries to solve this problem by constructing a generative convolutional neural network  $G_\phi : \mathbb{R}^k \rightarrow \mathbb{R}^n$  parameterized by a set of weights  $\phi$ , and searching for a set of weights  $\phi^*$  that satisfy

$$\phi^* = \arg \min_{\phi} \|x - G_\phi(z)\|_2, \quad (7)$$

for some randomly selected  $z \in \mathbb{R}^k$ , which is held constant throughout the optimization procedure. The final output is  $G_{\phi^*}(z)$ , which is a natural image close to the input, with respect to the  $\ell_2$  distance.

An important issue here is that the search over  $\phi$  is a gradient descent procedure that we terminate early: as observed by [47], if too many steps are performed,  $G_\phi(z)$  becomes too expressive and also reconstructs the adversarial noise. The number of steps was empirically tuned in our experiments and depends on the power of the adversary. We discuss the DIP INC in more detail in Section 3.1, and our experiments in Section 4.4.

As a final note, it should be emphasized that while all the previous methods of our paper also apply to non-image datasets, in order to apply our generator-free approach to non-image data we would have to develop an architecture, serving as the analog of DIP, for that type of data.

### 3 Implementation

We now describe Invert and Classify (INC) in detail. The algorithm is given in pseudocode 1, and the schematic is shown in Figure 1.

Given an input  $x$ , our strategy is to find a  $z^*$ , such that  $G(z^*)$  is close to  $x$  in  $\ell_2$  distance. This is achieved by sampling a  $z_0 \sim \mathcal{N}(0, 1)$ , and running  $T$  iterations of SGD, where the gradient at iteration  $t$  is given by  $(\nabla_z \|x - G(z)\|_2^2)|_{z=z_{t-1}}$ .

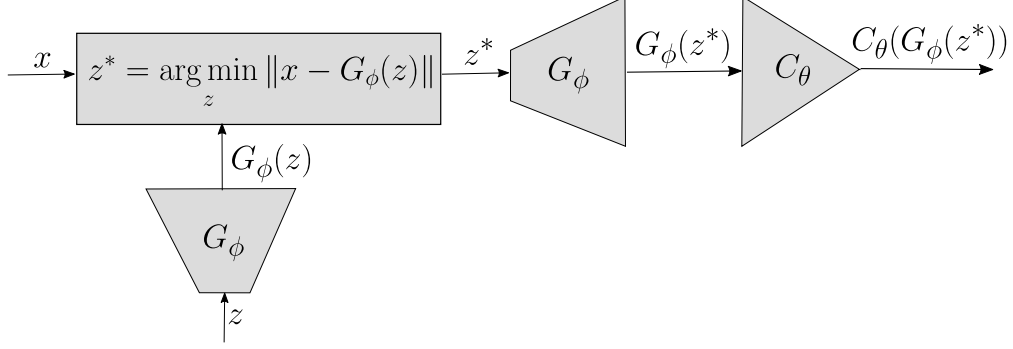


Figure 1: Schematic showing our proposed defense strategy, Invert-and-Classify. The trapezoid labeled  $G_\phi$  and triangle labeled  $C_\theta$  denote the generator and classifier models respectively. The rectangle denotes an optimization procedure that accepts an image  $x$  as input, and repeatedly queries  $G_\phi$  to find a  $z^*$  that minimizes  $\|x - G_\phi(z)\|_2$ . Once this  $z^*$  is found, the classifier makes its prediction based on  $G_\phi(z^*)$ . Note that in Invert-and-Classify,  $G_\phi$  refers to a generator that has been pretrained, and  $\phi$  is held constant.

Our intuition for why this strategy works is based on the observation that adversarial noise is very high dimensional, whereas the natural images form a low dimensional manifold in  $\mathbb{R}^n$ . Hence, searching for an image in the span of  $G$  that is close to  $x$  in  $\ell_2$  norm is equivalent to projecting  $x$  onto the manifold of natural images, assuming  $G$  has learned the true probability distribution over the training dataset.

### 3.1 Deep Image Prior Defense

The algorithm for Deep Image Prior INC is given in Algorithm 2, and the schematic is shown in Figure 2.

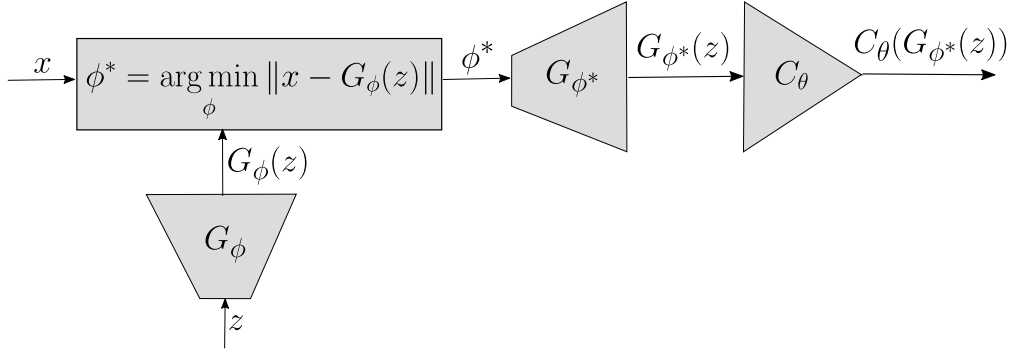


Figure 2: Schematic showing the Deep Image Prior INC. The trapezoid labeled  $G_\phi$  and triangle labeled  $C_\theta$  denote the generator and classifier models respectively. The rectangle denotes an optimization procedure that accepts an image  $x$  as input, and repeatedly queries  $G_\phi$  to find a  $\phi^*$  that minimizes  $\|x - G_\phi(z)\|$ . Once this  $\phi^*$  is found, the classifier makes its prediction based on  $G_{\phi^*}(z)$ . Note that in the Deep Image Prior defense,  $G_\phi$  refers to a generator that has been randomly initialized, and we search for optimal parameters  $\phi^*$ , while  $z$  is held constant throughout.



---

**Algorithm 2:** Deep Image Prior Invert-and-Classify

---

**Input** : Input image  $x$ , Classifier  $C_\theta$   
**Output:** Classifier prediction  $C_\theta(G_{\phi^*}(z))$   
**begin**  
     $z \sim \mathcal{N}(0, 1)$   
     $\phi_0$  randomly initialized  
    **for**  $t \leftarrow 1$  **to**  $T$  **do**  
         $\phi_t \leftarrow \phi_{t-1} - \eta_t(\nabla_\phi \|x - G_\phi(z)\|_2^2)|_{\phi=\phi_{t-1}}$   
     $\phi^* \leftarrow \phi_T$   
    **return**  $C_\theta(G_{\phi^*}(z))$

---

Given an input  $x$ , our strategy is to find a set of parameters  $\phi^*$ , such that  $G_{\phi^*}(z)$  is close to  $x$  in  $\ell_2$  distance. This is achieved by sampling a  $z$  and  $\phi_0$  at random, and running  $T$  iterations of SGD, where the gradient at iteration  $t$  is given by  $(\nabla_\phi \|x - G_\phi(z)\|_2^2)|_{\phi=\phi_{t-1}}$ . An important benefit of this method is that it alleviates the need for a pre-trained generator. In Section 4.4 we show how this method can be used to protect against attacks on Imagenet.

As observed by Ulyanov *et al.* [47], the Deep Image Prior can yield very accurate reconstructions if many iterations. We found that to use this method to remove adversarial noise we have to use early stopping which has to be carefully tuned.

## 4 Evaluation

### 4.1 Experimental Setup

#### Datasets

We evaluate the robustness of Invert-and-Classify for classification tasks on two datasets: hand-written digit classification on the MNIST dataset [27] and gender classification on the CelebA dataset [30]. To evaluate the robustness of the Deep Image Prior defense, we ran experiments on 1000 validation images randomly sampled from the Imagenet dataset [40].

The MNIST images were not pre-processed. The CelebA images were cropped using a bounding box of size  $128 \times 128$  (the top left corner of the bounding box was placed at coordinate (50, 25)) and were then resized to  $64 \times 64$  using bilinear interpolation. The pixel values of the images were normalized to lie in the range  $[-1, 1]$ . The Imagenet images were resized to  $224 \times 224$  and pixel values were normalized to lie in the range  $[-1, 1]$ .

#### Classifiers

For MNIST classification<sup>1</sup>, the classifier  $C : \mathbb{R}^{28 \times 28} \rightarrow \mathbb{R}^{10}$  has two convolutional layers, followed by a fully connected layer and a softmax layer. The convolutional layers have 32 and 64 filters respectively, and the filters are of size  $5 \times 5$ . The fully connected layer has 1024 nodes and the softmax layer has 10 nodes.

---

<sup>1</sup>Code borrowed from [https://www.tensorflow.org/get\\_started/mnist/pros](https://www.tensorflow.org/get_started/mnist/pros); architecture was kept the same.



For gender classification<sup>2</sup> on the CelebA dataset, the classifier  $C : \mathbb{R}^{64 \times 64 \times 3} \rightarrow \mathbb{R}^2$  has two convolutional layers followed by two fully connected layers and a softmax layer. The convolutional layers have 64 filters in each layer, and all filters are of size  $5 \times 5$ . The fully connected layers have 384 and 192 filters respectively, and the softmax layer has two nodes.

For classifying Imagenet images, we used a ResNet152 [21].

## Generative Models

We chose the decoder of a Variational Auto Encoder (VAE) [24] for generating images of digits from MNIST<sup>3</sup>. The encoder  $E : \mathbb{R}^{28 \times 28} \rightarrow \mathbb{R}^{20}$  has 2 fully connected hidden layers, each with 500 nodes; the output is 20-dimensional. The decoder  $G : \mathbb{R}^{20} \rightarrow \mathbb{R}^{28 \times 28}$  has 2 fully connected hidden layers, each with 500 nodes; the final output is of size  $28 \times 28$ .

To generate images of celebrities, we trained a BEGAN [6]<sup>4</sup> on the first 160,000 images in the CelebA dataset. The generator of the BEGAN,  $G : \mathbb{R}^{128} \rightarrow \mathbb{R}^{64 \times 64 \times 3}$  has 1 fully connected layer and 5 convolutional layers. The input to the generator is distributed according to  $\mathcal{N}(0, 4\mathbb{1}_{128 \times 128})$ , and the fully connected layer is of size  $8 \times 8 \times 128$ . The first 4 convolutional layers have 128 filters of size  $3 \times 3$ , and the final convolutional layer has 3 filters of size  $3 \times 3$ .

For image generation in the Deep Image Prior method, we use the skip-autoencoder architecture employed by Ulyanov *et al.* [47] for image denoising<sup>5</sup>.

## Optimizers

For the Invert-and-Classify experiments on MNIST and CelebA, we used a Tensorflow [1] implementation of an Adam optimizer with initial learning rate 0.1 and 0.01 respectively.

For the Deep Image Prior experiments, we used a PyTorch [38] implementation of an Adam optimizer [23] with initial learning rate 0.01.

## 4.2 Robustness against Attacks

We demonstrate the robustness of our method against standard attacks. Note that since inversion is non-differentiable gradient descent, first-order attacks against the end-to-end system are infeasible.

### 4.2.1 First-Order Classifier Attacks

Most methods to construct adversarial examples try to find perturbations that have small  $\ell_p$  norm. First-order attacks on the full system are intractable to generate; in this section, we demonstrate robustness to first-order attacks on the unprotected classifier.

We first focus on the case where the adversarial perturbation has low  $\ell_\infty$  norm, where we use the Fast Gradient Sign [18] method. We perform *untargeted attacks*—we only require that the classifier predicts some label other than the true label; the adversarial perturbation is given by

$$\delta = +\epsilon \cdot \text{sign}(\nabla_X L(y, C_\theta(X))|_{X=x}), \quad (8)$$

<sup>2</sup>The code was borrowed from <https://github.com/tensorflow/models/tree/master/tutorials/image/cifar10/> and minimal modifications were made to use it for CelebA.

<sup>3</sup>Code borrowed from <https://jmetzen.github.io/notebooks/vae.ipynb>; architecture was kept the same.

<sup>4</sup>Code borrowed from <https://github.com/carpedm20/BEGAN-tensorflow>; architecture was kept the same.

<sup>5</sup>See the supplementary material at [https://dmityulyanov.github.io/deep\\_image\\_prior](https://dmityulyanov.github.io/deep_image_prior) for the precise details

where  $x$  is the original image and  $y$  is its label;  $L(y; C_\theta(x))$  is the cross entropy loss between the label  $y$  and the classifier prediction on  $x$ . For the CelebA dataset, we additionally evaluated how robust Invert-and-Classify is against the Carlini-Wagner  $\ell_2$ ,  $\ell_0$ ,  $\ell_\infty$  attacks [10]. For the  $\ell_2$  attack, we set the confidence parameter to 5.

**MNIST** The accuracy of the MNIST classifier and the Invert-and-Classify method for varying  $\epsilon$  is shown in Figure 3.

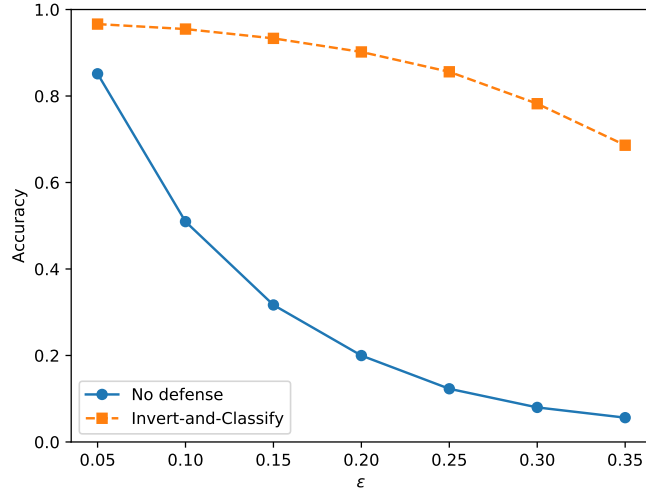


Figure 3: Accuracy of the classifier and INC protected classifier as we vary the  $\ell_\infty$  norm of the adversarial perturbation for **MNIST** using a FGSM attack.

**CelebA** We report the accuracy of the CelebA classifier and the Invert-and-Classify method against the FGSM attack and Carlini-Wagner attacks in Table 1.

$\epsilon$	No defense	Invert and Classify
Clean Data	97%	84%
FGSM ( $\epsilon = 0.05$ )	1%	82%
FGSM ( $\epsilon = 0.1$ )	0%	80%
FGSM ( $\epsilon = 0.2$ )	0%	73%
Carlini-Wagner $\ell_2$	0%	77%
Carlini-Wagner $\ell_0$	0%	65%
Carlini-Wagner $\ell_\infty$	0%	66%

Table 1: Accuracy of the **celebA** classifier and the Invert-and-Classify method under different adversarial attacks: FGSM [18] at various powers  $\epsilon$ , and the Carlini-Wagner attacks [10].

Figure 4 shows the reconstructions obtained by Invert-and-Classify for women and men in the CelebA dataset.

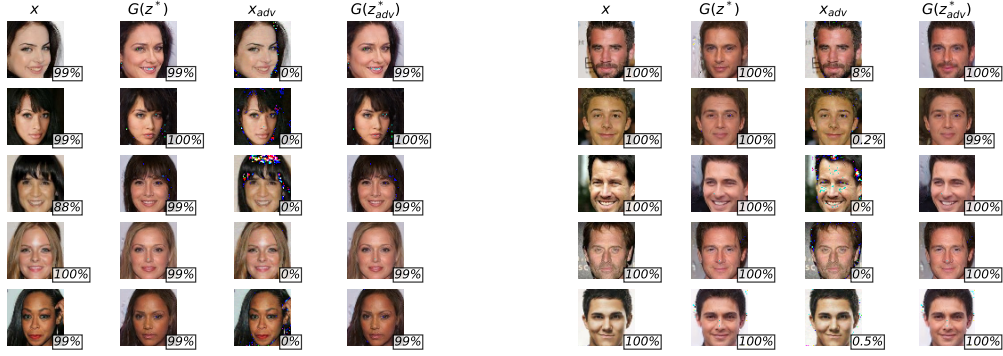


Figure 4: In this figure,  $x$  are the original images,  $G(z^*)$  are the images obtained by inverting  $x$ ,  $x_{adv}$  are the adversarial examples obtained from  $x$  ( $\epsilon = 0.05$ ), and  $G(z_{adv}^*)$  are the images obtained by inverting  $x_{adv}$ . The values at the corner of each image indicate the confidence with which the classifier predicts the image as the correct gender. As shown, the unprotected gender classifier has confidence 0 – 8% and the INC protected one confidence 99%.

#### 4.2.2 Substitute Model Attacks

We investigate the robustness of our end-to-end INC protected classifier. The gradient descent INC inversion procedure is non-differentiable and hence much harder to attack. One possible attack is to “unfold” the gradient descent steps and create a differentiable model that can be subsequently attacked. Since our gradient descent projection in INC involves thousands of iterations (typically more than 2000) we could not make such an attack work.

The second approach is to leverage the transferability of the adversarial examples [45] and design a black-box attack to evaluate the Invert-and-Classify model, and show that the new attack is also ineffective.

We train our end-to-end substitute network as in the typical black-box setting [36] using input-output pairs of the target model. We emphasize that our substitute model is differentiable and only attempts to approximate the decision boundaries of the non-differentiable Invert-and-Classify model. The architecture of the substitute model follows that of the inner classifier  $C_\theta$  described in Section 4.1. We train our network on images from the celebA dataset labeled by the output of the Invert-and-Classify model. Finally, to improve the substitute model even further we also train on inputs that are adversarial to the inner classifier  $C_\theta$ . This incorporates white box information since we provide to the adversary knowledge of the gradients of  $C_\theta$  on the training samples.

We craft first-order adversarial examples using our substitute model and feed them to the target INC model. We measure the percentage of the adversarial examples generated for the substitute model that are misclassified by the target model as well. Figure 5 shows the inability of first-order attacks, namely the Fast Gradient Sign Method [18] (FGSM) and Basic Iterative Method [26] (BIM) on the substitute model to transfer to Invert-and-Classify.

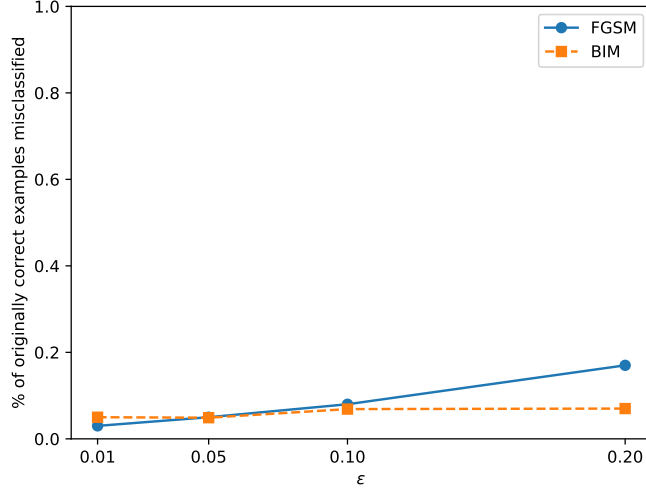


Figure 5: Non-transferability of attacks on substitute networks: The network is trained on natural and adversarial input-output pairs from celebA. The validation set consists of natural images that were correctly classified by both the target and substitute model. The images were adversarially perturbed with FGSM [18] and 20 steps of BIM [26] for  $\epsilon = 0.01, 0.05, 0.1, 0.2$ .

#### 4.2.3 Combined Attack

Another idea for attacking INC is to train a differentiable substitute model for the inversion step only, then combine this model with standard first-order attacks on the classifier to produce adversarial inputs. As mentioned, unfolding the gradient descent process is intractable, since thousands of iterations are run per projection. Thus, we model the inversion step using a convolutional neural network, effectively training a differentiable encoder and making  $G$  an autoencoder. As we show in Table 12 in Appendix A, it is indeed easy to attack the autoencoder but the attacks do *not* in typically transfer to the INC system.

### 4.3 Increasing Robustness with Overpowered Generator Attacks

As described in section 2.2, we perform an overpowered attack on the standard invert-and-classify architecture. We search for  $z$  and  $z'$  such that  $G(z)$  and  $G(z')$  are close but induce dramatically different classification labels. Recall that this involves solving the following min-max optimization problem:

$$\inf_{\lambda \leq 0} \sup_{z, z'} \|C_{\theta}(G(z)) - C_{\theta}(G(z'))\|_2^2 + \lambda \cdot (\|G(z) - G(z')\|_2^2 - \eta^2). \quad (9)$$

In practice, we set our  $\ell_2$  constraint to  $\eta^2 \approx 2.46$ , corresponding to an average squared difference of  $2 \cdot 10^{-4}$  per pixel-channel. We implement the optimization through alternating iterated gradient descent on both  $\lambda$  and  $\theta$ , with a much more aggressive step size for the  $\lambda$ -player (since its payoff is linear in  $\lambda$ ). The gradient descent procedure is run for 10,000 iterations. Because the  $\ell_2$  constraint



Figure 6: Pairs of images  $G(z)$  and  $G(z')$  generated by the Min-max overpowered attack. Since a reasonable person would disagree with these confidence changes, these pairs of images are adversarial attacks for this classifier, according to our definition. These adversarial attacks lie on the manifold of natural images so the classifier must be made robust.

was imposed through a Lagrangian, we consider two  $z, z'$  valid if the mean distance between the images is  $< 0.0005$ .

The optimization terminated with **93%** of the images satisfying the  $\ell_2$  constraint; within this set, the average KL-divergence between classifier outputs was **2.47**, with **57%** inducing different classifications. Figure 6 shows randomly selected successful results of the attack.

First, note that in contrast to the attacks found in Figure 4 on the unprotected classifier, the attacks found with this optimization tend to yield images with semantically relevant features from both classes, and furthermore often introduce meaningful (though minute) differences between  $G(z)$  and  $G(z')$  (e.g. facial hair, eyes widening, etc.). This suggests that the attack is exploiting the hard decision boundary introduced in classifier training. Secondly, as described in Section 2.2, none of these images actually induce different classifications on the end-to-end classifier, which we attribute to imperfections in the projection step of the defense (that is, since  $G(z^*) \neq x$  exactly). That said, we opt to robustify the classifier against this attack regardless. Recall the more complex min-max optimization proposed in Section 2.3:

$$\inf_{\theta} \mu \left( \sup_{z, z'} \|C_{\theta}(G(z)) - C_{\theta}(G(z'))\|_2^2 \right) + (1 - \mu) \left( \frac{1}{N} \sum_{i=1}^N f(y^{(i)}, C_{\theta}(x^{(i)})) \right), \quad (10)$$

$$\text{s.t. } \|G(z) - G(z')\|_2^2 \leq \eta^2. \quad (11)$$

We implement this through *adversarial training* [31]; at each iteration, in addition to sampling a cross-entropy loss from images from the dataset, we also sample an adversariality loss, where we generate a batch of “adversarial” inputs using 500 steps of the min-max attack, then add the final  $\ell_2$  distance between the classification outputs to the cross-entropy loss. As shown in Figure 7, the classifier eventually learns to minimize the adversary’s ability to find examples, most likely by learning and softening the decision boundaries being exploited by the generator. After robustifying the classifier using this adversarial training, we once again try the attack described earlier in this section for the same 10,000 iterations. Figure 8 shows the convergence of the attack against both the initial and adversarially trained classifier for two values of  $\eta^2$ , showing the inefficacy of the attack on the adversarially trained classifier. After 10,000 iterations, **100%** of the images were valid, but with **22%** of them inducing different classification, and an average KL divergence of **0.08**, showing that the classifier has indeed significantly softened its decision boundary.

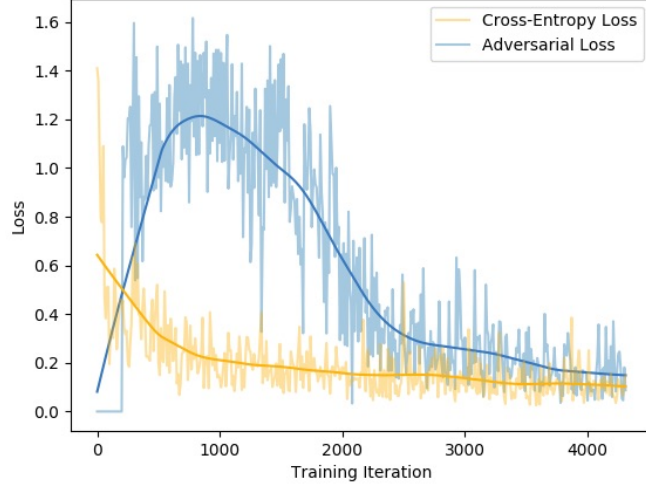


Figure 7: The cross-entropy and adversarial components of the loss decaying as training continues.

Though causing softer decision boundaries, the adversarial training does not significantly impact classification accuracy over the standard classifier: on normal input data, the model achieves the same **97%** accuracy undefended. We also feed the “adversarial” inputs generated by the min-max attack on the initial classifier into the adversarially trained classifier, and observe that the average classification divergence between examples drops to **0.007**, with only **18%** of the valid images being classified inconsistently. Figure 9 shows a randomly selected subset of these examples with their respective classifier output.

## 4.4 Deep Image Prior

### 4.4.1 Adversarial Attack

The adversarial attacks to the ResNet were constructed using 20 steps of the Basic Iterative Method [26]. This attack is given by

$$x_{t+1} = \Pi_{x_0, \epsilon}(x_t + 2 \cdot \epsilon \cdot \text{sign} \nabla_x L(y; C_\theta(x))|_{x=x_t}), \quad t \in \{0, 1, \dots, 19\},$$

where  $x_0$  is the original image and  $y$  is its label;  $L(y; C_\theta(x_t))$  is the cross entropy loss between the label  $y$  and the classifier’s prediction on  $x_t$ ;  $\Pi_{x_0, \epsilon}(x)$  is a clipping such that  $x$  lies in an  $\ell_\infty$  ball of radius  $\epsilon$ , centered at  $x_0$ .

We evaluated the robustness of the Deep Image Prior defense for  $\epsilon = 0.01, 0.05, 0.10$ . (*Note:* Our original images were rescaled such that each pixel lies in the range  $[-1.0, 1.0]$ ).

### 4.4.2 Early Stopping

For  $\epsilon = 0.01, 0.05$ , we run at most 500 iterations of DIP, while for  $\epsilon = 0.10$  we run at most 100 iterations. We also stop the optimization procedure if the Mean Square Error falls below a threshold of 0.005.

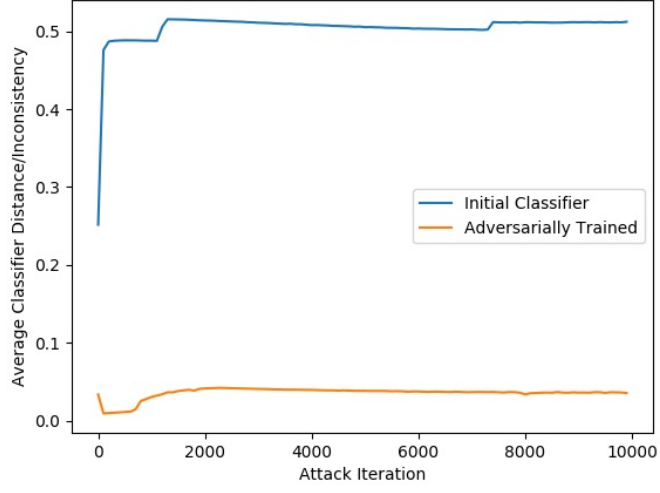


Figure 8: The average  $\|C(G(z)) - C(G(z'))\|_2$  for pairs  $(z, z')$  found by the attack.

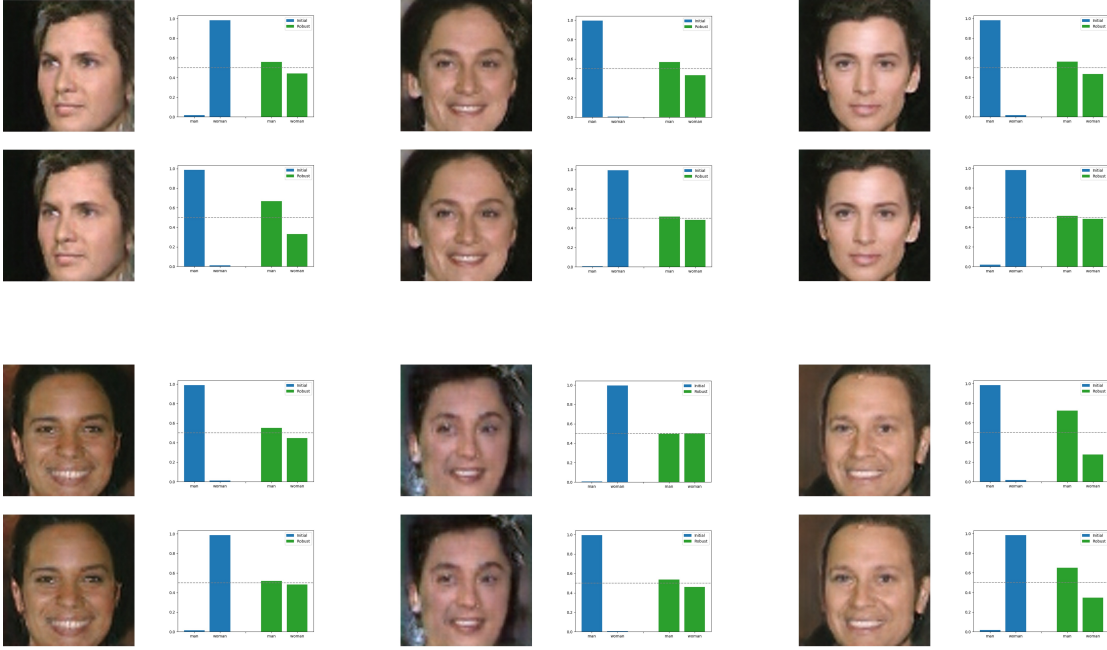


Figure 9: The softmax output of both the original (blue) and robust adversarially trained (green) classifier on the "borderline" images generated by the attack on the non-robustified classifier.



Note that running only 100 iterations leads to a decrease in accuracy for images that are *not adversarial*. If we run DIP for more than 100 iterations on adversarial images that have  $\epsilon = 0.1$ , then the adversarial perturbation is also reconstructed, and the reconstruction remains adversarial. This can be attributed to the low Signal to Noise Ratio that results from adding an adversarial perturbation with  $\epsilon = 0.1$ .

$\epsilon$	Clean Images		Adversarial Images	
	Unprotected Classifier	DIP Protected Classifier	Unprotected Classifier	DIP Protected Classifier
0.01	71%	52%	5%	49%
0.05	71%	52%	2%	40%
0.1	71%	35%	1%	30%

Table 2: This table shows the percentage of 1000 random images that were correctly classified by a ResNet152. The accuracy of the unprotected classifier and the Deep Image Prior protected classifier on clean images are reported in columns 2 and 3 respectively; the accuracy of the unprotected classifier and the Deep Image Prior protected classifier on adversarial examples are reported in columns 4 and 5 respectively. The adversarial images were constructed by performing the Basic Iterative Method[26] for  $\epsilon = 0.01, 0.05, 0.1$ , where  $\epsilon$  is the  $\ell_\infty$ -norm of the adversarial perturbation.

#### 4.4.3 Quantitative Results

Table 2 reports the accuracy of a ResNet152 against adversarial examples constructed using the Basic Iterative Method for varying  $\epsilon$ . DIP Protected Classifier refers to first reconstructing images via the Deep Image Prior method, followed by classification using a ResNet152.

#### 4.4.4 Qualitative Results

Figure 10 shows the reconstructions obtained using the Deep Image Prior method on original and adversarial images. The adversarial images were generated using 20 steps of the Basic Iterative Method with  $\epsilon = 0.05$ . The Deep Image Prior method was run for 500 iterations or until the Mean Square Error fell below 0.005.

## 5 Related work

There is currently a deluge of recent work on adversarial attacks and defenses. Common defense approaches involve modifying the training dataset such that the classifier is made more robust [20], [41], modifying the network architecture to increase robustness [11] or performing defensive distillation [37]. The idea of adversarial training [18] and its connection to robust optimization [41, 32, 43] leads to a fruitful line of defenses. On the attacker side, Carlini and Wagner [8], [10] show different ways of overcoming many of the existing defense strategies.

Our approach to defending against adversarial examples leverages the power of GANs [17]. The GAT-Trainer work by Lee et al. [28] uses generative models to perform adversarial training but in a very different way from our work and further without projecting on the range of a GAN. MagNet [34] and APE-GAN [42] have the similar idea of denoising adversarial noise using a generative model

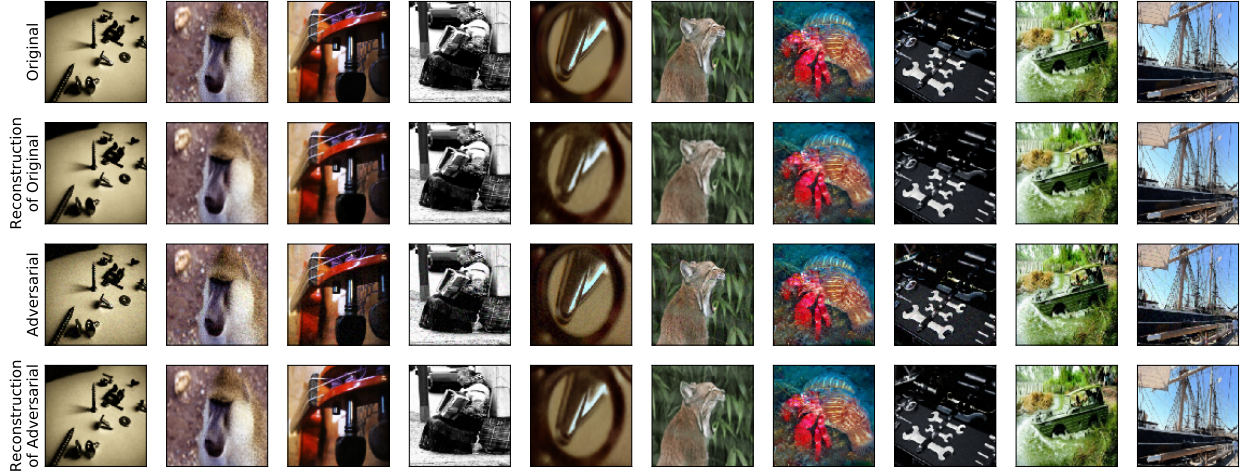


Figure 10: Figure showing the reconstructions obtained via the Deep Image Prior method. The top two rows show the original images and their reconstructions, while the bottom two rows show adversarially perturbed images and their reconstructions. The Deep Image Prior method was run for at most 500 iterations, and we stopped early if the MSE fell below  $5 \times 10^{-3}$ . The adversarial images were generated using 20 steps of the Basic Iterative Method with  $\epsilon = 0.05$ , which is the  $\ell_\infty$ -norm of the adversarial perturbation.

but use differentiable projection methods that have already been successfully attacked by Carlini and Wagner [9].

While we were writing this paper we found two related submissions appearing online: The most closely related concurrent work is the DefenseGAN [3], submitted to ICLR, that proposes a very similar method to INC, independently from our work. However, the current manuscript [3] only validates on MNIST, does not discuss Min-Max attack, the robust process or the Deep image prior method. The second related paper is PixelDefend [44]. The main difference of our work to this paper uses PixelCNN generators as opposed to GANs and hence the projection, attack and defense processes are different.

## 6 Conclusion

This work demonstrates the possibility of resisting adversarial attacks using a generative model. We propose the Invert-and-Classify (INC) algorithm, based on the idea of projecting inputs  $x$  into the range of a trained Generative Adversarial Network  $G$  before classification. The INC projection is performed using Gradient Descent in  $z$ -space which is a non-differentiable process.

We demonstrate the mechanism’s ability to resist both off-the-shelf and specifically designed first-order and black-box attacks. Then, through a crafted min-max optimization, we demonstrate that there are still adversarial images in the range of the GAN. These points are very close yet induce drastically different classifications. These points show that a classifier can be tremendously confident and disagree with human judgement. We show how to solve this problem by adversarially training on these inputs and obtain a robust INC model that displays natural uncertainty around decision boundaries.

Finally, for the cases when no pre-trained generative model is available, we propose the Deep Image Prior (DIP) INC defense. This relies on a structural prior given by an untrained generator to defend against adversarial examples. We show that this allows for a defense for the Imagenet dataset that is robust to first-order methods against the unprotected classifier.

## References

- [1] Martín Abadi, Ashish Agarwal, Paul Barham, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016.
- [2] Anonymous. Adversarial spheres. *ICLR Submission, available on OpenReview*, 2017.
- [3] Anonymous. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. *ICLR Submission, available on OpenReview*, 2017.
- [4] Anonymous. Spatially transformed adversarial examples. *ICLR Submission, available on OpenReview*, 2017.
- [5] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.
- [6] David Berthelot, Tom Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.
- [7] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. *arXiv preprint arXiv:1703.03208*, 2017.
- [8] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *arXiv preprint arXiv:1705.07263*, 2017.
- [9] Nicholas Carlini and David Wagner. MagNet and “efficient defenses against adversarial attacks” are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017.
- [10] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017.
- [11] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863, 2017.
- [12] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.
- [13] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.
- [14] Ethan R Elenberg, Alexandros G Dimakis, Moran Feldman, and Amin Karbasi. Streaming weak submodularity: Interpreting neural networks on the fly. *Advances in Neural Information Processing Systems (NIPS)*, 2017.

- [15] Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.
- [16] Ivan Evtimov, Kevin Eykholt, Earlene Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*, 1, 2017.
- [17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [19] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.
- [20] Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [22] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defenses: Ensembles of weak defenses are not strong. *arXiv preprint arXiv:1706.04701*, 2017.
- [23] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [25] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. *arXiv preprint arXiv:1702.06832*, 2017.
- [26] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.
- [27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [28] Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. Generative adversarial trainer: Defense to adversarial perturbations with gan. *arXiv preprint arXiv:1705.03387*, 2017.
- [29] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

- [30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *stat*, 1050:9, 2017.
- [33] Dongyu Meng and Hao Chen. MagNet: a two-pronged defense against adversarial examples. *CoRR*, abs/1705.09064, 2017.
- [34] Dongyu Meng and Hao Chen. MagNet: a two-pronged defense against adversarial examples. *arXiv preprint arXiv:1705.09064*, 2017.
- [35] Alan D Miller and Ronen Perry. The reasonable person. *New York University Law Review*, 87:323, 2012.
- [36] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*, 2016.
- [37] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016.
- [38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.
- [40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- [41] Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of neural nets through robust optimization. *arXiv preprint arXiv:1511.05432*, 2015.
- [42] S. Shen, G. Jin, K. Gao, and Y. Zhang. Ape-gan: Adversarial perturbation elimination with gan. *ICLR Submission, available on OpenReview*, 2017.
- [43] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

- [44] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017.
- [45] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [46] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.
- [47] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. *arXiv preprint arXiv:1711.10925*, 2017.
- [48] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. *arXiv preprint arXiv:1707.06728*, 2017.

## A Autoencode-and-Classify

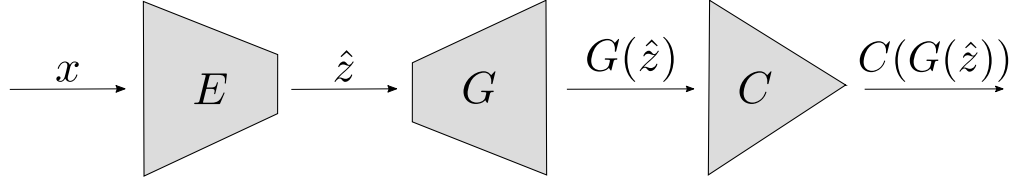


Figure 11: Schematic showing a strategy to detect the presence of adversarial perturbations. Given an input image  $x$  that may be adversarially perturbed, the image we input to the classifier is  $G(E(x))$ . If  $x$  has been adversarially perturbed, then  $\|x - G(E(x))\|_2$  is high.

We now introduce a strategy that can be useful in detecting the presence of adversarial perturbations. Certain generative models, like VAEs [24], BiGANs [12], ALI [13], have an autoencoder structure. In these models, an encoder  $E : \mathbb{R}^n \rightarrow \mathbb{R}^k$  and generator  $G : \mathbb{R}^k \rightarrow \mathbb{R}^n$  are learned jointly and are designed to be approximate inverses of each other.  $E$  maps an input image  $x$  to a seed value  $z = E(x)$ , and the generator  $G$  maps the seed value  $z$  to a reconstruction  $\hat{x}$ , such that  $\hat{x} \approx x$ .

MagNet [34] and APE-GAN [42] propose closely related ideas. Unfortunately, having a differentiable denoiser (e.g. an encoder) to protect from adversarial examples is problematic since an attacker can design new input attacks by back-propagating through the encoder. This was shown by the Carlini and Wagner [9] attacks on MagNet and APE-GAN. This approach also shows that one can attack an encoder to generate an image from a different class [25].

The final prediction on an input  $x$  is given by  $C(G(E(x)))$ . Hence we can view  $C_1 = C \circ G \circ E$  as a feedforward classifier for which we can construct an attack. In this case, equation 8 can be modified to

$$\delta = +\epsilon \cdot \text{sign}(\nabla_X L(y, C_1(X))|_{X=x}), \quad (12)$$

where  $x$  is the original image and  $y$  is its label;  $L(y; C_1(x))$  is the cross entropy loss between the label  $y$  and the classifier prediction on  $x$ .

We observe that adversarial attacks drive the encoding of images away from the typical set. If an adversarial example  $(x + \delta)$  is encoded to  $E(x + \delta)$ , then the distribution of  $E(x + \delta)$  differs significantly from the true distribution of the seed values. In most models, seed values to the generator are drawn from  $\mathcal{N}(0, 1)$ . Figure 14 shows the distribution of seed values produced by adversarial examples, and non-adversarial examples.

Figure 13 shows how the  $\ell_2$  distance between input images and images from the generator vary when the input is adversarial or natural.



## Quantitative Results

The accuracy is plotted in Figure 12. Figure 13 shows how the  $\ell_2$  distance between input images and images from the generator vary when the input is adversarial or natural.

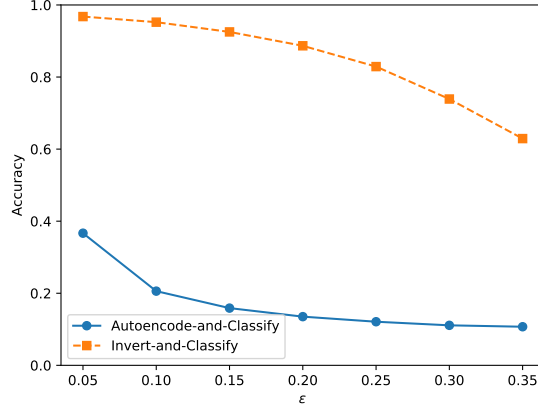


Figure 12: Accuracy of classifier with change in the  $\ell_\infty$  norm of the adversarial perturbation defined in 12. Notice that attacks found from Autoencode-and-classify do not typically transfer to INC.

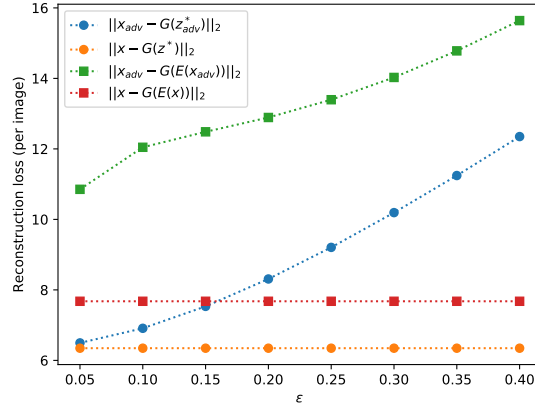


Figure 13: This figure plots the  $\ell_2$  norm of difference between the input and reconstructed image.  $x$  refers to the original image,  $x_{adv} = x + \delta$  is the adversarial example, where  $\delta$  is constructed according to 12.  $z^* = \arg \min_z \|x - G(z)\|_2$  is the inverse of  $x$ , and  $z^*_{adv} = \arg \min_z \|x_{adv} - G(z)\|_2$  is the inverse of  $x_{adv}$ .  $G(E(x))$  is the result when  $x$  is autoencoded by the VAE, while  $G(E(x_{adv}))$  is the result when  $x_{adv}$  is autoencoded by the VAE. Notice that a high  $\ell_2$  error can be used to detect whether an image is adversarial or not.

## Qualitative Results

If an adversarial example  $(x + \delta)$  is encoded to  $E(x + \delta)$ , then the distribution of  $E(x + \delta)$  differs significantly from the true distribution of the seed values, which in our case are distributed according to  $\mathcal{N}(0, 1)$ . Figure 14 shows the distribution of seed values produced by adversarial examples, and non-adversarial examples.

Figure 15 shows the images obtained by autoencoding  $x_{adv}$  versus  $G(z_{adv})$ , where  $z_{adv}$  is the inversion of  $x_{adv}$ .

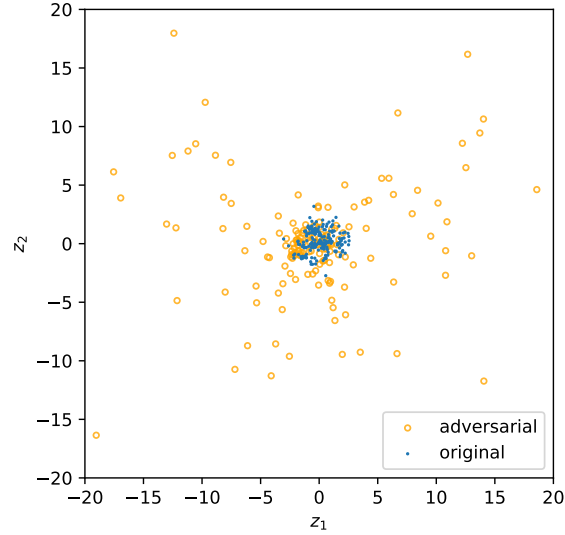


Figure 14: The blue dots denote how non adversarial examples are encoded by  $E$ . The orange dots denote how adversarial examples are encoded by  $E$ . In this case  $E$  and  $G$  are the encoder and decoder of a VAE such that the seed  $z$  has dimension 2.

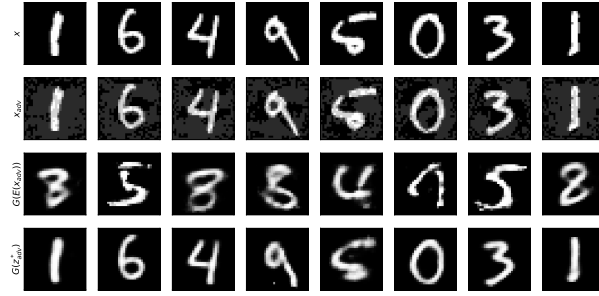


Figure 15:  $x$  are the original images,  $x_{adv}$  are the adversarial examples obtained from  $x$  (the adversarial perturbation has  $\|\delta\|_\infty = 0.1$ ),  $G(E(x_{adv}))$  are the results of autoencoding  $x_{adv}$ , and  $G(z_{adv}^*)$  are the images obtained by inverting  $x_{adv}$ . It is interesting to note here that the attack on Autoencode-and-classify is actually attacking the encoder: there actually exist better codes  $z^*$  that produce images  $G(z^*)$  that are closer to the input  $x_{adv}$ , compared to  $G(E(x_{adv}))$ . These images  $G(z^*)$  are closer and actually *of the correct class* but the adversarial noise fools the encoder in producing further images from a different class. Gradient descent projection used by INC does not have this problem.