**EE 381V: Special Topics on Unsupervised Learning** **Spring 2018**

# Attacking Robust Manifold Defense

*Professor: Alex Dimakis* *Students: Kwon Jeongyeol, Dany Haddad, Justin Lewis*

## 1 Synopsis

The success of adversarial attacks has brought increased scrutiny regarding the robustness of neural networks. A variety of defense methods have been proposed, some of the more successful techniques are closely related to the Invert and Classify (INC) approach. INC uses a well-trained generative model to project a potentially adversarial image onto the manifold of natural images. This sister-image can now be safely classified as it no longer contains adversarial components which are far from the manifold. This process is not directly differentiable and as a result is seemingly difficult to attack. In response, our proposal is the following. Given the well-trained GAN used for the Manifold Defense, train a matching encoder. Together, the GAN and encoder form an autoencoder which is now directly differentiable and susceptible to adversarial attack.

## 2 Background

The following subsections will introduce basic notation for deep learning frameworks. Additionally the notion of adversarial attack and training will be made clear.

### 2.1 Notation

Let our dataset be denoted by $\{\mathcal{X}, \mathcal{Y}\}$, a set of images $\mathcal{X}$ and corresponding class labels $\mathcal{Y}$

Classifier: $\mathcal{C}_\theta$

GAN: generator $\mathcal{G}_\phi$ and discriminator $\mathcal{D}_\psi$

AE: encoder $\mathcal{E}_\alpha$ and decoder $D_\beta$

$x$: a natural image

$x_{adv}$: an adversarial image

### 2.2 Adversarial Attacks and Training

The notion of adversarial attacks is a phenomenon brought to light in the context of deep learning by Goodfellow et al. [2]. In essence, they argue that deep learning classifiers are less robust than their celebrated performance assumes and are susceptible to attack from an adversarial user. Given image $x$ with correct label $y$ and trained classifier $\mathcal{C}_\theta$ with loss function $L(x, y; \theta)$, a canonical adversarial attack is conducted as follows:

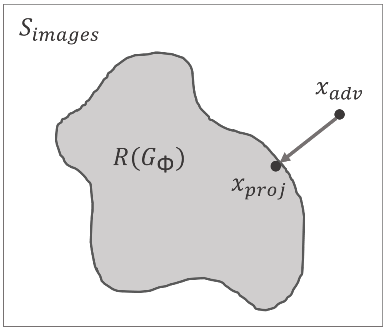
$$\begin{aligned} \underset{\delta}{\text{maximize}} \quad & L(x+\delta, y; \theta) \\ \text{subject to} \quad & \|\delta\|_p \leq \epsilon \end{aligned}$$

Plainly said, find adversarial image $x_{adv} = x + \delta$ within an $\epsilon$ sized $\ell_p$ ball around the original image which fools the classifier the most. Commonly, this optimization problem is approximately solved using one of two approaches: Fast Gradient Sign Method (FGSM) and projected gradient descent (PGD). It has been empirically verified that FGSM attacks are less effective at fooling a classifier than PGD approaches [4]. Note, a suitable loss function $L(x, y; \theta)$ might be cross-entropy loss for multi-label classification.

In order to defend against adversarial attacks, a number of defense strategies have been employed [1]. The most straightforward defense strategies introduce adversarial examples $\mathcal{X}_{adv}$ into the set of training images $\mathcal{X}$, which allows the network to learn to correctly classify even these examples. Other methods try to ”detect” adversarial examples. Essentially, these defense strategies give no clear guarantees and the process of attack-defend becomes an arms-race.

# 3 Invert and Classify Defense

The defense strategy of particular interest to our proposal is the Invert and Classify (INC) approach as introduced by [3] and a very similar technique developed by [5]. As mentioned before, INC leverages the strong representation power of GANs. The idea is to project an adversarial image onto the range of a GAN and classify this ’cleaned’ image. This is done using gradient descent. If the GAN has approximated the manifold of natural images well enough, and the gradient method projects onto the manifold well enough, then the INC strategy is feasible.



$S_{images}$ : the space of all images

$R(G_{\phi})$ : the range of the GAN

# 4 Attacking INC

Similar to the Backward Pass Differentiable Approximation (BPDA) method of Athalye et al., we propose to attack the INC defense mechanism by building a differentiable approximation for the process of projecting onto the manifold of natural images. Using this approximation, we can evaluate a gradient and perform gradient based attacks (FGSM and PGD). We learn this approximation by training an encoder that corresponds to the generative model used in the INC process. In this way, passing an image through the encoder and then the generative model will give us an image on the learned manifold which will serve as an approximation to the projection process. To learn this encoder, we minimize the following loss with respect to the parameters of the encoder, $E$:

$$\mathbb{E}[||E(G(Z)) - Z||_2^2] \tag{1}$$

### 4.1 Preliminary Results

We ran some experiments using the MNIST dataset. The adversarial images $\mathcal{X}_{\dashv\lceil\sqsubseteq}$ were generated through FGSM attacks on the unprotected classifier $\mathcal{C}_\theta$. The classification accuracy results are presented below.

- Clean test data: 98%
- Adversarial test set using FGSM and $\epsilon = 0.2$: 19%
- Adversarial test set against INC protected classifier: 87%
- Adversarial test set against fitted encoder protected classified: 75%

Using our BPDA style attack on an INC protected classifier (differentiating through an encoder fitted to the GAN used in the INC process) we did not arrive at the results that we had hoped for: the accuracy of the INC protected classifier remained at 93%. Shown below are the original images, the adversarial examples created via FGSM and the images resulting from the autoencoding process.

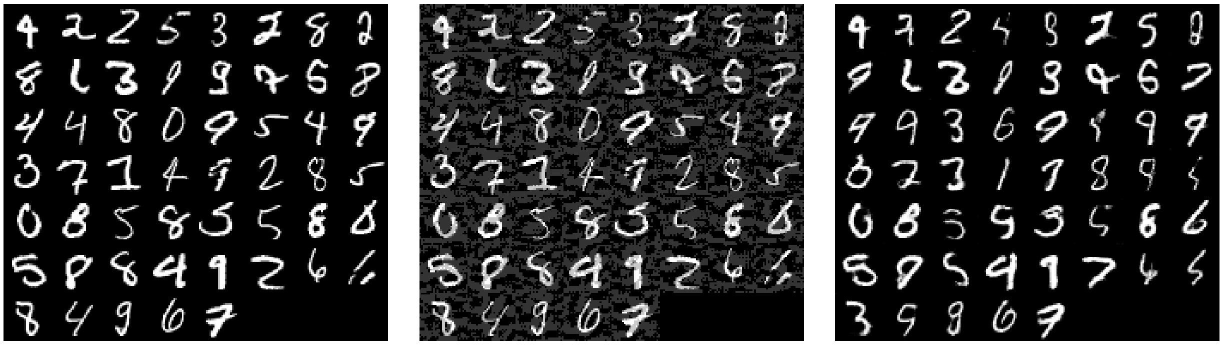


Figure 1: Some failure cases in MNIST test data, the adversarial examples created via FGSM and the images resulting from the autoencoding process

### 4.2 Paths Forward

It is likely that the reason this attack fails is that the encoder does not serve as a close enough approximation of the projection step. It would be interesting to see if for certain synthetic datasets our attack will be effective.

Additionally, we have also considered using the fitted encoder to bootstrap an attack which unrolls the projection process.

# 5 Using Fitted Encoder to Improve INC

Does the failure of the attack using the fitted encoder mean that the fitted encoder is useless? During experiment, we observed interesting phenomenon.

## 5.1 Learned Encoder vs. Projection

As stated earlier, projection step in INC solves the optimization problem that minimizes $||G(z) - X||_2$. However, the non-convex nature of this problem requires us to start over many randomly initialized points and pick the best one. It might be computationally very expensive to iterate over and over again until getting a satisfactory reconstruction, sometimes making this defense mechanism impractical to use, as suggested in their experiment with CIFAR-10 data.

From our experiment, we observed that a latent point mapped by fitted encoder becomes a very good starting point to launch the optimization procedure. We speculate that the learned encoder at least gives us very good reference for each test image and even for corrupted image.

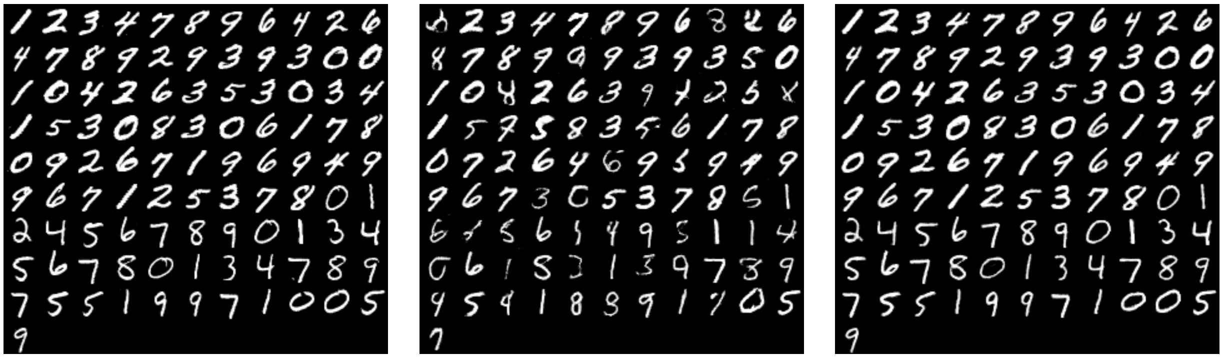


Figure 2: Some MNIST test images, reconstruction from random z, and encoded z.

# References

[1] A. Athalye, N. Carlini, and D. Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *ArXiv e-prints*, February 2018.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. *ArXiv e-prints*, December 2014.

[3] A. Ilyas, A. Jalal, E. Asteri, C. Daskalakis, and A. G. Dimakis. The Robust Manifold Defense: Adversarial Training using Generative Models. *ArXiv e-prints*, December 2017.

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. *ArXiv e-prints*, June 2017.

[5] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman. PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples. *ArXiv e-prints*, October 2017.