
AT-GAN: An Adversarial Generator Model for Non-constrained Adversarial Examples

Xiaosen Wang^{*1} Kun He^{*1} Chuanbiao Song¹ Liwei Wang² John Hopcroft³

Abstract

Despite the rapid development of adversarial machine learning, most adversarial attack and defense researches mainly focus on the perturbation-based adversarial examples, which is constrained by the input images. In comparison with existing works, we propose non-constrained adversarial examples, which are generated entirely from scratch without any constraint on the input. Unlike perturbation-based attacks, or the so-called unrestricted adversarial attack which is still constrained by the input noise, we aim to learn the distribution of adversarial examples to generate non-constrained but semantically meaningful adversarial examples. Following this spirit, we propose a novel attack framework called AT-GAN (Adversarial Transfer on Generative Adversarial Net). Specifically, we first develop a normal GAN model to learn the distribution of benign data, and then transfer the pre-trained GAN model to estimate the distribution of adversarial examples for the target model. In this way, AT-GAN can learn the distribution of adversarial examples that is very close to the distribution of real data. To our knowledge, this is the first work of building an adversarial generator model that could produce adversarial examples directly from any input noise. Extensive experiments and visualizations show that the proposed AT-GAN can very efficiently generate diverse adversarial examples that are more realistic to human perception. In addition, AT-GAN yields higher attack success rates against adversarially trained models under white-box attack setting and exhibits moderate transferability against black-box models.

1. Introduction

Deep Neural Networks (DNNs) have exhibited impressive performance on various computer vision tasks (Krizhevsky et al., 2012; He et al., 2016). However, DNNs are also found vulnerable to adversarial examples (Szegedy et al., 2014), which has raised serious concerns on the safety of deep learning models. For instance, Eykholt et al. (2018) and Cao et al. (2019) have found that adversarial examples can mislead the self-driving cars to make wrong decisions on traffic signs and obstacles.

In recent years, numerous works of exploring adversarial examples have been developed, including adversarial attack (Goodfellow et al., 2015; Carlini & Wagner, 2017), adversarial defense (Goodfellow et al., 2015; Kurakin et al., 2017; Song et al., 2019) and the properties of adversarial examples (He et al., 2018; Shamir et al., 2019). However, most studies mainly focus on the *perturbation-based adversarial examples* constrained by the input images. In contrast, Song et al. (2018) propose to generate *unrestricted adversarial examples* using Generative Adversarial Net (GAN) (Goodfellow et al., 2014) by searching adversarial noise around the input noise. However, their method is still constrained by the input noise, and time-consuming.

In this paper, we propose a novel attack framework called AT-GAN (Adversarial Transfer on Generative Adversarial Net), which aims to learn the distribution of adversarial examples and generate a new type of adversarial examples, called *non-constrained adversarial examples*. Unlike traditional perturbation-based attacks that search adversarial examples around the input image, or the unrestricted adversarial attack (Song et al., 2018) that searches adversarial noise around the input noise, AT-GAN is an adversarial generator model that could produce semantically meaningful adversarial examples directly from any input noise. This to our knowledge is the first work of this category.

Specifically, we first develop a normal GAN model to learn the distribution of benign data, then transfer the pre-trained GAN model to estimate the distribution of adversarial examples for the target model. Note that once our generator is transferred from generating normal images to adversarial images, it can directly generate non-constrained adversarial

^{*}Equal contribution ¹School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China ²School of Electronics Engineering and Computer Sciences, Peking University, Peking, China ³Department of Computer Science, Cornell University, NY, USA. Correspondence to: Kun He <brooklet60@hust.edu.cn>, Xiaosen Wang <xiaosen@hust.edu.cn>.

examples for any input random noise, leading to high diversity and efficiency. In general, AT-GAN is not only a new way of generating adversarial examples, but also a useful way to explore the distribution of adversarial examples.

To evaluate the effectiveness of AT-GAN, we develop AT-GAN on three benchmark datasets, MNIST, Fashion-MNIST and CelebA, and apply typical defense methods to compare AT-GAN with other perturbation-based or unrestricted attack methods. Empirical results show that the non-constrained adversarial examples generated by AT-GAN yields higher attack success rates, and state-of-the-art adversarially trained models exhibit little robustness against AT-GAN, indicating the high diversity of the generated examples. Besides, AT-GAN is more efficient than the gradient-based or query-based adversarial attacks in generating adversarial instances.

2. Preliminaries

In this section, we first give the definitions of different types of adversarial examples, then introduce Generative Adversarial Net (GAN) that will be used in our method, and provide more details of unrestricted adversarial examples which are most related to our work. The related works on typical adversarial attack and defense methods are introduced in Appendix A.

Let \mathcal{X} be the set of the legal images, $\mathcal{Y} \in \mathbb{R}$ be the output label space and p_z be an arbitrary probability distribution, such as Gaussian distribution. The classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$ takes an image x and predicts its label $f(x)$. Suppose p_x and p_{adv} are the distributions of benign and adversarial examples respectively.

2.1. Adversarial Examples

Assume we have an oracle $o : \mathcal{X} \rightarrow \mathcal{Y}$, which could always predict the correct label for any input $x \in \mathcal{X}$, we can define several types of adversarial examples as follows:

For perturbation-based adversarial examples, tiny perturbations are added to input images, which are undetectable by human but can cause the target classifier to make wrong prediction.

Definition 1. Perturbation-based Adversarial Examples. Given a subset (trainset or testset) $\mathcal{T} \subset \mathcal{X}$ and a small constant $\epsilon > 0$, the perturbation-based adversarial examples can be defined as: $\mathcal{A}_p = \{x_{adv} \in \mathcal{X} | \exists x \in \mathcal{T}, \|x - x_{adv}\|_p < \epsilon \wedge f(x_{adv}) \neq o(x_{adv}) = f(x) = o(x)\}$.

Song et al. (2018) define a new type of adversarial examples, called unrestricted adversarial examples, which add adversarial perturbation to the input noise of a mapping like GAN so that the output of the perturbed noise is an adversary for the target classifier.

Definition 2. Unrestricted Adversarial Examples. Given a mapping G from $z \sim p_z$ to $G(z) \sim p_\theta$, where p_θ is an approximation distribution of p_x , and a small constant $\epsilon > 0$, unrestricted adversarial examples can be defined as: $\mathcal{A}_u = \{G(z^*) \in \mathcal{X} | \exists z \sim p_z, z^* \sim p_z, \|z - z^*\|_p < \epsilon \wedge f(G(z^*)) \neq o(G(z^*)) = f(G(z)) = o(G(z))\}$.

In this work, we define a new type of adversarial examples, called non-constrained adversarial examples, in which we train a mapping like GAN to learn the distribution of adversarial examples so as to generate adversaries from any noise.

Definition 3. Non-constrained Adversarial Examples. Given a mapping G^* from $z \sim p_z$ to $G^*(z) \sim q_\theta$, where q_θ is an approximation distribution of p_{adv} , the non-constrained adversarial examples can be defined as: $\mathcal{A}_n = \{G^*(z) \in \mathcal{X} | o(G^*(z)) \neq f(G^*(z))\}$.

In summary, perturbation-based adversarial examples are based on perturbing an image $x \in \mathcal{X}$, and unrestricted adversarial examples are based on perturbing an input noise $z \sim p_z$ for an existing mapping G . In contrast, non-constrained adversarial examples are more generalized that we learn a mapping G^* such that for any input noise sampled from distribution p_z , the output can fool the classifier. It is clear that $\mathcal{A}_p \subset \mathcal{A}_u \subset \mathcal{A}_n$ from the above definitions.

2.2. Generative Adversarial Net

Generative Adversarial Net (GAN) (Goodfellow et al., 2014) consists of two neural networks, G and D , trained in opposition to each other. The generator G is optimized to estimate the data distribution and the discriminator D aims to distinguish fake samples from G and real samples from the training data. The objective of D and G can be formalized as a min-max value function $V(G, D)$:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x \sim p_x} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))]$$

Deep Convolutional Generative Adversarial Net (DCGAN) (Radford et al., 2016) is the convolutional version of GAN, which implements GAN with convolutional networks and stabilizes the training process. Auxiliary Classifier GAN (AC-GAN) (Odena et al., 2017) is another variant that extends GAN with some conditions by an extra classifier C . The objective function of AC-GAN can be formalized as follows:

$$\begin{aligned} \min_G \max_D \min_C V(G, D, C) = & \mathbb{E}_{x \sim p_x} [\log D(x)] \\ & + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z, y)))] \\ & + \mathbb{E}_{x \sim p_x} [\log(1 - C(x, y))] \\ & + \mathbb{E}_{z \sim p_z} [\log(1 - C(G(z, y), y))]. \end{aligned}$$

To make GAN more trainable in practice, Arjovsky et al. (2017) propose Wasserstein GAN (WGAN) that uses Wasserstein distance so that the loss function has more desirable properties. Gulrajani et al. (2017) introduce WGAN with gradient penalty (WGAN-GP) which outperforms WGAN in practice. Its objective function is formulated as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_x}[D(x)] - \mathbb{E}_{z \sim p_z}[D(G(z))] - \lambda \mathbb{E}_{\hat{x} \sim p_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2],$$

where $p_{\hat{x}}$ is uniformly sampling along straight lines between pairs of points sampled from the data distribution p_x and the generator distribution p_g .

2.3. Unrestricted Adversarial Examples

Song et al. (2018) propose to construct Unrestricted Adversarial Examples using AC-GAN. Specifically, given an arbitrary noise vector z^0 , their method aims to search a noise input z^* in the ϵ -neighborhood of z^0 for AC-GAN so as to produce an adversarial example $G(z^*, y_s)$ for the target model f , whereas $G(z^*, y_s)$ can be still classified as y_s for the extra classifier C . The objective function can be written as:

$$\begin{aligned} z^* = \arg \min_z & \left\{ \frac{\lambda_1}{m} \sum_{i=1}^m \max(|z_i - z_i^0| - \epsilon, 0) \right. \\ & - \lambda_2 \log C(y_s | G(z^*, y_s)) \\ & \left. - \max_{y \neq y_s} \log f(y | G(z^*, y_s)) \right\}, \end{aligned}$$

where $C(y|x)$ (or $f(y|x)$) denotes the confidence of prediction label y for the given input x on classifier C (or f).

Though $G(z^*, y_s)$ can be classified correctly by C , it cannot assure that $G(z^*, y_s)$ is a realistic image for human eye (Nguyen et al., 2015). Therefore, their method needs human evaluation for the generated adversarial examples.

3. The Proposed AT-GAN

In this section, we first introduce the estimation on the distribution of adversarial examples, then propose the AT-GAN framework to generate non-constrained adversarial examples, and provide further analysis that AT-GAN can learn the distribution of adversarial examples in the end.

3.1. Estimating the Adversarial Distribution

In order to generate non-constrained adversarial examples, we need to estimate the distribution of adversarial examples $p_{adv}(x_{adv}|y)$. Given the parameterized estimated distribution $q_\theta(x)$, we can define the estimation problem as

$$q_{\theta^*}(x_{adv}|y) = \arg \min_{\theta \in \Omega} KL(q_\theta(x_{adv}|y) \| p_{adv}(x_{adv}|y)), \quad (1)$$

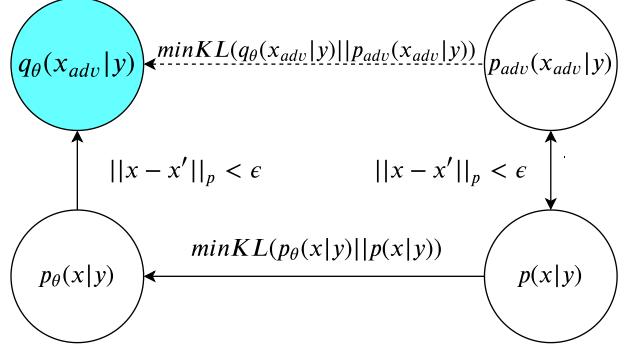


Figure 1. Estimating the distribution of adversarial examples q_θ in two stages: 1) estimate the distribution of benign data p_θ . 2) transfer p_θ to estimate q_θ .

where θ indicates trainable parameters and Ω is the parameter space.

It is hard to calculate Eq. 1 directly as $p_{adv}(x_{adv}|y)$ is unknown. Inspired by the perturbation-based adversarial examples, as shown in Figure 1, we postulate that for each adversarial example x_{adv} , there exists some benign examples x where $\|x - x_{adv}\|_p < \epsilon$. In other words, $p_{adv}(x_{adv}|y)$ is close to $p(x|y)$ to some extent. By Bayes' theorem, we have

$$p(x|y) = \frac{p(y|x) \cdot p(x)}{p(y)}.$$

Thus, we can approximately solve Eq. 1 in two stages: 1) Fit the distribution of benign data p_θ . 2) Transfer p_θ to estimate the distribution of adversarial examples q_θ .

Specifically, we propose a new adversarial attack model called AT-GAN to learn the distribution of adversarial examples. The overall architecture of AT-GAN is illustrated in Figure 2. Corresponding to the above two training stages, we first train the GAN model to get a generator $G_{original}$ to learn p_θ , then we transfer $G_{original}$ to attack f_{target} for the learning of q_θ .

3.2. Training the Original Generator

Figure 2 (a) illustrates the overall architecture of AC-WGAN-GP that we used as the normal GAN. AC-WGAN-GP is the combination of AC-GAN (Odena et al., 2017) and WGAN-GP (Gulrajani et al., 2017), composed by three neural networks: a generator G , a discriminator D and a classifier f . The generator G takes a random noise z and a label y as the inputs and generates an image $G(z, y)$. It aims to generate an image $G(z, y)$ that is indistinguishable to discriminator D and makes the classifier f to output label y . The loss function of G can be formulated as:

$$\begin{aligned} L_G(z, y) = & \mathbb{E}_{z \sim p_z(z)} [H(f(G(z, y)), y)] \\ & - \mathbb{E}_{z \sim p_z(z)} [D(G(z))]. \end{aligned}$$

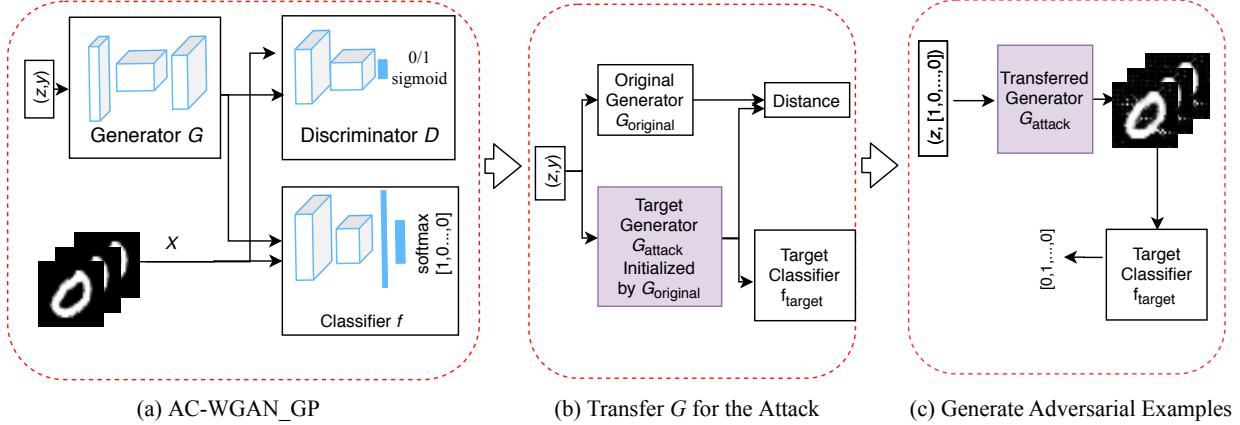


Figure 2. The architecture of AT-GAN. The first training stage of AT-GAN is similar to that of AC-WGAN_GP. After G is trained, we regard G as the original model $G_{original}$ and transfer $G_{original}$ to train the adversarial generator G_{attack} to fool the target classifier. After the second stage of training, AT-GAN can generate adversarial examples by G_{attack} .

Here $H(a, b)$ is the entropy between a and b . The discriminator D takes the training data x or the generated data $G(z, y)$ as the input and tries to distinguish them. The loss function of D with gradient penalty for samples $\hat{x} \sim p_{\hat{x}}$ can be formulated as:

$$\begin{aligned} L_D(x, z, y) = & -\mathbb{E}_{x \sim p_{data}(x)}[D(x)] \\ & + \mathbb{E}_{z \sim p_z(z)}[D(G(z, y))] \\ & + \lambda \mathbb{E}_{\hat{x} \sim p_{\hat{x}}(\hat{x})}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2]. \end{aligned}$$

The classifier f takes the training data x or the generated data $G(z, y)$ as the input and predicts the corresponding label. There is no difference from other classifiers and the loss function is:

$$\begin{aligned} L_f(x, y) = & \mathbb{E}_{x \sim p_{data}(x)}[H(f(x), y)] + \\ & \mathbb{E}_{z \sim p_z(z)}[H(f(G(z, y)), y)]. \end{aligned}$$

Note that the goal of this stage is to train a generator G which could output realistic samples and estimate the distribution of real data properly so that we could later on transfer the generator G to estimate the distribution of adversarial examples. So one could train a generator G using other methods as long as the generator G could learn a good distribution of the real data.

3.3. Transferring the Generator for Attack

After the original generator $G_{original}$ is trained, we transfer the generator $G_{original}$ to learn the distribution of adversarial examples in order to attack the target model. As illustrated in Figure 2b, there are three neural networks, including the original generator $G_{original}$, the attack generator G_{attack} to be transferred that has the same weights as $G_{original}$ in the beginning, and the classifier f_{target} to be

attacked. The goal of the second stage can be described as:

$$\begin{aligned} G_{attack}^* = & \arg \min_{G_{attack}} \|G_{original}(z, y) - G_{attack}(z, y)\|_p \\ s.t. \quad & f_{target}(G(z, y)) = y_t \neq y, \end{aligned} \quad (2)$$

where y_t denotes the target adversarial label, $\|\cdot\|_p$ denotes the ℓ_p norm and we focus on $p = 2$ in this work.

To optimize Eq. 2, we construct the loss function by L_a and L_d , where L_a aims to assure that f_{target} yields the target label y_t :

$$L_a(z, y) = \mathbb{E}_{z \sim p_z}[H(f_{target}(G_{attack}(z, y)), y_t)], \quad (3)$$

and L_d aims to assure that the adversarial generator G_{attack} generates realistic examples:

$$L_d(z, y) = \mathbb{E}_{z \sim p_z}[\|G_{original}(z, y) + \rho - G_{attack}(z, y)\|_p]. \quad (4)$$

Here ρ is a random noise constrained by both l_0 and l_∞ norm.

The objective function for transferring $G_{original}$ to G_{attack} can be formulated as:

$$L(z, y) = \alpha L_a(z, y) + \beta L_d(z, y), \quad (5)$$

where α and β are hyperparameters to control the training process. Note that in the case that $\alpha = 1$ and $\beta \rightarrow \infty$, the objective function is similar to that of the perturbation-based attacks (Goodfellow et al., 2015; Tramr et al., 2018; Madry et al., 2018).

For the untargeted attack, we can replace y_t in L_a with $\max_{y \neq y_s} C(G_{attack}(z, y_s))$, where $C(\cdot)$ is the logits of the target classifier f_{target} .

3.4. Theoretical Analysis of AT-GAN

In this subsection, we provide theoretical analysis why AT-GAN can generate as realistic and diverse non-constrained adversarial examples as real data. We will prove that under ideal condition, AT-GAN can estimate the distribution of adversarial examples which is close to that of real data as described in Section 3.1.

Suppose p_{data} is the distribution of real data, p_g and p_a are the distribution learned by the generator of AC-WGAN.GP and AT-GAN respectively. For the optimization of Eq. 2, L_d aims to constrain the image generated by G_{attack} in the ϵ -neighborhood of $G_{original}$. We prove that under the ideal condition that L_d guarantees $G_{attack}(z, y)$ to be close enough to $G_{original}(z, y)$ for any input noise z , the distribution of AT-GAN almost coincides the distribution of AC-WGAN.GP. Formally, we state our result for the two distributions as follows.

Theorem 1. Suppose $\max_{z,y} L_d(z, y) < \epsilon$, we have $KL(p_a \| p_g) \rightarrow 0$ when $\epsilon \rightarrow 0$.

The proof of Theorem 1 is in Appendix C. Samangouei et al. (2018) prove that the global optimum of WGAN is $p_g = p_{data}$ and we can show that the optimum of AC-WGAN.GP has the same property. We can formalize the property as follows:

Theorem 2. The global minimum of the virtual training of AC-WGAN.GP is achieved if and only if $p_g = p_{data}$.

The proof of Theorem 2 is in Appendix C. According to Theorem 1 and 2, under the ideal condition, we conclude $p_a \approx p_g = p_{data}$, which indicates that the distribution of adversarial examples learned by AT-GAN is very close to that of real data as discussed in Section 3.1, so that the non-constrained adversarial instances are as realistic and diverse as the real data.

4. Experiments

To validate the effectiveness of AT-GAN, we empirically evaluate our non-constrained adversarial examples on various datasets, and demonstrate that AT-GAN yields higher attack success rates against adversarial training and achieves moderate transferability with higher efficiency. Besides, AT-GAN can learn a distribution of adversarial examples which is close to the real data distribution, and generate realistic and diverse adversarial examples.

4.1. Experimental Setup

Datasets. We consider three standard datasets, namely MNIST (LeCun et al., 1989), Fashion-MNIST (Xiao et al., 2017) and CelebA (Liu et al., 2015). Specifically, MNIST is a dataset of hand written digit number from 0 to 9.

Fashion-MNIST is similar to MNIST with 10 categories of fashion clothes. CelebA is a dataset with more than 200,000 celebrity images, and we group them according to female/male and focus on gender classification as in Song et al. (2018). For all datasets, we normalize the pixel values into range $[0, 1]$.

Baselines. We compare the proposed AT-GAN with typical perturbation-based adversarial attacks, i.e., FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018), R+FGSM (Tramr et al., 2018) and unrestricted adversarial attack (Song et al., 2018).

Models. For MNIST and Fashion-MNIST, we adopt four models used in Tramr et al. (2018), denoted as Model A to D. For CelebA, we consider three models, namely CNN, VGG16 and ResNet. More details about the architectures are provided in Appendix C.

Evaluation Setup. To evaluate the performance of these attacks, we consider normal training and existing strong defenses, namely adversarial training (Goodfellow et al., 2015), ensemble adversarial training (Tramr et al., 2018) and iterative adversarial training (Madry et al., 2018).

All experiments are conducted on a single Titan X GPU. The hyper-parameters used in the experiments are described in Appendix C.

4.2. Evaluation Results

4.2.1. COMPARISON ON ATTACK SUCCESS RATE

To validate the efficacy, we compare AT-GAN with other baseline attack methods under white-box setting. Since Athalye et al. (2018) have shown that the currently most effective defense method is adversarial training, we consider adversarially trained models as the defense models. The attack success rates are reported in Table 1.

On MNIST, AT-GAN is the best attack method against all the defense models. As for Model A, B and C by normal training, AT-GAN gains the second highest attack success rates over 98%. On Fashion-MNIST, AT-GAN achieve the highest attack success rate on average. On CelebA, AT-GAN achieves the best attack performance almost on all models. The only exceptions are that Song et al. (2018) achieves the highest attack rate on VGG, and PGD achieves the highest attack rate on ResNet by normal training respectively. Under the two cases, the results of AT-GAN are still close to the best results.

On average, AT-GAN achieves the highest attack success rate on all defense models. As AT-GAN aims to estimate the distribution of adversarial examples, adversarial training with some specific attacks has little robustness against AT-GAN, raising a new security issue for the development of more generalized adversarially training models.

Table 1. Attack success rate (%) of adversarial examples generated by AT-GAN and the baselines attacks under white-box attack setting against models by normal training and various adversarial training methods. For each model, the highest attack success rate is highlighted in **bold**. Nor.: Normal training, Adv.: Adversarial training, Ens.: Ensemble adversarial training, Iter. Adv.: Iterative adversarial training.

(a) Comparison of AT-GAN and other attacks on MNIST.

Attack	Model A				Model B				Model C				Model D			
	Nor.	Adv.	Ens.	Iter.	Nor.	Adv.	Ens.	Iter.	Nor.	Adv.	Ens.	Iter.	Nor.	Adv.	Ens.	Iter.
Clean Error (%)	0.9	0.9	0.8	0.9	1.0	0.9	1.0	0.9	1.0	1.0	1.0	1.0	2.7	2.6	2.6	1.5
FGSM	68.8	2.9	8.7	4.0	88.0	9.5	18.0	9.2	70.7	4.3	7.8	4.7	89.6	23.5	34.6	26.7
PGD	100.0	92.6	85.1	5.9	100.0	42.4	98.2	36.2	100.0	76.9	96.4	9.6	91.7	96.8	99.5	81.1
R+FGSM	77.9	28.4	20.3	2.6	96.5	7.1	42.1	4.6	81.6	7.2	19.9	2.9	93.8	76.2	51.6	25.3
Song's	82.0	70.5	75.0	84.6	76.9	65.0	72.0	80.7	74.2	75.6	72.6	87.8	67.7	43.6	56.3	44.5
AT-GAN	98.7	97.5	96.7	91.4	99.5	97.7	99.3	95.6	99.3	95.8	96.9	90.0	99.9	99.5	99.7	

(b) Comparison of AT-GAN and other attacks on Fashion-MNIST.

Attack	Model A				Model B				Model C				Model D			
	Nor.	Adv.	Ens.	Iter.	Nor.	Adv.	Ens.	Iter.	Nor.	Adv.	Ens.	Iter.	Nor.	Adv.	Ens.	Iter.
Clean Error (%)	8.2	8.3	8.0	10.5	9.8	9.3	10.4	9.3	7.8	8.1	7.9	10.0	14.8	13.7	14.5	10.6
FGSM	82.7	14.6	36.0	23.2	82.3	24.2	30.0	23.8	82.7	11.0	47.8	22.3	77.2	33.8	47.5	22.9
PGD	99.9	82.6	90.8	30.2	96.0	69.9	95.3	34.7	100.0	88.3	99.9	28.8	85.4	54.7	76.8	30.2
R+FGSM	95.5	63.3	68.4	37.8	90.9	74.2	81.3	28.2	98.4	76.5	71.3	31.2	88.3	45.3	62.1	33.6
Song's	93.1	96.7	95.3	82.1	80.3	88.5	92.2	80.0	96.5	95.6	96.6	83.2	66.8	49.2	63.1	84.6
AT-GAN	96.1	92.7	95.4	93.5	98.5	91.1	93.6	91.6	98.0	88.9	93.9	91.6	99.9	99.4	99.1	93.1

(c) Comparison of AT-GAN and other attacks on CelebA.

Attack	CNN				VGG				ResNet			
	Nor.	Adv.	Ens.	Iter.	Nor.	Adv.	Ens.	Iter.	Nor.	Adv.	Ens.	Iter.
Clean Error (%)	2.7	2.7	3.2	2.4	2.1	3.0	2.9	2.4	2.3	3.0	3.0	2.5
FGSM	81.2	11.8	14.7	9.5	76.7	10.6	16.2	8.9	98.7	9.8	12.3	9.5
PGD	97.3	16.4	22.6	11.4	87.9	14.6	26.3	10.7	100.0	10.9	15.1	10.5
R+FGSM	68.7	9.5	11.3	7.9	68.4	8.7	13.2	7.3	97.5	7.9	9.5	7.8
Song's	90.6	83.4	85.7	89.8	98.7	87.5	95.7	81.6	99.2	93.4	91.0	90.6
AT-GAN	97.5	98.9	95.9	99.6	97.1	96.7	95.8	97.8	98.4	98.5	97.3	98.5

4.2.2. COMPARISON ON ATTACK EFFICIENCY

There are many scenarios where one needs large amount of adversarial examples, such as adversarial training or exploring the properties of adversarial examples. The efficiency for adversarial attacks to generate adversarial examples is significant, but is ignored in most previous works.

As an adversarial generation model, AT-GAN can generate adversarial examples very quickly. Here we evaluate the efficiency of each attack method to attack **Model A** on MNIST data as an example. The average time of generating 1000 adversarial examples is summarized in Table 2. Among the five attack methods, AT-GAN is the fastest as it could generate adversarial examples without the target classifier and gradient calculation once it is trained. Note that Song et al. (2018) needs much longer time than others

because it needs multiple searches and queries to generate one adversarial example.

Table 2. Comparison on the example generation time, measured by generating 1000 adversarial instances using **Model A** on MNIST.

	FGSM	PGD	R+FGSM	Song's	AT-GAN
Time	0.3s	1.8s	0.4s	$\geq 2.5\text{min}$	0.2s

4.2.3. VISUALIZATION OF ADVERSARIAL EXAMPLES

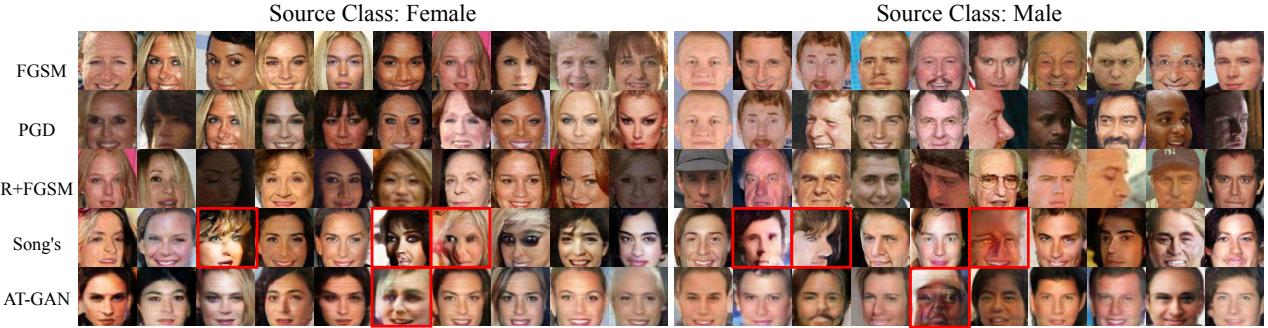
The goal of adversarial examples is to fool neural networks but not to fool humans. Thus, we illustrate some adversarial examples generated by different attacks for **Molde A** on MNIST and Fashion-MNIST, and CNN on CelebA.

FGSM	
PGD	
R+FGSM	
Song's	
AT-GAN	

(a) Adversarial examples for Model A on MNIST



(b) Adversarial examples for Model A on Fashion-MNIST)



(c) Adversarial examples for CNN on CelebA

Figure 3. Adversarial examples generated by various methods on three datasets (Zoom in for details). The red borders indicate unrealistic adversarial examples generated by Song's method or AT-GAN.

As shown in Figure 3, on MNIST, AT-GAN generates slightly more realistic images than Song et al. (2018), for example on 0 and 3. On both Fashion-MNIST and CelebA, some adversarial examples generated by Songs method are not as realistic as AT-GAN to human perception, for example on t-shirt/top (0), sandal (5) and the details of some faces. As for perturbation-based attack, their adversarial examples are not clear enough especially on MNIST and Fashion-MNIST datasets due to the adversarial perturbations. More non-constrained adversarial examples generated by AT-GAN for target attack are shown in Appendix C.

In general, AT-GAN can generate realistic and diverse adversarial examples as Eq. 1 forces the generated non-constrained adversarial examples to be close enough to the benign examples generated by the original generator.

4.3. Transferability of AT-GAN

One important feature for adversarial examples is the transferability across different models. To demonstrate the transferability of non-constrained adversarial examples, we use adversarial examples generated by attacking **Model A** (MNIST and Fashion-MNIST) and **CNN** (CelebA), to evaluate the attack success rates on **Model C** (MNIST and

Fashion-MNIST) and **VGG** (CelebA).

As shown in Table 3, non-constrained adversarial examples generated by AT-GAN exhibit moderate transferability although it is not always the best. Note that as shown in Section 4.2.3, adversarial examples generated by other attack methods are not as realistic as non-constrained adversarial examples and less realistic images could result in a higher transferability.

4.4. Visualization of Adversarial Distribution

As discussed in Section 3.4, we give a brief analysis that AT-GAN can learn a distribution of adversarial examples which is close to the distribution of real data. To identify this empirically, we randomly choose 5,000 benign images and 5,000 adversarial examples generated by different attack methods, and merge these images according to their real label for MNIST and Fashion-MNIST. Then we use t-SNE (Maaten & Hinton, 2008) on these images to illustrate the distributions in 2 dimensions. T-SNE models each high-dimensional object in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability. It indicates that, if the adversarial examples have different distribution

Table 3. Transferability of non-constrained adversarial examples and other traditional adversarial examples on the three datasets. For MNIST and Fashion-MNIST datasets, we attack **Model C** with adversarial examples generated on **Model A**. For CelebA dataset, we attack VGG using adversarial examples generated on CNN. Numbers represent the attack success rate (%).

	MNIST				Fashion-MNIST				CelebA			
	Nor.	Adv.	Ens.	Iter. Adv.	Nor.	Adv.	Ens.	Iter. Adv.	Nor.	Adv.	Ens.	Iter. Adv.
FGSM	46.7	4.2	1.7	4.6	68.9	23.1	20.8	14.8	15.6	4.3	3.3	4.1
PGD	97.5	6.5	4.1	4.1	84.7	27.6	39.6	14.6	18.3	4.3	3.1	4.1
R+FGSM	82.3	6.7	4.8	4.1	21.2	32.1	17.5	26.3	11.0	4.0	3.3	3.8
Song’s	23.8	20.8	20.6	20.1	39.2	34.0	31.5	30.3	9.6	31.8	21.5	38.8
AT-GAN	65.3	24.6	27.9	17.2	58.0	22.7	32.0	15.2	63.7	15.4	16.5	17.6

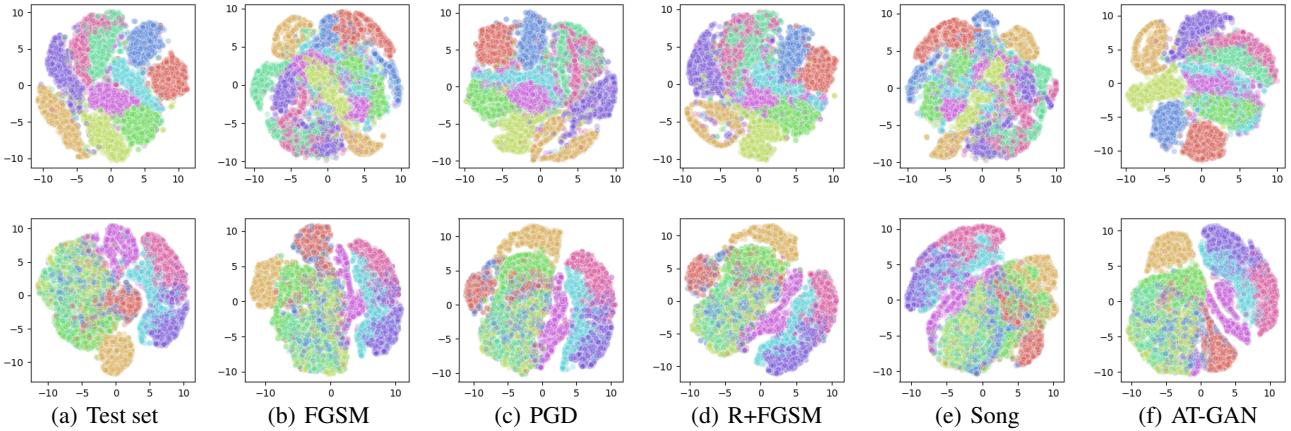


Figure 4. T-SNE visualizations for the combination of test set and adversarial examples generated by different adversarial attacks on MNIST (top) and Fashion-MNIST (bottom). For (a), we use 10,000 sampled real images in test set. For (b) to (f), we use 5,000 sampled images in test set and 5,000 adversarial examples generated by different attacks. The position of each class is random due to the property of t-SNE.

from that of the benign data, t-SNE could not deal well with them and the points with different categories will overlap with each other after dimension reduction, i.e., the results will be in chaos.

The results are illustrated in Figure 4. For AT-GAN, different categories are separated as that of the test set while those of other methods are mixed with each other. It indicates the distribution that AT-GAN learned is indeed very close to the distribution of real data.

5. Conclusion

We propose an adversarial generator model which could generate non-constrained adversarial example for any input noise sampled from a distribution such as Gaussian distribution. Unlike perturbation-based adversarial examples that perturbs the input image or unrestricted adversarial examples that perturbs the input noise, our non-constrained adversarial examples are generated entirely from scratch

without any constraint.

Specifically, we propose a novel attack framework called AT-GAN to estimate the distribution of adversarial examples that could be very close to the distribution of real data. Extensive experiments and visualizations show that AT-GAN can generate diverse and realistic non-constrained adversarial examples efficiently. Besides, AT-GAN achieves higher attack success rates under white-box setting and exhibits moderate transferability.

Our work of AT-GAN suggests that adversarial training, i.e., current strongest defense based on perturbation-based adversarial examples, could not guarantee the robustness against non-constrained adversarial examples. A possible reason is that AT-GAN learns a more complete version on the distribution of adversarial examples, which is much more diverse than that of the perturbation-based method. Our method also offers a new way of building adversarial attacks by designing an adversarial generator directly, which may inspire more researches of this category in future work.

References

- Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv Preprint arXiv:1701.07875*, 2017.
- Athalye, A., Carlini, N., and Wagner, D. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. *International Conference on Machine Learning*, 80:274–283, 2018.
- Bhagoji, A., He, W., Li, B., and Song, D. Exploring the Space of Black-box Attacks on Deep Neural Networks. *arXiv Preprint arXiv:1712.09491*, 2017.
- Buckman, J., Roy, A., Raffel, C., and Goodfellow, I. Thermometer Encoding: One Hot Way To Resist Adversarial Examples. *International Conference on Learning Representations*, 2018.
- Cao, Y., Xiao, C., Yang, D., Fang, J., Yang, R., Liu, M., and Li, B. Adversarial Objects Against LiDAR-Based Autonomous Driving Systems. *arXiv Preprint arXiv:1907.05418*, 2019.
- Carlini, N. and Wagner, D. Towards Evaluating the Robustness of Neural Networks. *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.
- Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *Neural Information Processing Systems*, pp. 2172–2180, 2016.
- Eykholz, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust Physical-World Attacks on Deep Learning Models. *Computer Vision and Pattern Recognition*, 2018.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Neural Information Processing Systems*, pp. 2672–2680, 2014.
- Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations*, 2015.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved Training of Wasserstein GANs. *Neural Information Processing Systems*, pp. 5767–5777, 2017.
- Guo, C., Rana, M., Cisse, M., and Maaten, L. Countering Adversarial Images using Input Transformations. *International Conference on Learning Representations*, 2018.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
- He, W., Li, B., and Song, D. Decision Boundary Analysis of Adversarial Examples. *International Conference on Learning Representations*, 2018.
- Krizhevsky, A., Sutskever, I., and Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*, pp. 1106–1114, 2012.
- Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial Machine Learning at Scale. *International Conference on Learning Representations*, 2017.
- LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.
- Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., and Zhu, J. Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser. *Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.
- Liu, Y., Chen, X., Liu, C., and Song, D. Delving into Transferable Adversarial Examples and Black-box Attacks. *International Conference on Learning Representations*, 2017.
- Liu, Z., Luo, P., Wang, X., and Tang, X. Deep Learning Face Attributes in the Wild. *International Conference on Computer Vision*, pp. 3730–3738, 2015.
- Maaten, L. and Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations*, 2018.
- Metzen, Hendrik, J., Genewein, T., Fischer, V., and Bischoff, B. On Detecting Adversarial Perturbations. *International Conference on Learning Representations*, 2017.
- Nguyen, A., Yosinski, J., and Clune, J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *Computer Vision and Pattern Recognition*, pp. 427–436, 2015.
- Odena, A., Olah, C., and Shlens, J. Conditional Image Synthesis With Auxiliary Classifier GANs. *International Conference on Learning Representations*, 2017.
- Radford, A., Metz, L., and Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *International Conference on Learning Representations*, 2016.

Samangouei, P., Kabkab, M., and Chellappa, R. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. *International Conference on Learning Representations*, 2018.

Shamir, A., Safran, I., Ronen, E., and Dunkelman, O. A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance. *arXiv Preprint arXiv:1801.02774*, 2019.

Shen, S., Jin, G., Gao, K., and Zhang, Y. APE-GAN: Adversarial Perturbation Elimination with GAN. *IEEE International Conference on Acoustics, Speech and SP*, pp. 3842–3846, 2017.

Simonyan, K. and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations*, 2015.

Song, C., He, K., Wang, L., and Hopcroft, J. Improving the Generalization of Adversarial Training with Domain Adaptation. *International Conference on Learning Representations*, 2019.

Song, Y., Shu, R., Kushman, N., and Ermon, S. Constructing unrestricted adversarial examples with generative models. *Neural Information Processing Systems*, pp. 8322–8333, 2018.

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

Tramr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble Adversarial Training: Attacks and Defenses. *International Conference on Learning Representations*, 2018.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

A. Related Work

A.1. Gradient-based Attacks

There are two types of attacks regarding how the attacks access the model. The *white-box attack* can fully access the target model, while the *black-box attack* (Carlini & Wagner, 2017) has no knowledge of the target model. Existing black-box attacks mainly focus on *transferability* (Liu et al., 2017; Bhagoji et al., 2017), in which an adversarial instance generated on one model could be transferred to attack another model.

We will introduce three typical adversarial attack methods. Here the components of all adversarial examples are clipped in $[0, 1]$.

Fast Gradient Sign Method (FGSM). FGSM (Goodfellow et al., 2015) adds perturbation in the gradient direction of the training loss J on the input x to generate adversarial examples.

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x J(\theta, x, y))$$

Here y is the true label of a sample x , θ is the model parameter and ϵ specifies the ℓ_∞ distortion between x and x_{adv} .

Projected Gradient Descent (PGD). PGD adversary (Madry et al., 2018) is a multi-step variant of FGSM, which applies FGSM iteratively for k times with a budget α .

$$\begin{aligned} x_{adv_{t+1}} &= \text{clip}(x_{adv_t} + \alpha sign(\nabla_x J(\theta, x_{adv_t}, y)), \\ &\quad x_{adv_t} - \epsilon, x_{adv_t} + \epsilon) \\ x_{adv_0} &= x, \quad x_{adv} = x_{adv_k} \end{aligned}$$

Here $\text{clip}(x, p, q)$ forces its input x to reside in the range of $[p, q]$.

Rand FGSM (R+FGSM). R+FGSM (Tramr et al., 2018) first applies a small random perturbation on the benign image with a parameter α ($\alpha < \epsilon$), then it uses FGSM to generate an adversarial example based on the perturbed image.

$$\begin{aligned} x_{adv} &= x' + (\epsilon - \alpha) \cdot sign(\nabla_{x'} J(\theta, x', y)) \\ \text{where } x' &= x + \alpha \cdot sign(\mathcal{N}(\mathbf{0}^d, \mathbf{I}^d)) \end{aligned}$$

A.2. Adversarial Training

There are many defense strategies, such as detecting adversarial perturbations (Metzen et al., 2017), obfuscating gradients (Buckman et al., 2018; Guo et al., 2018) and eliminating perturbations (Shen et al., 2017; Liao et al., 2018), among which adversarial training is the most effective method (Athalye et al., 2018). We list several adversarial training methods as follows.

Adversarial training. Goodfellow et al. (2015) first introduce the method of adversarial training, where the standard loss function f for a neural network is modified as:

$$\tilde{J}(\theta, x, y) = \alpha J_f(\theta, x, y) + (1 - \alpha) J_f(\theta, x_{adv}, y).$$

Here y is the true label of a sample x and θ is the model's parameter. The modified objective is to make the neural network more robust by penalizing it to count for adversarial samples. During the training, the adversarial samples are calculated with respect to the current status of the network. Taking FGSM for example, the loss function could be written as:

$$\begin{aligned} \tilde{J}(\theta, x, y) &= \alpha J_f(\theta, x, y) + (1 - \alpha) J_f(\theta, x + \\ &\quad \epsilon sign(\nabla_x J(\theta, x, y)), y) \end{aligned}$$

Ensemble adversarial training. Tramr et al. (2018) propose an ensemble adversarial training method, in which DNN is trained with adversarial examples transferred from a number of fixed pre-trained models.

Iterative adversarial training. Madry et al. (2018) propose to train a DNN with adversarial examples generated by iterative methods such as PGD.

B. Theoretical Analysis of AT-GAN

In this section, we provide the proof of theorems in Section 3.4.

Theorem 3. Suppose $\max_{z,y} L_d(z, y) < \epsilon$, we have $KL(p_a \| p_g) \rightarrow 0$ when $\epsilon \rightarrow 0$.

To prove this theorem, we first consider that given a distribution $p(x)$ in space \mathcal{X} , we construct another distribution $q(x)$ by selecting the points $p_\epsilon(x)$ in the ϵ -neighborhood of $p(x)$ for any $x \in \mathcal{X}$. Obviously, when $p_\epsilon(x)$ is close enough to $p(x)$, $q(x)$ has almost the same distribution as $p(x)$. Formally, we can have the following lemma:

Lemma 1. Given two distributions P and Q with probability density function $p(x)$ and $q(x)$ in space \mathcal{X} , if there exists a constant ϵ that satisfies $\|q(x) - p(x)\| < \epsilon$ for any $x \in \mathcal{X}$, we could get $KL(P \| Q) \rightarrow 0$ when $\epsilon \rightarrow 0$.

Proof. For two distributions P and Q with probability density function $p(x)$ and $q(x)$, we could get $q(x) = p(x) + r(x)$ where $\|r(x)\| < \epsilon$.

$$\begin{aligned} KL(P \| Q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \int p(x) \log p(x) dx - \int p(x) \log q(x) dx \end{aligned}$$

$$\begin{aligned}
&= \int (q(x) - r(x)) \log p(x) dx - \\
&\quad \int (q(x) - r(x)) \log q(x) dx \\
&= \int q(x) \log p(x) dx - \int q(x) \log q(x) dx \\
&\quad - \int r(x) \log p(x) dx + \int r(x) \log q(x) dx \\
&= \int r(x) \log \frac{q(x)}{p(x)} dx - KL(Q||P) \\
&\leq \int \epsilon \log(1 + \frac{\epsilon}{p(x)}) dx
\end{aligned}$$

Obviously, when $\epsilon \rightarrow 0$, we could get $\int \epsilon \log(1 + \frac{\epsilon}{p(x)}) dx \rightarrow 0$, which means $DL(P||Q) \rightarrow 0$. \square

Now, we come to proof of Theorem 1.

Proof. For two distributions p_a and p_g , $\max_{y,z} L_d < \epsilon$ indicates $\forall z \sim p_z, \|p_a(z) - p_g(z)\| < \epsilon$. According to Lemma 1, we have $KL(p_a||p_g) \rightarrow 0$ when $\epsilon \rightarrow 0$. This concludes the proof. \square

Theorem 4. The global minimum of the virtual training of AC-WGAN_GP is achieved if and only if $p_g = p_{data}$.

Proof. To simplify the analysis, we choose a category y of AC-WGAN_GP and denote $p_g(x|y)$ and $p_{data}(x|y)$ the distribution the generator learns and the distribution of real data respectively. Then for each category, the loss function is equivalent to WGAN_GP. We refers to Samangouei et al. (2018) to prove this property. The WGAN_GP min-max loss is given by:

$$\begin{aligned}
&\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [D(x)] - \\
&\quad \mathbb{E}_{z \sim p_z(z)} [D(G(z))] - \lambda \mathbb{E}_{\hat{x} \sim p_{\hat{x}}(\hat{x})} [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \\
&= \int_x p_{data}(x) D(x) dx - \int_z p_z(z) D(G(z)) dz \\
&\quad - \lambda \int_{\hat{x}} p_{\hat{x}}(\hat{x}) [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] d\hat{x} \\
&= \int_x [p_{data}(x) - p_g(x)] D(x) dx \\
&\quad - \lambda \int_{\hat{x}} p_{\hat{x}}(\hat{x}) [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] d\hat{x}
\end{aligned} \tag{6}$$

For a fixed G , the optimal discriminator D that maximizes $V(D, G)$ should be:

$$D_G^*(x) = \begin{cases} 1 & \text{if } p_{data}(x) \geq p_g(x) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

According to Eq. 6 and Eq. 7, we could get:

$$\begin{aligned}
V(D, G) &= \int_x [p_{data}(x) - p_g(x)] D(x) dx \\
&\quad - \lambda \int_{\hat{x}} p_{\hat{x}}(\hat{x}) [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] d\hat{x} \\
&= \int_{\{x | p_{data}(x) \geq p_g(x)\}} (p_{data}(x) - p_g(x)) dx \\
&\quad - \lambda \int_{\hat{x}} p_{\hat{x}}(\hat{x}) d\hat{x} \\
&= \int_{\{x | p_{data}(x) \geq p_g(x)\}} (p_{data}(x) - p_g(x)) dx - \lambda
\end{aligned} \tag{8}$$

Let $\mathcal{X} = \{x | p_{data}(x) \geq p_g(x)\}$, in order to minimize Eq. 8, we set $p_{data}(x) = p_g(x)$ for any $x \in \mathcal{X}$. Then, since both p_g and p_{data} integrate to 1, we could get:

$$\int_{\mathcal{X}^c} p_g(x) dx = \int_{\mathcal{X}^c} p_{data}(x) dx.$$

However, this contradicts Eq. 7 where $p_{data}(x) < p_g(x)$ for $x \in \mathcal{X}^c$, unless $\mu(\mathcal{X}^c) = 0$ where μ is the Lebesgue measure.

Therefore, for each category we have $p_g(x|y) = p_{data}(x|y)$ which means $p_g(x) = p_{data}(x)$ for AC-WGAN_GP. \square

C. More Details on Experiments

C.1. More Experimental Setup

Here we describe the details about the experimental setup, including the model architectures and attack hyper-paramters.

Model Architectures. We first describe the neural network architectures used in experiments. The abbreviations for components in the network are described in Table 4. The architecture of AC-WGAN_GP for MNIST and Fashion-MNIST is shown in Table 5 where the generator and discriminator are the same as in Chen et al. (2016), while the architecture of AC_WGAN_GP for CelebA is same as in Gulrajani et al. (2017). The details of Model A through D are described in Table 6. Both Model A and Model C have convolutional layers and fully connected layers. The difference is only on the size and number of convolutional filters. Model B uses dropout as its first layer and adopts a bigger convolutional filter so that it has less number of parameters. Model D is a fully connected neural network with the least number of parameters and its accuracy will be lower than other models. VGG in our experiments is the same as VGG16 in Simonyan & Zisserman (2015) and we adopt ResNet from Song et al. (2018).

Hyper-parameters. The hyper-parameters used in experiments for each attack method are described in Table 7.

Table 4. The abbreviations for network architectures.

Abbreviation	Description
Conv($m, k \times k$)	A convolutional layer with m filters and filter size k
DeConv($m, k \times k$)	A transposed convolutional layer with m filters and filters size k
Dropout(α)	A dropout layer with probability α
FC(m)	A fully connected layer with m outputs
Sigmoid	The sigmoid activation function
Relu	The Rectified Linear Unit activation function
LeakyRelu(α)	The Leaky version of a Rectified Linear Unit with parameter α
Maxpool(k, s)	The maxpooling with filter size k and stride s

Table 5. The architecture of WGAN-GP with auxiliary classifier.

Generator	Discriminator	Classifier
FC(1024) + Relu	Conv(64, 4×4) + LeakyRelu(0.2)	Conv(32, 3×3) + Relu
FC($7 \times 7 \times 128$) + Relu	Conv(128, 4×4) + LeakyRelu(0.2)	pooling(2, 2)
DeConv(64, 4×4) + Sigmoid	FC(1024) + LeakyRelu(0.2)	Conv(64, 3×3) + Relu
DeConv(1, 4×4) + Sigmoid	FC(1) + Sigmoid	pooling(2, 2)
		FC(1024)
		Dropout(0.4)
		FC(10) + Softmax

C.2. Non-constrained Adversarial Examples for Target Attack

Here we show some non-constrained adversarial examples generated by AT-GAN for target attack. The results are illustrated in Figure 5 and 6. Instead of adding perturbations to the original images, AT-GAN transfers the generator so that the generated adversarial instances are not in the same shape of the initial examples (in diagonal) generated by the original generator.

Table 6. The architectures of **Models A** through **D** and CNN used for classification. After each model's name we put the number of parameters of that model.

Model A (3,382,346)	Model B (710,218)	Model C (4,795,082)
Conv(64, 5 × 5)+Relu	Dropout(0.2)	Conv(128, 3 × 3)+Relu
Conv(64, 5 × 5)+Relu	Conv(64, 8 × 8)+Relu	Conv(64, 3 × 3)+Relu
Dropout(0.25)	Conv(128, 6 × 6)+Relu	Dropout(0.25)
FC(128)+Relu	Conv(128, 5 × 5)+Relu	FC(128)+Relu
Dropout(0.5)	Dropout(0.5)	Dropout(0.5)
FC(10)+Softmax	FC(10)+Softmax	FC(10)+Softmax
Model D (509,410)	CNN (17,066,658)	
$\left[\begin{array}{l} \text{FC}(300)+\text{Relu} \\ \text{Dropout}(0.5) \end{array} \right] \times 4$	Conv(32, 3 × 3)+Relu Conv(32, 3 × 3)+Relu Dropout(0.3)	
FC(10) + Softmax	Conv(64, 3 × 3)+Relu Conv(64, 3 × 3)+Relu Maxpool(2, 2) + Dropout(0.3) Conv(128, 3 × 3)+Relu Conv(128, 3 × 3)+Relu Maxpool(2, 2) + Dropout(0.3) FC(512) + Relu Dropout(0.3) FC(10) + Softmax	

Table 7. Hyper-paramters of different attack methods on the datasets.

Attack	Datasets				Norm
	MNIST	Fashion-MNIST	CelebA		
FGSM	$\epsilon = 0.3$	$\epsilon = 0.1$	$\epsilon = 0.015$		ℓ_∞
PGD	$\epsilon = 0.3, \alpha = 0.075, \text{epochs} = 20$	$\epsilon = 0.1, \alpha = 0.01, \text{epochs} = 20$	$\epsilon = 0.015, \alpha = 0.005, \text{epochs} = 20$		ℓ_∞
R+FGSM	$\epsilon = 0.3, \alpha = 0.15$	$\epsilon = 0.2, \alpha = 0.1$	$\epsilon = 0.015, \alpha = 0.003$		ℓ_∞
Song's	$\lambda_1 = 100, \lambda_2 = 0, \text{epochs} = 200$	$\lambda_1 = 100, \lambda_2 = 0, \text{epochs} = 200$	$\lambda_1 = 100, \lambda_2 = 100, \text{epochs} = 200$		N/A
AT-GAN	$\alpha = 2, \beta = 1, \text{epochs} = 100$	$\alpha = 2, \beta = 1, \text{epochs} = 100$	$\alpha = 3, \beta = 2, \text{epochs} = 200$		N/A

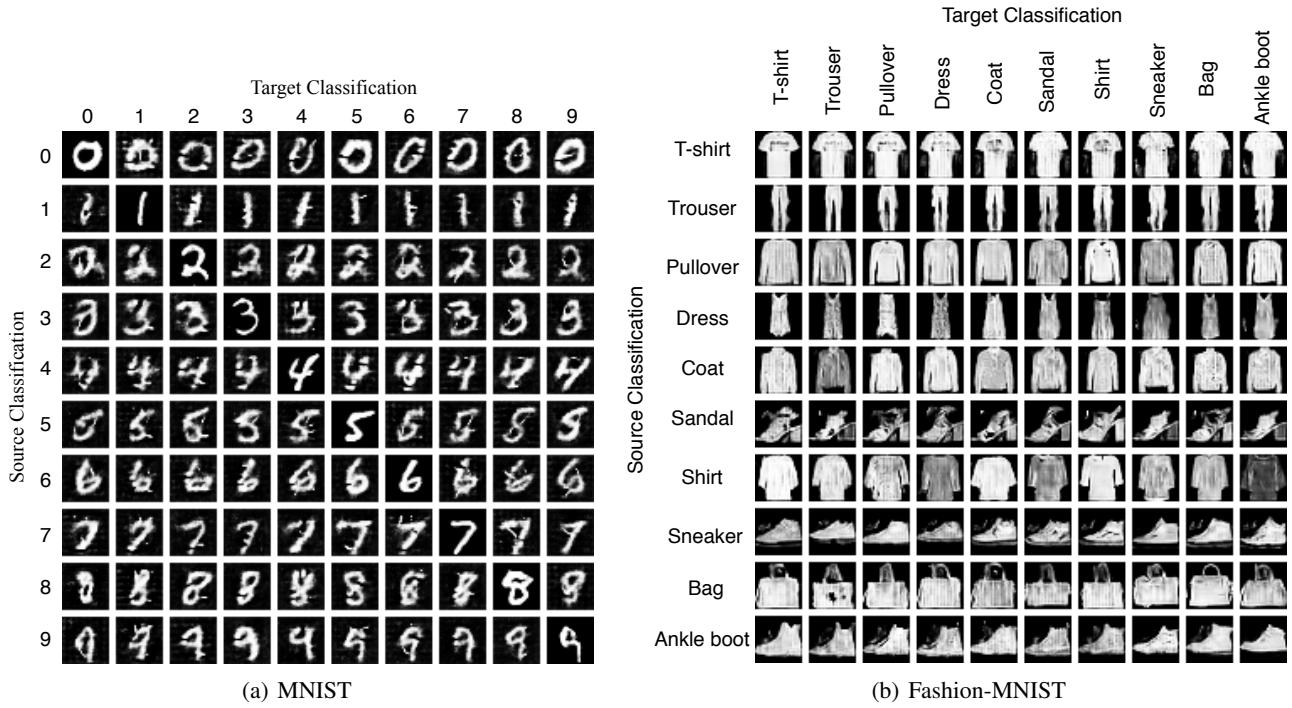


Figure 5. Adversarial examples generated by AT-GAN to different targets with the same random noise input for each row. The images on the diagonal are generated by $G_{original}$ which are not adversarial examples and treated as the initial instances for AT-GAN.



Figure 6. Adversarial examples generated by AT-GAN on CelebA dataset.