EVALUATING ROBUSTNESS OF GENERATIVE MODELS WITH ADVERSARIAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

With the advent of adversarial robustness as a research area, much novel work attempts to design creative defense mechanisms against adversarial vulnerabilities that arise. While classification models are the most common target of adversarial robustness research, generative models are often underestimated though they play essential roles in many applications. This work evaluates generative models for reconstruction tasks in terms of their adversarial robustness. We constructed two frameworks: a standard and a universal-attack framework. The standard framework requires an input to find its perturbation, and the universal-attack framework generates adversarial perturbation from the distribution of a dataset. Extensive experimental evidence discussed in this paper suggests that both frameworks can effectively alter how images are reconstructed and classified using classic generative models trained on MNIST and Cropped Yale Face datasets. Further, these frameworks outperform state-of-the-art adversarial attacks. Moreover, we showcase using the proposed framework to retrain a generative model to improve its resilience against adversarial perturbations. Furthermore, for the sake of generative models, an attack may desire not to alter the latent space. Thus, we also include the analysis of the latent space.

1 INTRODUCTION

The societal impact of machine learning applications is self-evident as it continues to solve many highly complex tasks (e.g., self-driving cars (Rao & Frtunikj, 2018)). However, many of these models are still susceptible to small perturbations that naturally occur in the world or are fabricated by learning to model the input space. In the latter case, models could be trained to be robust against the noise. Goodfellow et al. (2014) and Szegedy et al. (2013) demonstrated how to create small additive perturbations to mislead classification models and expose the issue. Their research shows that many state-of-the-art models were not robust against a perturbation that is currently known as an *adversarial example*. Naturally, many works (Goodfellow et al., 2014; Papernot et al., 2016b; Xu et al., 2017; Madry et al., 2017) started trying to determine how to defend models against adversarial examples effectively. While defense mechanisms were developed, others worked around this finding new adversarial examples against such defenses (Moosavi-Dezfooli et al., 2016; Papernot et al., 2016a; 2017; Carlini & Wagner, 2017). Note, however, that most of the existing work focuses on adversarial examples for classification models, even though generative models are widely used in practice (e.g., people in clothing (Lassner et al., 2017), music generation (Dhariwal et al., 2020) and molecule graphs generation (Samanta et al., 2020)). Therefore, our work attempts to construct frameworks for evaluating the adversarial robustness of generative models. In particular, we focus on models whose goal is to reconstruct the input. Adversarial examples in this setup try to make the targeted models generate an output that belongs to a different class than the corresponding input. We use well-known GANs (Goodfellow et al., 2020) to train a generator to produce small perturbations. Then, we add such perturbations to an image to create an adversarial example for a target generative model. This paper introduces two frameworks to demonstrate the general idea, i.e., a standard and a universal-attack framework. The standard framework trains a generator such that it requires input to produce perturbation during execution. On the contrary, the universal-attack framework trains a generator that does not need an input to produce a perturbation during execution. The proposed methodology produces generators effective on MNIST and Cropped Yale Face dataset (CYFD).

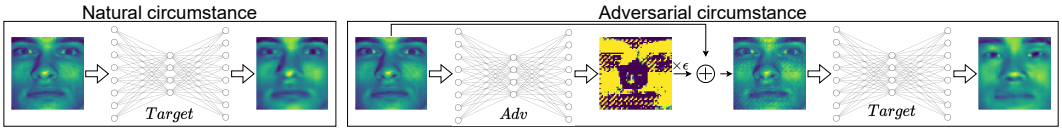


Figure 1: Example of our trained generator misleading a targeted generative model where *Target* is the targeted model and *Adv* is our generator. This shows what happens in a natural scenario (left) and an adversarial scenario (right).

Figure 1 shows how to apply our trained generator to mislead a target. The natural circumstance (left) shows that the target works fine with a clean image since the reconstructed image looks similar to the clean image. However, a clean image fed to our trained generator in the adversarial circumstance (right) produces a perturbation multiplied by ϵ (i.e., the perturbation bound). Then, it is added to the clean image to create an adversarial example. Note that when the adversarial example enters the target model, the output is a face from a completely different class than the input. We can retrain the target models with our framework to make them robust against these perturbations. Moreover, we analyze the latent spaces of the target models and found an interesting pattern. That is, the differences between latent spaces of the clean samples and their corresponding adversarial examples increase when the accuracy on the reconstructed images of the adversarial examples increases in MNIST. However, we did not find this pattern in CYFD. This paper showcases the positive effects of such retraining on MNIST and CYFD data with good results.

Our main contributions can be summarized as follows:

- We formally define a data reconstruction problem based on generative models introducing adversarial examples.
- We design and implement a standard and universal-attack framework to train adversaries for generative models, producing effective adversarial examples using the well-known approach by Goodfellow et al. (2020).
- We successfully use the proposed standard framework to retrain a generative model improving its adversarial robustness.

2 Related Works

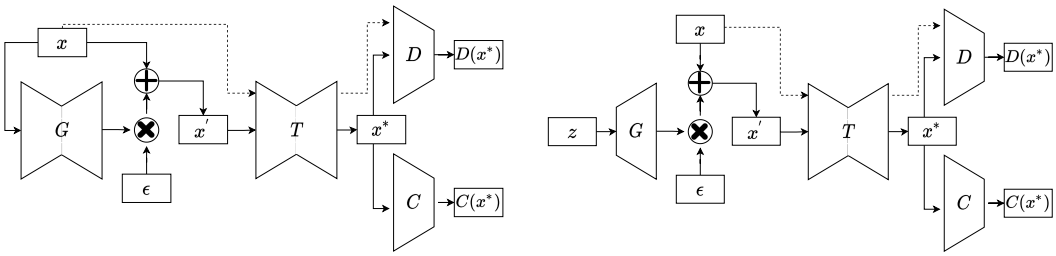
Because our framework applies GANs to evaluate generative models, this section briefly describes previous works that applied GANs to create adversarial examples for discriminative models in Section 2.1. We further go over existing works creating adversarial examples for generative models by using optimization-based attacks in Section 2.2.

2.1 GANs-Based Attacks

Several works apply GANs to create adversarial examples on classification models. However, the most related works are described as follows. Xiao et al. (2018) utilized GANs to train a generator to create perturbation that can be added to a clean image to generate an adversarial example for a classification model. They proposed the attack under both semi-whitebox and blackbox settings. Bai et al. (2021) applied the work of (Xiao et al., 2018) and proposed a solution to make it converge earlier during the training. Their solution is to train a generator with the existing adversarial examples in the first stage to make the generator know the direction to create perturbation in the same space as one of the adversarial examples. In the second stage, they followed the training in (Xiao et al., 2018). Intuitively, the existing works in this area do not consider the generative models.

2.2 Adversarial Examples on Generative Models

Only a few works focused on finding adversarial examples in generative models. Tabacof et al. (2016) tried to find an adversarial example of a clean image by minimizing the difference between the adversarial example's latent space and the target's image latent space where the perturbation added to the clean image was minimized. Kos et al. (2018) proposed three attacks against generative models:



(a) Standard framework (b) Universal-attack framework

Figure 2: The architecture of the standard and universal-attack frameworks where G is the generator, D is the discriminator, C is the pretrained classifier, T is the targeted generative model, x is an instance from $\mathcal{X}$, z is a random noise from the normal distribution and ϵ is the perturbation bound.

classifier attack, $\mathcal{L}_{VAE}$ attack and latent attack. They were all based on optimization problems and used Carlini and Wagner attack (Carlini & Wagner, 2017) to approximate the solutions. So far, there are only attacks based on optimization problems that take time to approximate a solution during testing. In 2020, Pope et al. (2020) applied the projected gradient descent (Madry et al., 2017) for evaluating generative models.

To the best of our knowledge, our work is the first to adopt GANs to find adversarial examples on generative models.

3 Our Approach

This section explains our attacks in detail and starts with defining the problem of adversarial examples in generative models. Then, we derive the loss functions of the standard framework and describe how to train the generator with the framework. Furthermore, we discuss the universal-attack framework.

3.1 Problem Definition

Suppose that there is a generative network T (e.g., a VAE) trained on clean dataset $\mathcal{X} \subseteq \mathbb{R}^n$, where n is the number of features. We also denote a class set as $\mathcal{Y}$ and instance i sampled from distribution P as (x_i, y_i) where $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$. Note that y_i is the ground truth class of x_i. The goal of an adversary is to create an adversarial example ($x_i^{'}$) from x_i such that $x_i^{'}$ looks like samples in class y_i; however, $T(x_i^{'})$ belongs to another class (not equal to y_i). Our approach generates perturbation which is bound by L_∞ and then adds it to a clean input to create an adversarial example. Note that $L_\infty(x_i, x_i^{'}) = \max|x_i - x_i^{'}|$.

3.2 Standard Framework (SF)

The main concept of our framework is to train a generator to create perturbation such that this perturbation can be added to clean input and generate an adversarial example for a target network (T). Figure 2a shows our framework consisting of a generator (G), a discriminator (D) and a pretrained classifier (C). In the following, we assume that x is an arbitrary member of $\mathcal{X}$. Each component in the framework has its own role in the framework as follows. D is used to remain x^* in the same distribution of $T(x)$. To achieve it, D needs to minimize

$$L_D = log(D(T(x)) + log(1 - D(x^*)),$$

where $T(x^{'}) = x^*$. Also, G needs to minimize

$$L_G = log(D(x^*)).$$

As a result, D should not be able to distinguish between $T(x)$ and x^*. Further, C directs G to update its parameters such that $C(x^*) \neq y$ where $C(x^*)$ is a class classified by C given x^*, and y is the ground truth label of x. Therefore, the loss regarding C is the opposite of categorical cross entropy

used for training a classification network, and we can define it as

$$L_C = \sum_{j=0}^{|\mathcal{Y}|-1} l[j] log(Z(x^*)[j]), \tag{1}$$

where l is the ground truth logit, $l[j]$ is the value at index j of l and $Z(x^*)[j]$ is the logit at class j of C given x^* (i.e., the output before softmax activation). Note that if $\mathcal{X}$ is single-class data, l has 1 only in one index. Nevertheless, if $\mathcal{X}$ is multi-class data, l can have 1 in multiple indexes. Moreover, we also want to maximize the difference between $x^{'}$ and x^* so that $x^{'}$ is not similar to x^*. We use the opposite of mean square error as

$$L_{MSE} = -\sum_{j=0}^{n-1} (x^{'}[j] - x^*[j])^2,$$

where $x^{'}[j]$ is feature j of $x^{'}$. Therefore, the total loss of G is

$$L_{Total} = L_{MSE} + \lambda_G L_G + \lambda_C L_C,$$

where λ_G and λ_C are hyperparameters to balance all the losses. Also, ϵ is the perturbation bound, and we use L_∞ as described earlier.

During the training time, the pretrained classifier (C) and the targeted generative model (T) are untrainable. Then, the framework processes as follows: a) feed x to G and obtain perturbation; b) multiply the perturbation with ϵ to fit the perturbation in the bound and add it to x to create $x^{'}$; c) feed $x^{'}$ to target T and obtain x^*; d) feed x to T and obtain $T(x)$; e) feed x^* and $T(x)$ to D to compute L_D and L_G and feed x^* to C to compute L_C f) compute L_{MSE} from $x^{'}$ and x^*; g) compute L_{Total} from the three losses and update the parameters of D and G by using a gradient descent method on L_D and L_{Total} respectively.

We can only feed an arbitrary input to the generator (G) during the testing time and obtain perturbation. Then, we multiply the perturbation with ϵ and add it to the input. Consequently, we obtain an adversarial example of the input. Note that the last layer of G has to be a *tanh* activation function to limit the adversarial example in the bound ϵ.

Noticeably, this framework is an untargeted attack because it tries to find adversarial example ($x^{'}$) such that $C(x^*) \neq y$. Hence, T can produce an instance belonging to any class except the correct class. Additionally, we can also adjust it to create a targeted attack as follows. Instead of using the logit of the ground-truth label in equation 1, we use the logit of the targeted label instead. Also, we use the categorical cross entropy loss instead for L_C. Thus, the loss is

$$\sum_{j=0}^{|\mathcal{Y}|-1} t[j] log(Z(x^*)[j]),$$

where t is the logit of the target class, and $t[j]$ is the value at index j of t.

3.3 Universal-Attack Framework (UF)

In the real world, some attacks may study only the distribution of a dataset and create adversarial perturbation without requiring inputs. Therefore, we design this framework to evaluate a target model with respect to the attacks described earlier. Therefore, we do not feed input to G during the testing time but feed noise z. Figure 2b demonstrates the framework. All the components and their roles are the same as in the previous framework. However, the training steps are slightly different.

z is sampled from the normal distribution during the training time and fed to G. Then, the rest of the process is the same as the previous framework. That is, the output of G is multiplied by ϵ and then added to x to create $x^{'}$. Then, $x^{'}$ enters the target model T, and it outputs x^*. After that, x^* is fed to the discriminator D and the pretrained classifier C. Note that we also perform the same process of $x^{'}$ on x. At last, we use the output from D, C and T to compute gradients and train the framework. As a result, in the testing time, G can generate perturbation later used for creating adversarial examples on samples from the distribution of $\mathcal{X}$ for T.

Table 1: Accuracy achieved by the classifier on the sets of samples concerning MNIST and CYFD. Note that adversarial examples were generated by our FGSM, PGD and standard framework (SF).

Model	X	R_X	FGSM		PGD		SF	
			A	R_A	A	R_A	A	R_A
MNIST								
VAE	99.47%	91.51%	98.78%	32.37%	99.39%	34.82%	99.22%	**4.84%**
VAEGAN		94.65%	98.72%	50.37%	99.41%	43.33%	99.15%	**5.63%**
CYFD								
VAE	99.39%	77.8%	70.88%	18.33%	99.98%	37.68%	74.95%	**18.13%**
VAEGAN		82.48%	77.19%	19.14%	99.39%	31.77%	76.37%	**6.92%**

4 Experiments and Results

In this section, we first evaluate our approach as untargeted and targeted attacks on MNIST (Deng, 2012) and the Cropped Yale Face dataset (CYFD) (Lee et al., 2005). MNIST is a hand-written-digit-image dataset consisting of 50000 training samples and 10000 test samples and thus has ten classes. CYFD is a human-face dataset consisting of 1960 training samples and 491 test samples and has 38 classes (i.e., 38 faces). Further, we also explore the universal-attack framework on the same datasets. We choose a variational autoencoder (VAE) (Kingma & Welling, 2013) and a VAEGAN (Yu et al., 2019) as our targeted generative models because they are well-known in this kind of model. For different datasets, we use different architectures of components in the frameworks. We used the test samples of the datasets unseen by our attacks to evaluate the efficiency of our attacks. Further, because crowdsourcing was costly, we used the pretrained classifiers in our frameworks to predict classes of the input and output of the targets for the evaluation. The architectures of our frameworks and the targets and how we preprocessed the framework are explained in Appendix A.

4.1 Baselines

We can use the attack in (Pope et al., 2020) as a baseline. However, the attack did not use the pretrained classifier, and this pretrained classifier had information to mislead the target model. Thus, it was not fair to use it as a baseline. Then, we have designed a framework on which we can apply state-of-the-art attacks. This framework includes only the pretrained classifier from our approach. First, an input is fed to the target (i.e., VAE or VAE-GAN), and the output is fed into the pretrained classifier. Then, the attack uses the classifier's output to compute the gradients with respect to the input and performs a state-of-the-art gradient-based attack. Further, in our framework, we have two baselines which are based on Pope et al. (2020) and our attack. The first one uses the fast gradient sign method (FGSM) (Goodfellow et al., 2014), and the second one uses projected gradient descent (PGD) (Madry et al., 2017).

4.2 Standard Framework

For MNIST, we trained our standard framework with the setting of 250 batch size and $\epsilon = 0.1$ for 100 epochs. For CYFD, we trained it with the setting of 140 batch size and $\epsilon = 0.05$. Further, we set λ_G to be 0.5 because most of the images from the targets were valid. For the untargeted attack, we set λ_C to be 1 since we wanted to change the class of the output of the targets. For the targeted attack, we set it to 2 because we would like to focus on changing the class of the reconstructed image, and the reconstructed images did not look like the inputs when increasing it (e.g., 3 and 4).

4.2.1 Untargeted Attack

After we trained the adversary and used our generator to create adversarial examples, then, we denote a set of clean test samples as X, a set of their reconstructed images as R_X, a set of their corresponding generated adversarial examples as A and a set of their reconstructed images as R_A. A is obtained by feeding X to G, multiplying G's output with ϵ and adding it to X. R_A is simply obtained by feeding A to the target model T. Note that we use these notations throughout Section 4 and 5.

Table 1 shows accuracy obtained from the MNIST and CYFD classifiers on those four sets of images according to two baselines and our standard framework (SF). Explicitly, all the attacks (i.e., FGSM, PGD and SF) work very well because the accuracy of the classifiers on the reconstructed images generated from our attack was significantly reduced while the ones on X and A were still high. Specifically, our attack explicitly outperformed the two baselines since the accuracy of the classifier on R_A of SF is significantly lowest. However, in CYFD, PGD did not perform well compared to the other two attacks because it did not find the correct direction to mislead the target. Note that PGD required a gradient direction multiple times. Then, it did not perturb the clean images so much since the classifier's accuracy on A was about the same.

4.2.2 TARGETED ATTACK

When we trained our generator, we excluded training samples that belong to the targeted class because it did not update any parameter for the targeted-class samples. For MNIST, we picked all the classes to be the targeted classes. Therefore, we had ten generators, each of which is for each targeted class. For CYFD, we randomly picked only three classes (i.e., 2, 10 and 28) to be the targeted classes. Thus, we had three generators.

Table 2: Results achieved by the classifier, the two baselines and our targeted attack performed by our standard framework on VAE in MNIST with each targeted class. Note that A R_X is the accuracy on R_X, TC R_X is the confidence on the targeted class on R_X, SR is the success rate and TC R_A is the confidence on the targeted class on R_A.

Class	A R_X	TC R_X	FGSM		PGD		SF	
			SR	TC R_A	SR	TC R_A	SR	TC R_A
0	91.04%	0.010	9.73%	0.10	28.02%	0.26	**73.46%**	0.72
1	90.63%	0.004	2.17%	0.02	11.31%	0.12	**28.76%**	0.29
2	91.73%	0.006	20.8%	0.21	41.29%	0.39	**81.79%**	0.78
3	91.94%	0.009	9.15%	0.09	22.45%	0.22	**75.16%**	0.74
4	91.78%	0.008	7.00%	0.07	23.26%	0.23	**47.75%**	0.47
5	92.59%	0.010	3.78%	0.40	11.58%	0.12	**48.54%**	0.47
6	91.17%	0.007	2.52%	0.30	13.70%	0.14	**36.12%**	0.35
7	91.86%	0.011	26.00%	0.26	42.77%	0.40	**76.57%**	0.74
8	92.40%	0.015	21.16%	0.21	44.01%	0.42	**83.55%**	0.82
9	91.95%	0.016	8.13%	0.08	31.42%	0.30	**75.90%**	0.74

We evaluated our generators and a generative model (i.e., VAE) on the test samples as shown in Table 2. Note that we excluded samples belonging to the targeted classes from the test samples to evaluate our generators fairly. Moreover, we consider that our generator successfully creates an adversarial example when its reconstructed image belongs to the targeted class.

According to Table 2, the classifier could achieve high accuracy on the reconstructed images of the clean samples on VAE with significantly low average confidences on the targeted classes. Further, FGSM achieved a very low success rate and average targeted confidence on every class. Then, PGD could achieve a slightly higher success rate and average targeted confidence. Essentially, our attack (SF) significantly outperformed those two baselines with a much higher success rate and average targeted confidence on every class. Nonetheless, it achieved low success rates and average targeted confidence on some targeted classes (i.e., class 1, 4, 5 and 6) (still higher than the baselines). Implicitly, those targeted classes are distinct from the other classes. We also found similar results on VAEGAN.

However, in the entire test samples of CYFD, the generators on those targeted classes cannot achieve high success rates, as seen in Table 3, especially the targeted class 28. The reason is that humans' faces are very similar to each other. Thus, it is tough for the targeted attack to transform reconstructed images into its target. Nevertheless, our attack could outperform the baselines. We found the similar results when the target was VAEGAN. Despite low success rates (still higher than the baselines), they can increase the classifier's confidence on the targeted classes by some significant amounts. Therefore, our generators are still effective on CYFD. Also, our attack could be effective in some classes, as demonstrated in Figure 7.

Table 3: Results achieved by the classifier, the two baselines and our targeted attack performed by our standard framework on VAE in CYFD with each targeted class. Note that A R_X is the accuracy on R_X, TC R_X is the confidence on the targeted class on R_X, SR is the success rate and TC R_A is the confidence (the final output after the softmax activation function) on the targeted class on R_A.

Class	A R_X	TC R_X	FGSM		PGD		SF	
			SR	TC R_A	SR	TC R_A	SR	TC R_A
2	76.31%	0.007	13.62%	0.12	36.69%	0.32	**41.51%**	0.38
10	75.83%	0.010	7.92%	0.07	23.96%	0.21	**33.33%**	0.31
28	76.53%	0.004	6.13%	0.06	20.72%	0.20	**21.78%**	0.2

4.3 Universal-Attack Framework

This section explores the performance of the universal-attack framework. We used the same setting as the standard framework during the training steps. The input size of the generator is 100 in our experiment. Note that the input of the generator in this framework is noise z.

Table 4: Accuracy achieved by the classifier on the sets of samples concerning MNIST and CYFD. Note that adversarial examples were generated by our universal-attack framework.

Model	MNIST				CYFD			
	X	R_X	A	R_A	X	R_X	A	R_A
VAE	99.47%	91.96%	99.28%	25.54%	99.39%	76.58%	73.93%	35.23%
VAEGAN	99.47%	91.89%	99.28%	25.46%	99.39%	84.73%	74.54%	24.24%

Although Table 4 shows that this universal-attack framework is highly effective in MNIST and CYFD, it is weaker than the standard framework when we compare their results in Table 1. The generator trained by the standard framework can reduce the accuracy more than the one trained by this framework. In addition, this universal-attack framework could outperform FGSM and PGD in MNIST even though these baselines required clean images as their inputs, as seen in Table 4 and 1. Also, in CYFD, it outperformed PGD. However, FGSM slightly outperformed this framework as seen in Table 4 and the column FGSM in Table 1.

4.4 Ablation Study

Since our framework consists of several components, we show what happens when we exclude the classifier and/or the discriminator. Then, we used the standard framework to find adversarial examples of VAE on MNIST to demonstrate the effect of lacking each part. Hence, we evaluated four frameworks: the complete standard framework (F_1), the standard framework without the discriminator (F_2), the standard framework without the classifier (F_3) and the standard framework without the discriminator and the classifier (F_4).

Table 5: Accuracy achieved by the classifier on the sets of samples from MNIST with VAE as the target. Note that adversarial examples were generated by our standard framework.

Framework	X	R_X	A	R_A
F_1	99.47%	91.51%	99.22%	4.84%
F_2			99.23%	4.04%
F_3			99.21%	76.65%
F_4			99.19%	18.72%

Table 5 shows the results after we trained and evaluated the frameworks (i.e., F_1, F_2, F_3 and F_4). Implicitly, by training the framework with the classifier, the generator could generate more effective adversarial examples than training it without the classifier due to the higher accuracy of the classifier on R_A. Further, training it with the discriminator slightly reduced the accuracy of the classifier on R_A. However, we found that training with the discriminator results in more valid reconstructed images than training without it. The visual comparison can be found in Appendix C.

Table 6: Accuracy achieved by the classifier on the sets of samples with retrained VAE as the target, attacked by the baselines and the universal-attack framework (UF)

Dataset	X	R_X	FGSM		PGD		UF	
			A	R_A	A	R_A	A	R_A
MNIST	99.47%	91.7%	99.05%	75.15%	99.41%	79.55%	99.12%	89.68%
CYFD	99.39%	89.21%	82.89%	48.07%	99.39%	49.49%	71.08%	70.67%

Table 7: Euclidean distance and cosine similarity between the latent spaces of VAE with standard dataset and the adversarial examples on MNIST and CYFD

Metric	No Retrain				Retrain			
	FGSM	PGD	SF	UF	FGSM	PGD	SF	UF
MNIST								
Euclidean	4972	4695	5855	5443	4689	4678	4618	4590
Cosine	0.72	0.75	0.67	0.71	0.76	0.76	0.76	0.76
CYFD								
Euclidean	267.58	236.95	267.53	262.35	283.86	264.73	256.57	270.89
Cosine	0.64	0.71	0.64	0.65	0.58	0.69	0.68	0.60

5 RETRAINING FOR ROBUSTNESS

In addition to finding adversarial examples, our framework can improve generative models' robustness by utilizing the trained framework from the previous section. Intuitively, suppose a generative model is robust against the untargeted attack. In that case, it will also be robust against the targeted attack since the untargeted attack will likely result in reconstructed images belonging to the easiest classes for their ground-truth classes. Similarly, suppose a generative model is robust against the standard framework. In that case, it will be robust against the universal-attack framework because it is much more difficult to find adversarial examples than the standard one. Therefore, we used only the trained standard framework for the untargeted attack to retrain a pretrained generative model to promote its robustness.

Note that we used only a pretrained VAE as our target for our experiment. For each epoch of our retraining, we first trained our framework for two epochs with the training mentioned in Section 3 and then trained the target for one epoch with the general training of VAE. However, the training samples of the target included the general training samples used in the previous section and the samples generated by our generator in the framework after feeding the general training samples. Hence, the target's training samples were twice more extensive than the general training samples.

To evaluate it, we used our untrained framework and trained it with the retrained target by following the instruction of the previous section. As a result, when we fed the samples created by our generators in the standard frameworks to the retrained target, we obtained the reconstructed images similar to the input. We achieved an accuracy of 89.5% in MNIST and 81.87% in CYFD on the reconstructed images produced by the retrained target in the standard framework. Furthermore, the classifiers also achieved high accuracy on the reconstructed images produced by the retrained target in both the universal framework and the baselines (i.e., FGSM and PGD), as demonstrated in Table 6. However, in CYFD, the robustness of the target against those baselines did not significantly improve in those baselines. They were still more robust than the original target. Moreover, we did not find any drawback in this retraining process because the accuracy of the classifier on R_X reconstructed by the retrained target was about the same as the one reconstructed from the natural target, as seen in Table 6.

Additionally, we also trained the standard frameworks with $\epsilon = 0.05$ and $\epsilon = 0.15$. Note that we used $\epsilon = 0.1$ for all the previous experiments on MNIST. We found that the framework with $\epsilon = 0.05$ did not work on the retrained VAE and the framework with $\epsilon = 0.15$ was less effective. However, the attack with $\epsilon = 0.15$ was too obvious.

6 LATENT SPACE ANALYSIS

This section experiments the latent spaces of a targeted model (i.e., VAE) on MNIST and CYFD. Table 7 shows the average differences between the latent spaces of VAE after passing the clean images and adversarial examples. Noticeably, the euclidean distance and cosine similarity follow the same pattern. That is, when the euclidean distance increases, the cosine similarity decreases. According to Table 1 and Table 7, even though SF can harm VAE the most, it also changes the latent spaces the most. Therefore, if an attacker needs the latent space to remain the same, we can add another factor to the loss function for it. Further, when we retrain VAE with our SF, as seen in Table 7, all the similarities increase.

For CYFD, in Table 7, the results of no-retrained VAE follows the same trend as in MNIST. Nonetheless, The results of the retrained VAE are interesting. Although the retrained VAE is more robust against those attacks than the no-retrained one, the cosine similarity achieved by FGSM, PGD and UF decrease from the no-retrained one. This phenomenon implies that the VAE for CYFD is not very organized since the latent vectors that are far away from the original ones can result in the similar reconstructed images.

7 DISCUSSION AND CONCLUSION

We proposed two GAN-based frameworks. We focused on attacking generative models that reconstructed images that looked like their corresponding inputs and mathematically defined the problem. We discussed their corresponding loss functions and how to train standard and universal-attack frameworks. After training a generator in the standard framework, the framework needs an image input to find how to perturb the image within the specified perturbation bound to generate an adversarial example for a targeted generative model. As a result, it is effective over MNIST on both untargeted and targeted attacks. Similarly, the strategy is also effective on CYFD for both attacks, although not as good as in MNIST. Furthermore, we created two baselines based on state-of-the-art adversarial attacks in the literature (e.g., FGSM and PGD) for generative models. Essentially, the proposed attack could outperform those baselines.

Moreover, we further experimented with the universal-attack framework. It trained a generator which is also effective in MNIST. Although it could not reduce the classification accuracy of the classifier as much as the one trained by the standard framework, it is perfect for attacking a target in real-time. Furthermore, from the ethical point of view, although our framework can be negatively used, we could successfully apply our framework to retrain generative models to improve their robustness and empirically show that our attacks can no longer harm the retrained generative models. Also, the retrained model could be robust against other attacks (i.e., FGSM and PGD).

Despite the success of our attack, it needs so much preprocessing time to train the whole framework. Hence, this preprocessing can be a significant limitation of this attack while FGSM and PGD do not require any training or preprocessing.

In addition to the generative models that reconstruct the inputs, these frameworks can be applied to evaluate other applications that use generative models (e.g., interpretable keypoints from videos (Jakab et al., 2020), anomaly detection (Zhou & Paffenroth, 2017), and graph prediction (Tran, 2018) by slightly adjusting the loss functions and their architectures. We leave this as future work.

REFERENCES

Tao Bai, Jun Zhao, Jinlin Zhu, Shoudong Han, Jiefeng Chen, Bo Li, and Alex Kot. Ai-gan: Attack-inspired generation of adversarial examples. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2543–2547. IEEE, 2021.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

François Chollet. Variational autoencoder. https://keras.io/examples/generative/vae/. Accessed: 2020-05-03.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8787–8797, 2020.

Amy Jang. Tensorflow: Mnist cnn tutorial. https://www.kaggle.com/amyjang/tensorflow-mnist-cnn-tutorial. Accessed: 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 ieee security and privacy workshops (spw)*, pp. 36–42. IEEE, 2018.

Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 853–862, 2017.

K.C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

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.

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.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387. IEEE, 2016a.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pp. 582–597. IEEE, 2016b.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Phillip Pope, Yogesh Balaji, and Soheil Feizi. Adversarial robustness of flow-based generative models. In *International Conference on Artificial Intelligence and Statistics*, pp. 3795–3805. PMLR, 2020.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Qing Rao and Jelena Frtunikj. Deep learning for self-driving cars: Chances and challenges. In *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*, pp. 35–38, 2018.

Bidisha Samanta, Abir De, Gourhari Jana, Vicenç Gómez, Pratim Kumar Chattaraj, Niloy Ganguly, and Manuel Gomez-Rodriguez. Nevae: A deep generative model for molecular graphs. *Journal of machine learning research. 2020 Apr; 21 (114): 1-33*, 2020.

Matthew Stewart. Gans vs. autoencoders: Comparison of deep generative models. https://towardsdatascience.com/gans-vs-autoencoders-comparison-of-deep-generative-models-985cf15936ea. Accessed: 2019-05-12.

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.

Pedro Tabacof, Julia Tavares, and Eduardo Valle. Adversarial images for variational autoencoders. *arXiv preprint arXiv:1612.00155*, 2016.

Phi Vu Tran. Learning to make predictions on graphs with autoencoders. In *2018 IEEE 5th international conference on data science and advanced analytics (DSAA)*, pp. 237–245. IEEE, 2018.

Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Xianwen Yu, Xiaoning Zhang, Yang Cao, and Min Xia. Vaegan: A collaborative filtering framework based on adversarial variational autoencoders. In *IJCAI*, pp. 4206–4212, 2019.

Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 665–674, 2017.