Adversarial Attacks Against Variational Autoencoders

Haowei Lin
Yuanpei College
Peking University
linhaowei@pku.edu.cn

Abstract

Adversarial examples have been extensively investigated for a variety of deep neural networks. However, most previous work like Szegedy et al.(2013) and Goodfellow et al.(2014) focuses on the application of adversarial examples to the task of classification. It is also revealed that deep generative models, like Variational Autoencoders(VAE), also suffer from adversarial attacks. This summary introduces three papers, Tabacof et al.(2016), Kos et al.(2018) and Gondim-Ribeiro et al.(2018) about adversarial attacks against VAE. Unlike classification scenarios, the research about attack algorithm and evaluation is much tougher and there hasn't been unified procedures in this task yet. On top of that, I will propose some sensible ideas for future exploration in section 4.

1. Introduction

Variational autoencoders (VAEs) are a powerful method for learning deep generative models, finding application in areas as image and language generation as well as represetnation learning. But like other deep neural networks, VAEs are susceptible to adversarial attacks, whereby small perturbations of an input can induce meaningful, unwanted changes in output.

However, the vulnerability of VAEs is undesirable. Shown in Kos et al.(2018) and Gondim-Ribeiro et al.(2018), VAEs can be seen as compression methods for transmission systems (Fig 1). Attacks on VAEs may potentially arise safety problem in information transmission, hypothetically.

In comparison to the extensive literature on adversarial attacks for classifiers, attacks for VAEs are unexplored. The reason relates to the difficulty of performing attack and assessment towards VAEs. VAEs are more robust to classifiers built upon deep architectures thus make attacking difficult. What's more, there are no clear-cut success criteria for VAE and, neither for the attack.

To overcome these difficulties, three papers design various promising attack methods and evaluation protocols for the task. The earliest research, Tabacof et al.(2016) comes up with the first adversarial attack on VAE. It also firstly reveals the different robustness properties between autoencoders and classifiers. The second work, Kos et al.(2018), proposes three different attack methods: attacking an extraneous classifier after the latent representation, attacking the latent representation directly with an ℓ_2 objective, and attacking the output of the decoder using the VAE loss function. They also introduces a quantitative evaluation. The last work, Gondim-Ribeiro et al.(2018) is an extended version of the first paper, with a advanced scheme to attack and quantitatively evaluate VAEs in more datasets.

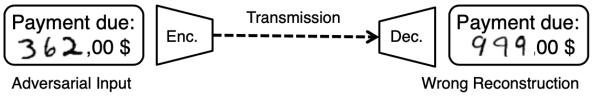

Figure 1. Depiction of the attack scenario. Motivated by [3]: an attacker may convince a sender to transmit a seemingly innocent document that is reconstructed in a malicious way.

2. Attack Methods

In conventional evasion attack, attacker minimize an adversarial loss to mislead the model, with distorting the input as little as possible. One can be even more strict to only allow a perturbation within a ϵ-ball. To generate adversarial examples for classification, the classical methods like FGSM proposed by Goodfellow et al.(2014) maximize the misdirection towards a certain wrong label or away from the correct one. In VAEs, there is not a single class output to misclassify, so the attack attempts to mislead the reconstruction: if a slightly altered image entered the VAE, the reconstruction is wrecked. On top of that, three papers proposed their methods for the goal respectively as follows.

2.1. Latent Attack

Tabacof et al.(2016) is the first to perfrom attacks towards VAE. Their attack consists in selecting an original

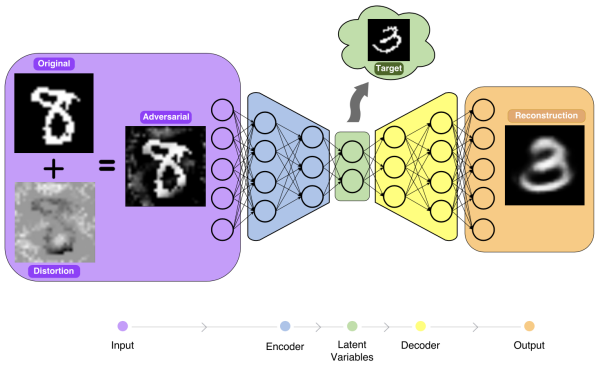

Figure 2. Adversarial attacks for VAEs add small distortions to the input aiming at making the VAEs reconstruct a different target. [5] attack the latent representation, attempting to match it to the target image's latent representation.

image and a target image, and then feeding the network the original image added to a small distortion, optimized to get an output as close to the target image as possible. But minimizing the distance to the target image fails so they attack the latent representation instead. They use the following adversarial optimization:

$$\begin{aligned} \min_{\boldsymbol{d}} \quad & \Delta(\boldsymbol{z}_a, \boldsymbol{z}_t) + C||\boldsymbol{d}|| \\ \text{s.t.} \quad & L \leq \boldsymbol{x} + \boldsymbol{d} \leq U \\ & \boldsymbol{z}_a = \text{encoder}(\boldsymbol{x} + \boldsymbol{d}) \end{aligned}$$

where $\boldsymbol{d}$ is the adversarial perturbation, $\boldsymbol{z}_a$ and $\boldsymbol{z}_t$ are the latent representations, respectively for the adversarial and the target images. $\boldsymbol{x}$ is the original image. $\boldsymbol{x} + \boldsymbol{d}$ is the adversarial image, L and U are the bounds on the input space, and C is the regularizing constant the balances reaching the target and limiting the distortion. The attack succeeds to fool the VAEs, and is also adapted by Kos et al.(2018) for GAN-VAE tasks.

Here we need to notice that, Δ is a kind of matric of distance. [5] and [3] take ℓ_2 distance and [1] chooses KL-divergence.

2.2. Classifier Attack

By adding a classifier f_{class}, Kos et al.(2018) turn the attack problem into previously solved attacks for classifiers. The attack procedures are as follows.

Step 1. Froze the weights of VAEs, and train a new classifier $f_{class}(\boldsymbol{z}) \rightarrow \hat{y}$ on top of encoder. We need to get the corresponding labels for latent representation $\boldsymbol{z}$.

Step 2. With the trained classifier, the attacker finds adversarial examples $x^\star$ using the methods on attacking classifiers.

This method does not always result in high-quality reconstructions, because f_{class} adds additional noise to the process.

2.3. $\mathcal{L}_{\text{VAE}}$ Attack

The second method proposed by Kos et al.(2018) generates adversarial perturbation using the VAE loss function. The adversary precomputes the reconstruction $\hat{\boldsymbol{x}}_t$ by

$$\hat{\boldsymbol{x}}_t = \text{decoder}(\text{encoder}(\boldsymbol{x}_\text{t}))$$

Then instead of computing the reconstruction loss between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, the loss is computed between $\hat{\boldsymbol{x}}^\star$ and $\hat{\boldsymbol{x}}_t$. Here $\hat{x}$ symbol denotes the reconstructed image, and $x^\star$ symbol denotes the adversarial example.

3. Evaluation

When attacking classifiers, there is a clear-cut criterion for success: the target class has higher probability than all others (targeted attacks), or the right class has lower probability than some other (untargeted attacks). When attacking VAEs, there seems no sharp criterion for success.

Fo targeted attacks, Tabacof et al.(2016) and Gondim-Ribeiro et al.(2018) proposed **Distortion-Distortion plots**. And the **AUDCC** (Area under Distortion-Distortion Curve) is used to evaluate the robustness of model or the success of attack. The Distortion-Distortion plots show, with each attempt, how much the original is distorted and how much the reconstructed reaches the target (measured by ℓ_2). The graph is normalized so that the distance between lines is 1. and AUDCC is the area under the curve given by the linear interpolation of the experiment points. The closer this area is to 1, the more resistant the model was to the attack (the less successful the attack was). See Fig **??** for an example.

Kos et al.(2018) provide a quantitative metric based on an ancillary classifier network. The classifier uses the latent representation as input, and is trained on labels related to the input. They compute two metrics: $AS_{ignore-target}$ (how often the reconstruction of the attack input mislead the classifier) and AS_{target} (how often the reconstruction of the attack input matched with the class of the target).

4. Conclusion and Ideas for Future Exploration

what did the papers do well? These papers are the first to explore the robustness of VAEs and successfully make it to perform attacks on it. They have tried various experiments and provided comprehensive views for robustness of VAEs. Also, the attack methods and evaluation criteria are natural and feasible to some extent. Exploration in attacking VAEs is rare, so they makes meaningful breakthrough.

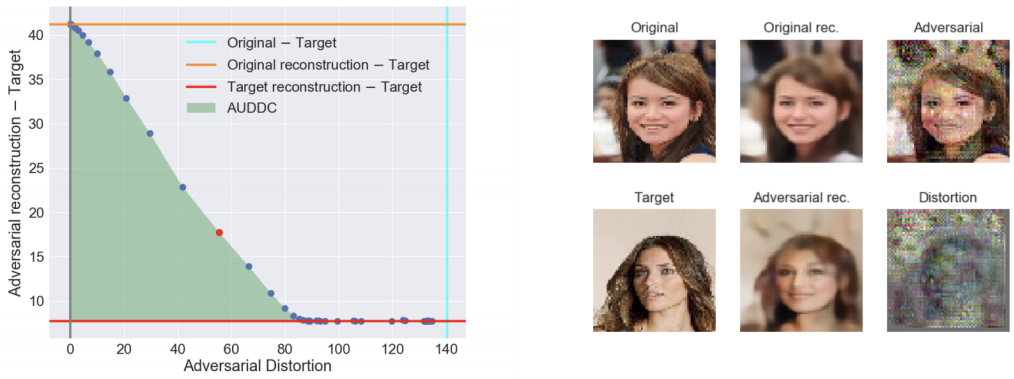

Figure 3. Left: the proposed metric: Area Under the Distortion-Distortion Curve (AUDCC). Right: visualization of a single point (red dot) of the left plot.

where did the papers fall short? Apparently, since adversarial attacks are new in deep learning research, the attacking methods used in three papers are naive. The attack scenarios are also restricted in only white box evasion attacks, and the optimization is just simple. Also, the evaluation protocols can only applied to targeted attack and sometimes labeled datasets.

Questions about the papers.[1] Neither of the papers tried untargeted attacks towards VAEs. I wonder why they didn't try it. Maybe the reconstructed image may be noisy, but somehow I think this attack is also meaningful. Also, why is the transferability of adversarial examples not explored? Since the transferability is proposed in Szegedy et al.(2013), they should have known the property.

For future exploration, we may focus on advanced attack methods proposed nowadays. And the various properties in attack scenarios like black-box attacks, backdoor attacks maybe probable. More importantly, the evaluation for the attack should be designed more properly for generative models. Only do we make the works solid in attack against VAEs, can we explore the robustness and defense for VAEs.

References

[1] George Gondim-Ribeiro, Pedro Tabacof, and Eduardo Valle. Adversarial attacks on variational autoencoders. *arXiv preprint arXiv:1806.04646*, 2018. 1, 2

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[3] Jernej Kos, Ian Fischer, and Dawn Song. Adversarial examples for generative models. In *2018 ieee security and privacy workshops (spw)*, pages 36–42. IEEE, 2018. 1, 2

[4] 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. 1, 3

[5] Pedro Tabacof, Julia Tavares, and Eduardo Valle. Adversarial images for variational autoencoders. *arXiv preprint arXiv:1612.00155*, 2016. 1, 2

[1] the required "what did you learn from these papers" is described as the whole summary.