Project Proposal – Attacking PuVAE

Nao Rho **Fan Yang** **Junran Yang**

1 Introduction

Deep learning is providing major breakthroughs in solving the problems that have withstood many attempts of machine learning and artificial intelligence community in the past. With the continuous improvements of deep neural network models, open access to efficient deep learning libraries, and easy availability of hardware required to train complex models, deep learning is fast achieving the maturity to enter into safety and security critical applications, e.g. self driving cars, surveillance, malware detection, and voice command recognition. However, the technology comes with a severe downfall. In 2014, Szegedy et al. discovered an intriguing weakness of deep neural networks in the context of image classification.[2] They showed that despite their high accuracies, modern deep networks are surprisingly susceptible to adversarial attacks in the form of small perturbations to images that remain (almost) imperceptible to human vision system. Such attacks can cause a neural network classifier to completely change its prediction about the image. Even worse, the attacked models report high confidence on the wrong prediction. For instance, a hacker can construct an image of a $100 check that looks harmless to a banker or a machine, but, with careful construction, a machine learning algorithm can recognize it as $999 with high confidence. Such attack is detrimental to industrial applications and must be dealt beforehand.

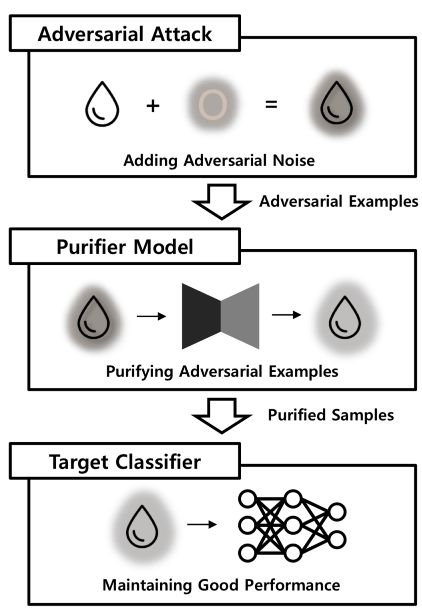

Figure 1: Overview of the defense mechanism using the purifier model.

2 Motivation

Yoon et al. at Seoul National recently proposed PuVAE (Purifying Variational Autoencoder) [1], which is built upon CVAE (Conditional Variational Autoencoder), for purifying adversarial examples. The proposed method eliminates an adversarial perturbation by projecting an adversarial example on the manifold of each class, and determines the closest projection as a purified sample. It's shown that the proposed method performs competitively with state-of-the-art defense methods against various attacks including FGSM (Fast Gradient Sign Method) and CW (The Carlini and Wagner attack), and the inference time is approximately 130 times faster than that of Defense-GAN, which is the state-of-the art purifier model [3].

Though the paper demonstrated that PuVAE can effectively prevent target classifiers from being attacked by adversarial examples and adversarial training, the authors did not conduct any experiments to attack the whole PuVAE. In other words, they did not conduct experiments to to prove the robustness of VAE part of PuVAE, but only the classifier. Thus, whether PuVAE itself is invulnerable to any sorts of attack remains a question.

3 Method

Compared to attacks on classifiers, attacks on autoencoders are much less explored [4,5,6]. Moreover, to our best knowledge, attacks on CVAE and its variation PuVAE have never been reported. Attacking autoencoders and their variations is a more involved procedure than attacking classifiers. In the latter we target a small output vector, often focusing at just one or two values on that vector. In the former we need to address a very high-dimensional output. Targeted attacks to autoencoders consist in adding (as small as possible) adversarial distortion to the original input in order make the reconstructed output as close as possible to the target.

Tabacof et al. introduced attacks on autoencoders and variational autoencoders, showing that they are possible, although much harder than attacks on classifiers [4]. They attacked the latent representation with a KL-divergence objective in both MNIST and SVHN. They showed that there is a linear trade-off between the intensity of the input distortion and the degree of success in the attack - frustrating the hope that a small change in the input could lead to drastic changes in the reconstruction. Kos et al. followed up with a work that attacked both the latent representation and the output of VAE–GAN autoencoders [6]. They proposed three modes of attack: attacking an extraneous classifier after the latent representation, attacking the latent representation directly with an l2 objective, and attacking the output of the decoder using the VAE loss function. They introduced a quantitative, although indirect, evaluation of attack inferred from success in fooling the extraneous classifier.

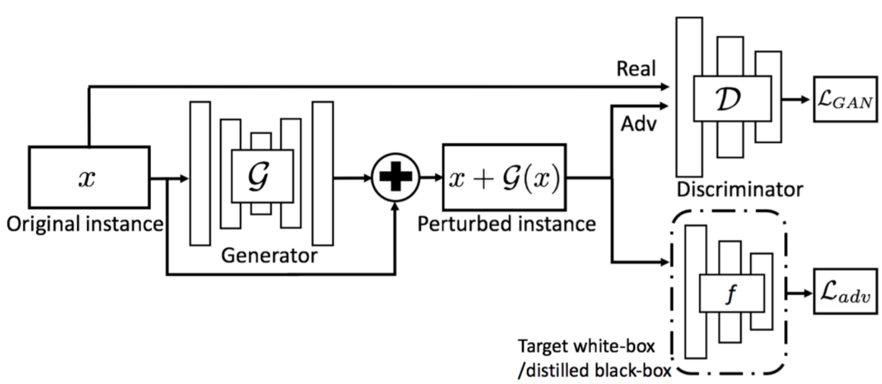

Figure 2: Overview of advGAN

In this project, our group will verify the efficiency of PuVAE against a list of attacking methods not tested in the paper (the paper used FGSM, CW and 2 of their variations). Some possible choices include Jacobian-based Saliency Map Attack (JSMA), UPSET and ANGRI, DeepFool, etc. A more important goal of the project is to crack PuVAE and consequently break the defense, by adapting the methods proposed in [4] and [6] for attacking regular VAEs to attack PuVAE. However, considering the how hard attacking VAEs is and that very few papers on this topic can be found, we are not sure if cracking PuVAE is possible. In fact, it remains as a question whether this PuVAE is resilient against attacks targeted at PuVAE itself. Another approach is to take advantage of advGAN, an adversarial network proposed in [5] for generating adversarial examples. The idea is to use PuVAE as the target white-box model, as shown in Figure 2, and train a GAN which can generate images indistinguishable from benign images but can fool the discriminator when added a small perturbation.

References

[1]Uiwon Hwang, Jaewoo Park, Hyemi Jang, Sungroh Yoon: “PuVAE: A Variational Autoencoder to Purify Adversarial Examples”, 2019; [http://arxiv.org/abs/1903.00585 arXiv:1903.00585].

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.

[3]Pouya Samangouei, Maya Kabkab: “Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models”, 2018; [http://arxiv.org/abs/1805.06605 arXiv:1805.06605].

[4]George Gondim-Ribeiro, Pedro Tabacof: “Adversarial Attacks on Variational Autoencoders”, 2018; [http://arxiv.org/abs/1806.04646 arXiv:1806.04646].

[5]Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu: “Generating Adversarial Examples with Adversarial Networks”, 2018; [http://arxiv.org/abs/1801.02610 arXiv:1801.02610]

[6] J. Kos, I. Fischer and D. Song, "Adversarial Examples for Generative Models," 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, 2018, pp. 36-42. doi: 10.1109/SPW.2018.00014