6.883 Final Project Plan: Adversarial Generation and Perturbation Elimination with GANs

Mucong Ding, Sirui Lu, Zhiwei Ding

Neural networks are vulnerable to adversarial examples: samples that are close to clean inputs but classified incorrectly. They pose potential security threats to practical applications of machine learning. Adversarial generation and perturbation elimination (i.e. defense by pre-processing the inputs to remove the adversarial perturbations) can be naturally formulated as a two-player game, which implies generative adversarial networks (GANs) might be a generic and effective solution to both of the problems. In the past several months, a few papers have proposed some novel approaches based on GANs for attack and defense, including [1] and [2]. However, each of them merely studies a half of the problem. In this project, we plan to re-implement three GAN-based attack and defense methods and analyze their advantages and shortcomings by testing them against the state-of-art attacks and defenses. Our ultimate goal is to improve them by combining the attack and defense models as one GAN and investigate the entire problem in terms of the min-max formulation.

I. The Problem

A. Inspiration

In the context of adversarial examples, designing attack and defense methods are the two sides of one problem. As a two-player game, we expect that we can get a deeper understanding of the problem if we investigate the two sides as a whole. However, most of the current designs do not take their opponents into consideration (i.e. a defense method which is attack-agnostic), and no conclusion can be drawn about their effectiveness before evaluating them against their opponents. In this project, we plan to study the whole problem in terms of a min-max formulation and obtain neural-network-based attacks and defenses using GANs.

B. Why it is important?

GAN provides a principled approach to understand the adversarial problem, as it directly models the distributions of the clean and adversarial samples. In this regard, the difference between the original and adversarial distributions is explicit, and this could deepen our understanding of the intrinsic characteristics of adversarial examples.

C. Applications

GAN-based attacks and defenses have the potential to outperform the start-of-art approaches, as they are trained against their opponents and know them better. We plan to evaluate the GAN-based methods against the state-of-art benchmarks to get an idea of their performance.

II. Prior Work

A. Adversarial Generation with GANs

A few attempts have been made to generate adversarial examples using GANs, for example, [1] and [3]. Some modifications of the GAN's architecture are proposed to make the generated sample adversarial for a given black-boxed classifier. However, the related papers did not provide comprehensive evaluations of their GAN-based attacks against the start-of-art defenses. In this project, we will carefully evaluate the AdvGAN proposed in [1].

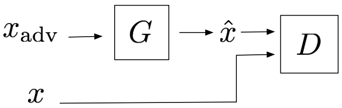

Fig. 1: Sketch of APE-GAN[2].

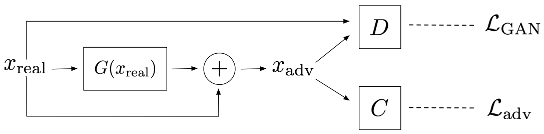

Fig. 2: Overview of AdvGAN[1].

B. Adversarial Perturbation Elimination with GANs

GANs are also used to eliminate the perturbations in adversarial examples and reconstruct the

Emails of the authors: mcding@mit.edu, sirui@mit.edu, dingzw@mit.edu

closet clean input. When it is applied before the classification, the purified input can be classified correctly again. Two recently proposed models are the APE-GAN in [2] and the Defense-GAN in [4][5]. However, in [6], it is shown that APE-GAN can be defeated by Carlini and Wagner (CW) attack with a special loss function. The Defense-GAN partially solve this problem as it follows an Invert and Classify (INC) approach and we cannot back-propagate to train CW attacks. Defense-GAN is shown to be effective against many attacks. [7].

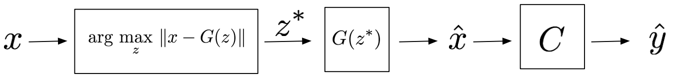

Fig. 3: The schematic of Defense-GAN[4].

III. Papers and Code

A. List of Papers

We will carefully investigate on [1] [4] [2] and [5] and re-implement their models, namely AdvGAN, Defense-GAN and APE-GAN. These papers describe the state-of-the-art techniques of adversarial generating and perturbation eliminating using GANs.

B. Released Code

The code of APE-GAN ([2]) is released. The code of AdvGAN ([1]) can probably be obtained by contacting the corresponding authors. The code of Defense-GAN ([4]) is also available in the repository of [7] which evaluate its performance against several attacks.

IV. Objectives & Road-map

A. Mid-Project Milestone

First, we aim to re-implement the GAN models in [1], [4] and [2] for **MNIST** and **CIFAR-10** datasets. We will evaluate the attacks obtained from AdvGAN against the state-of-art defenses including Adversarial training [8]. We will evaluate the Defense-GAN and APE-GAN against CW attack and the attack generated by AdvGAN. We will conduct a thoughtful analysis on their performances.

B. Final Goal

We will try to improve the existing work by combing the models used for adversarial generation and perturbation elimination into one GAN, for exmaple, combining the AdvGAN and APE-GAN as shown in Fig. . We will rigorously evaluate the effectiveness and robustness of attack and defense methods obtained from the combined-GAN by the approach proposed in [9].

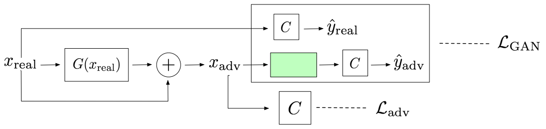

Fig. 4: A preliminary schematic of combining both attacks and defense. The green box here shares the same structure as part of Figure 1.

.

References

[1] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," *arXiv preprint arXiv:1801.02610*, 2018.

[2] S. Shen, G. Jin, K. Gao, and Y. Zhang, "Ape-gan: Adversarial perturbation elimination with gan," *ICLR Submission, available on OpenReview*, 2017.

[3] W. Hu and Y. Tan, "Generating adversarial malware examples for black-box attacks based on gan," *arXiv preprint arXiv:1702.05983*, 2017.

[4] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," 2018.

[5] A. Ilyas, A. Jalal, E. Asteri, C. Daskalakis, and A. G. Dimakis, "The robust manifold defense: Adversarial training using generative models," *arXiv preprint arXiv:1712.09196*, 2017.

[6] N. Carlini and D. Wagner, "Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples," *arXiv preprint arXiv:1711.08478*, 2017.

[7] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," *arXiv preprint arXiv:1802.00420*, 2018.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[9] N. Carlini, G. Katz, C. Barrett, and D. L. Dill, "Ground-truth adversarial examples," *arXiv preprint arXiv:1709.10207*, 2017.