A Review of IDSGAN: Generative Adversarial Networks for Attack Generation against Intrusion Detection

Rohit Das
ID No.: 11910230
M. Tech. (CSE)
Reviewer: Dr. Sk Subidh Ali

October 16, 2019

I. Summary

The manuscript by Lin, Shi and Xue outlines how a Generative Adversarial Network (GAN) was developed to generate adversarial malicious network data and successfully evade any black-box intrusion detection system (IDS). The manuscript aims to fill a void in research into attacks on IDS, and particularly the use of GANs in such attacks.

II. Outline

Listed below are the milestones covered by the paper:

- An overview of why IDSGAN is necessary to build robust intrusion detection systems, and highlights on the lack of work done in IDS evasion, specifically using GANs.
- A brief explanation of why Wasserstein GAN was preferred over other variants or the vanilla GAN, and outlines how the GAN architecture had been used to train the generator to generate adversarial data.
- Usage of a benchmark dataset, NSL-KDD, consisting of network data for both normal and malicious traffic, which had been divided to be separately fed into the black-box IDS and the GAN discriminator and generator.
- Visual representations of the modified architecture of Wasserstein GAN used for training the generator. It also specifies the layers used in the generator to add noise to the malicious data from the dataset.
- Detailed layout of the discriminator and generator, the role of a black-box IDS in helping to train the discriminator and eventually the generator to evade the black-box. Also provided are the training algorithms, pseudo code for the same, loss functions and formulae used.
- Extensive supporting data and graphs to throw light on how current IDSs are vulnerable to adversarial attacks and how the GAN framework developed was successful in evading most models used as a black-box.

III. Best Elements

Below mentioned are the best elements of the paper:

- The paper has been aimed to contribute to the field of intrusion detection and evasion, and has successfully presented that current IDSs are nowhere near being secure against adversarial attacks.
- All conclusions, facts and experiments have been well-supported with relevant data and graphs. The graphs clearly highlight how the IDSGAN was successful in evading all the models used for the black-box.
- Using the predicted output labels of the black-box as target labels for the discriminator in the Wasserstein GAN used was an innovative way to train the discriminator with parameters having values as close as possible to those of the black-box. This helped the generator to train and accordingly add noises to malicious examples that successfully evade the IDS.
- All equations, formulae and algorithms mentioned have been lucidly presented such that even with basic knowledge, the paper can be understood and appreciated.

IV. Limitations

The limitations of the paper are listed below:

- The paper has clearly not been proof-read adequately. It is filled with grammatical errors. Often the sentences are too garbled to make out the true meaning, and other times, some words have been morphed such that the actual meaning of the sentence changes.
- There are also informational errors, occurring while describing the values in NSL-KDD, errors in images and pseudo-code, which could have been avoided. Also, some parts of the text has been left in ambiguity.
- Since the dataset has less data under U2R and R2L categories of attacks, the paper's claim that the GAN has successfully evaded IDS with those attacks is trivial and weak.

V. Opinion

While the idea and solution presented here is novel and standard procedures for experiments and data representation have been used, the paper itself is of average quality filled with mistakes and errors, easily avoidable had it been proof-read properly.