Modeling and Generating Cyber Attack Data Using Generative Adversarial Networks

Bryan Quinn; Shanchieh Jay Yang; Bo Yuan; Matthew Wright
Rochester Institute of Technology

Abstract

Machine learning techniques to network-based intrusion detection systems (NIDS) require quality, labeled, training data and a balance between benign and malicious traffic.

This work investigates a solution to these two problems by applying Generative Adversarial Networks (GANs) to generate NetFlows that resemble real cyber-attack data.

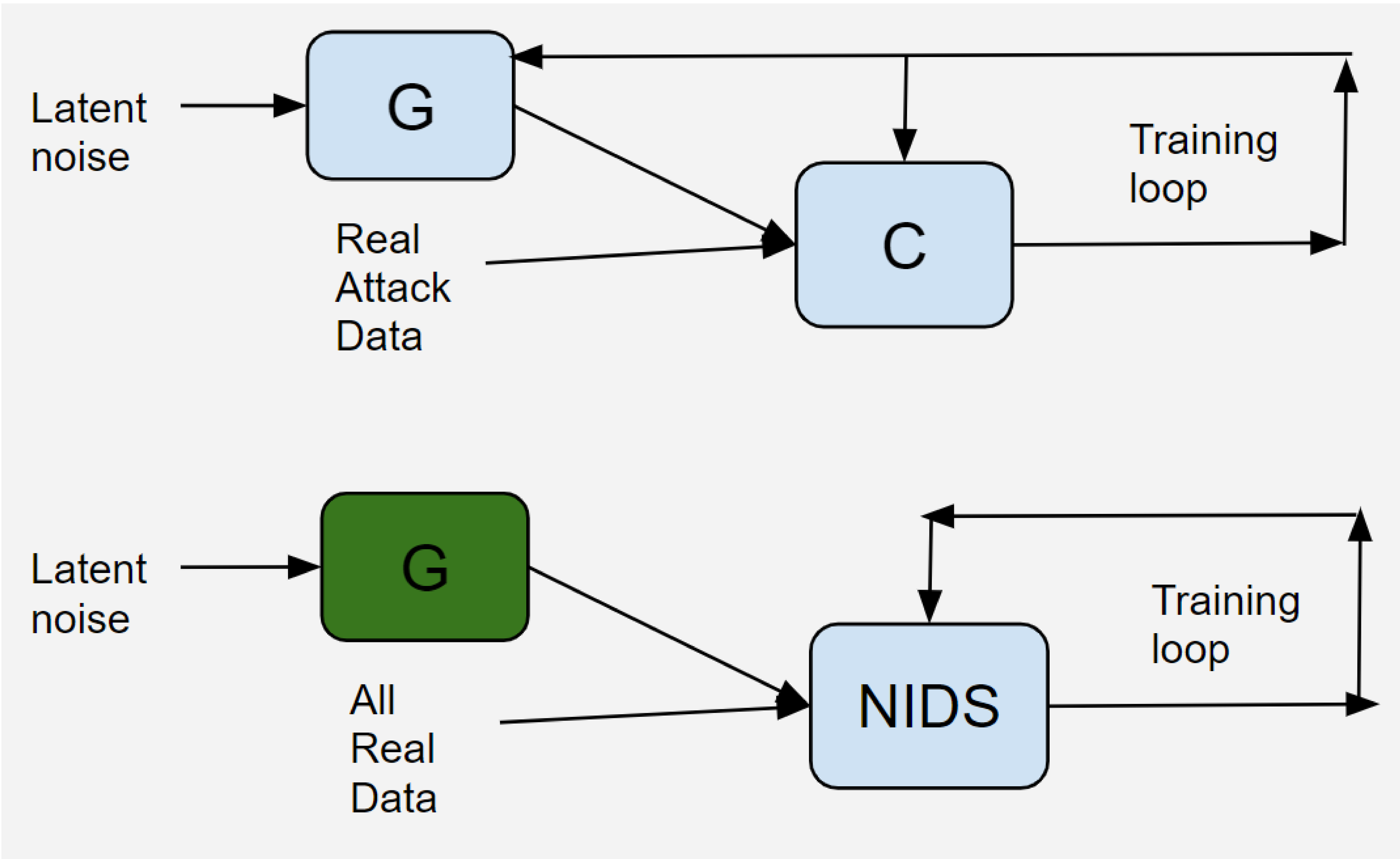
The resulting models show a reduction in error over a model trained on only the original data for each attack type. With a reduction in test error rate from 0.3% to 0.0% for SSH-Bruteforce and from 7.1% to 0.1% for Botnet.

Process

Train GAN on real attack data from the CSE-CIC-IDS2018 dataset.

Balance a dataset of real attack data between benign and malicious samples while also creating more robust examples.

Train a new classifier that shows improved performance on Network-based Intrusion Detection against unseen data.



Semi-Supervised Learning GAN training and generating

Methods

Specifically, we report on training GANs to generate data matching two types of attacks, SSH-Bruteforce and Botnet, from the CSE-CIC-IDS2018 dataset..

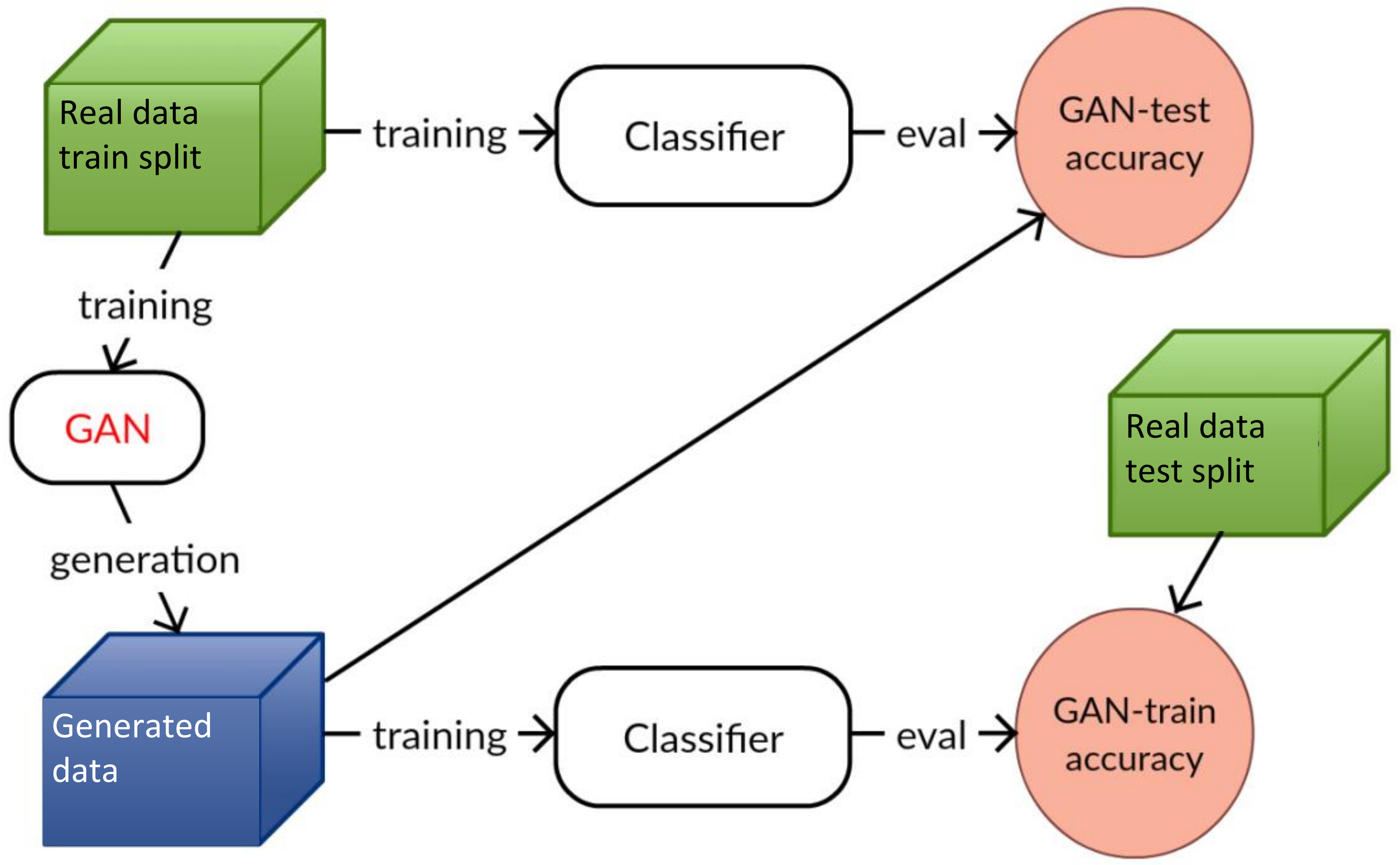
The GAN models learn the values of the features and relationships between them that would normally be seen in these attacks. These models can then be used to generate more examples of intrusion data for training purposes, allowing researchers to create more malicious examples needed to balance a training set against benign samples.

Evaluation

Two metrics -- GAN-test and GAN-train -- are used to analyze the validity of the generated data. To evaluate the utility of the generated data in a semi-supervised approach, I use the generated data together with the original training data to train classification algorithms for intrusion detection.

GAN-test: a classifier trained on real data and evaluated on generated malicious data.

GAN-train: a classifier trained on generated data and evaluated on real data.



GAN-train GAN-test evaluation process [1]

Results

TABLE I: WGAN Performance (Accuracy rounded)

Attack Type	Original Data	GAN-test	GAN-train	Balanced Data
SSH	0.997	1.000	0.890	1.000
Bot	0.929	0.964	0.995	0.999

Discussion

The current cyber attack models each provide a promising representation of the real data. Further attack types could be modeled on NetFlows resulting from other cyber attack types, either from the CIC-IDS dataset or other data sources.

To utilize this data generation in a semi-supervised approach, a classification algorithm was trained for intrusion detection using the generated data, as well as the original training data. The performance of this model was compared against the performance of a model which is trained on only the original data and showed an improvement for both attack types.

Handling of the categorical features that were dropped before training the GAN models, and how dropping those feature would affect the ability to construct attacks based on the GAN's generated NetFlow parameters. discuss how this may be accomplished in future works.

Future Work

Further work may be done to train models on more complex cyber attack types. A sequential GAN (LSTM-GAN) could be used to capture sequential stages of an attack.

We imagine a tool which would take the parameters of a NetFlow and generate an attack by manipulating the packets containing the malicious payload. The generator could produce new adversarial examples which are specifically designed to fool a NIDS system to believing they are benign.

Future works may explore other techniques using the GAN approach. A technique similar to the one used by Lin et. al. [2] may be incorporated into this work, to force the generator to produce new adversarial examples which are specifically designed to fool a NIDS system to believing they are benign.

Contact

Bryan Quinn
Rochester Institute of Technology
Email: bquinn@mail.rit.edu

References

1. Konstantin Shmelkov, Cordelia Schmid, and KarteekAlahari. How good is my gan? In Proceedings of the European Conference on Computer Vision (ECCV), pages 213–229, 2018.
2. Zilong Lin, Yong Shi, and Zhi Xue. Idsgan: Generativeadversarial networks for attack generation against intru-sion detection.arXiv preprint arXiv:1809.02077, 2018
3. Martin Arjovsky, Soumith Chintala, and L´eon Bottou.Wasserstein generative adversarial networks. volume 70ofProceedings of Machine Learning Research, pages214–223, International Convention Centre, Sydney, Aus-tralia, 06–11 Aug 2017. PMLR.