CSCE 698 - WI

Proto Prospectus

2d Lt Marvin Newlin

Proposal Abstract.

Machine learning is becoming an integral part of cybersecurity today, particularly in the area of network intrusion detection. However, machine learning techniques require large volumes of data to be effective. Although some real and semi-real data sets exist, many are outdated or do not contain enough useful information for training/classification of network intrusion detection systems (NIDS), particularly considering the new types and variants of communication protocols in today's environment. Generating sufficiently realistic synthetic network traffic is imperative for training effective NIDS. This research aims to outline a basic approach to developing synthetic data by leveraging machine learning and suggests opportunities to tailor the dataset for specific research objectives. We believe this work might have broad applicability to other researchers developing their own NIDS framework.

Annotated bibliography.

A Convolutional Neural Network for Modelling Sentences

This paper develops a model for modelling sentences using a Dynamic Convolutional Neural Network (DCNN). This approach takes in input sentences of different lengths and induces a feature graph that determines both the long and short term relationships between the words. They tested the approach in 4 different experiments and saw significant improvement over existing methods in their results. This is applicable since it deals with generating synthetic data, in this case synthetic sentences, but the process remains the same.

[1]

Improved Training of Wasserstein GANs

This paper discusses the stability improvements of a Wasserstein Generative Adversarial Network (GAN). They also discuss the limitations that it produces with weight clipping and demonstrate some of the flaws is produces. To alleviate this problem, they propose a new method of a gradient penalty to maintain the 1-Lipschitz property of the Wasserstein GAN without the instability produced by weight clipping. This is applicable since it is a modification of a GAN like I will be dealing with. Additionally, this method successfully performs on discrete data which is a performance increase over other GAN frameworks.

[2]

Complex log file synthesis for rapid sandbox-benchmarking of security- and computer network analysis tools

- Problem Statement: Generate synthetic log files that can be used for testing log based IDS.

- Applicable since this is exactly what we want to do with network data with a different approach

- Generation Method
  - Data Points are single log lines
  - Run Log Line Clustering to build clusters and then assign log lines to clusters
  - Select clusters from candidates
  - Model clusters as Markov chains
- Evaluation Method
  - Mean Coverage rate of clustering algorithm
  - Number of Outliers
  - Number of Clusters
  - These metrics evaluated on semi-synthetic data & reference logs

[3]

Generative Model for Category Text Generation

- Uses SGD as optimization method
- Mapped to real interval with SoftMax function
- Generation Method
  - Long Short-Term Memory (LSTM) w/classifier
  - RNN & RL as discriminator & classifier
- Evaluation Method
  - Negative Log-likelihood
  - Classification results on synthetically generated sentences

[4]

Adversarial Feature Matching for Text Generation

- Problem Statement: Address the mode collapsing issue of standard GAN with an LSTM RNN
- Optimization technique is to minimize a NN-based MMD distance of real data distribution & synthetic data distribution
- Evaluate with BLEU score & Kernel Density Estimation

[5]

Generating Synthetic Mobility Traffic Using RNNs

- Data points are trajectories of users (in this case trajectories is a collection of lat/longs)
- Develop a synthetic traffic generator based on an LSTM RNN
- Coordinates mapped to a grid then feature exploration is performed on the grid
- Extracted feature vectors then used for training

[6]

Toward Controlled Generation of Text

- Idea is to generate realistic (plausible sentences) with Machine Learning
- Model is a variational autoencoder fed into a generator/discriminator
- Generator is LSTM-RNN
- Variable Auto Encoder uses sleep-wake algorithm
- Applicable since we want to generate realistic network traffic

[7]

Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization

Discusses the lack of “good” available datasets for training Network Intrusion Detection Systems (NIDS). Comprehensively analyzes the available datasets and then presents a framework for evaluating the quality of datasets. They then develop their own dataset and analyze it according to their framework. Main contribution to work is the analysis of available datasets and the evaluation framework for datasets.

[8]

Generative Adversarial Nets

Seminal paper on the idea of a Generative Adversarial Network. Presents the idea of an adversarial game played between a generator and a discriminator in which the generator tries to generate data that will trick the discriminator into thinking it is real data. Solid idea that has become the basis for lots of data generation techniques with machine learning. Basic idea is the basis of the work I will be doing.

[9]

Improved Techniques for Training GANs

Improves on the framework of GANs by updating its framework and adding in semi-supervised learning. This reduces some of the mode collapse issues of the original GAN. Quality of generated data is significantly improved. Applicable to my work since I will most likely be utilizing the GAN framework.

[10]

Article Review.

Complex log file synthesis for rapid sandbox-benchmarking of security- and computer network analysis tools

This paper presents the idea of generating synthetic log files for training IDS. The approach they use is to analyze actual logs, find the clusters in the logs with a log clustering algorithm, and then selects the best clusters and from there generates assigns log lines to the best fit clusters. From there they generate synthetic log lines using a Markov Chain approach to fit each cluster and augment the existing data. The biggest assumption they make in their model is that the timestamp is independent of the content of the log line. This is not necessarily true in general but doesn’t seem to affect the results.

To evaluate the model, they generated several sets of logs and combined their generated logs with actual log files. They then fed the logs through an IDS and measured the rates that the IDS caught the issues present in the logs. On the statistical side, they utilized several measures like mean coverage rate to determine how well a cluster description matched the log line fed into it. Overall, their evaluation measures are very heavy on statistical measures which is to be expected since this is a new idea that hasn't yet been replicated. Additionally, a lack of alternatives for measuring the quality of synthetic data leaves statistical measures as the only viable option.

Additionally, to further evaluate the effectiveness of the synthetic logs, they tested on an IDS. For the evaluation, they combined their generated logs with some real logs and then compared their results against a reference log. The results looked at the false positive rates and the cluster coverage rates. Overall, the Markov Chain approach worked very well statistically. However, they don't include any examples of the generated logs which raises some questions for me.

[3]

Bibliography

[1] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A Convolutional Neural Network for Modelling Sentences," 2014.

[2] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," 2017.

[3] M. Wurzenberger, F. Skopik, G. Settanni, and W. Scherrer, "Complex log file synthesis for rapid sandbox-benchmarking of security- and computer network analysis tools," *Inf. Syst.*, vol. 60, no. C, pp. 13–33, Aug. 2016.

[4] Y. Li, Q. Pan, S. Wang, T. Yang, and E. Cambria, "A Generative Model for category text generation," *Inf. Sci. (Ny).*, vol. 450, pp. 301–315, 2018.

[5] Y. Zhang *et al.*, "Adversarial Feature Matching for Text Generation," 2017.

[6] V. Kulkarni and B. Garbinato, "Generating synthetic mobility traffic using RNNs," in *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery - GeoAI '17*, 2017, pp. 1–4.

[7] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward Controlled Generation of Text," Mar. 2017.

[8] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," *Proc. 4th Int. Conf. Inf. Syst. Secur. Priv.*, no. Cic, pp. 108–116, 2018.

[9] I. J. Goodfellow *et al.*, "Generative Adversarial Nets," 2014.

[10] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," Jun. 2016.