Deep Learning for Network Traffic Data

Manish Marwah
manish.marwah@microfocus.com
Micro Focus
Santa Clara, CA, USA

Martin Arlitt
martin.arlitt@microfocus.com
Micro Focus
Calgary, AB, Canada

ABSTRACT

Network traffic data is key in addressing several important cybersecurity problems, such as intrusion and malware detection, and network management problems, such as application and device identification. However, it poses several challenges to building machine learning models. Two main challenges are manual feature engineering and scarcity of training data due to privacy and security concerns. In this tutorial we provide a comprehensive review of recent advances to address these challenges through use of deep learning. Network traffic data can be cast as a multivariate time-series (sequential) data, attributed graph data, or image data to leverage representation learning architectures available in deep learning. To preserve data privacy, generative methods, such as GANs and autoregressive neural architectures can be used to synthesize realistic network traffic data. In particular, our tutorial is organized into three parts: 1) we describe network traffic data, applications to security and network management, and challenges; 2) we present different deep learning architectures used for representation learning instead of feature engineering of network traffic data; and, 3) we describe use of generative neural models for synthetic generation of network traffic data.

ACM Reference Format:
Manish Marwah and Martin Arlitt. 2022. Deep Learning for Network Traffic Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3534678.3542618

1 INTRODUCTION

In computer security and network management, several tasks can be cast as machine learning problems. A readily available, abundant data source for these problems is network traffic data, which is typically collected at edge routers in an organization. This voluminous data, which can run into TBs of data every hour for a mid-sized organization, mostly consists of data packets, or information extracted from data packets, exchanged between hosts within an organization and those outside. Machine learning models are built using this data source, among others, to handle various cybersecurity and network management related tasks, such as intrusion detection, malware detection, application identification and device identification. There are several challenges in building these models: messy feature engineering that requires deep domain expertise; data privacy; dearth of labelled data; long-tailed distribution of normal events resulting in a large number of false positives, which leads to operator alarm fatigue; adversaries actively trying to circumvent the detection models; difficulty in model validation, due to the absence and/or inaccuracy of ground truth labels; and, highly non-stationary nature of network traffic data [1]. While there are studies investigating most of these challenges, in this tutorial we will focus on two main challenges.

First, feature engineering has always been a key part of a model built using network traffic data [3]. In fact, for high model accuracy, constructing the correct features is just as important if not more than the specific machine learning method used. However, feature engineering is problematic since it is manual, ad hoc, dependent on specialized domain knowledge and tedious to use for highly non-stationary data. Inspired by deep learning architectures in computer vision [9] and natural language processing [21], recent work has attempted to move from feature *engineering* to feature or representation *learning* of network traffic data. These learned features can then be used for downstream machine learning tasks.

Secondly, although network traffic data is readily available and abundant, it is notoriously difficult to use for training models (even internally within an organization) due to privacy and security concerns. While methods exist for anonymization, they are not foolproof; differential privacy methods provide provable guarantees, however, addition of noise degrades data for machine learning tasks. A promising alternative is to generate synthetic network traffic data, similar in spirit to generative models in computer vision [17, 20] and speech [19], which have produced encouraging results.

In this tutorial, we present a comprehensive review of recent work in application of deep learning methods to modeling of network traffic data, especially related to representation learning and synthetic data generation. While we look at these topics from a data science lens, they are multi-disciplinary spanning both data science/AI/ML and computer systems, hence we present related work from venues in both disciplines. The tutorial is organized into three parts. The first part introduces network traffic data, its collection and processing and different formats (e.g., *pcap, netflow*[2]); cybersecurity and network management applications, such as intrusion detection and device identification, which use network traffic data; and the main challenges posed in casting these applications into machine learning problems. The second part discusses recent advances in modeling of network traffic data using deep learning for representation learning. Network traffic data can be cast as a multivariate time series (sequential) data [10, 13, 14], graph data [26], or image data [22, 23], to exploit different deep neural architectures, such as, transformers [21], graph neural networks (GNN) [5], convolutional neural networks (CNN), and generative adversarial

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3542618

networks (GAN) [4]. Finally, the third part of the tutorial covers methods for synthetic generation of network traffic data using neural generative models. Compared to other tabular data, network traffic data contains specialized fields such as IP addresses and port numbers, and furthermore rows are not independent since there is temporal dependency. Most recent work use GANs for synthesizing the data. Another challenge is validation of the generated data. A recent approach is to train a ML model on synthetically generated data, and use the accuracy of the resulting model on real data as a proxy for the quality of the generated data [16].

2 TUTORIAL OUTLINE

The tutorial is planned for three hours, organized into three parts:

- **Part 1: Introduction, Data and Challenges**
  - Introduction and motivation
  - Network traffic data - features [3], raw representations [6]
  - Data collection, formats, data hygiene
  - Cybersecurity and network management applications
  - Data Science challenges
- **Part 2: Network Data Representation Learning**
  - Network traffic representation learning [7, 10, 22, 23]
  - Network traffic data cast as a multivariate time series (sequential) data [10, 13–15]
  - Data cast as a graph - GNN [5], GNN-based anomaly detection [12, 26]
  - Data cast as an image [22, 23]
- **Part 3: Synthetic Network Traffic Generation**
  - Problem Statement
  - Challenges in generation of network traffic data
  - Methods - GAN-based architectures (DoppleGANger [11], WP-GAN [18], CT-GAN [24], ODDS [8]), other generative models, such as autoregressive neural models (e.g., STAN [25]) and variational autoencoders (e.g., TVAE [24])
  - Validation methods to evaluate the generated data

3 TUTORS' BIOGRAPHY

Manish Marwah is a principal research scientist at Micro Focus. His main research interests are in the broad areas of AI and data science, and their applications to cybersecurity and to cyberphysical systems. His research has led to over 65 refereed papers, several of which have won awards, including at AAAI, KDD, and IGCC. He has twice co-organized – Data Mining for Sustainability (SustKDD) – a workshop at KDD. He has been granted 52 patents. Manish received his Ph.D. in Computer Science from University of Colorado, Boulder. He has taught graduate data science courses at Santa Clara University as an adjunct faculty.

Martin Arlitt is a principal research scientist and research team manager at Micro Focus. His general interests are workload characterization of computer servers, performance evaluation of distributed computer systems, and analyzing network traffic to improve IT security. His 100 research papers have been cited over 13,000 times (per Google Scholar). He has 46 granted patents. He is an ACM Distinguished Scientist, a senior member of the IEEE, and an adjunct assistant professor at the University of Calgary.

REFERENCES

[1] Blake Anderson and David McGrew. 2017. Machine Learning for Encrypted Malware Traffic Classification: Accounting for Noisy Labels and Non-Stationarity. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. 1723–1732.

[2] Benoît Claise. 2004. Cisco Systems NetFlow Services Export Version 9. RFC 3954.

[3] Jonathan J. Davis and Andrew J. Clark. 2011. Data Preprocessing for Anomaly Based Network Intrusion Detection: A Review. *Comput. Secur.* 30, 6–7 (sep 2011), 353–375.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. MIT Press, Cambridge, MA, USA, 2672–2680.

[5] William L. Hamilton. 2020. Graph Representation Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 14, 3 (2020), 1–159.

[6] Jordan Holland, Paul Schmitt, Nick Feamster, and Prateek Mittal. 2020. nPrint: A Standard Data Representation for Network Traffic Analysis. *CoRR* abs/2008.02695 (2020). arXiv:2008.02695 https://arxiv.org/abs/2008.02695v1

[7] Jordan Holland, Paul Schmitt, Nick Feamster, and Prateek Mittal. 2021. New directions in automated traffic analysis. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 3366–3383.

[8] Steve T.K. Jan, Qingying Hao, Tianrui Hu, Jiameng Pu, Sonal Oswal, Gang Wang, and Bimal Viswanath. 2020. Throwing Darts in the Dark? Detecting Bots with Limited Data using Neural Data Augmentation. In *2020 IEEE Symposium on Security and Privacy (SP)*. 1190–1206.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.), Vol. 25.

[10] I-Ta Lee, Manish Marwah, and Martin F. Arlitt. 2020. Attention-Based Self-Supervised Feature Learning for Security Data. *CoRR* abs/2003.10639 (2020). arXiv:2003.10639

[11] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. 2020. Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions. In *Proceedings of the ACM Internet Measurement Conference*. 464–483.

[12] X. Ma, J. Wu, S. Xue, J. Yang, C. Zhou, Q. Z. Sheng, H. Xiong, and L. Akoglu. 2021. A Comprehensive Survey on Graph Anomaly Detection with Deep Learning. *IEEE Transactions on Knowledge & Data Engineering* 01 (oct 2021).

[13] Gonzalo Marín, Pedro Caasas, and Germán Capdehourat. 2021. DeepMAL-deep learning models for malware traffic detection and classification. In *Data Science–Analytics and Applications*. Springer, 105–112.

[14] Gonzalo Marín, Pedro Casas, and Germán Capdehourat. 2018. RawPower: Deep Learning Based Anomaly Detection from Raw Network Traffic Measurements. In *Proceedings of the ACM SIGCOMM 2018 Conference on Posters and Demos* (Budapest, Hungary) *(SIGCOMM '18)*. 75–77.

[15] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep Learning for Anomaly Detection: A Review. 54, 2 (March 2021).

[16] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. Data Synthesis Based on Generative Adversarial Networks. 11, 10 (jun 2018), 1071–1083.

[17] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[18] Markus Ring, Daniel Schlör, Dieter Landes, and Andreas Hotho. 2019. Flow-based network traffic generation using generative adversarial networks. *Computers & Security* 82 (2019), 156–172.

[19] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. *CoRR* abs/1609.03499 (2016). arXiv:1609.03499

[20] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. 2016. Conditional Image Generation with PixelCNN Decoders. *CoRR* abs/1606.05328 (2016). arXiv:1606.05328

[21] Ashish Vaswani, Noam Shazeer, and et al. 2017. Attention is All you Need. In *Advances in NeurIPS*, Vol. 30.

[22] Wei Wang, Yiqiang Sheng, Jinlin Wang, Xuewen Zeng, Xiaozhou Ye, Yongzhong Huang, and Ming Zhu. 2018. HAST-IDS: Learning Hierarchical Spatial-Temporal Features Using Deep Neural Networks to Improve Intrusion Detection. *IEEE Access* 6 (2018), 1792–1806.

[23] Wei Wang, Ming Zhu, Xuewen Zeng, Xiaozhou Ye, and Yiqiang Sheng. 2017. Malware traffic classification using convolutional neural network for representation learning. In *2017 International conference on information networking (ICOIN)*. IEEE, 712–717.

[24] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*. 7333–7343.

[25] Shengzhe Xu, Manish Marwah, Martin Arlitt, and Naren Ramakrishnan. 2021. STAN: Synthetic Network Traffic Generation with Generative Neural Models. In *International Workshop on Deployable Machine Learning for Security Defense*. Springer, 3–29.

[26] Li Zheng and et al. 2019. AddGraph: Anomaly Detection in Dynamic Graph Using Attention-based Temporal GCN. In *Proceedings of IJCAI-19*. 4419–4425.