42nd Brazilian Symposium on Computer Networks and Distributed Systems - SBRC 2024

Tutorial Proposal

Abstract. *This tutorial explores the role of generative artificial intelligence in the production of synthetic data, with a special emphasis on Generative Adversarial Networks (GANs) in the context of Computer Networks. It highlights the evolution of these networks, from their introduction in image synthesis to their use in capturing complex data distributions, including variants tailored for time series. The tutorial aims to place GANs within the machine learning paradigms, addressing their various applications in Computer Networks, such as data generation, system optimization, and classification, focusing on issues involving data imbalance and privacy preservation. Introducing a methodology for collecting, generating, and applying synthetically generated data, the tutorial proposes two use cases. The first one is training Reinforcement Learning agents on data plans using synthetic data. The second one is a synthetic generation of PCAP traces.*

1. Identification data

1.1. Title

Generation of Synthetic Datasets in the Context of Computer Networks using Generative Adversarial Networks

1.2. Authors

Thiago Caproni Tavares
Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais
IFSULDEMINAS - Campus Poços de Caldas
Av. Dirce Pereira Rosa, 300, Jardim Esperança
thiago.caproni@ifsuldeminas.edu.br

Leandro C. de Almeida
Instituto Federal de Educação, Ciência e Tecnologia do Paraíba
IFPB - Campus João Pessoa
Av. Primeiro de Maio, 720 - Jaguaribe
leandro.almeida@ifpb.edu.br

Washington Rodrigo Dias da Silva
Departamento de Computação (DComp)
Universidade Federal de São Carlos (UFSCar)
Rodovia João Leme dos Santos, km 110, Sorocaba/SP

washingtonrds@estudante.ufscar.br

Ariel Góes de Castro
Departamento de Engenharia de Computação e Automação Industrial (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (UNICAMP)
a272319@dac.unicamp.br

Christian Esteve Rothenberg
Departamento de Engenharia de Computação e Automação Industrial (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (UNICAMP)
chesteve@dca.fee.unicamp.br

Fábio Luciano Verdi
Departamento de Computação (DComp)
Universidade Federal de São Carlos (UFSCar)
Rodovia João Leme dos Santos, km 110, Sorocaba/SP
verdi@ufscar.br

1.3. Indication of the authors who will present the tutorial

Thiago Caproni Tavares
Ariel Góes de Castro
Fábio Luciano Verdi

2. General data

2.1. Main goal

Currently, the use of Machine Learning (ML) models in computer networks has experienced a significant increase within the specialized scientific community. However, obtaining suitable datasets for training these models is often challenging for several reasons. First, network data may contain sensitive or private information of the users or organizations, which raises ethical and legal issues for sharing or publishing them. Second, network data may be scarce or outdated [Holland et al. 2021], especially for emerging or evolving network scenarios (e.g., 5G, IoT, SDN). Third, network data may be biased or incomplete [Bühler et al. 2022, Dai and Wang 2021], which limits the generalization and robustness of the network analysis models.

Therefore, this tutorial aims to present the essential principles related to generating synthetic time series data using Generative Adversarial Networks (GANs). Additionally, practical applications of GANs for creating telemetry data originating from a Programmable Data Plane (PDP) and generation of PCAPs will be discussed. We will demonstrate how these synthetic data can be employed in different use cases as training a Reinforcement Learning (RL) agent (aiming to optimize resources in the context of PDP) and synthetic PCAP generation.

2.2. Target audience profile

This tutorial is designed for both undergraduate and postgraduate students, encompassing master's and doctoral levels, along with professionals who exhibit an interest in synthetic data generation models within the realm of computer networks. A foundational understanding of network concepts and ML is assumed as a prerequisite. Also, we assume the participants must have basic Python language knowledge and a basic understanding of Tensorflow – i.e., an open-source ML platform. The tutorial will delve into pertinent topics of interest, including PDP, GANs, RL, and the generation of Synthetic PCAP.

3. Tutorial structure

The tutorial will be organized into distinct sections as follows:

1. Introduction and motivation
2. Fundamentals of Generative Adversarial Networks
3. In-band Network Telemetry and Programmable Data Planes
4. Synthetic data generation
    (a) Generation of telemetry data
    (b) Generation of synthetic PCAP
5. Use cases for synthetic data application
6. Presentation and artifacts
7. Conclusions and tendencies

This delineation ensures a logical and systematic progression through the essential aspects covered in the tutorial, providing a clear roadmap for the audience.

4. Summary of the content to be covered

I - Introduction and motivation (3 pages)

This chapter will discuss and present the advantages of GAN's utilization and how this sort of generative model can be applied in the context of computer networks. Generative artificial intelligence, including ChatGPT, produces various data types such as images, text, and media. It has gained popularity beyond the academic environment, focusing on language understanding and generation, evolving from traditional models with enhanced learning capabilities due to an increase in parameters.

Within generative models, GANs stand out. Initially introduced for image synthesis, these networks have gained prominence in capturing high-dimensional data distributions. They rely on two neural networks: the generator and the discriminator. Both are engaged in an interactive game as outlined in [Goodfellow et al. 2014]. The generator creates synthetic data to deceive the discriminator, which acts as a judge to distinguish real data from synthetic data. The goal is to find an equilibrium where the generator produces packets that the discriminator cannot distinguish from the real ones.

At the end of this chapter, the reader will be motivated to learn a methodology that can be used to create, expand, and balance network datasets. Additionally, they will understand that synthetic data created by a GAN can be used as an environment simulator. This is because the generated data carries the characteristics and distributions of data collected from the real environment. It is also important to highlight that the use of synthetic data generated by GAN accelerates experiments, eliminating the need to run the entire application every time a change is made. GAN also enables the creation of diverse datasets and improves statistical data variation.

II - Fundamentals of Generative Adversarial Networks (5 pages)

This chapter will provide the fundamentals of ML and GANs, especially TimeGANS. It is important to delineate that machine learning encloses three main paradigms: supervised, unsupervised, and reinforcement learning. For example, GAN falls under unsupervised learning, with the generator replicating data distributions and the discriminator playing a role similar to supervised learning by comparing generated and real data. GANs have various applications in computer networks, including data generation, system optimization, and data classification, as documented in existing literature [Navidan et al. 2021, Zou et al. 2023]. The choice of GAN variant depends on specific objectives, and recent adaptations extend GANs to temporal series network data. There are many GAN models, such as Conventional GAN, BIGAN, CGAN [Mirza and Osindero 2014], InfoGAN [Chen et al. 2016], CycleGAN [Zhu et al. 2020], EBGAN [Zhao et al. 2017], and LSGAN [Lee et al. 2022]. However, there is a trend towards including temporal series data [Zou et al. 2023]. Specific GANs, such as CTGAN [Xu et al. 2019], DoppelGANger [Lin et al. 2020], and TimeGAN [Yoon et al. 2019], are designed to learn the complexities of this kind of data [Naveed et al. 2022].

The reader of this chapter will have a good overview of the versatility of GANs in data generation, addressing issues such as data imbalance and privacy preservation. GANs can rectify biased datasets, impute missing values, and suppress sensitive information, promoting secure data exchange. These data generators play a crucial role in creating new datasets, often combined with complementary machine learning methods, such as RL models, as demonstrated in [Hua et al. 2019].

III - In-band Network Telemetry and Programmable Data Planes (5 pages)

In this chapter, we will show how recent progress in the field of Programmable Data Plane has paved the way for a new fine-grained monitoring paradigm: In-band Network Telemetry (INT). In this context, network devices autonomously report the network's state, eliminating the need for direct control plane intervention [Arslan and McKeown 2019]. That is, packets incorporate telemetry instructions within their header fields, facilitating the fine-grained collection and recording of network data.

This substantial volume of data could prove useful for network management strategies that leverage Machine Learning (ML). These strategies can learn the network state from in-line rate INT measurements, opening new horizons for solutions not yet addressed in various fields, such as congestion control, rerouting, anomaly detection, QoS/QoE prediction, and others. By reading this chapter, it is hoped that the reader will understand how fine-grained measurements provided by INT can powerfully guide ML models for addressing computer network problems.

IV - Synthetic Data Generation (15 pages)

This chapter will delineate the process of training and generating synthetic data through GANs. Starting with an exposition of the theoretical support of this technique for synthetic data generation using GAN for temporal series, it will subsequently clarify the complexities inherent in applying GAN within the context of computer networks. These complexities arise from the layered and multi-protocol characteristics of computer networks.

It is important to emphasize that this chapter will be structured into two sections: the creation of telemetry data and the generation of PCAP. The synthetic data generation process for each will leverage GAN models for telemetry data and PCAP generation. Each use case will be addressed in a dedicated subsection, following the chapters outlined in the tutorial's structure.

V - Use cases for synthetic data application (10 pages)

Within this tutorial, synthetic data will find specific applications in two distinct use cases:

1. ***Training a reinforcement learning agent for Resource Optimization in the Data Plane:*** This involves utilizing synthetic data to train a Reinforcement Learning agent geared towards optimizing resources within the data plane.
2. ***Generation of Synthetic PCAP Traces for Authentic Experiments:*** Synthetic data will also be employed to generate PCAP traces, providing datasets for real experiments.

The practical implementation of synthetic data within PDP will be comprehensively examined, with a specific example serving as an illustrative guide. Within this context, a detailed use case will unfold, focusing on the training of a Reinforcement Learning agent. RL is a machine learning paradigm where an autonomous agent optimizes decision-making in an unknown environment. This involves the agent interacting with the environment, taking actions, and receiving rewards or penalties based on the outcomes. Rewards follow positive outcomes, and penalties follow detrimental outcomes, leading the agent to optimize state-action pairs through iterative exploration and trial and error.

In conclusion, this chapter will introduce another application scenario, concentrating on the generation of synthetic PCAP traces. Readers will be guided through a methodology to create PCAP traces and instructed on leveraging them in diverse contexts. Examples include deploying these synthetic traces for the training of alternative Machine Learning models or incorporating them to simulate real workload scenarios in experimental settings.

VI - Presentation and artifacts (1 page)

This chapter will provide the artifacts essential for the practical segment of this tutorial, explaining their deployment. A notebook hosted on Google Colab, along with the requisite datasets necessary for the training and generation of synthetic data using GANs, will be available. Additionally, a complementary set of code will be accessible on GitHub, facilitating extended experimentation for the audience.

During the tutorial presentation at SBRC 2024, we intend to divide it into two parts: the first one ($\approx$2 hrs) will be dedicated to the theoretical content, presenting the basic concepts behind GAN and its fundamentals. The second part of the tutorial ($\approx$2hrs) will focus on the hands-on, providing simple examples of GAN usage for two use cases, as mentioned above: generation of synthetic network telemetry data and generation of synthetic PCAP traces.

All the artifacts (Google Colab, Github, source codes, etc.) used for the presentation will be available beforehand publicly for the attendees.

VII - Conclusions and research outlook (1 page)

This chapter will conclude the tutorial by elucidating the merits associated with the application of synthetic data and delineating the challenges inherent in its utilization. The application of GANs to the computer networking domain represents an area under significant development, with numerous open research opportunities to be examined and explored.

Expected total pages: 40 - 45 pages

5. Main bibliography used in the tutorial

References

Arslan, S. and McKeown, N. (2019). Switches know the exact amount of congestion.

Bühler, T., Schmid, R., Lutz, S., and Vanbever, L. (2022). Generating representative, live network traffic out of millions of code repositories.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets.

Dai, E. and Wang, S. (2021). Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets.

Holland, J., Schmitt, P., Feamster, N., and Mittal, P. (2021). New directions in automated traffic analysis.

Hua, Y., Li, R., Zhao, Z., Zhang, H., and Chen, X. (2019). Gan-based deep distributional reinforcement learning for resource management in network slicing.

Lee, C.-K., Cheon, Y.-J., and Hwang, W.-Y. (2022). Least squares generative adversarial networks-based anomaly detection. *IEEE Access*, 10:26920–26930.

Lin, Z., Jain, A., Wang, C., Fanti, G., and Sekar, V. (2020). Using gans for sharing networked time series data: Challenges, initial promise, and open questions.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets.

Naveed, M. H., Hashmi, U. S., Tajved, N., Sultan, N., and Imran, A. (2022). Assessing deep generative models on time series network data. *IEEE Access*, 10:64601–64617.

Navidan, H., Moshiri, P. F., Nabati, M., Shahbazian, R., Ghorashi, S. A., Shah-Mansouri, V., and Windridge, D. (2021). Generative adversarial networks (gans) in networking: A comprehensive survey & evaluation.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan.

Yoon, J., Jarrett, D., and van der Schaar, M. (2019). Time-series generative adversarial networks.

Zhao, J., Mathieu, M., and LeCun, Y. (2017). Energy-based generative adversarial network.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2020). Unpaired image-to-image translation using cycle-consistent adversarial networks.

Zou, C., Yang, F., Song, J., and Han, Z. (2023). Generative adversarial network for wireless communication: Principle, application, and trends. *IEEE Communications Magazine*, pages 1–7.

6. Curriculum Vitae

Thiago Caproni Tavares. He holds a bachelor's degree in Computer Science from the Pontifical Catholic University of Minas Gerais (2006), a master's, and a Ph.D. in Computer Science and Computational Mathematics from the University of São Paulo (ICMC/USP, 2009 and 2014). Additionally, he served as a visiting researcher (post-doctorate) at the Royal Institute of Technology - KTH in Stockholm, Sweden. Currently, he is a post-doctoral researcher at the University of São Carlos, in the Sorocaba Campus, and is a professor at the Federal Institute of Education, Science, and Technology of Southern Minas Gerais - Campus Poços de Caldas. His professional journey encompasses experiences in Computer Science, with a special focus on Computer Networks and Distributed Systems. His main areas of expertise include computer networks, artificial intelligence (generative models), distributed systems, web services, sensor networks, the Internet of Things, performance evaluation, replication control, and aspect-oriented programming.

Leandro C. de Almeida. He holds a degree in Computer Networks from the Federal Institute of Education, Science, and Technology of Paraíba - IFPB (2007) and a master's degree in Informatics from the Federal University of Paraíba (2013). Currently, he is a doctoral student in the postgraduate program at the Federal University of São Carlos and a professor at the Federal Institute of Education, Science, and Technology of Paraíba - IFPB. He has experience in the field of Computer Science, with emphasis on Computer Networks, focusing primarily on the following topics: routing, security, machine learning, and network programmability..

Washington Rodrigo Dias da Silva. He holds a degree in Systems Analysis and Development from the Itu Technology College (2016) and a master's degree in Computer Science with a focus on Image and Signal Processing from the Federal University of São Carlos - Sorocaba Campus (2021). His research interests include Image Processing and Analysis, Deep Learning, Computer Vision, Reinforcement Learning, and FPGA programming for Artificial Intelligence model inference. Currently, he is a regular doctoral student in the Computer Science Postgraduate Program at the Federal University of São Carlos.

Ariel Góes de Castro. Ariel has a degree in Computer Science from the Federal University of Pampa (UNIPAMPA). In addition, he has a master's degree in Software Engineering from the same university. He is currently seeking a PhD degree from the Faculty of Electrical and Computer Engineering at the State University of Campinas (UNICAMP). His interests include In-band Network Telemetry (INT), Network Orchestration, Software-defined Networks (SDNs) and he is currently working with Machine Learning and Generative Networks.

Christian Esteve Rothenberg. He is Associate Professor and head of the Information & Networking Technologies Research & Innovation Group (INTRIG) at the School of Electrical and Computer Engineering (FEEC) of the University of Campinas (UNICAMP), where he received his Ph.D. in Electrical and Computer Engineering in 2010. From 2010 to 2013, he worked as Senior Research Scientist in the areas of IP systems and networking, leading SDN research at CPQD R&D Center in Telecommunications, Campinas, Brazil. He holds the Telecommunication Engineering degree from the Technical University of Madrid (ETSIT – UPM), Spain, and the M.Sc. (Dipl. Ing.) degree in Electrical Engineering and Information Technology from the Darmstadt University of Technology (TUD), Germany, 2006. His research activities span multiple layers of distributed systems and network architectures and are often carried in collaboration with academia and industry (e.g., Ericsson, Samsung, CPQD, Padtec, RNP) around the world, leading to multiple open-source networking projects (e.g., RouteFlow, libfluid, ofsoftswitch13, Mininet-WiFi) in the areas of SDN and NFV among other scientific results. Christian has contributed to several international patents, co-authored three books, and over 200 scientific publications, including top-tier scientific journals and networking conferences such as SIGCOMM and INFOCOM, altogether featuring 10,000+ citations (h-index: 35+, i10-index: 80+). His experience as co-author of national and international short courses / tutorials include: SBC SBRC (2010, 2012, 2014), ACM SIGCOMM (2016), IEEE IM (2016), IEEE NetSoft (2017), ENUCOMP (2017), ERIPI (2018), JAI (2023). Recently, Christian has served as Tutorial Co-Chair of ICIN 2022 and IEEE NetSoft 2022. Currently, he is the PI of the FAPESP Engineering Research Center SMARTNESS (SMART NEtworks and ServiceS for 2030) co-funded by Ericsson and expected duration until 2033.

Fábio Luciano Verdi. He is an Associate Professor in the Department of Computer Science at UFSCar, Sorocaba, SP, leading the LERIS Research Laboratory. He has a bachelor's degree in Computer Science from the Regional University of the Northwest of the State of Rio Grande do Sul (1999), a master's degree in Computer Science (2002), and a Ph.D. in Electrical Engineering with a focus on Computer Engineering (2006), both from the State University of Campinas (Unicamp). He has completed two postdoctoral fellowships, one at Unicamp in 2009 and another one at KTH Royal Institute of Technology in 2022. He has been actively working as TPC and General Chair as well as TPC member of many important conferences such as SBRC, IEEE NetSoft, IEEE NOMS, IEEE CNSM, IEEE SCC, and ACM Sigmetrics SCR. His main areas of research include Data Centers, Cloud Computing, Routing, SDN, Network Programmability, Hardware Acceleration and Network Monitoring.