NetDiffusion: Network Data Augmentation Through Protocol-Constrained Traffic Generation

Xi Jiang
xijiang9@uchicago.edu
University of Chicago
Chicago, Illinois, USA

Shinan Liu
shinanliu@uchicago.edu
University of Chicago
Chicago, Illinois, USA

Aaron Gember-Jacobson
agemberjacobson@colgate.edu
Colgate University
Hamilton, New York, USA

Arjun Nitin Bhagoji
abhagoji@uchicago.edu
University of Chicago
Chicago, Illinois, USA

Paul Schmitt
pschmitt@hawaii.edu
University of Hawaii, Manoa / Invisv
Honolulu, Hawaii, USA

Francesco Bronzino
francesco.bronzino@ens-lyon.fr
Univ Lyon, EnsL, UCBL, CNRS, LIP
Lyon, France

Nick Feamster
feamster@uchicago.edu
University of Chicago
Chicago, Illinois, USA

ABSTRACT

Datasets of labeled network traces are essential for a multitude of machine learning (ML) tasks in networking, yet their availability is hindered by privacy and maintenance concerns, such as data staleness. To overcome this limitation, synthetic network traces can often augment existing datasets. Unfortunately, current synthetic trace generation methods, which typically produce only aggregated flow statistics or a few selected packet attributes, do not always suffice, especially when model training relies on having features that are only available from packet traces. This shortfall manifests in both insufficient statistical resemblance to real traces and suboptimal performance on ML tasks when employed for data augmentation. In this paper, we apply diffusion models to generate high-resolution synthetic network traffic traces. We present *NetDiffusion*[1], a tool that uses a finely-tuned, controlled variant of a Stable Diffusion model to generate synthetic network traffic that is high fidelity and conforms to protocol specifications. Our evaluation demonstrates that packet captures generated from NetDiffusion can achieve higher statistical similarity to real data and improved ML model performance than current state-of-the-art approaches (e.g., GAN-based approaches). Furthermore, our synthetic traces are compatible with common network analysis tools and support a myriad of network tasks, suggesting that NetDiffusion can serve a broader spectrum of network analysis and testing tasks, extending beyond ML-centric applications.

CCS CONCEPTS

• **Networks → Network simulations**; • **Computing methodologies → Neural networks**.

KEYWORDS

Network traffic, synthesis, diffusion model

ACM Reference Format:
Xi Jiang, Shinan Liu, Aaron Gember-Jacobson, Arjun Nitin Bhagoji, Paul Schmitt, Francesco Bronzino, and Nick Feamster. 2024. NetDiffusion: Network Data Augmentation Through Protocol-Constrained Traffic Generation. In *Abstracts of the 2024 ACM SIGMETRICS/IFIP PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS/PERFORMANCE Abstracts '24), June 10–14, 2024, Venice, Italy.* ACM, New York, NY, USA, Article 11, 2 pages. https://doi.org/10.1145/3652963.3655071

1 INTRODUCTION

Modern networks are increasingly reliant on machine learning (ML) techniques for a wide range of management tasks, ranging from security to performance optimization. A central impediment when training network-focused ML models is the scarcity of labeled network datasets, as their collection and sharing are often associated with high costs and privacy concerns, particularly when data is collected from real-world networks [2]. Unfortunately, existing public datasets rarely receive updates, making them static and unable to reflect evolving network behaviors. These limitations hinder the ability to train robust ML models that accurately reflect evolving real-world network conditions.

These challenges can be addressed through the creation of new synthetic network traces based on existing datasets. This approach aims to preserve the inherent characteristics of network traffic while introducing variations, thereby enhancing dataset size and diversity. Unfortunately, current state-of-the-art synthetic trace generation methods, particularly those based on Generative Adversarial Networks (GANs)-based methods [9, 10], are not always sufficient for producing high-quality synthetic network traffic. Specifically, these approaches tend to focus on a limited set of attributes or statistics, as

[1] We open source our sample datasets, pipeline, and results in https://github.com/noise-lab/NetDiffusion_Generator.

SIGMETRICS/PERFORMANCE Abstracts '24, June 10–14, 2024, Venice, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0624-0/24/06
https://doi.org/10.1145/3652963.3655071

early machine learning for network tasks often relied on basic flow statistics for classification. With recent ML advancements utilizing detailed raw network traffic to achieve enhanced classification accuracy [5, 11], there is a clear need for synthetic traffic generation that includes the intricate, potentially unforeseen patterns present in full network traces. Existing traffic generation methods face two main issues: (1) a lack of statistical similarity with real data due to the limited attributes in existing methods, making the synthetic data highly sensitive to variations, and (2) unsatisfactory classification accuracy when synthetic statistical attributes are used to augment existing datasets. Moreover, their simplistic attribute focus and disregard for transport and network layer protocol behaviors prevent their use with traditional networking tools such as tcpreplay [3] or Wireshark [1].

Fortunately, the general increase in available computational power and the breakthroughs in high-resolution image generation techniques, particularly diffusion models [7, 8], offer a promising avenue to overcome these challenges. Specifically, we harness the capabilities of text-to-image diffusion models, which execute conditioned generation based on descriptive text prompts. These models are adept at creating detailed, accurate visual representations from textual descriptions. By translating the intricate characteristics of network traffic into an appropriate image format, we can tap into the unique advantages offered by these models. In contrast to GANs, diffusion models are able to capture both broad patterns and detailed dependencies. This inherent generative quality makes them an ideal choice for producing network traces with high statistical resemblance to real traffic and full packet header values. By incorporating conditioning techniques, diffusion models can generate structured data that conforms to specific network properties, which ensures the desired sequential inter-packet characteristics and rough protocol dependencies. Moreover, the gradient dynamics of the training process in diffusion models is much more stable than GANs. We discuss the technical details and benefits of diffusion models in depth in Section **??**. These attributes collectively position diffusion models as a compelling choice for advancing the state-of-the-art for synthetic network trace generation, addressing the extant limitations of current methodologies.

In this paper, we introduce NetDiffusion, an approach to synthetic raw network traffic generation for producing packet headers leveraging fine-tuned, controlled stable diffusion models. Our contributions are as follows:

(1) **Generation of synthetic network traces with high resemblance to real traffic:** Using stable-diffusion techniques, we propose a two-fold strategy: (1) a conversion process for transforming raw packet captures to image representations (and vice versa), and (2) fine-tuning a text-to-image diffusion model based on packet capture-converted images for generating synthetic packet captures. To improve resemblance to real network traffic, we employ controlled generation techniques to maintain fidelity to the protocol and header field value distributions observed in real data and, post generation, use domain knowledge-based heuristics to finely check and adjust the generated fields, ensuring their semantic correctness in terms of compliance with transport and network layer protocol rules.

(2) **Improved classification accuracy in ML scenarios with synthetic network traffic augmentation:** We conduct a case study evaluation on a curated traffic classification dataset. By integrating NetDiffusion-generated network traffic into the real dataset at varying proportions during training and testing, we observe a general increase in accuracy compared to the state-of-the-art generation method. This improvement is attributed to our synthetic data's significantly high statistical resemblance to the real dataset. Additionally, our method shows promise in addressing class imbalance issues, enhancing the accuracy of ML models in such cases.

(3) **Extended applicability of synthetic network traffic for network analysis and testing beyond ML tasks:** NetDiffusion-generated network traffic can be converted into raw packet captures suitable for traditional network analysis and testing tasks. We validate this compatibility through tests with tools such as Wireshark and Scapy [6], as well as tcpreplay for retransmission. More importantly, we show that critical statistical features for various network operations can be effectively extracted from the generated network traffic.

The full version of this paper appears in [4].

ACKNOWLEDGMENTS

We thank the anonymous associate editors and reviewers for their constructive feedback and suggestions. This research is partly supported by grants from ANR Project No ANR-21-CE94-0001 (MINT), the Fédération Informatique de Lyon through the INTERFERE project, the France and Chicago Collaborating in the Sciences program, and NSF CNS-2124393.

REFERENCES

[1] Jay Beale, Angela Orebaugh, and Gilbert Ramirez. 2006. *Wireshark & Ethereal network protocol analyzer toolkit.* Elsevier.

[2] François De Keersmaeker, Yinan Cao, Gorby Kabasele Ndonda, and Ramin Sadre. 2023. A Survey of Public IoT Datasets for Network Security Research. *IEEE Communications Surveys & Tutorials* (2023).

[3] Tcpreplay Developers. 2023. Tcpreplay. https://tcpreplay.appneta.com/.

[4] Xi Jiang, Shinan Liu, Aaron Gember-Jacobson, Arjun Nitin Bhagoji, Paul Schmitt, Francesco Bronzino, and Nick Feamster. 2024. NetDiffusion: Network Data Augmentation Through Protocol-Constrained Traffic Generation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 8, 1 (2024), 1–32.

[5] Mohammad Lotfollahi, Mahdi Jafari Siavoshani, Ramin Shirali Hossein Zade, and Mohammdsadegh Saberian. 2020. Deep packet: A novel approach for encrypted traffic classification using deep learning. *Soft Computing* 24, 3 (2020), 1999–2012.

[6] R Rohith, Minal Moharir, G Shobha, et al. 2018. SCAPY-A powerful interactive packet manipulation program. In *2018 international conference on networking, embedded and wireless systems (ICNEWS).* IEEE, 1–5.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 10684–10695.

[8] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning.* PMLR, 2256–2265.

[9] Shengzhe Xu, Manish Marwah, Martin Arlitt, and Naren Ramakrishnan. 2021. Stan: Synthetic network traffic generation with generative neural models. In *Deployable Machine Learning for Security Defense: Second International Workshop, MLHat 2021, Virtual Event, August 15, 2021, Proceedings 2.* Springer, 3–29.

[10] Yucheng Yin, Zinan Lin, Minhao Jin, Giulia Fanti, and Vyas Sekar. 2022. Practical gan-based synthetic ip header trace generation using netshare. In *Proceedings of the ACM SIGCOMM 2022 Conference.* 458–472.

[11] Weiping Zheng, Jianhao Zhong, Qizhi Zhang, and Gansen Zhao. 2022. MTT: an model for encrypted network traffic classification using multi-task transformer. *Applied Intelligence* (2022), 1–16.