Paper Summary
NetDiffus: Network Traffic Generation by Diffusion Models through Time-Series Imaging

Sarthak | 2020CS10379

June 2024

Problem Statement

Motivation

- Having access to appropriate network data traces is increasingly becoming challenging due to :
  - Complexity of modern networks and the sheer volume of data being transferred, deploying data collection tools requires significant expertise and cost
  - Privacy and regulatory constraints have made many types of network data inaccessible or restricted in use for other purposes such as network management
- GANs suffer from mode-collapse, vanishing gradients, and instability unless the hyper-parameters are properly selected
- The ability to control the generated output makes DMs ideally suited for synthetic data generation for training ML models as it allows the generation of balanced datasets

Key Idea

To leverage Diffusion Models (DM) to generate synthetic network traffic data by converting one-dimensional network traffic data into two-dimensional images using Grammian Angular Summation Field (GASF) and fine tuning a Diffusion Model to synthesize high-fidelity traffic data.

Framework & Methodology

- The one dimensional time-series network traffic is normalized to $X = \{x_1, x_2, ...x_n\}$, $x_i \in [0, 1]$ and then converted to polar form : $\theta_i = arccos(x_i)$, $r_i = \frac{t_i}{C}$, where t_i is the timestamp of the i-th sample and C is constant to regularize the radius
- The GASF is formed using the trigonometric sum between each point to identify the temporal correlation within different time intervals: $GASF(i,\ j) = cos(\theta_i + \theta_j)$ and in matrix notation :

$$GASF = X^T \cdot X - \sqrt{I - X^2}^T \cdot \sqrt{I - X^2}$$

- The GASF images are resized to a fixed smaller resolution using OpenCV's resize() method with INTER_AREA interpolation, normalized to [0,1] to speed up training and add bijective properties
- Gaussian noise is gradually added in the forward step to the original GASF images over multiple steps, progressively degrading the images
- A U-Net model having five layers (2 downsampling and 2 upsampling) with Residual connections between corresponding down and upsampling layers is learnt to reconstruct the original GASF images by minimizing the loss function (crossentropy loss). All these layers contain conv2d, relu and maxpool as internal operations.
- Post training, a random sample from the same distribution as noise can be denoised to images by the U-Net model

Contributions

- First attempt at utilizing Diffusion Models for network data generation
- Developed an end-to-end framework NetDiffus that first converts one-dimensional time-series network traffic into two- dimensional images, and then synthesizes representative images for the original data
- Demonstrated that NetDiffus outperforms the state-of-the-art traffic generation based on GANs by providing 66.4% increase in fidelity and 18.1% increase in downstream machine learning tasks
- Evaluated NetDiffus on seven diverse traffic traces and showed that utilizing synthetic data significantly improves traffic fingerprinting, anomaly detection and traffic classification

Strengths

- Innovative use of Diffusion Models for network traffic generation providing significant improvements in data fidelity and ML task performance compared to existing methods
- GASF images can encode features such as packet sizes, inter-packet times and most importantly the correlation among the 1D time-series samples onto an image in 2D space making it a rich source of information for ML models
- Several simple image processing techniques such as contrast adjustment and image resizing can be applied to reduce complexity and improve feature learning
- Even without combining with original data, NetDiffus can achieve almost the same accuracy of original data or improved accuracy of 1–57% in ML tasks like traffic fingerprinting, anomaly detection

Weaknesses & Limitations

- Limited to generating traffic features in GASF format, not meta-data
- Dependency on image processing techniques may introduce biases or errors
- Deals with fixed lower image size to reduce the DM training time and improve scalability
- The privacy aspect is not evaluated in the paper
- Can extend to multivariate time-series data and incorporate differentially private noise to ensure privacy in synthetic datasets