Failure Data Augmentation for Optical Network Equipment Using Time-series Generative Adversarial Networks

Cheng Xing[1], Chunyu Zhang[1,*], Bing Ye[1], Danshi Wang[1], Yinqiu Jia[1], Jin Li[1], and Min Zhang[1]

[1] State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications (BUPT), Beijing, 100876, China;

[2] State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen, 518055, China

*Email: *chunyuzhang@bupt.edu.cn*

Abstract: We propose a failure data augmentation scheme based on time-series generation adversarial networks with real equipment performance data of optical networks and verify that the augmented failure sample data is similar to real failure data. © 2022 The Author(s)

1. Introduction

Recently, machine learning (ML) technology has been widely used in the field of optical network failure management [1], including failure prediction, failure detection, failure identification, failure location, etc. To effectively train the performance of the ML model, some real failure samples are needed. However, due to the robustness of optical network system, failures rarely occur and few failure samples are obtained, which leads to the problem of data imbalance in the collected data and brings great challenges to ML modeling [2]. Therefore, it is necessary to find a suitable method to augment the failure data.

Typical data augmentation methods include synthetic minority over-sampling technique (SMOTE), generative adversarial network (GAN) [3], etc. However, these methods do not make full use of the inherent characteristics of time-series data, while optical network failure data is time series data, so the failure data augmentation model should maintain the temporal relationship between the cross-time variables. Time-series generation adversarial networks (TimeGAN) retains the time dynamics by combining the dynamic control provided by GAN unsupervised training and autoregressive supervised training, and it has been used for time series data augmentation of industrial field [4].

In this paper, we propose a time-series failure data augmentation scheme based on TimeGAN, and the data used in the experiment are from the OTN board performance data managed by a network operator. To evaluate the augmentation effect of failure data, PCA and t-SNE algorithms are introduced to visualize the data augmentation effect of TimeGAN. Moreover, classical statistical metrics (mean absolute error, root mean square error, mean, variance) and KL divergence were used to evaluate the distribution of TimeGAN augmented and true failure data. The results show that the augmented time-series failure data is similar to the true data in terms of feature distribution, and the TimeGAN-based scheme can effectively generate augmented failure data for optical networks.

2.TimeGAN-Based OTN Data Augmentation

The schematic of TimeGAN-based optical network failure data augmentation scheme is shown in Fig.1. Firstly, the performance data of optical network equipment are collected from the network management system. Then, the failure data is augmented based on the TimeGAN scheme, and TimeGAN consists of four network components: embedding function, recovery function, generator and discriminator, as shown in Fig.1(b). Moreover, the embedding function and the recovery function form the embedding network, and the generator and the discriminator form the adversarial network. The joint training of embedded network and adversarial network enables it to learn coding features, generate representations and iterate across time simultaneously.

The embedded network provides a mapping between features and latent space, allowing the GAN network to learn the latent temporal dynamics of the data through a low-dimensional representation. Embedding function transforms the static and temporal features $s,x_{1:T}$ into their latent code h_s, $h_{1:T}$; recovery function transforms static and temporal latent codes h_s, $h_{1:T}$ into feature representations $\tilde{s}$, $\tilde{x}_{1:T}$. The adversarial network consists of generator and discriminator. Where the generating function generates H_s, H_t from z_s, $z_{1:T}$, then the static and time random vector elements are combined into the latent code $\tilde{h}_s$, $\tilde{h}_{1:T}$; the discriminator receives static and temporal codes, and returns the classification $\tilde{y}_s$, $\tilde{y}_{1:T}$. Where z_s is the vector space of static features, and z_t is the vector space of current temporal features; $\tilde{h}_s$ may represent real data or synthetic data.

Moreover, Fig.1(c) shows the loss function transfer during the joint training of the embedding network and adversarial network, as shown in equations (1), (2) and (3). Where, L_R is the reconstruction loss of feature after embedding and recovery function; L_U is the unsupervised loss generated by a generative adversarial network consisting of a generator and a discriminator; L_S is a stepwise supervised loss of the original data as a supervised

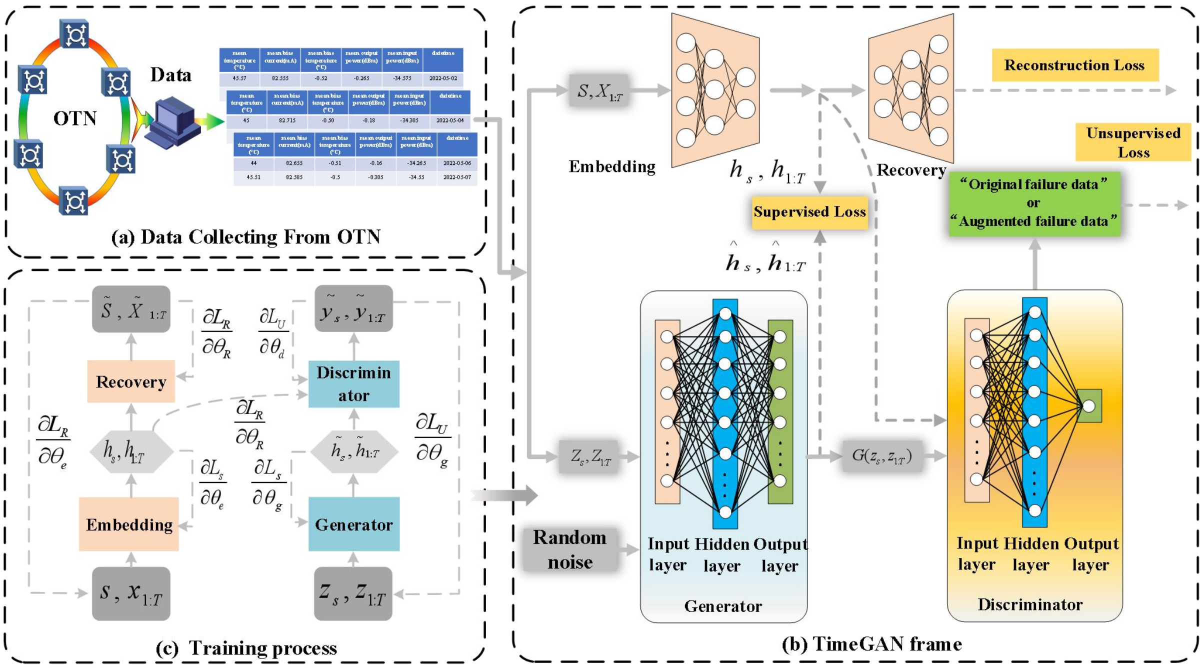


Fig. 1. (a) Data collecting, (b) TimeGAN Frame, (c) Training process. The solid line represents the forward propagation of data, and the dashed line represents the back propagation of gradient and loss.

term, whose purpose is to learn temporal features, and it can evaluate the difference between the actual and synthetic next step latent vectors. Therefore, by jointly training the embedding network and the generator network, the supervised loss can be minimized, and the potential space can not only improve the parameter efficiency, but also facilitate the generator to learn the time relationship through specific conditions.

$$L_R = E_{s,X1:T\sim P}[\| s - \tilde{s} \|_2 + \sum_t \| x_t - \tilde{x}_t \|_2] \tag{1}$$

$$L_R = E_{s,X1:T\sim P}[\| s - \tilde{s} \|_2 + \sum_t \| x_t - \tilde{x}_t \|_2] \tag{2}$$

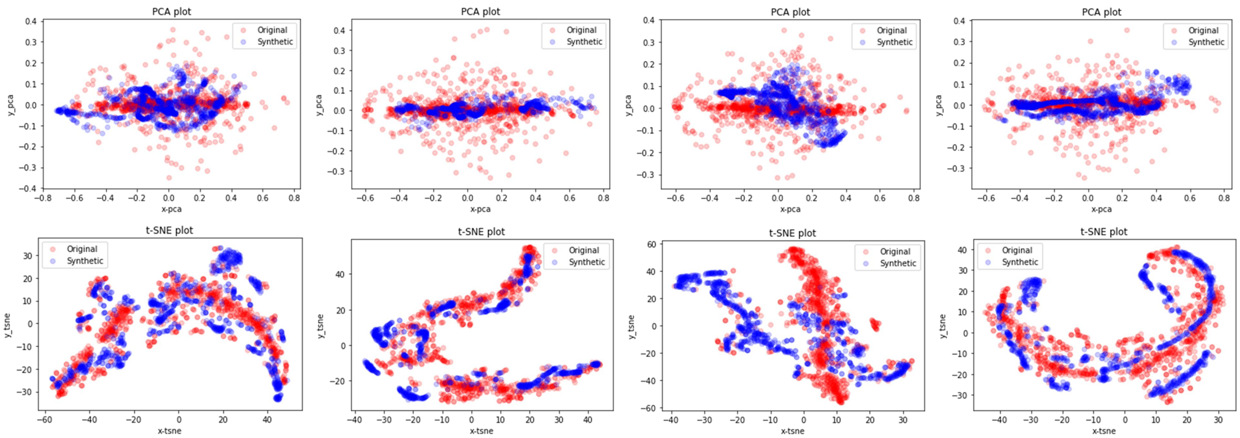
$$L_S = E_{s,x1:T\sim P}[\sum_t \| h_t - g_x(h_s, h_{t-1}, z_t) \|_2] \tag{3}$$

3.Experimental Results and Analysis

In this section, the performance of TimeGAN-based failure data augmentation of optical networks is evaluated and analyzed. The data used in the experiment came from the OTN board performance data managed by a network operator. A total of 38,467 samples were collected in 47 days, including 30295 normal samples and 8172 failure samples, which were collected in 24 hours (24h). Each sample contains date, unavailable time, environmental temperature, laser bias current, laser temperature offset, input optical power, output optical power. If the unavailable time is higher than 80,000, the operating state of the board is assumed to be faulty, which means that the board is unavailable that day.

To evaluate the effect of failure data generation based on TimeGAN scheme, the typical data augmentation methods were compared and analyzed, including TimeGAN-LSTM and TimeGAN-GRU, GAN and RCGAN [5]. The t-SNE and PCA algorithms were introduced on the original (real) and generated failure data to observe the generated samples and the similarity of the original failure samples, as shown in Fig. 2. It can be concluded that the number of failure data increases significantly after data augmentation based on TimeGAN. Moreover, the synthetic dataset generated by TimeGAN-LSTM has significantly better similarity with the original failure data than the other benchmarks visualized using t-SNE.

To evaluate the similarity of the distribution of the failure sample data augmented by the TimeGAN method with the real failure data, the classical statistical metrics (MAE, RMSE) and KL divergence were used to evaluate the performance of the proposed scheme, as shown in Table 1. It can be seen from Table 1 that TimeGAN-LSTM is superior to other proposed methods in MAE, RMSE and KL divergence scores Among them, the MAE of the failure samples generated by TimeGAN-LSTM reaches 0.1691, which is 24% lower than 0.2091 of GAN method, this is a significant improvement.



(a) TimeGAN-LSTM (b) TimeGAN-GRU (c) GAN (d) RCGAN

Fig. 2. PCA visualization on OTN failure data (1st row), t-SNE visualization on OTN failure data (2nd row).

Table 1. Result on Failure Time-series Datasets

Method	MAE	RMSE	KL divergence
TimeGAN-LSTM	0.1691	0.1822	0.4514
TimeGAN-GRU	0.1824	0.1940	0.4586
GAN	0.2091	0.2280	0.5322
RCGAN	0.1986	0.2131	0.5547

Moreover, in order to further compare the failure sample data generated based on TimeGAN with the real failure data, two key statistical parameters (mean and variance) were evaluated, as shown in Table 2. The results show that the difference between the mean of the augmented failure sample data and the original one is about 0.003-0.02, and the difference of the variance between the augmented failure sample data and the original failure sample data is less than 0.01. Therefore, the mean and variance of the augmented failure sample data and the original failure data are almost the same in terms of data performance. These results show that TimeGAN can effectively generate augmented failure data for optical networks.

Table 2. The Difference between the Augmented and Original Failure Data in Mean, Variance (TimeGAN-LSTM). Aug: augmented Ori: original

Feature	Mean temperature		Laser bias current		Mean bias temperature		Mean output optical power		Mean input optical power	
Data category	Aug	Ori	Aug	Ori	Aug	Ori	Aug	Ori	Aug	Ori
Mean	0.2557	0.2492	0.5550	0.5390	0.0452	0.0429	0.9764	0.9713	0.6271	0.6090
Variance	0.0111	0.0120	0.0380	0.0382	0.0022	0.0067	0.0067	0.0067	0.0340	0.0349

4.Conclusion

We proposed a TimeGAN-based failure data augmentation scheme for optical networks, which can effectively generate augmented failure sample data of optical networks. Moreover, we visualized the data augmentation effect of TimeGAN scheme by introducing PCA and t-SNE method, and the distribution of TimeGAN generated failure data and real failure data was evaluated by comparing the classical statistical indicators (mean absolute error, root mean square error, KL divergence, mean, variance).

Acknowledgements: This work was supported in part by National Natural Science Foundation of China (No. 61975020, 62171053).

5. References

[1] F. Musumeci, et al., “A Tutorial on Machine Learning for Failure Management in Optical Networks,” JLT, vol. 37, no. 16, 15 Aug.15, 2019.
[2] J. Pesic et al., "Transfer Learning from Unbiased Training Data Sets for QoT Estimation in WDM Networks." ECOC, Mo2K-2 2020.
[3] N. -T. Tran, et al., “On Data Augmentation for GAN Training,” IEEE Transactions on Image Processing, vol. 30, 2021.
[4] J. Yoon, et al., “Time-series Generative Adversarial Networks”, NeurIPS 2019, Sep. 2019.
[5] C. Esteban, et al., “Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs,” arXiv:1706.02633,Jun. 2017.