Research Article | Vol. 27, No. 13 / 24 June 2019 / Optics Express 18831

Optics EXPRESS

Deep learning based adaptive sequential data augmentation technique for the optical network traffic synthesis

JIN LI, DANSHI WANG,* SHUAI LI, MIN ZHANG, CHUANG SONG, AND XUE CHEN

State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China
*danshi_wang@bupt.edu.cn

Abstract: The lack of the sufficient and diverse training data is one of the main challenges limiting performances of the machine learning enabled applications in optical networks. Here, we propose a deep learning based sequential data augmentation technique for the aggregate traffic data augmentation for diverse optical network scenarios. A generative adversarial network (GAN) model is trained with the experimental traffic data to automatically extract the substantial characteristics of the experimental traffic data through the zero-sum game theory and then augment the traffic data adaptively. The statistical evaluation parameters of the augmented traffic are mean, variance and Hurst exponent. To add comparisons, two other classical generative models including the statistical parameter configuration (SPC) model and the variational autoencoder (VAE) model are also adopted to generate the traffic data that are similar to the actual traffic data. The comprehensive comparisons among the proposed GAN, the SPC and VAE show that the performances of the GAN exceed those of the SPC and the VAE obviously. The mean and the variance of the augmented traffic data from the GAN are almost equal to those of the experimental traffic data, where the average deviations are both within 2%. The Hurst exponent of the augmented traffic data from the GAN is respectively near 90% and 96% of those of the experimental traffic data in the access network and the core network. To estimate the similarity intuitively, the well-known *k*-mean algorithm is used to cluster the augmented traffic data according to the centroids determined by the corresponding experimental traffic data and the clustering accuracies are all higher than 95% for 6 kinds of typical traffic types in the optical networks. These results demonstrate that the proposed GAN is able to effectively generate the traffic data that is very close to the experimental traffic data and is difficult to be distinguished for diverse traffic types. Moreover, a relatively small dataset with a few hundred pieces of experimental traffic data is required and the amount of the augmented traffic data from the GAN is unlimited in theory, which can be augmented as much as we need. The proposed traffic data augmentation technique also has the potential to be utilized in other sequential data augmentation applications for the optical networks.

1. Introduction

Recently, machine learning (ML) enabled techniques have been widely used in optical communications to comply with the ever-increasing complexity and flexibility of optical transmission systems and networks. For the optical transmission systems, the deployment of the reconfigurable optical add-drop multiplexer (ROADM) and the coherent technology introduces complicated path dependent impairments and diverse adjustable configuration parameters including the modulation schemes, the symbol rate, the probabilistic shaping and the forward error correction overhead [1–2]. To address those problems, many ML based techniques are developed for the adaptive impairment compensation, the optical performance monitoring (OPM), the modulation format

#362972 https://doi.org/10.1364/OE.27.018831
Journal © 2019 Received 21 Mar 2019; revised 6 Jun 2019; accepted 11 Jun 2019; published 20 Jun 2019

Optics EXPRESS

recognition (MFR) and the optical amplifier control [3–8]. For the optical networks, due to the development of the software-defined networking (SDN), flexible and scalable data analyzers are crucial for the increasing network size and network openness. To improve the capabilities of the data analyzers for the optical networks, a plenty of ML applications are specially designed for the quality of transmission estimation, the routing and spectrum assignment (RSA), the virtual topology design and reconfiguration (VTDR) and the failure localization and prediction (FLP) [9–11].

In particular, numerous ML techniques trained with the traffic data have been successfully applied in the optical networks and have drawn a lot of attention. The traffic classification, traffic prediction and traffic anomaly detection techniques based on the ML are exploited for the resource reservation, the traffic grooming and the load balancing [12–14], which are capable of enhancing the network resource utilization efficiency and the network dependability dramatically. To guarantee the performances of the ML, the sufficient and diverse training data, the advanced algorithms and the strong computing power are necessary. However, the sufficient actual traffic data is not easy to be obtained. Even though amounts of actual traffic data is gathered from optical networks for a long time, the diversity and the instantaneity are not available in general. The data diversity is essential to improve the robustness and generalization performances of the ML enabled techniques and it is common to utilize the data augmentation to increase the size and diversity of the training dataset. Therefore, it is significant to propose sequential data augmentation techniques to provide plenty of diverse synthesized traffic data. Meanwhile, with the advent of emerging network services including the cloud computing, the virtual reality and the social applications [15–16], an increasing variety of the traffic data means that the high adaptivity and flexibility are needed in the traffic data synthesis for the optical networks.

For the network traffic data synthesis, the classic traffic models consist of the ON/OFF model, the fractional Brownian motion model (FBM), the fractional autoregressive integrated moving average (FARIMA) model and the wavelet based model [17–18] have been developed, where the important traffic self-similarity has been successfully described. However, in the reported classic traffic models, numerous experts are required to design the traffic models carefully. Moreover, those classical traffic simulation models usually act as the service sources for the routing and wavelength assignment and aim to general traffic to verify network system, not to generate traffic similar to the actual data. What's more, the actual traffic data is almost not used to be approximated for the traffic data synthesis and the quantitative similarities between the synthesized traffic and the actual traffic in the optical networks are generally not taken into consideration. Therefore, it is crucial to propose more adaptive traffic data augmentation techniques to provide sufficient and diverse traffic data that has consistent properties with the actual traffic data for various network scenarios in the dynamic optical networks.

On the other hand, generative adversarial network (GAN), one of the core members in the deep learning community, has shown remarkable ability in the image data augmentation, image style transfer and the video generation [19–21] and has attracted a lot of interest since it was proposed in 2014. Benefited from the zero-sum game theory, the competition between the generator and the discriminator in the GAN drives the corresponding optimization procedures until the augmented data is indistinguishable from the actual data [22]. This enlightens us to attempt to introduce the GAN to realize the effective sequential data augmentation. To the best of our knowledge, few work about the GAN based sequential data augmentation has been reported, not to mention the aggregate traffic data augmentation for optical networks.

In this paper, we propose an adaptive traffic data augmentation technique based on the GAN and utilize the experimental traffic data to augment the traffic dataset adaptively for various networks scenarios including 6 kinds of typical traffic types in the access networks and the core networks. Further, the augmented traffic date from the GAN are evaluated through the statistical evaluation parameters including the mean, the variance and the Hurst exponent. The

Optics EXPRESS

similarity between the augmented traffic data and the experimental traffic data is then verified by the well-known clustering algorithms named k-means algorithm. To add comparisons, two other classical generative models, i.e. the statistical parameter configuration (SPC) model [23,24] and the variational autoencoder (VAE) model [25], are also adopted to generate the traffic data. Results demonstrate that the proposed GAN is effective to augment the traffic data for diverse network scenarios. Taken into account of the mean, variance, Hurst exponent and the clustering accuracy, the generated traffic from the GAN is more similar to the corresponding actual traffic data than those from the SPC and AVE. In the traffic data augmentation for the school area (SA) in the access network, the average deviation of the mean, the variance and the Hurst exponent is about 0.001, 0.0002 and 0.026 accordingly, which is only 12.5%, 3.3%, 86.7% of those in the SPC and 9.1%, 10.0%, 68.4% of those in the VAE. Moreover, the clustering accuracy is 97.3%, 92.8% and 93.6% in the GAN, the SPC and the VAE respectively. The proposed GAN based technique is capable of adaptively augmenting the traffic data on demand for diverse optical network scenarios with small amounts of experimental traffic data, which also has the potential to be used for other sequential data augmentation applications.

2. Operational principle

2.1. Concept of the generative adversarial network

Generative adversarial network is specially designed for the data augmentation, which has been successful applied in the image data augmentation and the video generation. The key behind the GAN is the unique framework enlightened from the zero-sum game theory, where the discriminator and the generator are pitted against each other. In the GAN based traffic data augmentation, the objective of the discriminator is to correctly determine whether the data is from the actual traffic dataset or the augmented traffic dataset. Contrarily, the generator transfers the random noise into the augmented data and tries to make the characteristics of augmented data close to those of the actual data. After the intense competition, the discriminator and the generator are improved by each other and the augmented data eventually cannot be differentiated from the actual data.

The specific structure of the GAN based traffic data augmentation technique is illustrated in the Fig. 1(a). In the proposed GAN, the 24×1 actual traffic data, corresponding to 24 hours in one day, from different network scenarios is firstly normalized as the traffic data ranging from 0 to 1.The normalized actual traffic data is then sent into the discriminator D. Meanwhile, the 24×1 normalized random noise is generated and further fed into the generator G, where the noise data is transformed as the augmented traffic data gradually. As depicted in the Fig. 1(a), the discriminator D and the generator G are both constructed by the classic artificial neural network (ANN) in the proposed GAN. In the GAN, the discriminator D ties to differentiate whether the input traffic data belongs to the actual traffic group or the augmented traffic group, while the generator G intends to confuse the discriminator D and make it classify the augmented traffic data into the actual traffic data. Therefore, the objectives of the D and G are adversarial and should be taken into consideration comprehensively in the loss function of the GAN, which is expressed in the Eq. (1) as below:

$$L(D,G) = \frac{1}{N}\sum_{i=1}^{N}\Big(\log D(\boldsymbol{x}^i) + \log(1 - D(G(\boldsymbol{g}^i)))\Big). \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{g}$ respectively denotes the normalized actual traffic data and normalized random noise data and i represents the i-th traffic data sample. The output of discriminator is $D(\boldsymbol{x})$ and the augmented traffic data is generated by the $G(\boldsymbol{g})$. What's more, N denotes size of the training sample dataset. Seen from the loss function $L(D, G)$, the D is trained to maximize the probability of allocating the correct label to the actual traffic data and the augmented traffic data. Meanwhile,

Optics EXPRESS

the G is trained to minimize the term: $\log(1 - D(G(\boldsymbol{g})))$. Thus, the total goal of the GAN is to optimize the loss function $L(D, G)$ according to the zero-sum game theory.

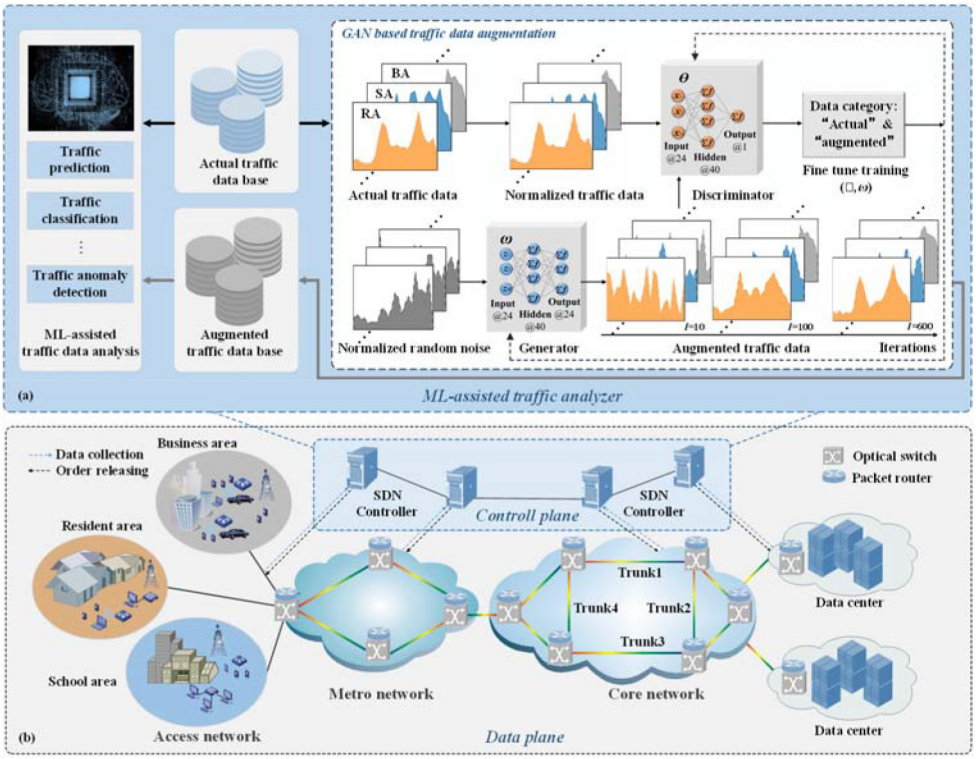

Fig. 1. Schematic diagram of (a).the specific structure of the GAN used to augment the traffic data in the ML-assisted traffic analyzer (b).the data plane in the optical networks. ML: machine learning; GAN: generative adversarial network.

During one iteration in the training procedure of the GAN, the parameters in the generator G are fixed at first and the discriminator D is updated through the ascending its stochastic gradient.

$$\boldsymbol{\theta}^{m+1} = \boldsymbol{\theta}^{m} - \eta \frac{\partial}{\partial \boldsymbol{\theta}} \left(\frac{1}{N} \sum_{i=1}^{N} \{\log D(\boldsymbol{x}^i) + \log(1 - D(G(\boldsymbol{g}^i)))\} \right). \tag{2}$$

where $\boldsymbol{\theta}$ represents the weight vector in the discriminator D and m denotes the m-th iteration. After the updating of the D, the discriminator is then frozen and the generator G is optimized by descending its stochastic gradient.

$$\boldsymbol{\omega}^{m+1} = \boldsymbol{\omega}^{m} - \eta \frac{\partial}{\partial \boldsymbol{\omega}} \left(\frac{1}{N} \sum_{i=1}^{N} \{\log(1 - D(G(\boldsymbol{g}^i)))\} \right). \tag{3}$$

where $\boldsymbol{\omega}$ is the weight vector in the generator. After iterations, the loss of the GAN will be convergent gradually and the characteristics of the augmented traffic data will be close to those of the actual traffic data.

2.2. Principle of the k-means algorithm

After the traffic data augmentation of the GAN, it is significant to evaluate the quality of the augmented data quantitatively. Besides the statistical evaluation parameters and the well-known

Optics EXPRESS

clustering algorithm, i.e. k-means algorithm, are adopted to estimate the similarity between the augmented traffic data and the corresponding actual traffic data intuitively.

As shown in the Fig. 2, during the training stage of the k-means algorithm, k denotes the number of clusters, the centroids of different clusters of the actual traffic date are updated until them are stable gradually. During the testing stage, the distances between the augmented traffic data and the centroids are calculated and the augmented traffic data will be assigned with a predicted cluster label corresponding to the traffic cluster with the nearest centroid. By comparing the predicted cluster label with the true cluster label of the augmented traffic data, the clustering accuracy is measured. When the characteristics of the augmented traffic data are close to those of the actual traffic data, the same cluster label will be allocated to the corresponding augmented traffic date. Therefore, it is straightforward to evaluate whether the augmented traffic data is similar to the accordingly actual traffic data or not. In the k-means algorithm, the computation method of the centroids is described in the Eq. (4):

$$\boldsymbol{\mu}_t = \frac{1}{|C_t|} \sum_{i \in C_t} \boldsymbol{x}^i. \tag{4}$$

where $\boldsymbol{x}^i$ denotes the i-th actual traffic data and the C_t represents the traffic data samples in the t-th cluster. Moreover, $\boldsymbol{\mu}_t$ is the centroid of the t-th traffic data cluster. The predicted cluster label of the augmented traffic data is determined according to the Euclidean distance between the augmented traffic data and different centroids.

$$P_i = index(\min_t ||\boldsymbol{x}^i - \boldsymbol{u}_t||^2). \tag{5}$$

where P_i denotes predicted cluster label of the augmented traffic data. The function $index(\cdot)$ can return the index, corresponding to the cluster label, of the centroid corresponding to the minimum Euclidean distance among all centroids and the augmented traffic data. During the training of the k-means, the initial centroids of different clusters are chosen randomly. After iterations, the updated centroids are replaced by the mean of the samples belonging to the same cluster until the centroids are stable. Once the centroids of different clusters of the actual traffic data are obtained, the predicted cluster label of the augmented traffic data will be allocated according to the Eq. (5). Comparing the predicted cluster label with the true cluster label, the clustering accuracy can be calculated, which is able to evaluate the similarity between the actual traffic data and the corresponding augmented traffic data.

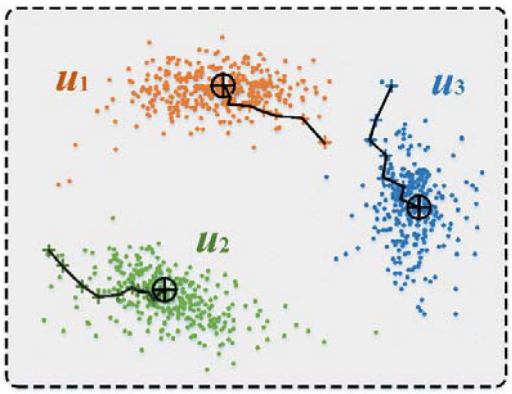

Fig. 2. Concept illustration of the k-means algorithm. $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ denotes the optimized centroid of the cluster1, cluster2 and cluster3 respectively. At first, the initial centroids are chosen randomly. After iterations, the centroids are convergent and stable eventually.

Optics EXPRESS

3. Applications in optical networks

In this section, the potential applications and value of the proposed sequential data augmentation method based on the GAN in specific optical network scenarios are discussed. Recently, the artificial intelligent (AI) techniques have been widely used in the cognitive optical networks. The AI-enabled cognitive optical network is promising to support the stringent quality of service from future 5G networks [26]. However, one of the crucial limitations in the researches on AI-based cognitive optical network is the lack of actual monitoring data for the targeted networking scenarios [27]. Especially for the inception 5G service, where sufficient monitoring data collected from the existed optical network are not available and less monitoring data are gathered from the testbed. To overcome the lack of actual data in AI-based cognitive optical networks, it is necessary to propose effective data augmentation approaches to provide sufficient data for machine learning applications. The proposed GAN has the potential to overcome this problem and augment high-quality data that are similar to the actual monitoring data from the cognitive optical network. One of the significant use cases of the proposed GAN for the cognitive optical network is to augment sufficient traffic data to train numerous AI-enabled applications including the reconfiguration of virtual network topologies according to the traffic changes [23], the dimensioning of next planning steps based on the traffic prediction [28,29], implementing of the load balancing and the resource reservation through the traffic classification [30]. The actual AI-enabled applications listed above in the cognitive networks require sufficient traffic data to train the machine learning algorithms.

Besides the adaptive traffic data augmentation, it should be noted that the proposed GAN can also provide a kind of general sequential data augmentation approach. In the cognitive optical network, the time-varied monitoring elements are widely distributed, including the optical power changes, optical signal noise ratio (OSNR) fluctuations, bit error ratio (BER) variations and the tendencies of running parameters in various optical network equipment such as the reconfigurable optical add-drop multiplexer (ROADM), optical cross-connect (OXC) devices, optical transceivers, etc. Those sequential monitored data are collected and further sent into the AI-enabled analysis modules in the control plane in the cognitive network, which can be analyzed to predict the optical component fault according to the variation of the temperature and the optical power in the optical transmission network (OTN) devices, find the specific location of network faults following the OSNR/BER degrading tendencies [31] and reduce the number of network alarms on the basis of the device condition parameters changes [32]. These time-varied monitored data in the cognitive optical network can also be augmented by the proposed GAN to generate sufficient data to train the AI-based models. Thus, the proposed GAN has the potential to augment diverse sequential monitoring data, not only the traffic data, for the cognitive optical network to enable the numerous intelligent AI-based network applications.

4. Experimental results and analysis

In this section, the performances of the GAN based traffic data augmentation technique are investigated in detail. Firstly, the experimental traffic data collected from three common scenarios in the access networks, i.e. the residential area (RA), the school area (SA) and the business area (BA), are utilized to train the GAN respectively and then the statistical evaluation parameters are adopted to compare the augmented traffic data with the corresponding experimental traffic data. Moreover, the k-means algorithm is selected to estimate their similarities intuitively. Further, the application of the adaptive traffic augmentation technique based on the GAN is extended to traffic data augmentation for the core network and the augmented traffic data in three kinds of trunks is evaluated. The types of the traffic data in three trunks include the morning-peak traffic, the evening-peak traffic and the multi-peak traffic. To add comparisons, two other classical generative models including the SPC model and the VAE model are also adopted to generate

Optics EXPRESS

the traffic data that are similar to the actual traffic data. All of the experimental traffic data is gathered from the optical networks of the China Mobile Communications Corporation (CMCC).

4.1. Traffic data augmentation for the access networks

To display the adaptive learning procedure of the proposed GAN based traffic augmentation technique, taking the traffic data augmentation for the SA for example, the relationship curves between training losses of the GAN and the number of iterations are shown in the Fig. 3(a). During the training stage of the GAN, the weights in the discriminator are updated to discriminate whether the category of the traffic data is the actual traffic or the augmented traffic. The total loss of the discriminator consists of the real loss, i.e. the loss of judging the actual traffic data as the augmented traffic data, and the fake loss, i.e. the loss of classifying the augmented traffic data as the actual traffic data. After iterations, the real loss and the fake loss of the discriminator in the GAN reach to the balance point, where both the real loss and the fake loss cannot be decreased and the corresponding loss curves are almost overlapped. This means that the discriminator is not able to differentiate whether the traffic data is actual or augmented by the generator. Moreover, the loss of the generator, i.e. the loss of generating the traffic data that is recognized by the discriminator as the augmented traffic data, fluctuates at the beginning of the training and further becomes more and more convergent, which indicates that the quality of the augmented traffic data is gradually stable. After the training stage, the augmented traffic is indistinguishable from the

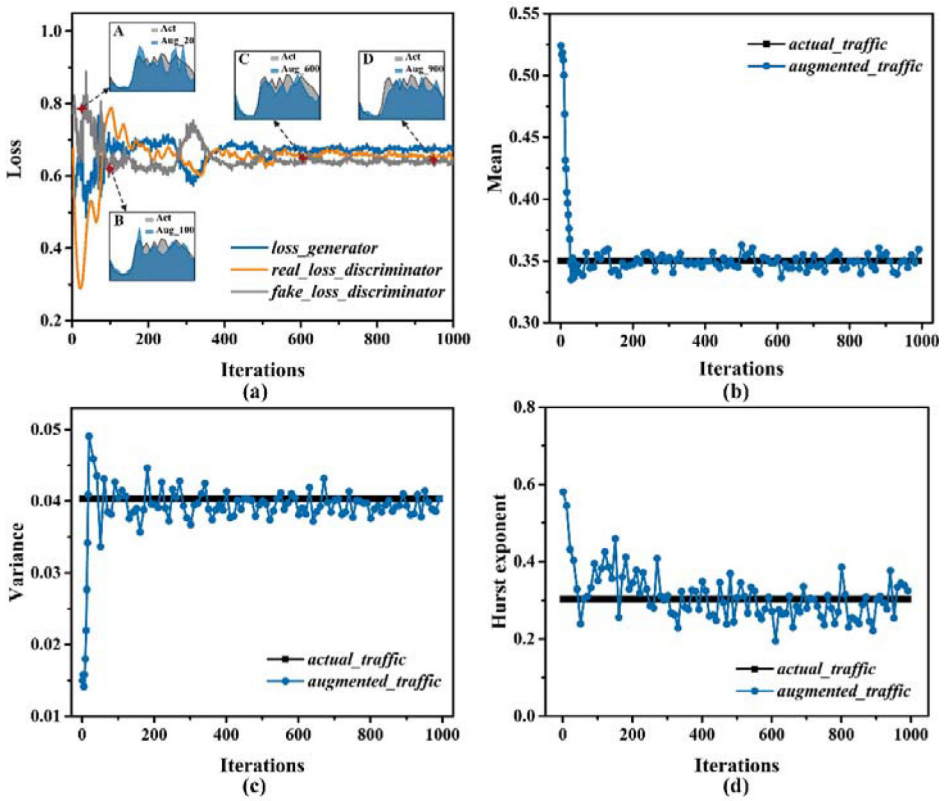

Fig. 3. (a).the training loss curves of the GAN in terms of the number of the iterations; the comparisons between (b).the mean, (c).the variance and (d) the Hurst exponent of the augmented traffic data and those of the experimental traffic data from the SA in the access networks. SA: the school area.

Optics EXPRESS

actual traffic and the probabilities of judging the traffic data as the actual data or the augmented data are both near 0.5.

Further, the augmented traffic data in the SA is evaluated and analyzed according to the statistical evaluation parameters of the traffic data. The Hurst exponent is recognized as the most significant statistical characteristic of the traffic data, which is used to describe the self-similarity of the traffic data and the classic R/S method [33,34] is adopted to calculate the Hurst exponent in this work. The comparisons between these statistical characteristics of the augmented traffic data and those of the actual traffic data from the SA in the access networks are shown in the Figs. 3(b)–3(d). It is observed that the mean, the variance and the Hurst exponent of the augmented traffic data are similar to those of the actual traffic data. After the convergence of the GAN, the average deviation of the mean, the variance and the Hurst exponent is about 0.001, 0.002 and 0.026 respectively.

Moreover, it is significant to select the parameters and architecture of the GAN to guarantee the convergence speed and the performances of the augmented traffic data from the GAN. The proposed GAN is constructed by a discriminative network and a generative network, where the classic single-hidden-layer artificial neural networks (ANN) are adopted. The reasons why we choose the single-hidden-layer ANN are that three-layer ANN is capable of fitting every complicated function and the sequential traffic data in the form of one dimension (1D) is convenient to be processed in the ANN. There are three important parameters required to be selected: the number of iteration times, the number of neurons in the hidden layer and the type of activation function. Moreover, the effect of different size of the training sample dataset is also investigated. To simplify the discussion, we choose a typical network scenario where the 126-day actual traffic data under the monitoring period T of 60 min are collected from the SA as the training sample dataset. The normalization processing is carried out for the collected actual traffic data to accelerate the training procedure of the GAN. To evaluate the convergence performances of the GAN, the Hurst exponent is adopted as the observation index for the reason that the Hurst exponent is recognized as the most significant statistical characteristic of the traffic data.

In the Fig. 4, the effects of different iteration times and diverse types of activation functions on the Hurst exponent of the generated traffic are illustrated. In the ANN, there are three classical types of activation functions: the rectified linear units (relu) activation function, the hyperbolic tangent (tanh) activation function and the sigmoid activation function. As shown in the Fig. 4, with the increase of the number of iteration times (above 600), the Hurst exponent of the generated traffic from the GAN with different activation functions gradually converges to relatively stable values. From the Fig. 4(a), we can see that the convergence values of the ANN with the relu activation function are closest to those of the actual traffic data. It can be found

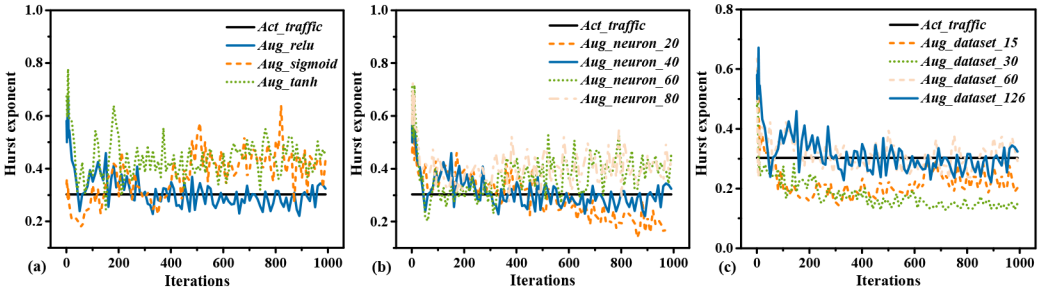

Fig. 4. the effects of (a).diverse types of activation functions, (b).various number of neurons in the hidden layer and (c) different size of the training sample dataset on the Hurst exponent of the generated traffic from the GAN.

Optics EXPRESS

from the Fig. 4(b) that the convergence performance the GAN is more stable when 40 hidden neurons are adopted. Finally, with the increase of the size of the training sample, the Hurst exponent difference between the generated traffic and the actual traffic decreases accordingly and there are less difference reduction when the size of the training sample dataset is larger than 60. According to Figs. 4(a)–4(c), we select the relu activation function and the number of the hidden neuron is set as 40. To guarantee the training performance, 126 pieces of traffic data are used as the training dataset for the traffic data augmentation.

Besides the SA in the access network, the GAN based traffic data augmentation for other two common scenarios including the business area (BA) and the resident area (RA) is also researched. In the Fig. 5, every row displays the comparisons between the actual traffic data and the corresponding augmented traffic data for three kinds of traffic scenarios. We can see that the distribution of the augmented traffic flow data from the GAN is similar with those of the actual traffic data, which indicates that the GAN based traffic data augmentation technique is feasible in diverse traffic scenarios for the access networks. For each class of traffic data, ~120 pieces of experimental traffic data are collected as the dataset to train the GAN. During the training stage, the generative neural network in the GAN gradually learns the intrinsic characteristics of the actual traffic data in the specific network scenarios and transfers the random noise into the

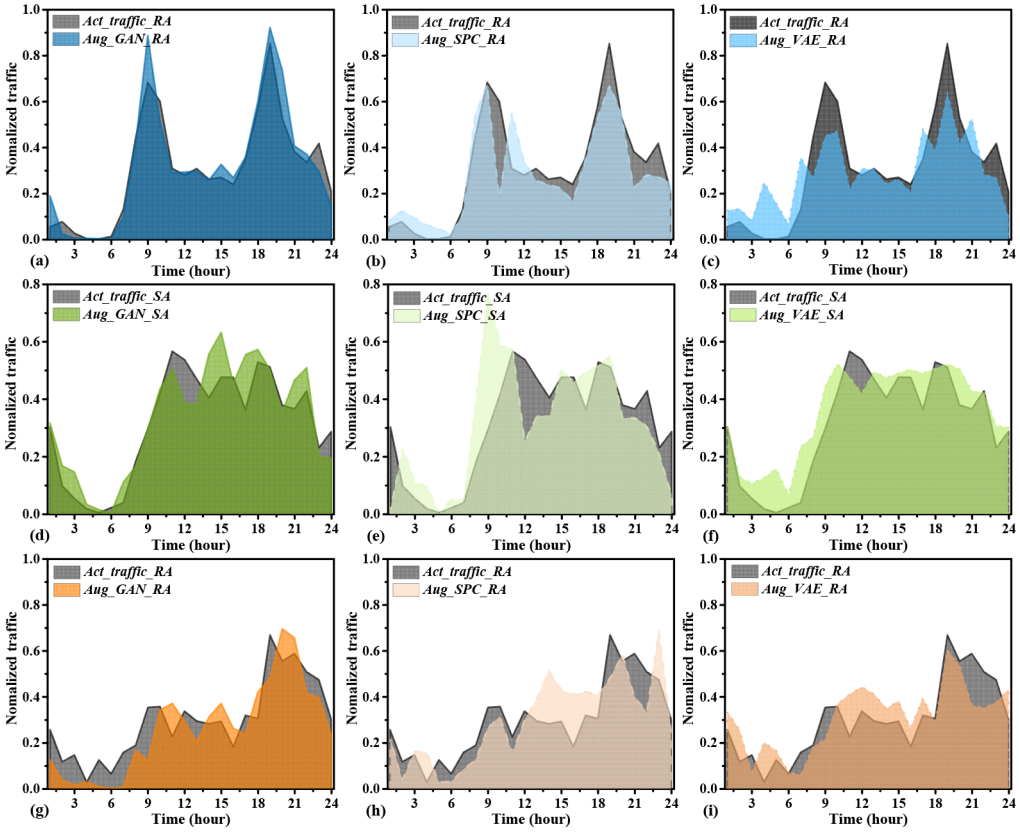

Fig. 5. the comparisons between the augmented traffic data from the generative models and the corresponding actual traffic data from (a)-(c).the RA, (d)-(f).the SA and (g)-(i).the BA in the access networks. RA: resident area; SA: school area; BA: business area; GAN: generative adversarial network; SPC: statistical parameter configuration; VAE: variational autoencoder.

Optics EXPRESS

augmented traffic data that the discriminative neural network cannot recognize the difference from the corresponding actual traffic data.

After the training of the GAN, 300 pieces of traffic data are augmented for every traffic scenario and the mean, variance and the Hurst exponent of the augmented data are calculated. What's more, the similarity between the actual traffic data and the augmented data is evaluated intuitively by the *k*-means algorithm. The reasons why *k*-means algorithm is selected are that it is capable of clustering different types of data according to the data distribution automatically and it is easy to understand. During the training stage of the *k*-means algorithm, the centroids of different clusters of the actual traffic date are upgraded until them are stable. During the testing stage, the augmented traffic data is assigned with a predicted cluster label. By comparing the predicted cluster label with the true cluster label of the augmented traffic data, the clustering accuracy is measured. When the characteristics of the augmented traffic data are close to those of the actual traffic data, the same cluster label will be allocated to the corresponding augmented traffic date. Thus, it is straightforward to evaluate whether the augmented traffic data is close to the accordingly actual traffic data or not. As shown in the Fig. 6(a), the clustering accuracy of three kinds of traffic data in the access network is 98.0%, 97.3% and 97.7% respectively.

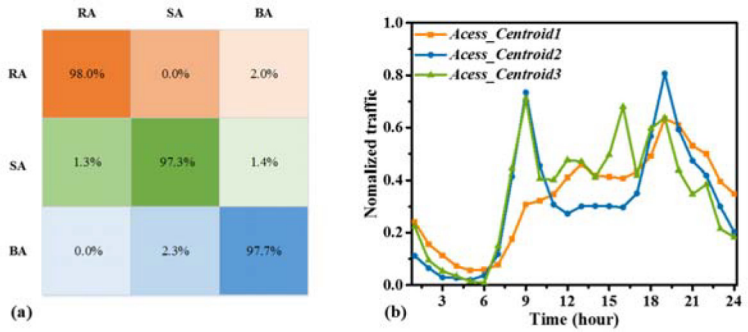

Fig. 6. (a).the clustering proportion of the *k*-means algorithm at each traffic types including the traffic data in the RA, SA and BA in the access network; (b).the centroid curves of different clusters identified in experimental traffic data from the access network.

To show the differences among the traffic data from various clusters, the centroid curves of diverse clusters recognized in the actual traffic data are selected to illustrate for the reason that the traffic data will be assigned into the cluster where the minimum distance between the traffic data and the corresponding centroid reaches in the k-means algorithm. In the other word, if the traffic data belongs to certain one cluster, the traffic data will be closest to the centroid of this cluster. We can see from the Fig. 6(b) that three centroid cluster curves are different from each other obviously in the access network, which means that the clusters identified in the experimental traffic data are also distant.

To add comparisons, two other classical generative models including the SPC model and the VAE model are also adopted to generate the traffic data. The reasons why we choose them are that the SPC model is intuitive to use the statistical parameters to configure the stochastic process and then generate traffic data with the similar statistical properties with actual traffic data and the VAE is the one of widely-used generative models in the machine learning community. As shown in the Table 1, after the training of the GAN, the average deviation of the mean, the variance and the Hurst exponent is about 0.001, 0.0002 and 0.026 in the traffic data augmentation for the SA in the access network accordingly, which is only 12.5%, 3.3%, 86.7% of those in the SPC and 9.1%, 10.0%, 68.4% of those in the VAE. Moreover, the clustering accuracy is 97.3%, 92.8% and 93.6% in the GAN, the SPC and the VAE respectively. For the traffic data augmentation in the BA and the RA, the average difference between the generated traffic data and the actual traffic

Optics EXPRESS

data is respectively 0.2%, 0.3% and 11.3% in terms of the mean, variance and the Hurst exponent in the GAN while the corresponding average difference is 9.4%, 4.6% and 28.1% in the SPC and 2.2%, 3.6% and 18.7% in the VAE. In the traffic data augmentation for the access network, the mean, variance, Hurst exponent and the clustering performances of GAN obviously exceeds those of the SPC and VAE, which is consistent with the traffic curve comparisons in the Fig. 5.

Table 1. The average performances of the augmented traffic data from three generative models trained with the actual traffic dada with the 60-minute interval for different traffic scenarios in the access network.

Performance parameter		Mean value		Variance value		Hurst exponent		Clustering accuracy
Traffic type	Model	Act	Aug	Act	Aug	Act	Aug	Aug
SA	GAN	0.350	0.349	0.040	0.040	0.303	0.329	97.3%
	SPC	0.350	0.342	0.040	0.046	0.303	0.273	92.8%
	VAE	0.350	0.361	0.040	0.038	0.303	0.341	93.6%
BA	GAN	0.331	0.331	0.041	0.043	0.307	0.321	97.7%
	SPC	0.331	0.324	0.041	0.031	0.307	0.391	96.3%
	VAE	0.331	0.332	0.041	0.038	0.307	0.354	97.9%
RA	GAN	0.300	0.301	0.046	0.046	0.470	0.385	98.0%
	SPC	0.300	0.250	0.046	0.043	0.470	0.335	79.7%
	VAE	0.300	0.312	0.046	0.049	0.470	0.366	87.0%

4.2. Traffic data augmentation for the core networks

In this section, the feasibility of the GAN based traffic data augmentation technique for the traffic data in the core network is investigated. Firstly, we analyze the effects of different time interval of the actual traffic data on the performances of the GAN. With the decrease of the time interval, the size of the traffic data increases exponentially. When the time interval is set as 60 minutes, 30 minutes, 10 minutes and 5 minutes respectively, the size of the traffic data is valued in 24×1, 48×1, 144×1 and 288×1 accordingly. As shown in the Fig. 7(a), the average clustering accuracy for three kinds of augmented traffic data increases from 98.3% to 99.8% when the size of the traffic data varies from 24×1 to 288×1. With the decrease of the time interval, more detailed information of the traffic data can be exploited and then the convergent clustering accuracy is improved gradually. Finally, the clustering accuracy increases by a small margin and achieves the saturation. To accelerate the training procedure of the GAN, the time interval of the traffic data is set as 24×1.

Further, the specific clustering accuracy of augmented traffic data for different types of traffic in the core network is also analyzed. As demonstrated in the Fig. 7(b), the clustering accuracy of the augmented traffic data with size of 24×1 for three categories of trunks, i.e. trunk1: morning-peak traffic, trunk2: multi-peak traffic and trunk3: evening-peak traffic, is 100.0%, 95.6% and 99.4% respectively. The clustering accuracy of the augmented traffic data with multi peaks is about 4% less than those with single peak because more complicated fluctuations are needed to be captured for the GAN for the multi-peak traffic data.

Moreover, the specific clustering results with the different number of iterations of the GAN are displayed in the Fig. 7(c). With the increase of the number of iterations, the clustering accuracies are improved gradually. After ~700 iterations, the clustering accuracy of the augmented traffic data is convergent. All of clustering accuracies of augmented traffic data for three kinds of trunks in the core network are more than 95%. It should be noted that centroids of these clusters are obtained automatically from the experimental traffic data by the k-means algorithm without the manual parameter selection, where k is set as 3 corresponding to three clusters. After 700

Optics EXPRESS

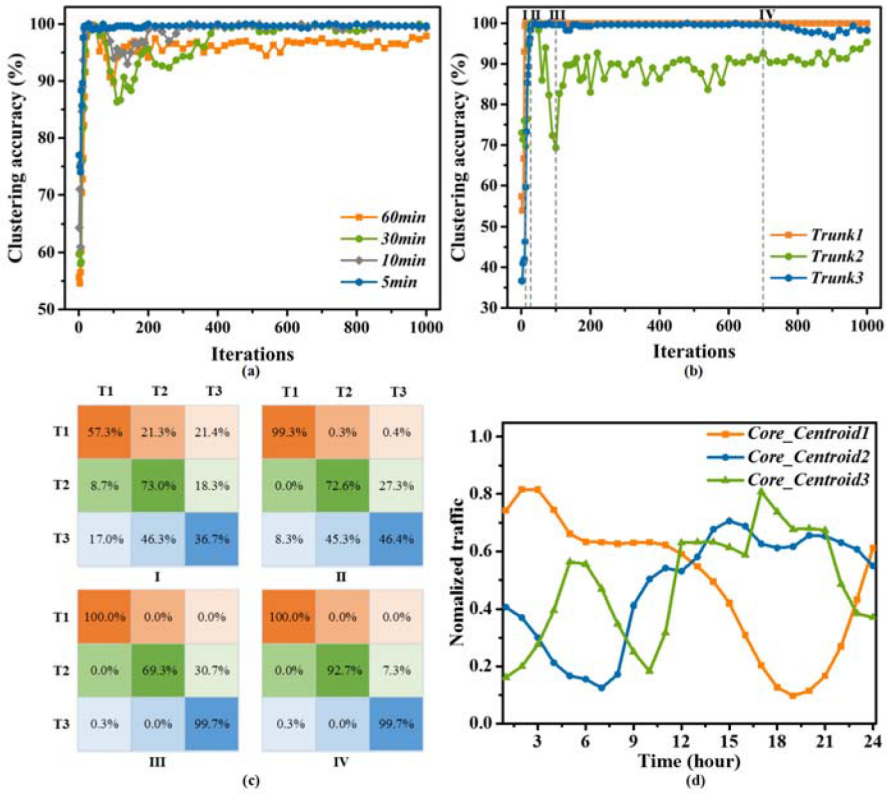

Fig. 7. (a).the average clustering accuracy curves of the augmented traffic data from the GAN for the core network; (b).the specific clustering accuracy of the 60-minute-interval augmented traffic data for three types of trunks; (c).the clustering results shown in the format as confusion matrixes when the number of iterations of the GAN is set as (I).1, (II).10, (III).100 and (IV). 700 respectively; (d).the centroid curves of different clusters identified in experimental traffic data from the core network.

iterations, the specific augmented traffic data for the trunk1, the trunk2 and the tunck3 is displayed in the Fig. 7. It is observed that the high similarity between the augmented traffic data and the actual traffic data is in accordance with the high clustering accuracy. In the Fig. 7(d), three centroid cluster curves are different from each other in the core network, which indicates that the clusters identified in the actual traffic data have obvious differences.

Moreover, the comparisons among three traffic generative models in the traffic data augmentation for the core network are also investigated. We can see from the Table 2 that the maximum difference of the mean and the deviation of the variance between the augmented traffic data and the actual traffic data is accordingly 1.7% and 1.5% in the GAN, which is respectively 2.7% and 4.3% in the SPC and 1.2% and 12.9% in the VAE. What's more, the average difference between the actual traffic and the generated traffic in terms of the Hurst exponent is respectively valued in 3.5%, 32.9% and 15.1% in the GAN, SPC and VAE. The average clustering accuracy of the augmented traffic from the GAN also outperforms than those of augmented traffic data from the SPC and the VAE, which is consistent with the traffic profile comparisons in the Fig. 8.

Taken into account of the mean, variance, Hurst exponent and the clustering accuracy, the generated traffic data from the GAN are more similar to the corresponding actual traffic data than those from the SPC and AVE, which also confirms to the traffic profile comparisons between the

Optics EXPRESS

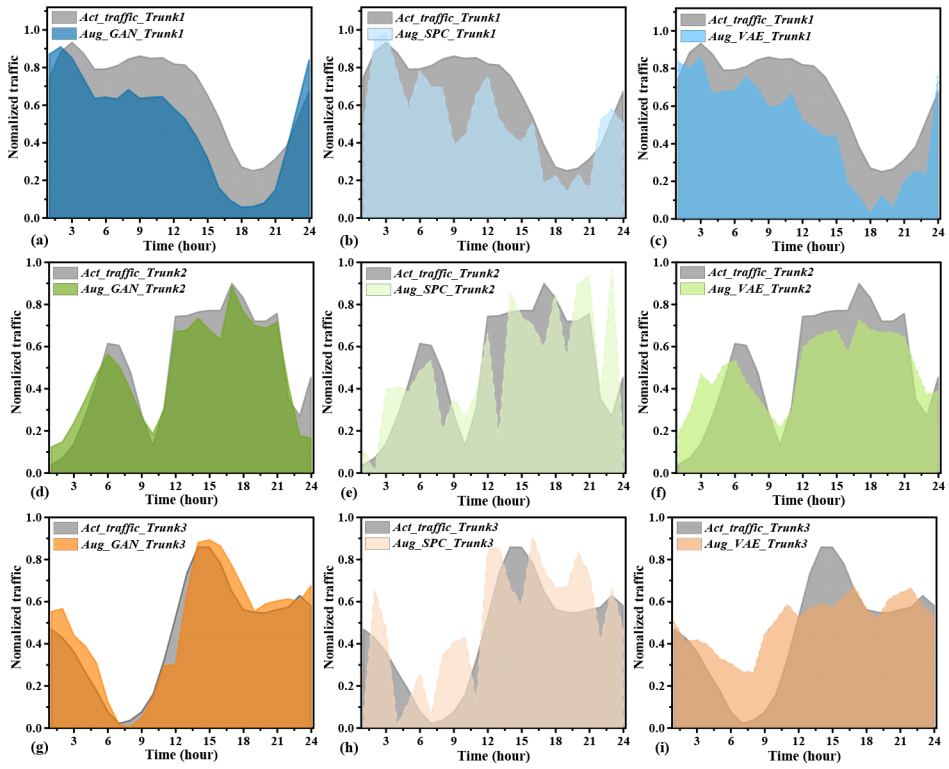

Fig. 8. the comparisons between the augmented traffic data from the GAN, SPC and VAE and the corresponding actual traffic data from (a)-(c).the Trunk1, (d)-(f).the Trunk2 and (g)-(i).the Trunk3 in the core networks. Trunk1: morning-peak traffic; Trunk2: hybrid traffic; Trunk3: evening-peak traffic.

Table 2. The average performances of the augmented traffic data from different generative models trained with the actual traffic dada with 60-minute interval for different traffic scenarios in the core network. T1: trunk1 (morning-peak traffic); T2: trunk2 (hybrid traffic); T3: trunk3 (evening-peak traffic).

Performance parameter		Mean value		Variance value		Hurst exponent		Clustering accuracy
Traffic type	Model	Act	Aug	Act	Aug	Act	Aug	Aug
T1	GAN	0.496	0.496	0.067	0.066	0.659	0.631	100.0%
	SPC	0.496	0.501	0.067	0.068	0.659	0.349	97.9%
	VAE	0.496	0.494	0.067	0.064	0.659	0.774	99.0%
T2	GAN	0.483	0.475	0.057	0.057	0.749	0.711	95.6%
	SPC	0.483	0.476	0.057	0.058	0.749	0.538	86.3%
	VAE	0.483	0.477	0.057	0.051	0.749	0.595	93.0%
T3	GAN	0.484	0.485	0.070	0.069	0.591	0.598	99.4%
	SPC	0.484	0.497	0.070	0.073	0.591	0.451	97.9%
	VAE	0.484	0.483	0.070	0.079	0.591	0.633	99.0%

Optics EXPRESS

generated traffic and the actual traffic. Therefore, the proposed GAN is more suitable to generate diverse traffic data that is close to the actual traffic data for the optical networks. In the SPC, the theory is intelligible and the traffic generative model is easy to be implemented, but the traffic data distribution is assumed as the certain distribution (the normal distribution is chosen generally). For the increasing complex and diverse traffic types from emerging network applications, the performances of the SPC for the traffic data augmentation are limited by the finite description capability of fixed traffic data distribution models. Contrastively, the data distribution model is learned from the actual traffic data in the GAN. The GAN is specialized in approximating the actual data distribution adaptively through the zero-sum gaming theory without the traffic distribution assumption. Compared with the VAE, one of the classic generative models in the machine learning community, the GAN is capable of automatically extracting eccentric features to improve the robust of the trained network and avoid over-fitting. After the adversarial learning procedure, the essential features of the actual traffic data are discovered by the well-trained GAN. Therefore, the GAN can be more insensitive to the variation among the same traffic types and generate the traffic data that conforms to the characteristics of the actual traffic robustly.

Moreover, the clustering accuracy, the mean, the variance and the Hurst exponent of the augmented traffic data with different intervals in the core networks are also calculated in the Table 3. When the time interval of the traffic data is set as 60 minutes, the average deviation of the mean, the variance and the Hurst exponent between 900 pieces of augmented traffic data and 360 pieces of experimental traffic data in the core network is 1.2%, 1.0% and 3.5% respectively. For the traffic data with the other time intervals, the similarities between the augmented traffic data and the experimental traffic data are also very high. In the telemetry context, traffic monitoring data can be measured with sub-second intervals. The feasibility of the proposed GAN for the traffic data augmentation where the granularity of the traffic data is 0.5s is also investigated. Owing to the monitoring granularity limitation, the minimum traffic monitoring granularity of the actual traffic data we have collected is 5 minutes. To investigate the feasibility of the proposed GAN for the traffic data augmentation with the sub-second interval, the polynomial interpolation is implemented for the actual traffic data with 5-minute interval and the traffic data is then resampled with the 0.5-second granularity. The resampled traffic data are further sent in to the GAN to generate the augmented traffic. As shown in the Table 3, the trained GAN is able to augment the traffic data where the traffic profiles are consistent with those of the traffic data with the 0.5-second interval. The average deviation between the augmented traffic data and the actual traffic data is respectively valued in 4.0%, 6.3% and 8.7% on the mean, variance and the Hurst exponent. Thus, the proposed GAN is also suitable for the fine-granular traffic data augmentation where the sub-second-interval traffic data served as the training data.

The comprehensive results show that the GAN-based traffic augmentation technique is able to capture the major features of different traffic types in the core network. Moreover, the detailed comparisons between the augmented traffic data from the GAN and the actual traffic data about the mean, the variance and the Hurst exponent for the access networks and the core networks are concluded in the Table 4. On the one hand, in the access network, the difference of the mean and the deviation of the variance between the augmented traffic data and the experimental traffic data is near 0.2% and 1.6% respectively, which is 1.2% and 1.0% in the core network accordingly. On the other hand, the Hurst exponent of the augmented traffic data is about 90% and 96% of the corresponding actual traffic data in the access network and the core network respectively. As we can see from the Table 2, the mean, the variance and the Hurst exponent of the augmented traffic data are all very close to those of the experimental traffic data. The reasons why the general performances of the GAN for the core network traffic augmentation are slightly better than those for the access network traffic augmentation are that the traffic data is more stable and regular in the core network, where the unpredictable traffic fluctuations from diverse services in the access network are converged and may be canceled each other out. Moreover, the clustering accuracies

Optics EXPRESS

Table 3. The average performances of the augmented traffic data from the GAN trained with the actual traffic data with 60-minute, 30-minute, 10-minute, 5-minute and 0.5-second intervals in the core network. Act: actual traffic data; Aug: augmented traffic data.

Performance parameter		Mean value		Variance value		Hurst exponent		Clustering accuracy
Traffic type		Act	Aug	Act	Aug	Act	Aug	Aug
60min	T1	0.496	0.496	0.067	0.066	0.659	0.631	100.0%
	T2	0.483	0.475	0.057	0.057	0.749	0.711	95.6%
	T3	0.484	0.485	0.070	0.069	0.591	0.598	99.4%
30min	T1	0.499	0.501	0.607	0.063	0.671	0.617	100.0%
	T2	0.484	0.483	0.055	0.058	0.752	0.723	100.0%
	T3	0.483	0.475	0.066	0.066	0.596	0.521	99.1%
10min	T1	0.483	0.487	0.057	0.060	0.720	0.652	100.0%
	T2	0.475	0.477	0.053	0.059	0.740	0.731	100.0%
	T3	0.465	0.467	0.064	0.076	0.632	0.551	98.0%
5min	T1	0.499	0.469	0.060	0.053	0.671	0.628	100.0%
	T2	0.484	0.482	0.055	0.053	0.752	0.636	100.0%
	T3	0.483	0.452	0.066	0.065	0.596	0.568	99.6%
0.5s	T1	0.461	0.464	0.044	0.043	0.662	0.631	100.0%
	T2	0.455	0.441	0.041	0.047	0.047	0.622	100.0%
	T3	0.450	0.461	0.053	0.054	0.589	0.554	99.8%

for 6 kinds of typical traffic types are all above 95%. The comprehensive results demonstrate that the proposed GAN is capable of extracting the intrinsic characteristics of traffic data in 6 kinds of network scenarios and providing sufficient and diverse augmented traffic data as we need for the dynamic optical networks.

Table 4. The average performances of the augmented traffic data from the GAN trained with the actual traffic dada with 60-minute interval for different traffic scenarios in the optical network. SA: school area; BA: business area; RA: resident area.

Performance parameter		Mean value		Variance value		Hurst exponent		Clustering accuracy
Traffic type		Act	Aug	Act	Aug	Act	Aug	Aug
Access network	SA	0.350	0.349	0.040	0.040	0.303	0.329	97.3%
	BA	0.331	0.331	0.041	0.043	0.307	0.321	97.7%
	RA	0.300	0.301	0.046	0.046	0.470	0.385	98.0%
Core network	T1	0.496	0.496	0.067	0.066	0.659	0.631	100.0%
	T2	0.483	0.475	0.057	0.057	0.749	0.711	95.6%
	T3	0.484	0.485	0.070	0.069	0.591	0.598	99.4%

5. Conclusion

The performances of the machine learning based applications are usually limited by the lack of the diverse training data in practice. By employing the excellent data augmentation and adversarial learning capability of the GAN, an adaptive aggregate traffic data augmentation technique based on deep learning is proposed for 6 kinds of typical network scenarios. The statistical evaluation parameters are adopted to evaluate the augmented traffic data from the trained GAN. After being trained with the experimental traffic data, the deviations between the augmented traffic data and the actual traffic data in the mean and the variance are both less than 1.7%. The Hurst exponent

Optics EXPRESS

of the augmented traffic data is about 90% and 96% of the corresponding actual traffic data in the access network and the core network respectively. To be more intuitive, the k-mean algorithm is used to estimate the similarity between the augmented traffic data and the actual traffic data. The clustering accuracies are all above 95% for different traffic categories. The comprehensive comparisons among the proposed GAN, the SPC and VAE show that the performances of the GAN exceed those of the SPC and the VAE. The proposed GAN is capable of learning the intrinsic features of various traffic types through the zero-sum game theory and transferring the random noise into the augmented traffic data that is indistinguishable from the corresponding actual traffic data. The proposed traffic data augmentation technique is able to generate the diverse augmented traffic data on demand with less experimental traffic data and shows great potentials in practical applications, such as the training dataset augmentation and the optical communication system modeling. The feasibility of the proposed GAN for other sequential data augmentation is also interesting to be investigated.

Funding

National Natural Science Foundation of China (NSFC) (61705016); Beijing University of Posts and Telecommunications (BUPT) (No.CX2019313); the Fundamental Research Funds for the Central Universities (2019RC12).

Acknowledgments

This work has been supported by the State Key Laboratory of Advanced Optical Communication Systems and Networks, China. We gratefully acknowledge the China Mobile Communications Corporation (CMCC) for the permission to use their network traffic data in this work.

References

1. Z. Dong, F. N. Khan, Q. Sui, K. Zhong, C. Lu, and A. P. T. Lau, "Optical Performance Monitoring: A Review of Current and Future Technologies," J. Lightwave Technol. **34**(2), 525–543 (2016).
2. F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini, and M. Tornatore, "An Overview on Application of Machine Learning Techniques in Optical Networks," IEEE Commun. Surv. Tut., Early Access, **21**(2), 1383–1408 (2019).
3. D. Wang, M. Zhang, J. Li, Z. Li, J. Li, C. Song, and X. Chen, "Intelligent constellation diagram analyzer using convolutional neural network-based deep learning," Opt. Express **25**(15), 17150–17166 (2017).
4. D. Wang, M. Zhang, Z. Li, J. Li, M. Fu, Y. Cui, and X. Chen, "Modulation Format Recognition and OSNR Estimation Using CNN-Based Deep Learning," IEEE Photonics Technol. Lett. **29**(19), 1667–1670 (2017).
5. D. Wang, M. Zhang, Z. Li, C. Song, M. Fu, J. Li, and X. Chen, "System impairment compensation in coherent optical communications by using a bio-inspired detector based on artificial neural network and genetic algorithm," Opt. Commun. **399**, 1–12 (2017).
6. J. Li, M. Zhang, D. Wang, S. Wu, and Y. Zhan, "Joint atmospheric turbulence detection and adaptive demodulation technique using the CNN for the OAM-FSO communication," Opt. Express **26**(8), 10494–10508 (2018).
7. J. Li, M. Zhang, and D. Wang, "Adaptive Demodulator Using Machine Learning for Orbital Angular Momentum Shift Keying," IEEE Photonics Technol. Lett. **29**(17), 1455–1458 (2017).
8. Y. Huang, C. L. Gutterman, P. Samadi, P. B. Cho, W. Samoud, C. Ware, M. Lourdiane, G. Zussman, and K. Bergman, "Dynamic mitigation of EDFA power excursions with machine learning," Opt. Express **25**(3), 2245–2258 (2017).
9. L. Barletta, A. Giusti, C. Rottondi, and M. Tornatore, "QoT Estimation for Unestablished Lighpaths using Machine Learning," in *Proceedings of Optical Fiber Communication Conference (OFC 2017)*, paper Th1J.1.
10. D. Rafique and L. Velasco, "Machine Learning for Network Automation: Overview, Architecture, and Applications [invited tutorial]," J. Opt. Commun. Netw. **10**(10), D126–D143 (2018).
11. Z. Wang, M. Zhang, D. Wang, C. Song, M. Liu, J. Li, L. Lou, and Z. Liu, "Failure prediction using machine learning and time series in optical network," Opt. Express **25**(16), 18553–18565 (2017).
12. F. Morales, M. Ruiz, L. Gifre, L. M. Contreras, V. López, and L. Velasco, "Virtual Network Topology Adaptability Based on Data Analytics for Traffic Prediction," J. Opt. Commun. Netw. **9**(1), A35–A45 (2017).
13. M. Wang, Y. Cui, X. Wang, S. Xiao, and J. Jiang, "Machine Learning for Networking: Workflow, Advances and Opportunities," IEEE Network **32**(2), 92–99 (2018).
14. W. Lin, S. Ke, and C. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," Knowledge-Based Systems **78**, 13–21 (2015).

Optics EXPRESS

15. L. Velasco, A. Castro, D. King, O. Gerstel, R. Casellas, and V. Lopez, "In-operation network planning," IEEE Commun. Mag. **52**(1), 52–60 (2014).
16. A. S. Thyagaturu, A. Mercian, M. P. McGarry, M. Reisslein, and W. Kellerer, "Software Defined Optical Networks (SDONs): A Comprehensive Survey," IEEE Commun. Surv. Tut. **18**(4), 2738–2786 (2016).
17. T. Karagiannis, M. Molle, and M. Faloutsos, "Long-range dependence ten years of Internet traffic modeling," IEEE Internet Comput. **8**(5), 57–64 (2004).
18. J. Kolbusz, S. Paszczyski, and B. M. Wilamowski, "Network Traffic Model for Industrial Environment," in *Proceedings of IEEE International Conference on Industrial Informatics (INDIN 2005)*, paper 10-12.
19. I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Proceedings of Advances in Neural Information Processing Systems (NIPS 2014)*.
20. X. Mao, Q. Li, H. Xie, R.Y.K. Lau, Z. Wang, and S. P. Smolley, "Least Squares Generative Adversarial Networks," in *Proceedings of IEEE International Conference on Computer Vision (ICCV 2017)*, paper 2794-2802.
21. X. Huang and S. Belongie, "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization," in *Proceedings of IEEE International Conference on Computer Vision (ICCV 2017)*, paper 2380-7504.
22. X. Wang and A. Gupta, "Generative Image Modeling Using Style and Structure Adversarial Networks," in *Proceedings of European Conference on Computer Vision (ECCV 2016)*, paper 318-335.
23. F. Morales, L. Gifre, F. Paolucci, M. Ruiz, F. Cugini, P. Castoldi, and L. Velasco, "Dynamic core VNT adaptability based on predictive metro-flow traffic models," J. Opt. Commun. Netw. **9**(12), 1202–1211 (2017).
24. F. Morales, M. Ruiz, and L. Velasco, "Metro-Flow Traffic Modelling for Cognitive Adaptation of Core Virtual Network Topologies," *In Proceedings of International Conference on Transparent Optical Networks (ICTON 2018)*.
25. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv preprint arXiv: 1312.6114, 2013.
26. R. Borkowski, R. J. Durán, C. Kachris, D. Siracusa, A. Caballero, N. Fernández, D. Klonidis, A. Francescon, T. Jiménez, J. C. Aguado, I. Miguel, E. Salvadori, I. Tomkos, R. M. Lorenzo, and I. T. Monroy, "Cognitive Optical Network Testbed: EU Project CHRON [Invited]," J. Opt. Commun. Netw. **7**(2), A344–A355 (2015).
27. M. Ruiz, F. Coltraro, and L. Velasco, "CURSA-SQ: A Methodology for Service-Centric Traffic Flow Analysis," J. Opt. Commun. Netw. **10**(9), 773–784 (2018).
28. L. Velasco, F. Morales, L. Gifre, A. Castro, O. González de Dios, and M. Ruiz, "On-demand incremental capacity planning in optical transport networks," J. Opt. Commun. Netw. **8**(1), 11–22 (2016).
29. M. Wang, Y. Cui, X. Wang, S. Xiao, and J. Jiang, "Machine Learning for Networking: Workflow, Advances and Opportunities," IEEE Network **32**(2), 92–99 (2018).
30. W. Lin, S. Ke, and C. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," Knowledge-Based Systems **78**, 13–21 (2015).
31. F. N. Khan, Q. Fan, C. Lu, and A. P. Tao Lau, "An Optical Communication's Perspective on Machine Learning and Its Applications," J. Lightwave Technol. **37**(2), 493–516 (2019).
32. Y. Zhao, B. Yan, D. Liu, Y. He, D. Wang, and J. Zhang, "SOON: self-optimizing optical networks with machine learning," Opt. Express **26**(22), 28713–28726 (2018).
33. W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," IEEE/ACM Trans. Netw. **5**(1), 71–86 (1997).
34. M. Grossglauser and J. C. Bolot, "On the Relevance of Long-Range Dependence in Network Traffic," IEEE/ACM Trans. Netw. **7**(5), 629–640 (1999).