DAGAT: Data Augmentation and Generation for Anomalous Time Series Signals

1st Thasorn Chalongvorachai
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
62606065@kmitl.ac.th

2nd Kuntpong Woraratpanya*
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
kuntpong@it.kmitl.ac.th

***Abstract*—Anomaly detection tasks using learning methods are still challenging. Especially, in deep learning, a model requires a high amount of training data, which are opposed to a limited number of anomalous data that are rare events, such as high privacy concerns, secret records, and difficult gathering data. One of the efficient solutions for this problem is data augmentation to increase the number of synthesized data for training. However, data augmented with simple mathematical techniques cannot provide a high variety of patterns for learning purposes. Hence, this paper proposed a novel framework of data augmentation and generation for anomalous time series signals (DAGAT). This framework was applied on a Secure Water Treatment (SWaT) dataset, recorded from various sensors, containing anomalous events occurred on the test bed of secure water treatment system. The performance of DAGAT shows that, from only one sample of the rare case, it can be generated up to almost three thousands of reliable samples with a high variety of patterns for deep learning.**

***Index Terms*—data augmentation, anomaly detection, deep learning, anomalous time series signals**

I. Introduction

In real world problems, anomaly detection in time series signals is one of the most challenging tasks. The most anomalous incidents, unlikely to happen often, typically offer a small number of accessible records. The causes of this were from the following reasons: privacy and secret concerns, such as medical records, high value machinery information, or trade secrets. This becomes an obstacle for anomalous detection since there is no information to observe in classifying between normal and abnormal situations.

In the past, to solve anomalous detection, an unbalanced learning approach, one of the machine learning techniques, was used to classify the problem of normal and abnormal situations. This approach allows a model to learn from normal data only, which are the large number of samples available, and assumes that any data are much different from the model learned as anomalous events [1]. Although this learning approach can be powerful to detect the anomaly events, there are some flaws, which were unable to learn the anomalous data from its characteristics, thus making it inadequate to learn in classifying types of anomalous events. Moreover, using the limited number of data in learning process, which has no generalization, may result an unsatisfactory model in deep learning [1].

Many researchers recognized these weaknesses and attempted to solve them by using data augmentation techniques to increase synthetic data for training models [2]. However, those data augmentation techniques for time series signals are still limited and needs to be extended the space to possible produced samples.

Therefore, this paper proposed Data Augmentation and Generation for Anomalous Time Series Signals (DAGAT) that leverage data augmentation to an advanced step. We successfully generated numerous, reliable, and diverse samples, for training a deep learning model, by applying typical augmentation methods with upsampling and downsampling, fast Fourier transform, and time series decomposition approaches for time series data. These methods produce different augmented types. Then, using the power of Variational Autoencoder (VAE) to generate samples from latent space based on previous learning-augmented data. In collaboration with VAE, data picker selects samples in between different augmented types, and then signal fragment assembler divides those samples and randomly reassemble to produce a large number of time series records. Finally, to ensure the quality of produced samples, quality classifier identifies duplicate data as well as poor quality data that look different from the existing anomalous dataset.

II. Related Works

A. Time Series Data Augmentation

When dataset has limited in numbers, especially the small number of available anomalous event records, data augmentation is a technique that can be used to increase numbers of data by applying slight modification or generating synthetic samples based on existing data [2]. Here is an example of the successful case. In 2017, Um *et al.* [3], proposed data augmentation of wearable sensors for parkinson's disease monitoring. Augmentation techniques were consisted of jittering, scaling, cropping, rotating, permutating, magnitude-warping, and time-warping methods to improve performance of convolutional neural networks.

B. Deep Learning

Deep learning is a state-of-the-art technology that uses a computational method, consisting of multiple processing layers with various functions in learning process, for developing models for various purposes [4]. Some usages of deep learning are facial recognition, stock-price forecasting, object detection, and anomaly detection. This technology requires a decent amount of data for deep models to achieve good performance on their tasks. In other words, the more data they learned, the better outcomes we get.

*Corresponding author: kuntpong@it.kmitl.ac.th

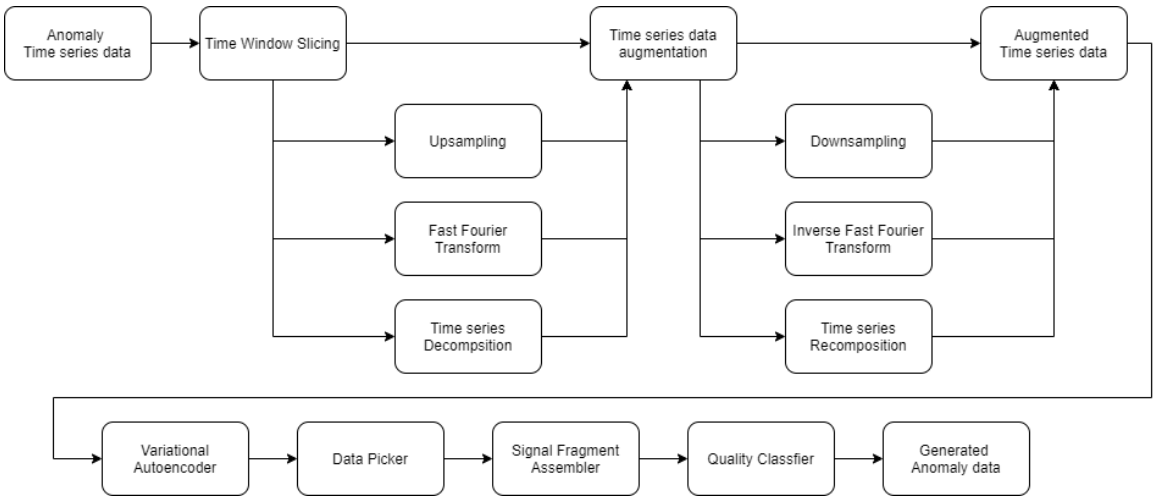

Fig. 1: A framework of DAGAT.

A small number of datasets for training process in deep learning can cause an overfitting problem in an output model [1]. It is happened when training samples having no generalization, thus resulting in poor efficiency of detection and prediction, since the model is not robust against the change of input that naturally does not have an exact same form.

C. Anomaly Detection for Time Series Data

Anomaly detection for time series data has been interested research topics in recent year, since it is a real world problem that not yet has been fully resolved. These unusual incidents can cause system failure, loss of machinery or life, if it is a medical issue [5]. However, these events are hard to detect and predict, since they do not happen often, which lead to small amount of data available.

Some examples of study on anomaly detection for time series data to prevent a problem of limited data are learned via normal data and assume that any data, not fit to the normal class, are classified as abnormal [6], or find the outlier from the cluster of data distribution [7]. This approach can be powerful for unsupervised learning since they do not require abnormal data to learn. Nonetheless, these solutions have not been profoundly examined yet; they cannot classify type of anomalous data for further purposes. In other words, though the system can detect anomalous incidents, it cannot identify what type of anomaly is.

III. Proposed Method

We proposed a framework of Data Augmentation and Generation for Anomalous Time Series Signals (DAGAT) that can synthesize up to a thousand of samples, with high quality and variation, from a few samples of abnormal incidents on time series data. The novel aspects of this framework are applying typical augmentation methods on various domains and the use of VAE to generate more possible samples based on latent space that contain vary traits of augmented data.

For better understanding, our framework is illustrated in Fig. 1 and explained as follows:

A. Time Window Slicing

Time window slicing is a simple technique for data augmentation by slightly moving the position of interested data area. It can be implemented by determining a window size and time step to move the window on a whole dataset to capture and create new samples based on the original data. Here, the time window slicing is the first step of our framework in initiating synthetic data before feeding them into other augmentation methods.

B. Vanilla Time Series Data Augmentation

As mentioned in section II, we have chosen some techniques from Um *et al.* [3] for our work.

a) Jittering: Adding random Gaussian noise into time series signals.

b) Scaling: Adjusting the magnitude of time series signals based on random scalar.

c) Magnitude Warping: Randomly applying synthetic curve to each sample to change the magnitude of time series signals.

Data from the time window slicing step flow into vanilla augmentation directly or transform to different domains as explained in sections III-C, III-D and III-E to expand the augmentation possibility before transforming them back to their original domains again.

C. Upsampling and Downsampling

Upsampling is a method for providing interpolated data based on the approximation of existing samples [8]. The upsampling in augmentation process is to widen the space of samples that can be permuted. After upsampling the data, time series augmentation from section III-B will be applied randomly. Once the process is completed, we have to downsampling data back to the original form before proceeding to the next step.

D. Fast Fourier Transform and Inverse Fast Fourier Transform

Most of augmentation methods were done on time domain, but augmentation methods on frequency domain have not been deeply investigated yet. We augmented our data by transforming samples with one-dimensional fast Fourier transform (FFT) [9], which can be defined in (1).

$$Y(k) = \sum_{n=0}^{k=1} x(j) W_n^{(j-1)-(k-1)} \tag{1}$$

where $W_n = e^{(-2\pi i/n)}$, n is the length of data, $x(j)$ is the data in time domain, and $Y(k)$ is the data in frequency domain. When data in frequency domain flowed into (1) are augmented completely, we have to transform them back to their usual form, time domain, by taking an inverse FFT as defined in (2),

$$x(j) = \frac{1}{n} \sum_{n=0}^{k=1} Y(k) W_n^{-(j-1)-(k-1)} \tag{2}$$

E. Time Series Decomposition and Time Series Recomposition

Time series decomposition is a statistical technique to extract signals into different components [10]. The decomposition methods can be additive and multiplicative models as expressed by (3) and (4), respectively.

$$y_{t_a} = T_t + S_t + R_t \tag{3}$$

$$y_{t_m} = T_t \times S_t \times R_t \tag{4}$$

where y_{t_a} and y_{t_m} are the original time series of additive and multiplicative models, respectively, T_t is the trend component of time series, S_t is the seasonal component of time series and R_t is the residual component of time series.

Note that we have to recompose each component back to the time series signal again after finishing augmentation individual elements to continue the process of generation.

F. Variational Autoencoder

Kingma and Welling [11] proposed a neural network architecture that can explicit the latent space of samples that are learned for data generation purposes. This kind of neural network architecture is called Variational Autoencoder (VAE). The objective function of VAE can be given by (5).

$$\begin{aligned} &logP(X) - D_{KL}[Q(z|X)||P(z|X)] = \\ &E[logP(X|z)] - D_{KL}[Q(z|X)||P(z)] \end{aligned} \tag{5}$$

where D_{KL} is the KL divergence function, X are the samples, z is the latent space, $P(X)$ is the probability distribution of X, $P(z)$ is the probability distribution of z, $Q(z|X)$ is the projection of X onto z and $P(X|z)$ is the projection of X given by z.

After applying all previous methods, as described in sections III-A, III-B, III-C, III-D, and III-E, to create augmented data, we trained VAE to learn the latent space of those augmented data, so that we can randomly generate a variety of patterns from that space.

G. Data Picker

In order to choose a sample, that has a slightly different characteristic, from the latent space of augmented data, we randomly picked that sample in the area of radius as expressed by (6), (7), (8), and (9).

$$\theta = 2\pi a \tag{6}$$

$$r_l = r_d \sqrt{b} \tag{7}$$

$$x_l = r \times cos(\theta) + x_s \tag{8}$$

$$y_l = r \times sin(\theta) + y_s \tag{9}$$

where x_s and y_s are the coordinates of samples on the latent space that has been trained on VAE, a and b are the random numbers within determined range, r_d is the specify radius randomed within a circle of samples, and given x_l and y_l as a random coordinate result of the new generated data.

H. Signal Fragment Assembler

After generating data from latent space, we divided each sample into fragments, after that we shuffled and randomly concatenated the first and second half of those samples. In this way, we can combine two different characteristics into the new generated sample, n_s, as defined in (10).

$$n_s = i_s || j_s \tag{10}$$

where i_s and j_s are the first and second half of random generated samples, respectively, and $||$ is the concatenation operator.

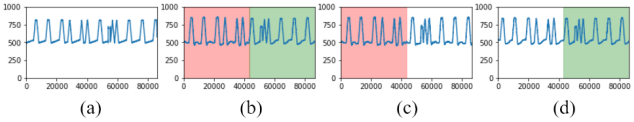

Fig. 2: An example of generated data using signal fragment assembler. (a) is original anomalous data, (b) is a sample of signal fragment assembler data after concatenation, (c) is left side fragment from random sample of generated data, and (d) right side fragment from random sample of generated data.

I. Quality Classifier

To ensure that our generated data look alike to the original dataset, yet not too similar, since that means generated data does not have generalization; therefore, we evaluate our samples that have been generated with histogram by measuring an overlapping area of the original time series and synthesized samples. The qualified generated data have to fall in an range of acceptable score of evaluation as defined in section IV-B.

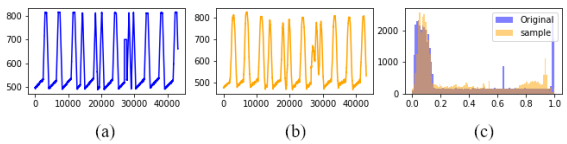

Fig. 3: An example of qualified generated data: (a) is the original data, (b) is a sample of generated data, and (c) is a histogram chart that shows an overlapping area between the original data and sample of generated data.

IV. Experiments and Results

A. Dataset Explanation

We used Secure Water Treatment (SWaT) dataset [12] which was collected samples from the test bed system. Those samples contained attack incidents, generated by researchers. In order to illustrate anomalous events, the actuator of motorized valve, that controls water flow to

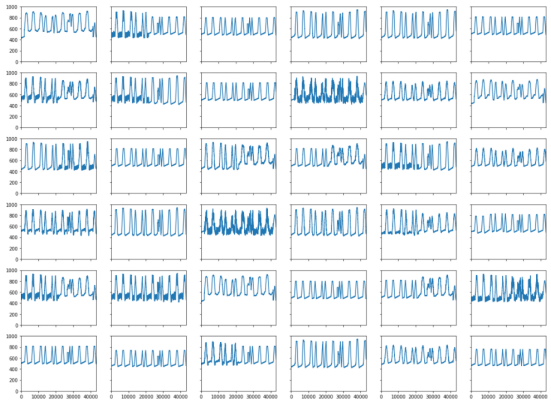

Fig. 4: Example of generated anomalous time series data.

the raw water tank, has been selected to apply in our framework, since it has obvious characteristics and easy to understand in human perception, when plotting into time series. The original data shows that, by keeping the motorized valve continuously, the raw water tank was overflowed.

B. Experimental Setup

Our DAGAT framework was implemented for experiments by setting up parameters as follows:

- Experiment was run on Windows 10 64-bits operating system, Intel(R) Core(TM) i7-8750H CPU @ 2.20GHz 2.21GHz, and RAM 8 GB.
- SWaT dataset with 43,200 records or 12 hours per sample. 500 samples with time window slicing.
- Variational Autoencoder had 4 layers. Encoder had 2 layers with 200 and 100 nodes, while decoder had 2 layers with 100 and 200 nodes. We divided the dataset into 50 % of training and 50% of testing, 100 epoches in training process. ReLU was an activation function and latent space was in 2 dimensions.
- Randomly choose a generated dataset and divide it into two parts in process of Signal Fragment Assembler.
- Randomly choose samples within 5 units of radius from randomly selected samples in latent space.
- Acceptable scores of quality classifier are in range of 75.00% - 95.00%

Noted that these parameters are adjustable since data characteristics are vary by its own nature. Moreover, each task requires different sensitivities of generalization of data. Generated samples can almost look like original dataset or contain a few attribute of augmented samples.

C. Experimental Results

Result illustrated in Fig. 4 showed that, from only one original attack incident with DAGAT framework, we can create up to 3,998 samples, with a variety of patterns and conform to characteristics of original anomalous events.

V. Conclusion

In this paper, we have proposed a novel framework, called DAGAT, for anomalous time series signal augmentation. The achievement of this framework is of applying vanilla augmentation methods on different domains to increase a number of samples. Four cascade procedures, VAE, data picker, signal fragment assembler, and quality classifier, screen good quality of augmented samples. The VAE plays a role in creating latent space representation after learning augmented signals, and then data picker explores the representation from that latent space to expand the interval of possible samples. By exploring latent space, we can generate samples in between different augmented types, deriving from sections III-A, III-B, III-C, III-D, and III-E. Next, signal fragment assembler breaks data into parts and reassemble them to multiple augmented characteristics in a single form. Lastly, quality classifier certifies quality of time series signals to prevent the duplication of original data by using statistical method to filter generated data before yielding a final anomaly augmented dataset.

In future work, we aim to extend DAGAT by reconstructing the flow of data generation framework and test our performance method with more open source anomalous datasets. We also have an assumption that a DAGAT framework might be able to implement on images and high dimensional data, that will be powerful for deep learning purposes, similar to time series domain that the current DAGAT does.

References

[1] Ying, Xue. “An overview of Overfitting and its solutions.” Journal of Physics: Conference Series. Vol. 1168. No. 2. IOP Publishing, 2019.

[2] A. Sakai, Y. Minoda and K. Morikawa, “Data augmentation methods for machine-learning-based classification of bio-signals,” 2017 10th Biomedical Engineering International Conference (BMEiCON), Hokkaido, 2017, pp. 1–4, doi: 10.1109/BMEiCON.2017.8229109.

[3] T. T. Um et al., “Data augmentation of wearable sensor data for parkinson’s disease monitoring using convolutional neural networks,” in Proceedings of the 19th ACM International Conference on Multimodal Interaction, ser. ICMI 2017. New York, NY, USA: ACM, 2017, pp. 216–220.

[4] Tariq M. Arif, Introduction to Deep Learning for Engineers: Using Python and Google Cloud Platform , Morgan & Claypool, 2020, doi: 10.2200/S01029ED1V01Y202007MEC028.

[5] S. M. A. Karim, N. Ranjan and D. Shah, “A Scalable Approach to Time Series Anomaly Detection & Failure Analysis for Industrial Systems,” 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2020, pp. 0678–0683, doi: 10.1109/CCWC47524.2020.9031262.

[6] F. Lüer, D. Mautz and C. Böhm, “Anomaly Detection in Time Series using Generative Adversarial Networks,” 2019 International Conference on Data Mining Workshops (ICDMW), Beijing, China, 2019, pp. 1047–1048, doi: 10.1109/ICDMW.2019.00152.

[7] P. Gogoi, D. K. Bhattacharyya, B. Borah and J. K. Kalita, “A Survey of Outlier Detection Methods in Network Anomaly Identification,” in The Computer Journal, vol. 54, no. 4, pp. 570–588, Apr. 2011, doi: 10.1093/comjnl/bxr026.

[8] A. Youssef, “Analysis and comparison of various image downsampling and upsampling methods,” Proceedings DCC ’98 Data Compression Conference (Cat. No.98TB100225), Snowbird, UT, USA, 1998, pp. 583–, doi: 10.1109/DCC.1998.672325.

[9] W. T. Cochran et al., “What is the fast Fourier transform?,” in Proceedings of the IEEE, vol. 55, no. 10, pp. 1664–1674, Oct. 1967, doi: 10.1109/PROC.1967.5957.

[10] V. Prema and K. U. Rao, ”Time series decomposition model for accurate wind speed forecast,”Renewables: Wind, Water, and Solar, vol. 2, p. 18, 2015.

[11] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” in 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings, 2014.

[12] Goh J., Adepu S., Junejo K. N., and Mathur A., “A Dataset to Support Research in the Design of Secure Water Treatment Systems,” The 11th International Conference on Critical Information Infrastructures Security.