Spacecraft anomaly detection via transformer reconstruction error

Hengyu Meng[1], Yuxuan Zhang[1], Yuanxiang Li[1*], Honeghua Zhao[2]

1.School of Aeronautics and Astronatics, Shanghai Jiao Tong University

2. Eastern Airlines Technic Co, China Eastern Airlines

{LULVHP, yuxuanzhang, yuanxli }@sjtu.edu.cn , hhzhao@ceair.com

Abstract— Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. Nowadays anomaly detection method deployed to production is based on reconstruction error generated by LSTM sequence modeling. Recently, the remarkable improvement achieved by BERT model in language translation demonstrated that the self- attention-based transformer is superior to LSTM models, due to its ignoring distance. In this paper, we continue the research on transformer and propose a transformer-based architecture, Masked Time Series Modeling, applying transformer in data stream, which has two novel components 1) the attention mechanism used for updating timestep in parallel, 2) the mask strategy used to detect the anomaly in advanced time. We compared the performances of our method with state-of-the-art AD methods on challenging public NASA telemetry dataset. The experiment results demonstrated our method saves about 80% time cost because of parallel computing compared with LSTM methods and achieves 0.78 F1 point-based score, moreover achieving a better score on range-based indicators.

Keywords—spacecraft, anomaly detection, transformer, mask, deep learning, attention mechanism

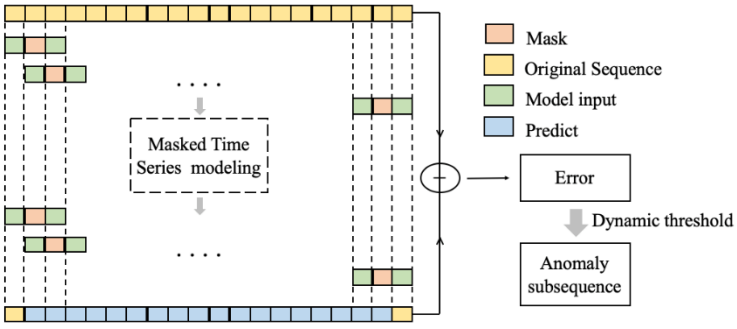

Fig 1. Framework of masked time series modeling. We intercept subsequences from the original sequence. The front and back parts of the subsequence are used as input to the model, and the middle part of the subsequence is output of reconstruct model. Detect anomalies based on the error between reconstructed sequence and the original.

Introduction

Anomaly detection (AD) is the process of identifying non-conforming items, events, or behaviors [1, 2]. Efficient detection of anomalies can be useful in many fields. Examples include quantitative transaction, threat detection for cyber-attacks [3, 4], or safety analysis for self-driving cars [5]. Many real-world anomalies can be detected due to promotion of various industrial sensors. Especially for in-orbit spacecraft, failure to detect hazards could cause serious or even irreparable damage since spacecrafts are expensive and complex system. In the absence of remedial measures, anomaly detection is important and necessary to warn operation engineer of anomalies.

Current anomaly detection methods for spacecraft telemetry primarily can be grouped into two types: expert-based and reconstruction-based. Expert based. Experts are required to define appropriate ranges for various monitoring indicators, or to provide clear definitions of anomalies. Clustering or density-based approaches has been implemented for a small number of spacecraft [6, 7]. However, as systems become more complex, these approaches are expensive and cannot handle increasing and complex anomalies.

Reconstruction-based anomaly detection is the most popular one and has been deployed into spacecraft [8, 9]. The main ideas of reconstruction-based anomaly detection methods are as follows: 1) What the "normal" sequence should look like, which means reconstructing sequence via RNN models trained by normal sequences. 2) Use the same model to reconstruct the sequences with anomalies and compare the reconstructed sequence with the input. 3) Set the error function and threshold. Where an abnormality occurs in the entire sequence, the reconstruction may be unideal, so classification or clustering can segment the anomalies [7, 10].

However, there exist two problems on reconstruction-based anomaly detection:1) LSTM depends on uni-directional sequential propagation, which makes the computation of LSTM low-effective. 2) one-direction propagation will delay the anomaly detection time due to the sparse unexpected data in the early observed anomaly subsequences.

Contributions. In this paper, we propose a masked time series modeling method based on transformer, as shown in Fig.1, which has two novel components 1) the

attention mechanism used for updating timestep in parallel, 2) the mask strategy used to detect the anomaly in advanced time. In this way, the reconstruction in the front of anomalies are able to be affected by the anomaly data, resulting in anomaly early detection. Once model reconstruction are generated, we apply a dynamic thresholding approach for evaluating reconstruction error. Experiments show that due to the characteristic of transformer encoder, the modeling greatly reduces time consumption without significant drop in accuracy compared with LSTM reconstruction. And using bi-directional data makes model to capture anomalies better for range-based precision and recall indicator. This work is tested on NASA spacecraft datasets, but can be applied to other anomaly detections task.

The structure of the article is as follows: Section 2 introduces the related work, the anomaly detection based on LSTM reconstruction and the transformer encoder generally used for NLP tasks. Section 3 presents our method, showing the inputs using contextual information, and gives the model reconstruction process. In Section 4, experiments were carried out on the NASA spacecraft dataset. The article compared the relationship between model consumption time and detection accuracy, and compared it with the previous methods on point-based and range-based indicators. Section 5 summarizes the full paper, puts forward the defects and shortcomings, and looks forward to the future work.

Related work

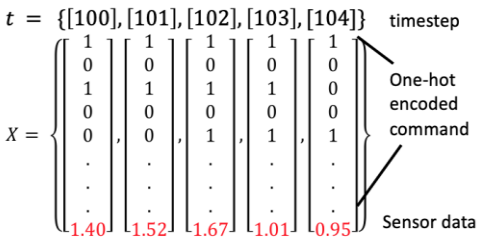

Fig 2. Each vector contains 2 parts: one-hot encoded command and continuous sensor data

Quite part of anomalies is caused by improper operations, especially in complex and stressful work of aerospace. Therefore, when considering abnormality detection, not only the sensor signal but also the operator's operation command should be considered. Since the one-hot coding representation (shown in Fig 2) of the operation command in the time dimension is discrete, which is similar to the discrete word vector in natural language processing (NLP), methods proven effective in NLP such as transformer can be considered.

Anomaly detection based on LSTM reconstruction

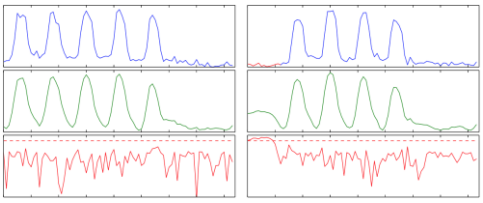

Fig. 3. Examples of anomalies detection via reconstruction

Fig.3 visually shows the anomaly detection based on LSTM reconstruction [11]. The researchers trained the reconstructed model using normal sequences. When a model reconstructs a sequence with anomalies (blue lines in the first column), there is a difference between the input and output (green lines in the second column) sequences. Anomalous subsequences can be segmented applying a simple threshold segmentation or probability distribution model on reconstructed error (red line in the third column).

For some anomaly detection studies, the size limit of the data set is an inevitable topic. For other reconstruction tasks, such as GAN based image generation, at least thousands of images are often required. However, some anomaly detection data sets often have only tens of hundreds of data, some specific anomalies even only have a little labelled data. And as the sampling density increases, length of the sequence also increases, making it increasingly difficult to generate the entire sequence.

Some anomaly detection researchers use time series local reconstruction to train

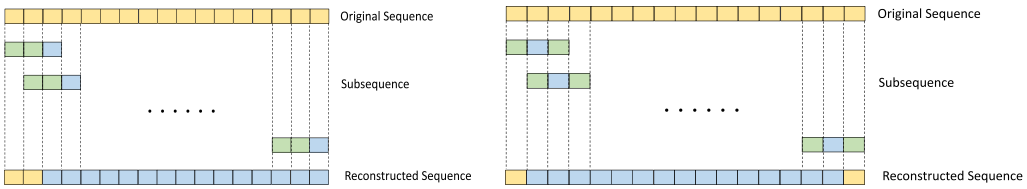

Fig. 4. Data interception process comparison. The left shows common way :The model can use the previous 180 timesteps as input and the tail 20 as output. The right is ours.

models. In details, researchers intercept fixed-length subsequences from the time series data, using the front part as the input of the time series model and the tail part as the prediction object. This process is shown in Fig.4.

Currently LSTM (Long short-term memory) reconstruction error based system instead of costly expert system has successfully identified several confirmed

anomalies since deployed to the Soil Moisture Active Passive satellite (SMAP) and the Mars Science Laboratory rover (MSL), Curiosity.

Transformer encoder

Self-attention is a special attention mechanism. In self-attention, query is equal to key equal to value.

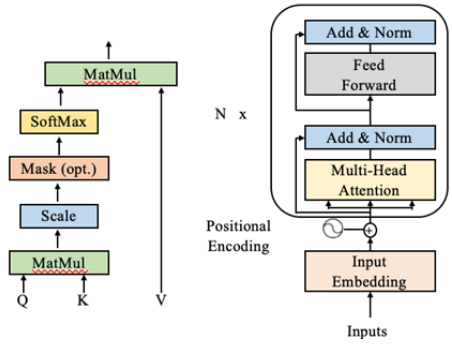

Fig. 5. Dot-Product Attention(left) and Transformer Encoder(right)

Transformer, a self-attention mechanism that learns contextual relations between words (or sub-words) in a text, was proposed by [12]. As shown in Fig.5, transformer includes two separate mechanisms—an encoder that reads the text input and a decoder that produces a prediction for the task. Since transformer based BERT model has dominated effect in language modeling task, which shows transformer's strong ability to extract features.

As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left) [13], the Transformer encoder reads the entire sequence of words at once. Therefore, it can be considered bidirectional, though it would be more accurate to say that it's non-directional. This characteristic allows the model t learn the meaning of a word based on its surroundings (left and right of the word). For deep transformer based network, researchers propose two pre-train task, one of which is masked language modeling. The main process of masked Bi-directional language model: randomly select 15% of the words in the corpus, and mask it, which means replacing the original word with the [Mask] mask, and then train the model to correctly predict the word that was discarded.

Method：masked time series modeling

In order to solve the shortcomings of the LSTM network, Transformer encoder is applied to the spacecraft time series data modeling. Inspired by the masked

language model, we propose a method using masked state prediction to reconstruct the time series.

Model input and Mask

Since LSTM can only propagate in one direction, the reconstruction process is actually more similar to prediction: predicting current data from historical data, and not using future data. In fact, from other time series tasks, the comprehensive use of information in both directions will greatly enhance the capabilities of the model. The input of masked time series model is intercepted from sequences like Fig.4. The subsequence of this model input includes past and future information compared to the input of the LSTM model. In other word, in LSTM-based reconstruction, L_i0 is the length of whole input, and L_i1 is 0. In our method, the two are greater than 0(usually equal to each other). So there exits problems about online

Table1. Notation

Notation	Description
L_{i0}, L_{i1}	Length of input subsequences in front of and behind model output
L_o	Length of reconstructed subsequences/model output
R, R_i, N_r	Set of annotated subsequences, i^{th} annotated subsequence, Total number of points in R
P, P_j, N_p	Set of predicted subsequences, i^{th} predicted subsequence, Total number of points in P
α, z	Weight of existence reward, Weight of standard deviation
$\gamma(.), \omega(.), \delta(.)$	overlap cardinality function, overlap size function, positional bias function
w_i, O_i, w_i'	Original timestamp, Output of transformer encoder, Reconstructed timestamp
$e, e_i, \varepsilon, \varepsilon_i$	Error sequence and its i^{th} timestamp; Threshold sequence and its i^{th} timestamps

Mask operation is an operation that blocks specific timesteps from the network. Specifically, in the transformer encoder, the score of the masked timesteps in the relation matrix is set to infinity so that it can only accept the reconstruction via the remaining time steps.

This article discusses the location and length for the mask operation. In fact, there are three variables that determine the mask, which are the length of time series before the mask, the length of time series after the mask, and the length of the mask itself. When the post length is 0, the model degenerates to be the same as LSTM,

and only uses uni-directional information for prediction. Different mask lengths also have an impact on accuracy as shown in Section IV.

Reconstruction Model structure

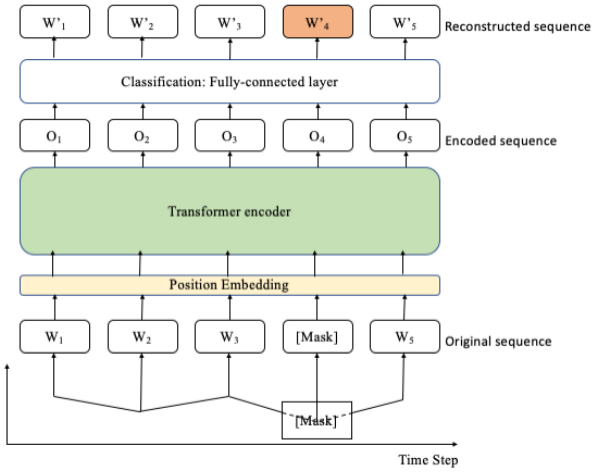

Fig. 6. Process of masked time series reconstruction

The model structure of masked time series modeling is shown in Fig.6. The input to the model is the normalized raw signal w_i. After masking the middle part which will be reconstructed, the position encoding layer is added to avoid self-attention mechanism ignores the position information. Next is the transformer encoder. After obtaining the features O_i extracted by the transformer encoder, the fully connected layer is used to obtain the final prediction of the masked state w_i'.

The model itself is similar to the masked language modeling. Here in this task, the input and output are continuous vectors. So we use regression instead of classification in the top layers. Moreover, so researchers use multi-head mechanism to map word vectors to different subspaces for parallel computation. But in the data set we use, the signal has only 25 or 55 dimensions, so using too many heads makes no difference. The experiments also verify this.

Anomaly detection via non-dynamic threshold [8]

After obtaining the reconstructed sequencew_i', we can calculate the reconstruction error e_i compared with the original sequence. Then we segment anomalies via dynamic threshold $\boldsymbol{\varepsilon}$ which is defined as: $\varepsilon = \mu(e_i) + z\sigma(e_i)$

where ε is determined by:

$$\varepsilon_i = argmax(\varepsilon) = \frac{\Delta\mu(e_i)/\mu(e_i) + \Delta\sigma(e_i)/\sigma(e_i)}{|e_a| + |P_j|^2}$$

Such that (μ is expectation and σ is standard deviation):

$$\Delta\mu(e_i) = \mu(e_i) - \mu(\{e_i \in e_i | e_i < \varepsilon\})$$

$$\Delta\sigma(e_i) = \sigma(e_i) - \sigma(\{e_i \in e_i | e_i < \varepsilon\})$$

$$e_a = \{e_i \in e_i | e_i < \varepsilon\} \quad P_j = sequence\ of\ e_a$$

Experiments

This section demonstrates the process and result of experiments. First we introduce open source datasets provided by NASA. Then we give comparison of the two indicators: point-based F1 score and time consumption with the traditional methods. Then, based on the shortcomings of the traditional accuracy indicator, we compare results on the range-based indicator.

Datasets

This article uses real-world, expert labeled data derived from Incident Surprise, Anomaly (ISA) reports for the Mars Science Laboratory (MSL) rover, Curiosity, and the Soil Moisture Active Passive (SMAP) satellite [8].

The dataset contains 82 normalized telemetries channels of data, which means we have to train 82 models—one model for one channel, and includes 105 anomalies.

Point-based precision and recall.

After obtaining the reconstruction sequence of the model output, we use the dynamic threshold method to segment the subsequence of the anomalies. The anomaly subsequence is compared with the annotated data, and the point-based precision and recall is defined as follows:

- True positive is defined as for any R_i in R ,there exists P_j that the intersection of R_i and P_j is not none. In other words, R_i overlaps with P_j . If there exist several P_j overlapping with R_i, True positive set only count once.
- False negative means when there is no P_j having an intersection with R_i, this R_i is ignored by our model so FN(false negative)+1.
- False positive means if a P_j don't overlap with any R_i, we count it as FP.

The effect of different hyper parameters on the experimental results is TABLE II.

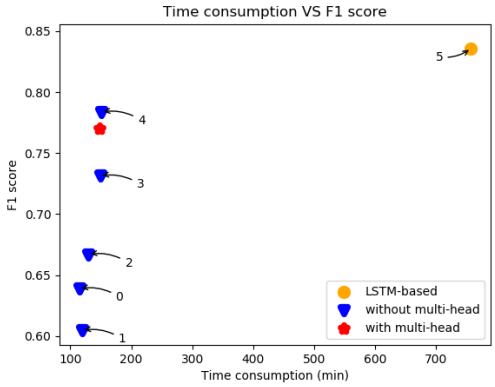

Fig. 8. Time consumption VS accurcy. The annotation numbers near the scatters are the experimental serial number. And the red one shows transformer with multi-head don't work better.

We also explore the balance between accuracy and time performance and whether multi-head mechanism would affect results. Fig 8 shows our method consume less time during training. At the same time, for the point-based precision and recall indicators, our method reaches the almost same precision and recall as the state of the art. And multi-head mechanism makes no sense in our task since the dimensions of time series is not complicated enough.

Table II. Experiments results comparison between different hyper parameters

Serial number	0	1	2	3	4	5
Stack of transformer	6	6	8	12	12	LSTM-based
Length of front input	100	200	100	100	100	200
Length of post input	100	0	100	100	100	0
Length of output	5	5	5	5	1	1
Precision	64.9%	62.0%	68.7%	78.3%	85.4%	87.5%
Recall	62.9%	59.0%	64.8%	68.6%	72.4%	80%
Time consumption(min)	114.3	118.8	129.3	138.5	150.7	757.3

Range-based precision and recall

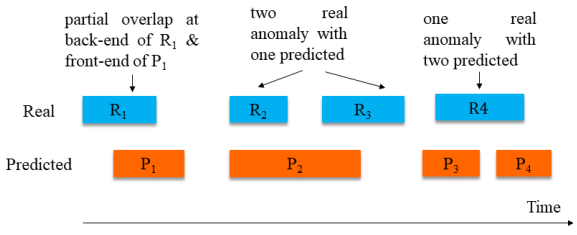

Fig.8 The three cases are all counted as true positive, however the first case detects anomalies delayed from the true value, the second case two anomalies are identified as one, and the third case counts one as two. Physical meanings are different

[14] presented a new mathematical model called range-based Precision and Recall metrics to evaluate the accuracy of time series classification algorithms. They expand the well-known Precision and Recall metrics to measure ranges. Fig.8 visualizing motivation of range-based.

$$Recall(R_i, P) = \alpha \times ExistenceReward(R_i, P) + (1-\alpha) \times OverlapReward(R_i, P)$$

The definition of existence reward is the approximately same as defined in point-based while overlap reward makes the differences. Overlap reward depends on three functions $\gamma(), \omega(), \delta()$, each of which captures the cardinality, overlap range size and position bias of overlap

$$OverlapReward(R_i, P) = [\sum_{j=1}^{N_p} \omega(R_i, R_i \cap P_j, P_j)] \times CardinalityFactor(R_i, P)$$

$$CardinalityFactor(R_i, P) = \begin{cases} 1 & if\ R_i\ overlap\ with\ at\ most\ one\ P_j \\ \gamma(R_i, P) & otherwise \end{cases}$$

Table III. Results evaluated on range-based indicator

method	Our method	LSTM-based
precision	0.64	0.62
recall	0.50	0.49

In this experiment, α is set to 0.5 because we want to pay equal attention to ExistenceReward and OverlapReward. $\gamma(.)$ is set to the reciprocal of the number of overlaps. $\delta(.)$ is set to front-end bias since we want to detect anomalies as early as possible. And range-based precision only consists OverLapReward. The parameters are set to the same in range-based recall.

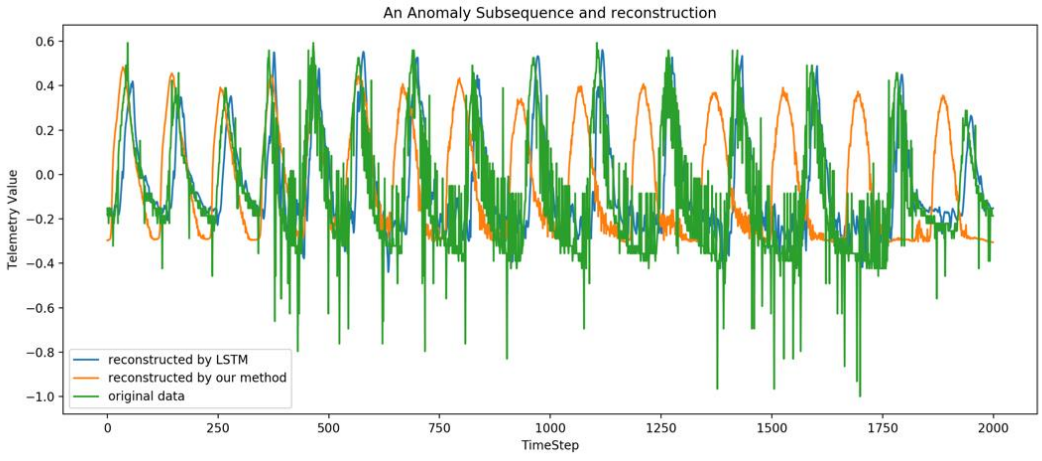

Fig. 9. Comparison of reconstruction results. This is one of anomaly subsequence.

Even for point-range recall, our method is slightly worse than the baseline. But on range-based recall, our approach is slightly better as shown in Table III. Fig.9 is a

comparison of the reconstructed sequence of one of the channels. The sequence reconstructed by our method has smaller gradients and less fluctuations due to the use of bidirectional information. On the one hand, the sequence reconstructed by our method reflects the abnormality of the input signal more quickly; on the other hand, the smaller the fluctuation means that it is not easy to be recognized as two anomalies [15].

Summary: according to the experimental results, our method has greatly reduced the consumption time while the accuracy has not dropped significantly. At the same time, our approach has slightly improved the range-based indicator due to the use contextual information.

Conclusion

LSTM-based anomaly detection method has been deployed into production currently. This paper proposes a reconstruction algorithm for time series data anomaly detection - masked time series modeling. This algorithm is based on the transformer model and is significantly faster than RNN. At the same time, the advantages of using both front and back information make it better highlight the abnormal point. The experimental results verify our vision.

In the future, we will continue to explore how contextual information reacts to anomalies. In particular, visualize the relationship matrix in the self-attention mechanism to explore the relationship of anomalous points in the full sequence. On this basis, the end-to-end algorithm will be developed to better detect outliers on range-based indicators.

Acknowledgment

Thank NASA for making a so cool dataset open-source.

References

1. C. C. Aggarwal. Outlier Analysis. Springer, 2013.
2. V. Chandola, A. Banerjee, and V. Kumar. Anomaly Detection: A Survey. ACM Computing Surveys, 41(3): 15:1–15:58, 2009.

3. AlEroud and G. Karabatis. A Contextual Anomaly Detection Approach to Discover Zero-Day Attacks. In International Conference on CyberSecurity (ICCS), pages 40–45, 2012.
4. V. Chandola, V. Mithal, and V. Kumar. Comparative Evaluation of Anomaly Detection Techniques for Sequence Data. In IEEE International Conference on Data Mining (ICDM), pages 743–748, 2008.
5. O. Anava, E. Hazan, and A. Zeevi. Online Time Series Prediction with Missing Data. In International Conference on Machine Learning (ICML), pages 2191–2199, 2015.
6. B. Zong, Q. Song, Min M R. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. 2018.
7. M. Sölch, J. Bayer, M. Ludersdorfer. Variational inference for on-line anomaly detection in high-dimensional time series. arXiv preprint arXiv:1602.07109, 2016.
8. H. Kyle,V. Constantinou,C. Laporte, I. Colwell, I. Soderstrom. Detecting Spacecraft Anomalies Using LSTMs and Nonparametric Dynamic Thresholding. 2018.
9. F, T. Li and D. Chana. Multi-level Anomaly Detection in Industrial Control Systems via Package Signatures and LSTM Networks. 2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Denver, CO, 2017, pp. 261-272.
10. Z. Mingyi, Ye. Kejiang, W. Yang, X. Cheng-Zhong. A Deep Learning Approach for Network Anomaly Detection Based on AMF-LSTM. 15th IFIP International Conference on Network and Parallel Computing (NPC'2018), pp.137–141, 2018.
11. P. Malhotra, L. Vig, G. Shroff, P Agarwal. Long short term memory networks for anomaly detection in time series. Proceedings. Presses universitaires de Louvain, 2015.
12. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, et al. Attention is all you need. In Advances in Neural Information Processing Systems(pp. 5998-6008). 2017.
13. K.Tae-Young, Cho.Sung-Bae. Web traffic anomaly detection using C-LSTM neural networks. Expert Systems with Applications,Volume 106,2018,Pages 66-76.N. Tatbul, T.J. Lee, S. Zdonik. M. Alam. J. Gottschlich. Precision and Recall for Time Series. Neural information processing systems (NIPS) 2018.
14. L,.Qi, K.Rudy, C.Chao. Unsupervised detection of contextual anomaly in remotely sensed data. Remote Sensing of Environment, 2017, 202: 75-87.