IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 16, 2022, accepted May 23, 2022, date of publication May 27, 2022, date of current version June 3, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3178373

A Novel Outlier Detection Model for Vibration Signals Using Transformer Networks

RUIHENG ZHANG, QUAN ZHOU, LULU TIAN, LIBING BAI, (Associate Member, IEEE), AND JIE ZHANG

School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China

Corresponding author: Quan Zhou (quanzhou@uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant U2030205, Grant U1830207, Grant 62003075, Grant 61903065, Grant 62003074 and Grant U1830133.

ABSTRACT Outlier detection in vibration signals can play an important role in addressing the issue of structural or environmental changes during vibration testing. In this study, a transformer-based model for outlier detection is proposed. Unlike previous statistical and regression outlier detection methods, the proposed model can identify the outlier location in a high dimensional observation space using the self-attention mechanism. The location of outliers within the vibration observation is marked by a combination of a spatial label and a temporal label. The outlier detection performance of the model is verified by a numerical study of the plane wave and an experimental study of the vibrating plate. These two studies show that the proposed model has good label prediction accuracies (all above 85%) toward the outlier location within the plane wave and vibrating plate observations.

INDEX TERMS Outlier detection, transformer network, vibration testing.

I. INTRODUCTION

Outlier analysis of vibration signals has been studied by many researchers to identify novel data caused by environmental or structural variability [1]–[3]. The measured data can deviate from its normal condition when the vibration response carries information about the structural changes or the environmental factor such as temperature and load. Both types of deviation need to be handled very carefully, lest they lead to false alarms. Therefore, analysis of structural-induced or environmental-induced outliers is of interest because it facilitates valid variation detection in measured responses before using any further data processing techniques for revealing the structural condition.

Detecting outliers in multivariate vibration observations often proves to be more difficult than in univariate data because of the additional dimensionality [4]. Several attempts have been made to extract outliers concealed in the vibration observation. The most commonly used outlier detection method is the Mahalanobis squared-distance (MSD) [5]–[7] which characterises the normal vibration observation as a mean vector and a covariance matrix. A discordance test follows to evaluate whether a new observation has outliers. Despite the accessibility of MSD, there remains a paucity of evidence on its performance over inclusive data (the data with outliers). Furthermore, [8] proposed minimum volume enclosing ellipsoid (MVEE) and minimum covariance determinant (MCD) method to improve the detection of the inclusive outlier. Another outlier detection method is dimension reduction [9]. Typically, principal component analysis (PCA) [10] is used to retain information related to outliers. By substituting a group of correlated variables into a new smaller group of principal components, PCA can find the component relevant to the variability caused by outliers. In addition, the regression method also achieved notable results in outlier analysis. Unlike the statistical or dimension reduction method, the regression method [3] focuses on predicting the next time step of the measured response. By discriminating outliers from the reference regression model, the regression method is advantageous for online monitoring. However, the above conventional methods have some disadvantages in detecting outliers in high dimensional space. For example, in statistical method, the masking or swamping effects [11] due to the dominant

The associate editor coordinating the review of this manuscript and approving it for publication was Qichun Zhang.

normal component of the high dimensional data can make the variation contributed by outliers invisible. Moreover, if observations that represent the normal condition are inconsistent, they will become dispersed across the feature space. As a result, dimension reduction techniques like PCA may be infeasible for outlier detection due to the masking effects caused by normal variation components. Regarding regression methods, few studies have been able to draw on any systematic research into the correlation of data points within the high dimensional observation, which may carry the high dimensional feature related to the outlier. In short, detecting outliers becomes challenging as the dimensionality of the observation space increases.

More recently, artificial neural network (ANN) [12] is utilized for outlier detection on account of its nonlinear approximation capability. ANN can approximate nonlinear features or classify groups of features divided by nonlinear boundaries. Multilayer perceptron, convolutional and recurrent neural networks (MLP, CNN, RNN) [13]–[15] are the most popular ANNs for outlier detection. [16] proposed a CNN-based model to identify or eliminate abnormal data. The CNN is used to extract temporal features in the vibration time series for abnormal data classification. But the outliers discussed in this paper fairly exceed the mean and variance of the normal vibration observations. Accordingly, the outlier approximation potential of the model in this paper is not fully investigated. [7] proposed an RNN model with long short-term memory (LSTM) cells to approximate the Mahalanobis distances of normal conditions. By subtracting the predicted distances from that of the measured observations, one can monitor the variation caused by outliers. However, the approximation performance of this model is limited by the statistical distance metrics it applies. So far, there has been little discussion about exploiting ANN capacity for locating outliers. CNN uses convolution windows or filters to transform data into feature maps and RNN relies on recurrent cells for sequential feature extraction. CNN has shown state-of-the-art performance in local feature extraction but remains highly sensitive to adversarial noise. For outlier analysis, this means that the CNN-based model has weak robustness against inconsistent normal conditions. Furthermore, RNN has been firmly established as the dominant approach in sequence modelling and prediction. The sequence processing mechanism of RNN based model is inherently sensitive to the input sequence order, which makes the generalization task of outliers at random sequential positions difficult to achieve. Although ANN-based model has achieved significant improvements in approximation capability for outlier features, the challenge of outlier detection in high dimensional space and the fundamental constraint of CNN and RNN architecture remains.

Transformer architectures, in recent work, have demonstrated impressive performance in the fields of natural language processing and computer vision [17], [18]. This type of architecture relies entirely on an attention mechanism to draw global dependencies between input and output.

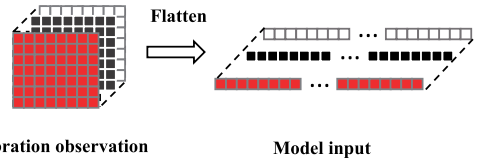

FIGURE 1. The input of the transformer network.

Previous research has shown that transformer architecture is highly robust to severe occlusions, perturbations, and domain shifts in images [19]. In terms of outlier analysis, the transformer network is considered as the promising ANN candidate for outlier detection in high dimensional space. It has the potential of revealing the location of outliers within a vibration observation with the help of positional embedding and attention mechanism. In this study, we tried to address the challenges of locating outliers in high dimensional space through a transformer-based machine learning model. The proposed model shows notable outlier detection performance in both a numerical study of plane wave propagation and an experimental study of a vibrating plate. The main contributions of this paper are:

1. A novel transformer-based model for outlier detection is proposed.

2. A multi-output layer is implemented in the proposed model to smooth the outlier location labelling in high dimensional space and exploit the learning capacity of the transformer.

3. Numerical and experimental studies are presented to showcase the performance of the proposed model on outlier detection.

The remainder of this paper is organized as follows. Section II introduces the fundamentals of the transformer architecture. In Section III, the outlier labelling and simulation process, as well as model training and evaluation are described. The numerical and experimental studies of the proposed model are presented in Section IV. Finally, the conclusions are given in Section V.

II. RELATED WORK

In this study, a machine learning model based on a transformer architecture is proposed for outlier detection in high dimensional space. Specifically, the proposed model uses the transformer encoder to replace conventional outlier detection procedures to directly identify outlier features from vibration observations. Assuming the input of the transformer network consists of patches of flattened representations of the vibration observation at each time step (Fig. 1). The input is first processed by adding position tokens to flattened vectors of each time frame using the positional embedding layer. The embedded input is then fed into the transformer encoder. Finally, the features extracted by the encoder are processed by a multi-output classification layer to perform outlier identification. The overview of the transformer network is shown in Fig. 2.

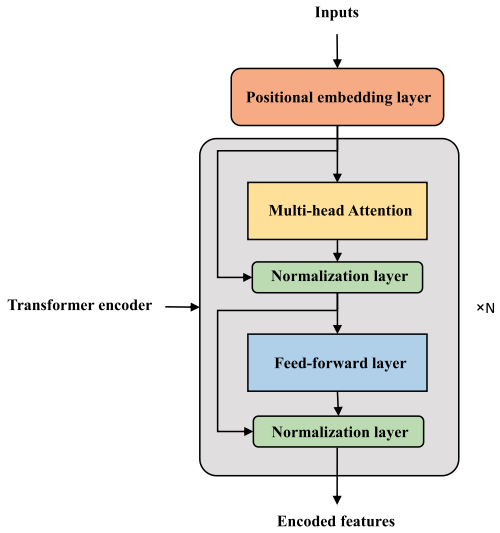

FIGURE 2. The overview of the transformer network.

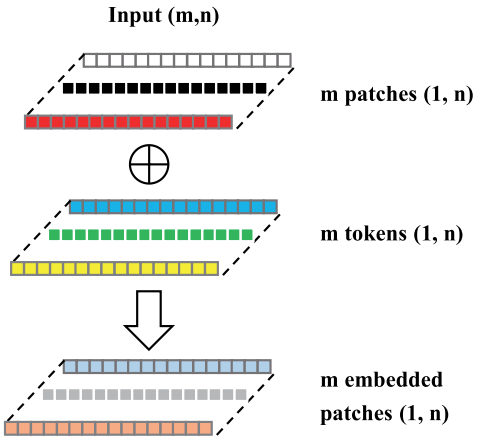

FIGURE 3. The positional embedding process.

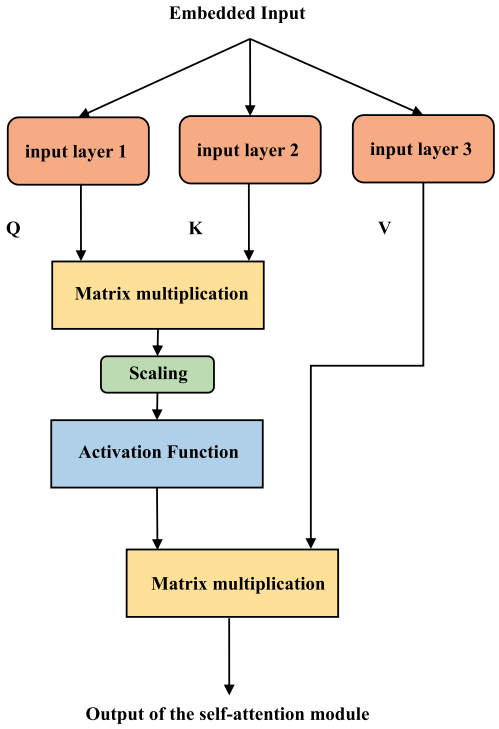

FIGURE 4. The architecture of the self-attention module.

A. POSITIONAL EMBEDDING LAYER

The positional embedding layer plays an important role in retaining the positional information of the input. In this study, the learned 1-dimensional positional embedding method is adopted, which means the weights of the embedding layer are trainable and the dimensionality of embeddings added to the initial vector representation is 1-dimensional. The shift of the vector representation provides necessary information for the following transformer encoder to identify the order of input patches representing different time steps. Details of the embedding process are shown in Fig. 3 and the number and size of embeddings are determined by the number of flattened patches and the size of each patch. For input of size (m, n), the embedding layer would generate m unique positional tokens of size $(1, n)$ where m is the number of time steps of the vibration observation and n is the flattened patch size of each time step. The weights of tokens are updated by backpropagation [20] during model training. More details about positional embedding or encoding can be found in [21], [22].

B. TRANSFORMER ENCODER

In this section, we discuss the fundamentals of the transformer encoder consisting of a self-attention module, a feed-forward layer, normalization layers, and residual connections [23]. The self-attention module (Fig. 4) has three input layers, namely, the query, key, and value layer. These three layers are linear layers that project each group of embedded vectors into the query, key, and value matrix respectively. The weights of each input layer are updated independently and the projection process can be formulated as:

$$Q = XW_q \tag{1}$$
$$K = XW_k \tag{2}$$
$$V = XW_v \tag{3}$$

where $X \in R^{m\times n}$ is the input for all three input layers of the self-attention module, $Q, K, V \in R^{m\times l}$ are the query, key, and value matrix, $W_q, W_k, W_v \in R^{n\times l}$ are the query, key, and value projection weights for X. Usually, the size $m \times l$ would be smaller than $m \times n$ to reduce the computation cost. Moreover, an activation function is applied to the scaled dot product of Q and K to obtain weights on V. The output of the self-attention module is computed by:

$$Att(Q, K, V) = softmax(\frac{QK^T}{\sqrt{ml}})V \tag{4}$$

where Att is the self-attention function and *softmax* is the activation function [24]. In addition, the multi-head self-attention mechanism is implemented by concatenating outputs of several self-attention modules, which can be expressed as:

$$MH(Q, K, V) = W_m C(Att_1, Att_2, \ldots, Att_h) \tag{5}$$

where MH is the multi-head self-attention function with an output of size $hm \times l$, C refers to concatenate, h is the number of heads, and $W_m \in R^{m \times hm}$ is the projection matrix for the concatenated outputs of self-attention functions.

In terms of the feed-forward layer, it provides similar projection functionalities as the input layer in the self-attention module. The major difference between the input layer and the feed-forward layer is that the input layer has no activation function applied while the feed-forward network has ReLU [25] activation functions in its hidden layer. More specifically, the feed-forward network is fully connected network and consists of projection layers with several hidden layers in between. Furthermore, there are two residual connections (also known as shortcut connections). One connection is between the input layer and the normalization layer after the self-attention module, and the other connection is between the normalization layer after the self-attention module and the normalization layer after the feed-forward layer. These residual connections are implemented to optimize the mapping process of both the self-attention module and the feed-forward network.

III. METHODOLOGY

A. ARCHITECTURE OF THE PROPOSED MODEL

In this study, a transformer-based supervised learning model is designed for outlier detection in high dimensional vibration observations. The architecture of the proposed model is shown in Fig 5. It mainly consists of two parts: a transformer encoder and a multi-branch output layer. Among these, the transformer encoder performs location feature extraction of outliers, and the multi-branch output layer performs the classification of both the spatial location and the temporal location of the detected outlier in the high dimensional vibration measurement.

B. TRAINING LABEL

According to [26], previous works on object detection usually takes a classifier for the target and evaluates it at various locations and scales in the observation space. By sliding the classification window in the observation space or using the divide and conquer strategy to decompose the observation space [27], [28], these detection methods can accomplish the object detection task at the cost of time and optimization difficulty. Conversely, you only look once (YOLO) system reframes object detection as a single regression problem to speed up the detection process. YOLO divides the 2-dimensional input data into grids and predicts the bounding box of each grid as well as the existence of the target in that box. The label for the object in one grid is a combination of the bounding box label with 5 predictions and the conditional class label with 1 prediction. Although YOLO has advantages in model training speed and can reasons globally about the input data when making predictions, the labelling process for training and testing samples is tedious.

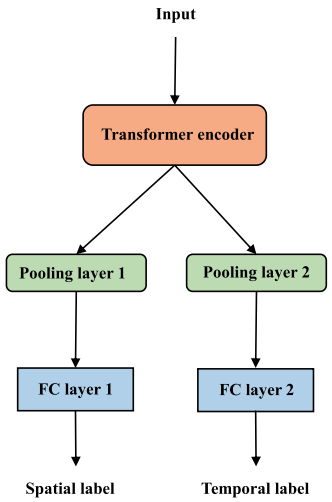

FIGURE 5. The architecture of the proposed model.

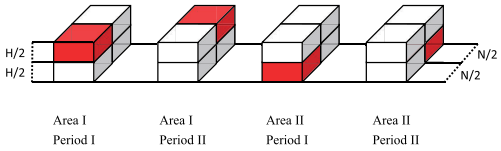

FIGURE 6. The labelling process.

Inspired by the above works, this study proposed an efficient labelling method for outlier detection. The desired output of the proposed model is the spatial and temporal labels of outliers within vibration observations. In this study, it is assumed that outliers occur in certain areas of the testing points and at a certain period of time throughout one observation. One high dimensional vibration observation with outliers can be represented as an $H \times W \times N$ tensor. For illustration, the vibration observation is divided into 2×2 areas of size $\frac{H}{2} \times W \times \frac{N}{2}$ as shown in Fig 6. Consequently, the related spatial and temporal labels according to one-hot labelling would be: test point area I($[1, 0]$) and time period I($[1, 0]$); test point area I($[1, 0]$) and time period II($[0, 1]$); test point area II($[0, 1]$) and time period I($[1, 0]$); test point area II($[0, 1]$) and time period II($[0, 1]$).

C. SIMULATION OF OUTLIERS

Previous studies mostly defined an outlier as a data point that falls far away from the overall points from a statistical point of view. However, in this study, the outlier is specified as the data deviate from its normal conditions, namely the sampling point strays away from the vibrating behaviour. A deviation ratio is implemented to introduce outliers in the labelled

area or subset of high dimensional vibration measurement. This is achieved by multiplying a percentage of randomly selected vibration vectors in the labelled area by the following deviation ratio:

$$\mathbf{g} \sim U\{0, 3\} \tag{6}$$

$$r_i = s * (g_i + 1) \tag{7}$$

where $s = 0.1$ is the scaling constant, r_i is the ith element of $\mathbf{r}$ and the deviation ratio for the ith element of the selected vibration vector, and $\mathbf{g}$ is the probability vector of the same length as $\mathbf{r}$ following the discrete uniform distribution.

D. MULTI-OUTPUT LAYER IN THE MODEL

The proposed model has two output branches after the transformer encoder (Fig 5). One is for spatial label learning, and the other is for temporal label learning. This multi-output architecture forces the transformer encoder above to maintain both spatial and temporal information in the encoded feature. During the training process, pooling layers are employed on these two branches to extract the spatial and temporal location of outliers respectively from the encoded feature produced by the transformer encoder. There are, of course, many other pooling configurations for the multi-output layer (e.g., max-pooling/average-pooling, average-pooling/max-pooling). Validating these configurations is essential to quantify how the pooling configuration of the multi-output layer influences the model prediction. In section IV, a detailed discussion about the configuration of the multi-output layer is presented. After the pooling layer in each branch, the learned feature from prior layers is flattened into a 1-dimensional tensor and passed into a fully-connected classification layer for the final prediction. The architecture of FC layers in the two branches is two stacked dense layers. The upper layer has ten neurons and the lower has four neurons.

E. LOSS FUNCTION IN THE MODEL

Both output branches utilize the categorical crossentropy [29] function as the loss function, which can be formulated as:

$$L_s = -\frac{1}{n}\sum(g_s \ln(p_s) + (1 - g_s)\ln(1 - p_s)) \tag{8}$$

$$L_t = -\frac{1}{n}\sum(g_t \ln(p_t) + (1 - g_t)\ln(1 - p_t)) \tag{9}$$

where L_s and L_t are losses of the spatial label branch and the temporal label branch respectively, n represents the number of training samples, g_s and g_t represent the ground truth spatial and temporal location of outliers within the training sample, p_s is the spatial label branch prediction of the sample and p_t is the temporal label branch prediction of the sample. In addition, two output branches have the same loss weights, which means the contribution of L_s and L_t to the loss of the model is balanced.

F. TRAINING AND EVALUATION

According to the area division of the input data, as described in section III-B, a possible outlier location in high-dimensional space is represented by a combination of a spatial label and a temporal label. Samples with simulated outliers are fed into the proposed model for training and a fraction of samples is used as validation data, which would be used for the evaluation at the end of each epoch. By monitoring the evaluation result of each epoch, the model that reaches the best evaluation performance during training is selected as the best model.

The evaluation metric for the proposed model is categorical accuracy. This metric computes the frequency with which the ground truth of the input matches the predicted label pair or probability pair. If the index of a maximal ground truth value is equal to the index of a maximal predicted value, it is counted as a successful prediction for the model being evaluated. In order to evaluate the proposed multi-output model, the categorical accuracy metric is applied to both output branches and the categorical accuracy of a single branch represents the performance of the corresponding outlier detection task. Additionally, the optimizer used for model training is adaptive moment estimation (Adam) and the loss weights of the two branches are 0.5 and 0.5.

IV. RESULTS AND DISCUSSION

The proposed model is validated with a numerical study and an experimental study: the 2-dimensional plane wave and the plate structure. The vibration data from both studies can be represented in 3-dimensional form with two spatial dimensions and one temporal dimension. Moreover, there is a notable difference between the vibration pattern of a plane wave and a plate, which helps to verify whether the proposed model is capable of detecting outliers within different vibration patterns. The plane wave vibration involves no shear force and its vibrational behaviour is predictable. Conversely, the composite plate has nonlinear characteristics (the discontinuity of mass) and can not be described using the analytic method. Tensorflow is used in the implementation of the proposed model. The detailed software environment is as follows: Tensorflow-gpu 1.14.0, CUDA 10.0, cuDNN 7.4, Keras 2.2.5, Python 3.7.3.

A. OUTLIER DETECTION IN PLANE WAVE

The governing equation of the 2-dimensional plane wave [30] is

$$F(x, y, t) = A(\omega_0)e^{i(k_x x + k_y y - \omega_0 t)} \tag{10}$$

where $F(x, y, t)$ is the value of the plane wave field at time t and location (x, y), $A(\omega_0) = 1$ is the amplitude of the wave at frequency ω_0, $k_x = 1$ is the wave number along the x axis, and $k_y = 1$ is the wave number along the y axis. F within the observation space of size 20×20 is calculated in the range of 0 to 1s at a sampling frequency of 20Hz. Consequently, the observation of the plane wave has a

TABLE 1. The categorical accuracy of the model for plane wave outlier detection using different pooling configurations (S/T accuracy: the categorical accuracy of the spatial/temporal label branch; MAX: max-pooling; AVG: average-pooling).

Pooling Configuration	S Accuracy	T Accuracy
MAX/AVG	84.3%	77.5%
MAX/MAX	73.7%	71.8%
AVG/MAX	64.3%	81.2%
AVG/AVG	51.2%	71.4%

TABLE 2. The categorical accuracy of the MAX/AVG model for plane wave outlier detection using different self-attention head numbers (S/T accuracy: the categorical accuracy of the spatial/temporal label branch).

Head Numbers	S Accuracy	T Accuracy
1	84.3%	77.5%
4	88.7%	84.3%
6	85.6%	93.1%
8	76.8%	78.7%

size of 20 × 20 × 20. This vibration observation was divided into 4 × 4 areas as described in section III-B for outlier location labelling. The vibration sequences within labelled area are modified by the deviation ratio mentioned in section III-C. With outlier simulation, 800 plane wave samples were generated for model training and 160 samples were generated for model evaluation. Both training and evaluation datasets have balanced outlier locations (50/10 samples for each possible location in the training/evaluation dataset)

The performances of the proposed model under different pooling configurations are compared. Table 1 provides the corresponding prediction result of each configuration and the transformer encoder in this comparison uses only one self-attention head.

Although the MAX/AVG configuration achieves the best overall performance, no evidence suggests that this configuration is optimal for the spatial and temporal label output branch. For example, the AVG/MAX configuration has better temporal label prediction accuracy than that of the MAX/AVG configuration. One possible implication of this is that the encoded feature from the single head transformer is insufficient for the following label prediction tasks. Therefore, the influence of the self-attention head number was investigated as well. The performance of the proposed model using different numbers of self-attention head is tabulated in table 2 and the MAX/AVG configuration was adopted by the model. The model attains reasonably good S and T accuracy (all above 85%) by increasing the self-attention head number to 6. However, as the self-attention head number reaches 8, prediction accuracies of the model drops below 80%. A likely explanation is that the deterioration of the model performance is caused by the overcomplicated features from the 8 head transformer encoder.

In summary, it has been shown in this numerical study that the proposed model is capable of the outlier detection task in the simulated plane wave observation. Additionally, it is evident that better prediction accuracy of the model can be achieved through the tuning of pooling configuration of the multi-output layer and self-attention head number of the transformer encoder.

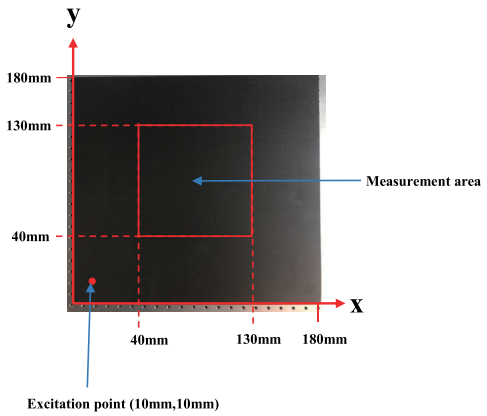

FIGURE 7. The vibration observation of the plate.

TABLE 3. The categorical accuracy of the model for vibrating plate outlier detection using different pooling configurations (S/T accuracy: the categorical accuracy of the spatial/temporal label branch; MAX: max-pooling; AVG: average-pooling).

Pooling Configuration	S Accuracy	T Accuracy
MAX/AVG	95.0%	81.25%
MAX/MAX	93.75%	94.38%
AVG/MAX	99.37%	87.5%
AVG/AVG	87.5%	68.75%

TABLE 4. The categorical accuracy of the MAX/MAX model for vibrating plate outlier detection using different self-attention head numbers (S/T accuracy: the categorical accuracy of the spatial/temporal label branch).

Head Numbers	S Accuracy	T Accuracy
1	93.75%	94.38%
4	93.12%	85.62%
6	99.9%	99.9%
8	95.63%	99.9%

B. OUTLIER DETECTION IN VIBRATING PLATE

As shown in Fig. 7, the vibration observation of the plate was collected from the measurement area. There are 10 × 10 testing points within this area and the interval between every two points is 10*mm*. The plate was excited by a hand-held exciter (B&K type 5961) at the excitation point and the vibration signal of each testing points was collected by an accelerometer (B&K type 8309). In this experimental study, the vibration data of size (10 × 10 × 100) was divided into 4 × 4 areas for the simulation of outliers at different locations. Moreover, 50 training samples and 10 evaluation samples were prepared for every possible outlier location.

Like the previous numerical study, the outlier detection performance of the proposed model in this experimental

study was examined by 4 pooling configurations. The prediction accuracy of the single self-attention head model is shown in Table 3. The AVG/AVG configuration brings the most biased and the best overall prediction performance is achieved using MAX/MAX configuration. According to the numerical study, it seems that the self-attention head number has positive impact on the prediction accuracy of the proposed model under certain conditions. Therefore, it is expected to obtain a further performance improvement of the MAX/MAX model by increasing its self-attention head number. However, as shown in Table 4, the increase of the self-attention head number not necessarily improve the prediction result. This inconsistency may be caused by the change of the vibration pattern of the model input. Nevertheless, the proposed model achieves good outlier detection performance using MAX/MAX configuration and 6 attention heads.

V. CONCLUSION

This study proposes a transformer based model for outlier detection. The multi-output layer in the model relieves the outlier labelling complexity in high dimensional space by separating spatial and temporal labels of outliers apart. During model training, this multi-output layer urges the transformer encoder to enclose the necessary spatial and temporal location of outliers in its encoded output for the following prediction tasks. The proposed model can locate outliers within pre-divided areas of simulated plane wave and vibrating plate observations with accuracies up to 85.6%/93.1% and 99.9%/99.9% respectively. A limitation of the proposed model is that the resolution of the outlier location is fixed, which hinders its application in detecting outliers with irregular distribution. A further study on improving the resolution of the outlier location prediction together with outliers clustering will be considered.

REFERENCES

[1] J. R. Casas and J. J. Moughty, "Bridge damage detection based on vibration data: Past and new developments," *Frontiers Built Environ.*, vol. 3, p. 4, Feb. 2017.
[2] C. Kim, S. Kitauchi, K. Chang, P. Mcgetrick, K. Sugiura, and M. Kawatani, "Structural damage diagnosis of steel truss bridges by outlier detection," in *Proc. 11th Int. Conf. Structural Saf. Rel.*, 2014, pp. 4631–4638.
[3] N. Dervilis, K. Worden, and E. J. Cross, "On robust regression analysis as a means of exploring environmental and operational conditions for SHM data," *J. Sound Vibrat.*, vol. 347, pp. 279–296, Jul. 2015.
[4] R. E. Bellman, *Adaptive Control Processes*. Princeton, NJ, USA: Princeton Univ. Press, 2015.
[5] G. J. McLachlan, "Mahalanobis distance," *Resonance*, vol. 4, no. 6, pp. 20–26, 1999.
[6] K. Worden, G. Manson, and N. R. J. Fieller, "Damage detection using outlier analysis," *J. Sound Vib.*, vol. 229, no. 3, pp. 647–667, 2000.
[7] M. Mousavi and A. H. Gandomi, "Structural health monitoring under environmental and operational variations using MCD prediction error," *J. Sound Vibrat.*, vol. 512, Nov. 2021, Art. no. 116370.
[8] N. Dervilis, E. J. Cross, R. J. Barthorpe, and K. Worden, "Robust methods of inclusive outlier analysis for structural health monitoring," *J. Sound Vibrat.*, vol. 333, no. 20, pp. 5181–5195, Sep. 2014.
[9] L. A. Bull, K. Worden, R. Fuentes, G. Manson, E. J. Cross, and N. Dervilis, "Outlier ensembles: A robust method for damage detection and unsupervised feature extraction from high-dimensional data," *J. Sound Vibrat.*, vol. 453, pp. 126–150, Aug. 2019.
[10] M. D. Ulriksen, D. Tcherniak, and L. Damkilde, "Damage detection in an operating vestas V27 wind turbine blade by use of outlier analysis," in *Proc. IEEE Workshop Environ., Energy, Structural Monitor. Syst. (EESMS)*, Jul. 2015, pp. 50–55.
[11] S. M. Bendre and B. K. Kale, "Masking effect on tests for outliers in exponential models," *J. Amer. Stat. Assoc.*, vol. 80, no. 392, pp. 1020–1025, Dec. 1985.
[12] K. Singh and S. Upadhyaya, "Outlier detection: Applications and techniques," *Int. J. Comput. Sci. Issues*, vol. 9, no. 1, p. 307, 2012.
[13] B. Karlik and A. Vehbi, "Performance analysis of various activation functions in generalized MLP architectures of neural networks," *Int. J. Artif. Intell. Expert Syst.*, vol. 1, no. 4, pp. 111–122, 2011.
[14] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[15] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D, Nonlinear Phenomena*, vol. 404, Mar. 2020, Art. no. 132306.
[16] Y. Zhang and Y. Lei, "Data anomaly detection of bridge structures using convolutional neural network based on structural vibration signals," *Symmetry*, vol. 13, no. 7, p. 1186, Jun. 2021.
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[18] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, "ViViT: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6836–6846.
[19] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–11.
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.
[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[22] Z. Raisi, M. A. Naiel, P. Fieguth, S. Wardell, and J. Zelek, "2D positional embedding-based transformer for scene text recognition," *J. Comput. Vis. Imag. Syst.*, vol. 6, no. 1, pp. 1–4, Jan. 2020.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[24] R. A. Dunne and N. A. Campbell, "On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function," in *Proc. 8th Aust. Conf. Neural Netw.*, Melbourne, VIC, Australia, vol. 181, 1997, p. 185.
[25] Y. Li and Y. Yuan, "Convergence analysis of two-layer neural networks with ReLU activation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.
[27] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[28] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[29] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.
[30] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. New York, NY, USA: Academic, 1999.

RUIHENG ZHANG received the B.S. and M.Sc. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree with the Laboratory of Measurement and Control Technology and Instrument. His current research interests include vibration analysis and machine learning.

QUAN ZHOU received the B.S. degree in measurement and control technology and instrument and the M.Sc. and Ph.D. degrees in measurement technology and automatic equipment from the University of Electronic and Science of China (UESTC), Chengdu, China, in 2008, 2011, and 2017, respectively. He is currently a Research Associate with UESTC, with a focus on sensor testing and reliability analysis.

LULU TIAN was born in Hubei, China. He received the Ph.D. degree from the University of Electronic Science and Technology of China (UESTC), China, in 2018. He is currently with the School of Automation Engineering, UESTC. His current research interests include crack detecting, neural networks, structural health monitoring, and non-destructive testing. He was a recipient of the Best Paper Award at the International Instrumentation and Measurement Technology Conference, in 2015.

LIBING BAI (Associate Member, IEEE) received the B.S. and Ph.D. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2008 and 2013, respectively. He was a Visiting Research Scholar with Newcastle University, Newcastle upon Tyne, U.K. He is currently a Professor with the School of Automation Engineering, UESTC. His current research interests include novel sensing and precision measurement, nondestructive testing and prognostics, and health management of structure and electronic equipment.

JIE ZHANG received the B.S. degree in measurement and control technology and instrument and the M.Sc. and Ph.D. degrees in measurement technology and automatic equipment from the University of Electronic and Science of China (UESTC), Chengdu, China, in 2010, 2013, and 2018, respectively. He is currently a Research Associate with the School of Automation Engineering, UESTC. His current research interests include nondestructive testing and prognostics and health management of electronics.