Dual-Path Transformer For Machine Condition Monitoring

Jisheng Bai, Mou Wang and Jianfeng Chen
School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
E-mail: {baijs, wangmou21}@mail.nwpu.edu.cn, chenjf@nwpu.edu.cn

***Abstract*—Anomalous sound detection (ASD) aims to detect anomaly sounds for earlier warning. Recently, with the development of machine learning methods for automatically detecting anomalous situations, ASD has attracted much attention. Previous work mainly focuses on finding acoustic patterns by deep neural networks, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs can catch local characteristics, but are difficult to model long sequences. RNNs are able to model long-time dependencies, but are time-consuming for training. In this paper, we propose DPTrans, a novel dual-path Transformer-based neural network for machine condition monitoring in ASD. DPTrans consists of several encoders, in which acoustic features are modeled on time-dimension and frequency-dimension by Transformer encoders. DPTrans can learn temporal and frequency dependencies and model interactive information effectively by taking advantages of self-attention module. Finally, we evaluated DPTrans on the dataset of DCASE2021 task2, the averaged AUC score was improved by 12% compared with official baseline systems.**

I. Introduction

Anomalous sound detection (ASD) is developed to detect anomalous cases before causing damage. For example, a pump suffering from a small leakage might not been inspected visually, but it can be detected acoustically through distinct sound patterns. Therefore, ASD has attracted much attention and been implemented in several scenarios, such as machine condition monitoring (MCM) and surveillance of buildings in recent years. The early detection of machinery with a reliable ASD system can prevent problems and reduce the cost of surveillance.

Machine learning based ASD methods aims to extract acoustic patterns and automatically detect anomalous sound. In general, those methods can be categorized into two classes, i.e., supervised and unsupervised learning. In supervised methods, normal and abnormal sounds should be available and annotated. But in fact, the abnormal samples are rare and usually difficultly collected. On the contrary, unsupervised methods can distinguish abnormal samples when only normal samples are available.

Autoencoders (AEs) are a typical unsupervised learning algorithm which have been applied for MCM [1], [2]. AEs can learn the characteristics of input data by minimizing the distance between reconstructed data and original data. The acoustic features are usually presented as two-dimension matrices, but AEs take one-dimension data as input. Therefore, the acoustic features have to be reshaped into one dimension and the local time-frequency information may be lost. Convolutional neural networks (CNNs) are able to extract local invariant time-frequency features and improved the performance of urban sound tagging [3] and MCM [4]. But CNNs can not handle the problem of long-time dependencies. Therefore, recurrent neural networks (RNNs) are used for catching temporal dependencies. However, the disadvantage of RNNs is that it needs many hidden cells to model sequence. Recently, attention mechanism has been proposed to solve the above problems. The multi-head self-attention layers are able to catch local and global dependencies and process information in parallel. Transformer based architectures achieved state-of-the-art performance in computer vision and natural language processing tasks [5], [6].

In the fields of speech separation on single-channel time-domain audio, dual-path networks have been proposed [7], [8]. The original dual-path architecture achieved state-of-the-art performance, since it can catch long sequence dependencies by alternate blocks. In this paper, we develop a novel dual-path Transformer-based neural network for MCM. However, due to the complexities of background noise and machine types, a time-frequency representation is more distinguishable than time-domain signal. Therefore, the proposed DPTrans takes log-Mel spectrogram as sound representation. The input spectrogram is modeled alternately by stacked DPTrans encoders. In each DPTrans encoder, Transformer-based encoders model the spectrogram on time-dimension and frequency-dimension in turn. We trained DPTrans using machine section IDs to distinguish the section of input signal. The network outputs the softmax anomaly score for each section, which is calculated as the averaged negative logarithm of predicted probabilities for the correct section. The main contribution of this work is that we proposed a dual-path Transformer-based network to model the sound in time-frequency domain and firstly applied it to MCM.

The rest of this paper is organized as follows: Section II introduces the proposed DPTrans. Section III describes the details of experiments. Section IV gives the results and discussion. Section V concludes this paper.

II. Proposed DPTrans for MCM

Given a recording x, we transforms x into a time-frequency matrix $\mathbf{X} \in \mathbb{R}^{T\times F}$ of T frames and F frequency bins. Let us assume the input of DPTrans is $\mathbf{Z}_t = (\mathbf{X}_t, ..., \mathbf{X}_{t+P-1}) \in \mathbb{R}^{P\times F}$, where $\mathbf{X}_t$ is the tth frame of $\mathbf{X}$. $\mathbf{Z}_t$ is obtained by concatenating consecutive P frames from $\mathbf{X}$.

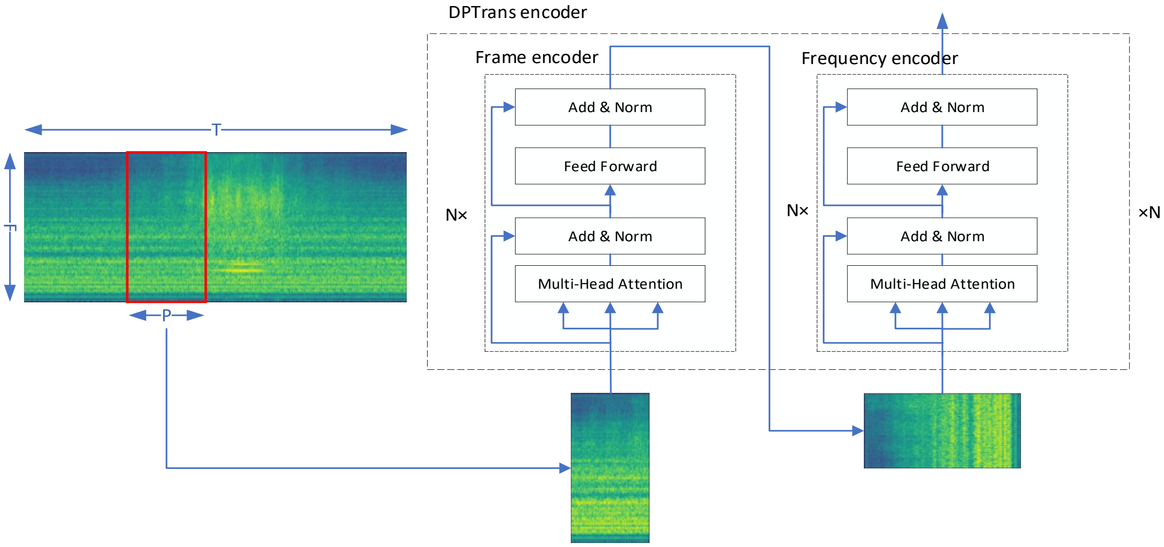


Fig. 1. The overview architecture of DPTrans. DPTrans stacks several DPTrans encoders. In a single DPTrans encoder, the input spectrogram is modeled by frame encoders and frequency encoders.

The overview architecture of DPTrans is shown in Fig.1. The procedure of proposed DPTrans is described in the following.

A. Transformer Encoder

Transformer is a sequence to sequence model which usually contains an encoder and a decoder. A Transformer encoder consists of positional encoding, multi-head attention (MHA) and position-wise feed-forward network (FFN).

We used two types of Transformer encoders in DPTrans to model the input spectrogram on different dimension. Therefore, positional encoding is not suitable in DPTrans, and only MHA and FFN are kept.

In each MHA module, we apply multiple scaled dot product attention modules. The attention of all heads are linearly concatenated and computed on the elements of sequence, and we employed residual connections and layer normalization (LN) [9] on the output of MHA. We fed the output of MHA into FFN followed by residual connections and LN to get the final output of transformer encoder. We formulate the processing as:

$$\mathbf{Z}'_t = LN\left(FFN\left(LN\left(MHA\left(\mathbf{Z}_t\right)\right)\right)\right), \tag{1}$$

where $\mathbf{Z}'_t \in \mathbb{R}^{P\times F}$ is the output of a transformer encoder.

B. DPTrans Encoder

DPTrans stacks several DPTrans encoders, each of them consists of two types encoders, i.e., frame encoder and frequency encoder. In a single DPTrans encoder, the input $\mathbf{Z}_t$ is modeled on frames by the frame encoder with embedding size of F and sequence length of P. Then the output of the frame encoder is transposed and modeled on frequency bins by the frequency encoder with embedding size of P and sequence length of F. Moreover, the frame and frequency encoder can be repeated for many times in the DPTrans encoder. The processing of a DPTrans encoder can be expressed as:

$$\overline{\mathbf{Z}}_t = E_f\left(E_t\left(\mathbf{Z}_t\right)^{\mathrm{T}}\right) = E_n\left(\mathbf{Z}_t\right), \tag{2}$$

where $E_t(\cdot)$ and $E_f(\cdot)$ presents the frame and frequency encoders, and $\overline{\mathbf{Z}}_t$ is the output of the nth DPTrans encoder $E_n(\cdot)$, $n \in \{1, ..., N\}$. After that, $\overline{\mathbf{Z}}_t$ is fed into the next DPTrans encoder.

At the end of DPTrans, a fully-connected (FC) layer is applied and the time dimension is reduced on the output of DPTrans encoders to get the final output:

$$\tilde{z} = FC\left(E_N\left(...\ E_1\left(\mathbf{Z}_t\right)\right)\right), \tag{3}$$

where $\tilde{z} \in \mathbb{R}^S$ is the probability vector predicted by DPTrans, and S is number of machine sections.

C. Loss Function

Cross entropy is used to calculate the classification loss, the loss function L can be formulated as follows:

$$L = CrossEntropy\left(\tilde{z}, l\right), \tag{4}$$

where $l \in \mathbb{R}^S$ is the real one-hot label of machine section IDs. By shifting the P by L frames, we can get $B = (T - P)/L$ spectrograms from $\mathbf{X}$. The anomaly score of the time-frequency matrix $\mathbf{X}$ is calculated as:

$$A(\mathbf{X}) = \frac{1}{B}\sum_{b=1}^{B}\log\{\frac{1-p(\mathbf{Z}_t)}{p(\mathbf{Z}_t)}\}, \tag{5}$$

where $p(\cdot)$ is the softmax output for the correct section.

III. Experiments

A. Dataset

The dataset used for evaluating DPTrans for MCM is the official dataset of Detection and Classification of Acoustic Scenes and Events (DCASE) 2021 task2. This dataset contains seven types of machines, including toyCar, toyTrain, fan, gearbox, pump, slider rail and valve. In addition, more data are provided to solve the problem of domain shift [10], [11].

The development dataset consists of three available sections for each machine. In each section, around 1,000 normal recordings in source domain and 3 normal recordings in target domain are provided for training, and around 100 clips of normal and anomalous recordings in source and target domain are provided for testing. Each recording is a 10-second audio that records the running sounds of a machine and its environmental noise. The overview of development dataset is shown in Fig.2.

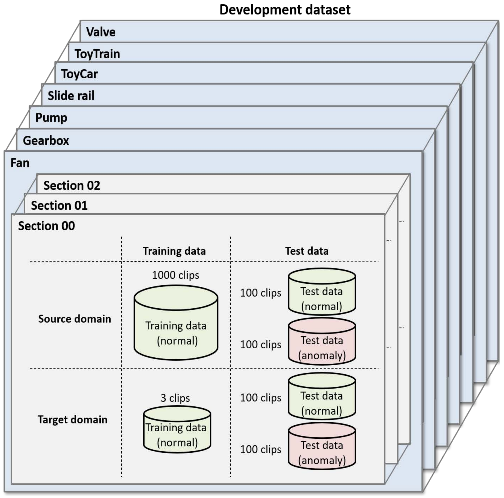


Fig. 2. The overview of the development dataset.

B. Experimental Setups

We loaded a recording x with the default sample rate of 16000Hz and applied short time Fourier transform (STFT) on x with a Hanning window size of 1024 and hop length of 512 samples. Mel filters with bands of 128 are used to transformed STFT spectrogram to Mel spectrogram. Mel spectrogram are calculated by the logarithm to get log spectrogram $\mathbf{Z}_t$. In our experiments, we extracted features using frequency bins F of 128 and different consecutive frames P of 64, 128 and 256.

In the training stage, we used the data of source and target domain and trained DPTrans for 20 epochs using Adam [13] as the optimizer. Moreover, we used a dynamic strategy to adjust the learning rate during training with an initial learning rate of 0.125. The learning rate increases linearly within the warm-up steps, and then decays by 0.98 per every two epochs.

We used 3 DPTrans encoders with 1 Transformer encoder layer and the head of MHA is set to 8. Moreover, DPTrans takes log-Mel spectrograms as input, the embedding size and sequence length of Transformer encoders is equal to P or F.

C. Baseline systems

To verify the performance of DPTrans, we compared the following methods:

Baseline-1 [12]: The organizers provide an AE-based baseline system. As shown in Table I, in this baseline, 5 consecutive frames of log-Mel spectrogram with bands of 128 are reshaped and taken as the input. The architecture of AE contains total 9 dense layers, the number of units of the first and last four layers is 128 and the number of units of the fifth layer is 8. AE is trained to minimize the minimum reconstruct error for normal sound, and the anomaly score is calculated as the mean reconstruct error of the observed sound.

Baseline-2 [12]: The organizers also provide a MobileNetV2-based baseline. As shown in Table I, this baseline takes 64 consecutive frames of log-Mel spectrogram with bands of 128 to identify from which section the observed signal was generated.

D. Data Augmentation Methods

Mixup: Data augmentation is an effective way to improve generalization and prevent overfitting of neural networks. In our system, we employ mixup as the data augmentation method in training stage [14]. The mixup operations on the training samples are expressed as follows:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j \tag{6}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j, \tag{7}$$

where x_i and x_j are the input features, y_i and y_j are the corresponding target labels and $\lambda \in [0, 1]$ is a random number drawn from the beta distribution.

SpecAugment: SpecAugment [15] is a simple but effective method which was proposed for augmenting speech data for speech recognition. SpecAugment contains frequency masking and time masking applied on spectrogram. The frequency bins and time frames are randomly masked by random number of masks with random width.

E. Evaluation

The evaluation metric used in our experiments is Area Under Curve (AUC) of the receiver operating characteristic curve. The AUC for each machine type and domain of all sections are calculated to compare the performance.

To determine the anomaly detection threshold, we assume that $A(\cdot)$ follows a gamma distribution. The parameters of the gamma distribution are estimated from the histogram of $A(\cdot)$, and the anomaly detection threshold is determined as the 90th percentile of the gamma distribution.

IV. Results and Discussion

Experimental results are given in the following Table II. In Table II, we present AUC results for each machine type of baseline and the proposed DPTrans.

TABLE I
SETTINGS OF EXPERIMENTAL METHODS.

Methods	Frames (P)	Frequency bins (F)	Network	Encoders	Encoder layers	Heads
Baseline-1	5	128	AE	-	-	-
Baseline-2	64	128	MobileNetV2	-	-	-
DPTrans-1	64	128	DPTrans	3	1	8
DPTrans-2	128	128	DPTrans	3	1	8
DPTrans-3	256	128	DPTrans	3	1	8

TABLE II
AUC SCORES OF EXPERIMENTAL METHODS.

Machine	ToyCar		ToyTrain		Fan		Gearbox		Pump		Slide rail		Valve		Averaged
Domain	source	target	source	target	source	target	source	target	source	target	source	target	source	target	
Baseline-1	**0.68**	**0.58**	0.72	**0.54**	0.66	0.62	0.63	0.71	0.71	0.56	0.78	0.6	0.55	0.53	0.63
Baseline-2	0.60	0.6	0.68	0.5	0.65	0.64	0.71	0.65	0.68	0.6	0.74	0.51	0.56	0.58	0.62
DPTrans-1	0.49	0.56	0.79	0.44	0.64	0.67	**0.79**	0.72	0.72	**0.66**	**0.88**	**0.65**	0.83	0.71	0.68
DPTrans-2	0.55	0.55	0.79	0.50	**0.7**	0.67	0.75	0.71	**0.8**	0.65	0.81	0.58	**0.9**	**0.81**	**0.70**
DPTrans-3	0.59	**0.58**	**0.82**	0.47	0.68	**0.71**	0.70	**0.73**	0.71	**0.66**	0.81	0.54	0.86	0.76	0.69

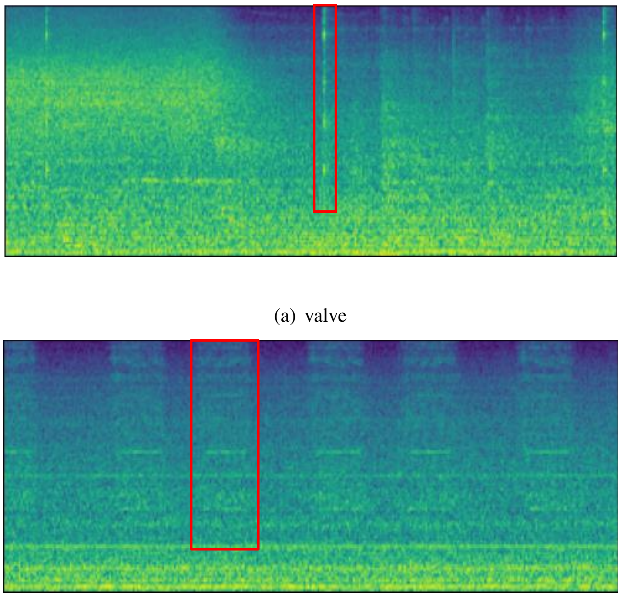
A. Comparison of Methods

The proposed method is compared with two official baseline systems in Table II, our DPTrans achieves state-of-the-art performance and significant improvements for most of the machine types. For toyCar, AE based baseline system achieves the best AUC scores in source domain, AE-based baseline and DPTrans-3 achieves the best AUC scores in target domain. For toyTrain, DPTrans-3 achieves the best AUC scores in source domain, AE-based baseline achieves the best AUC scores in target domain. For other machine types, DPTrans based methods perform better than two baseline systems. Compared with AE and MobileNetV2 based methods, the AUC score averaged on all machine types is improved by 11.1% and 12.9%, respectively.

B. Comparison of Frame Length

From Table II, longer frame length, which also equals to embedding size of the first Transformer encoders, can significantly improve the performance of toyTrain, fan, pump and valve. Specially, the AUC scores of valve are improved to 0.9 and 0.81 in source and target domain. This can be interpreted that valve sound happens in extreme short frames. If shorter P is used for generating feature, there will be more irrelevant spectrograms in training data. Therefore, longer frame length may contain more relevant sounds and it is better for extracting distinguishable acoustic patterns.

However, shorter frames length achieves better performance for the machine types of gearbox and slider rail. The sound of slider rail could happen in relative short frames, it is explained that proper frame length is critical for recognizing machine sounds. Two example spectrograms are presented in Fig.3 to show the characteristics of valve and slider rail sound.



(a) valve

(b) slider rail

Fig. 3. Two example spectrograms of valve and slider rail sound.

V. CONCLUSIONS

In this paper, we proposed DPTrans, a novel dual-path Transformer-based neural network, for anomalous machine sounds monitoring. In our approach, the log-Mel spectrogram of normal sound is modeled by consecutive DPTrans encoders, where Transformer encoders are implemented for sequentially modeling the spectrogram on frames and frequencies. We averaged the predicted probabilities for the correct section across all spectrograms generated from the log-Mel spectrogram to get the anomaly scores. Experiments of comparing methods are conducted on the latest development dataset of DCASE2021 task2, and our proposed method outperformed the two official baseline systems. It can be seen from the experimental results that DPTrans can catch local and global interactive relationship on time-dimension and frequency-dimension. Moreover, the length of frame is important for detecting anomaly sounds of different machine types.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant No.62071383 and Key research and development plan of Shaanxi Province (2021NY-036).

References

[1] J. Bai, C. Chen, and J. Chen, “Bai_lfxs_nwpu_dcase2020_submission,” DCASE2020 Challenge, Tech. Rep., July 2020.

[2] R. Giri, S. V. Tenneti, K. Helwani, F. Cheng, U. Isik, and A. Krishnaswamy, “Unsupervised anomalous sound detection using self-supervised classification and group masked autoencoder for density estimation,” DCASE2020 Challenge, Tech. Rep., July 2020.

[3] J. Bai, C. Chen, and J. Chen, “A multi-feature fusion based method for urban sound tagging,” in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1313–1317.

[4] P. Primus, “Reframing unsupervised machine condition monitoring as a supervised classification task with outlier-exposed classifiers,” DCASE2020 Challenge, Tech. Rep., July 2020.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *arXiv preprint arXiv:1706.03762*, 2017.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[7] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[8] J. Chen, Q. Mao, and D. Liu, “Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation,” *arXiv preprint arXiv:2007.13975*, 2020.

[9] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.

[10] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, “MIMII DUE: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions,” *In arXiv e-prints: 2006.05822, 1–4*, 2021.

[11] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, “ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions,” *arXiv preprint arXiv:2106.02369*, 2021.

[12] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, “Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions,” *In arXiv e-prints: 2106.04492, 1–5*, 2021.

[13] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[14] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.

[15] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.