DPTRANS: DUAL-PATH TRANSFORMER FOR MACHINE CONDITION MONITORING

Technical Report

Jisheng Bai

LianFeng Acoustic Technologies Co., Ltd.
Xi'an, China
baijs@mail.nwpu.edu.cn

Zejian Wang, Mou Wang, Jianfeng Chen

School of Marine Science and Technology,
Northwestern Polytechnical University,
Xi'an, China
329786619@qq.com, wangmou21@mail.nwpu.edu.cn,
cjf@nwpu.edu.cn

ABSTRACT

Anomaly detection has a wide range of application scenarios in industry such as finding fraud cases in financial industry or finding network intrusion in network security. Finding anomaly condition of machines in factories can prevent causing damage. Previous works mainly focus on finding local and deep features from spectrograms of anomaly sounds. Most importantly, deep features are always obtained after deep convolutional and pooling layers. However, the details of spectrogram, which present potential anomaly information, may be lost by these operations. In this paper, we introduce DPTrans, a novel dual-path Transformer-based neural network for DCASE 2021 challenge Task2 (Unsupervised Anomalous Sound Detection for Machine Condition Monitoring under Domain Shifted Conditions). DPTrans learns temporal and frequency dependencies through self-attention blocks, and achieves great performance. Moreover, DPTrans takes advantages of Transformer, which provide faster training speed and less GPU demand than comparative methods. Finally, we take different settings of Transformer train several models and make a fusion of them.

Index Terms— Anomaly detection, Transformer, model fusion

1. INTRODUCTION

Anomaly detection has been widely used for video surveillance, monitoring of critical situations. However, a pump suffering from a small leakage might not been inspected visually, it can be detected acoustically through distinct sound patterns. Further, acoustic monitoring system is cheaper and more easily developed. Therefore, anomaly detection in industrial applications has attracted much attention in recent years. The early detection of malfunctioning machinery with a reliable acoustic anomaly detection system can prevent greater damages and reduce the cost of surveillance.

The purpose of anomaly detection algorithm is to find a boundary between normal data and anomalous data. The challenge of anomaly detection is the lack of anomaly data and the uncertain types of anomaly data. In general, anomaly detection contains supervised and unsupervised learning algorithms. In supervised learning algorithm, normal and abnormal sounds should be available and annotated. But in fact, the abnormal samples are rare and usually difficultly collected. In unsupervised learning algorithm, only normal samples are available, and unsupervised learning algorithm have to distinguish abnormal samples.

Autoencoders (AEs) are one of the normal unsupervised learning algorithms which were applied for machine condition monitoring of DCASE2020 task2 in last year [1]. AEs are usually trained in an unsupervised way, by minimizing the distance between decoded data and initial data (reconstruction error), AEs can learn the characteristic of the input. But AEs task one-dimension data as input, which may loss the time-frequency features and can only model on one-dimension. Convolutional neural networks (CNNs) are able to extract local invariant acoustic features and show great performance in sound detection. But anomaly information has long-time dependencies, due to which the recurrent neural networks (RNNs) are suitable for catching temporal dependencies. However, one weakness of RNNs is that training the network is time-consuming. Recently, attention mechanism has achieved state-of-the-art performance in computer vision and natural language processing tasks, and Transformer based architectures have been widely used and performs much better than CNNs or RNNs [2, 3]. Transformer are able to catch long-time dependencies due to its multi-head self-attention, which can process parallel and provide less training time.

In this report, we develop a novel dual-path Transformer-based neural network for machine condition monitoring. We trained DPTrans using machine section IDs to distinguish the section of observed signal. The proposed DPTrans tasks STFT and log-Mel spectorgrams as sound representations. The network outputs the softmax anomaly score for each section, which is calculated as the averaged negative logit of the predicted probabilities for the correct section.

This paper is organized as follows: Section 2 introduce the proposed DPTrans. Section 3 describes the details of experiments. Section 4 gives the results and discussion.

2. PROPOSED METHOD

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

Given a recording x of length N. 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 $\mathbf{Z}_t = (\mathbf{X}_t, ..., \mathbf{X}_{t+P-1}) \in \mathbb{R}^{P \times F}$, which is obtained from $\mathbf{X}$ by concatenating consecutive P frames.

DPTrans consists of several DPTrans encoders, each of them consists of two Transformer encoders. Inside each DPTrans encoder, the input $\mathbf{Z}_t$ is modeled sequentially on frames by the first Transformer encoder, the the output of the first Transformer en-

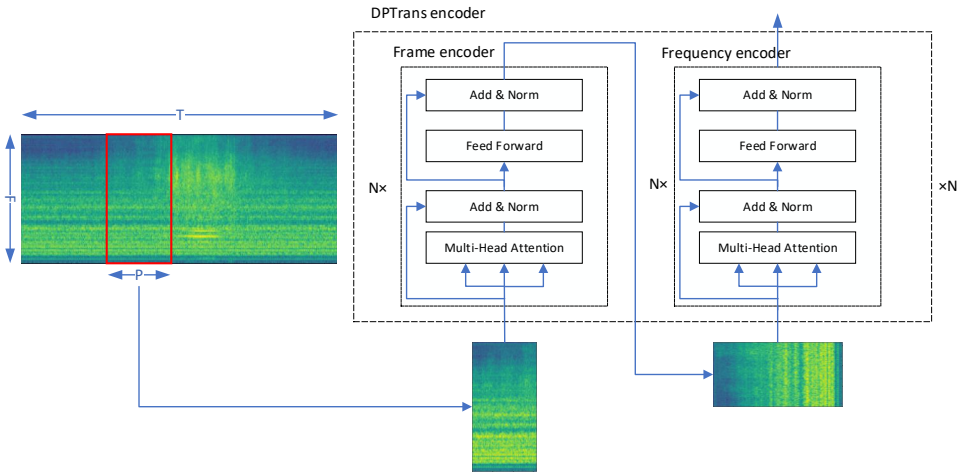

Figure 1: The overview architecture of DPTrans.

coder is transposed and modeled sequentially on frequency bins by the second Transformer encoder, which can expressed as:

$$\overline{\mathbf{Z}}_t = E_f\ (E_t\ (\mathbf{Z}_t)) = E_n\ (\mathbf{Z}_t)\,, \tag{1}$$

where $E_t(\cdot)$ and $E_f(\cdot)$ are the first and second Transformer encoder, and $\overline{\mathbf{Z}}_t$ is the output of the nth DPTrans encoder $E_n(\cdot)$. $\overline{\mathbf{Z}}_t$ is fed into the next DPTrans encoders, a linear layer is applied and the output of frame-axis is reduced on the output of the DPTrans encoders to get the final output:

$$\widetilde{z} = f\ (E_n\ (\ldots\ E_1\ (\mathbf{Z}_t)))\,, \tag{2}$$

where $\widetilde{z} \in \mathbb{R}^S$ is the probability vector predicted by DPTrans $f(\cdot)$, S is number of machine sections. We use *CrossEntropy* to calculate the classification loss, the loss function L_c can be formulated as follows:

$$L_c = CrossEntropy\,(\widetilde{z}, l)\,, \tag{3}$$

where l is the real one hot label of machine section IDs.

By shifting the P by L frames, $B = (\lfloor \frac{T-P}{L} \rfloor)$ images are extracted. The anomaly score is calculated as:

$$A(\mathbf{X}) = \frac{1}{B}\sum_{b=1}^{B} \log\{\frac{1-p(\mathbf{Z}_t)}{p(\mathbf{Z}_t)}\}, \tag{4}$$

where p is the softmax output of DPTrans for the correct section.

3. EXPERIMENTS

3.1. Dataset

The dataset of task2 consists of seven types of machines, including toyCar, toyTrain, fan, gearbox, pump, slider and valve. Not only the machine type is changed compared to the task of DCASE2020, but also more data are provided to solve the problem of domain shift for machine condition monitoring [4–6].

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

The additional training dataset provides the other three sections for each machine type. Each section consists of around 1,000 normal recordings in source domain and three normal recordings in a target domain for training. Around 100 clips each of normal and anomalous recordings in the source and target domain from the three sections will be used as evaluation dataset. The overview of the task2 dataset is shown in Figure 2

3.2. Features

A recording x is loaded with default sample rate and applied short time Fourier transform (STFT) 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. STFT and log-Mel spectrograms are calculated by the logarithm to get log spectrograms $\mathbf{Z}_t$. We extract consecutive frames P of 64, 128 or 256, and frequency bins F 128 or 320 for generating features. 128 is the number of Mel filters for generating log-Mel spectrogram, and 320 is the bins of frequency between 1k to 6k Hz to get STFT spectrogram.

3.3. Experimental methods

We conducted our experiments using the DCASE 2021Challenge Task 2 dataset. To verify the performance, we compared the following models:

Baseline: The organizers provide a MobileNetV2-based baseline. This baseline tasks log-Mel spectrogram with bands of 128 to identify from which section the observed signal is generated.

DPRNN: The original dual-path RNN used for time-domain speech separation [7]. DPRNN tasks log-Mel spectrogram with bands of 128 as input. We use three RNN encoders with 1 layer, and inside the encoder we use one directional GRU as RNN layers.

DPTrans: The proposed DPTrans is introduced in Section 2. We use 3 DPTrans encoders with 1 Transformer encoder layer and the head of self-attention is set to 8. Moreover, DPTrans takes log-Mel and STFT spectrograms as input. The P of log-Mel spectro-

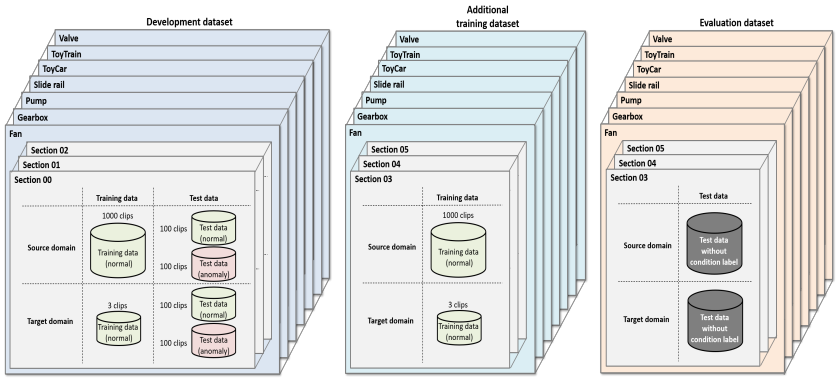

Figure 2: Overview of task2 datasets.

gram is set to 64, 128 and 256 and F is 128. The P of STFT spectrogram is set to 64 and F is 320.

Settings of experimental methods are listed in Table 1. 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.

3.4. Data augmentation methods

Mixup: Data augmentation is an effective way to improve generalization and prevent overfitting of the neural networks. In our system, we employ mixup as the data augmentation method in the training stage [8]. The mixup operations on the training samples are as follows:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j \tag{5}$$

$$\tilde{y} = \lambda y_i + (1-\lambda)y_j, \tag{6}$$

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.

Specaugmentation: SpecAugment [9] 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.

4. RESULTS AND DISCUSSION

Experimental results are given in the following Table 2. In Table 2, we present AUC results of each machine type of baseline and eight experiments .

4.1. Comparison with frame length

From Table.2, we can see that ToyTrain, gearbox, pump and valve benefit from different larger consecutive frames P, but the performance of ToyCar, fan and slider rail are worse with bigger value of P. The sound of valve happens in short frames, and the sound of ToyTrain only occurs in the middle of a recording. The above facts lead to lots of irrelevant spectrograms, which may effect extracting the acoustic pattern of these machines. Thus, longer frame length may contain more relevant sounds and it is better for extracting distinguishable acoustic pattern. But the sound of slider rail could happen in relative short frames, which may reveal that proper frame length is critical for recognizing machine sounds.

4.2. Comparison with features

Comparing the results of STFT and log-Mel spectrogram, STFT achieves better performance than log-Mel on most of the machines. Most of the machines have detailed differences in relative higher frequencies, STFT provides higher resolution in high frequencies and achieve better performance.

4.3. Comparison with encoder settings

The AUC results of exp3 and exp4 show that the increasing of encoders' number provides better performances on most of the machine types.

4.4. Comparison with methods

Compared with baseline and DPRNN in Table.2, DPTrans achieves much better results of each machine type. Moreover, it is faster for training DPTrans than baseline and DPRNN, and the need of GPU memories is much less than baseline. The computational performance of different methods are shown in 3.

4.5. Submissions

The final submission are conducted on the evaluation dataset of task2, in which the sections of each machine are changed. The experiment methods of submission are listed in Table.4, in which the submission3 is seperately trained and tested on the data of source and target domain.

5. CONCLUSION

In this paper, we present DPTrans, a novel dual-path Transformer-based neural network, for anomalous machine sounds monitoring. In our approach, the time-frequency acoustic representation is modeled by consecutive DPTrans encoders. In each DPTrans encoder, the acoustic representation is modeled sequentially on frames and then on frequencies by Transformer encoders. Finally, we averaged

	frame(P)	frequency bins(F)	net	feature	encoders	layers	heads
baseline	64	128	MobileNetV2	log-Mel	/	/	/
exp1	64	128	DPTrans	log-Mel	3	1	8
exp2	128	128	DPTrans	log-Mel	3	1	8
exp3	256	128	DPTrans	log-Mel	3	1	8
exp4	64	128	DPTrans	log-Mel	4	1	8
exp5	64	320	DPTrans	STFT	3	1	8
exp6	256	320	DPTrans	STFT	3	1	8
exp7	64	128	DPRNN	log-Mel	3	1	/
exp8	128	128	DPRNN	log-Mel	3	1	/

Table 1: Settings of experimental methods.

	ToyCar	ToyTrain	fan	gearbox	pump	slide rail	valve
baseline	0.5958	0.5916	0.6466	0.6824	0.642	0.6262	0.5707
exp1	0.5988	0.5652	0.7046	0.7188	0.6885	0.7420	0.7224
exp2	0.5749	0.6221	0.6974	0.6737	0.7164	0.7411	0.7592
exp3	0.5774	0.6125	0.6787	0.7246	0.6337	0.7234	0.6790
exp4	**0.6355**	0.6013	0.7361	**0.7531**	0.7043	0.7282	0.7308
exp5	0.5942	0.6523	**0.7426**	0.7197	**0.7443**	**0.7500**	0.7683
exp6	0.5845	**0.6864**	0.7249	0.5955	0.7217	0.6883	**0.8161**
exp7	0.5485	0.4873	0.6273	0.6663	0.6205	0.6241	0.6178
exp8	0.5365	0.6094	0.6359	0.6145	0.6271	0.6237	0.6203

Table 2: AUC scores of experiments.

	time(per epoch)	memory(GPU)
baseline	67s	10951MB
DPTrans(exp1)	53s	1943MB
DPRNN(exp7)	57s	1113MB

Table 3: Computational performance of different methods.

submission1	exp1
submission2	exp5
submission3	exp1-sep
submission4	ensemble of exp1, exp2 and exp5

Table 4: Submissions of task2.

the negative logit of the predicted probabilities for the correct section to get the anomaly scores. AUC results of comparing methods are calculated on the task2 development dataset of DCASE2021. It can be seen that DPTrans can improve the performance of each machine type, and the resolution of spectrogram in high frequencies is important for recognizing anomaly sound. Moreover, the computational performance of DPTrans are superior to the comparing methods.

6. REFERENCES

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[9] 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.