UNSUPERVISED ANOMALOUS SOUND DETECTION USING FEATURE EXTRACTOR AND ANOMALY DETECTOR

Technical Report

Jiacheng Gou[1], *Chuang Shi*[2], *Huiyong Li*[3]

University of Electronic Science and Technology of China, Chengdu, China
[1] jcgou@std.uestc.edu.cn
[2] shichuang@uestc.edu.cn
[3] hyli@uestc.edu.cn

ABSTRACT

This report proposes an anomalous sound detection method based on feature extraction and anomaly detection for DCASE 2022 task 2. In order to recognize the anomaly sound when only the normal sound is used as the training data, we use the clip of the spectrogram and the corresponding section name to train a feature extractor to generate the features of the normal sound. Then the anomaly detector is used to calculate the intensity of anomaly between the test sound features and the normal sound features, to provide the anomaly score of the test sound. In view of the domain generalization, the source domain and target domain select different shifts when clipping spectrum, and select different anomaly detectors based on whether the sound belongs to the source domain or target domain.

Index Terms— Anomalous sound detection, domain generalization, feature extraction, anomaly detection

1. INTRODUCTION

The purpose of unsupervised anomalous sound detection is to determine whether a given sound sample is similar to the training data, that is, the anomaly sound is substantially different from the training data [1]. In DCASE 2020, unsupervised methods are used for anomalous sound detection. In DCASE 2021, domain shift is added on the basis of DCASE 2020, so only a small amount of target domain training data is needed to detect anomaly sound in source domain and target domain, respectively. This year DCASE 2022 introduces domain generalization while retaining the features of the two previous years.

In the test sound provided by DCASE 2022 [2, 3], samples not affected by domain shift (source domain) and samples affected by domain shift (target domain) are mixed, and the domain is not specified. Therefore, the anomaly detector must detect anomaly sound for specific machine types and specific sections without knowing the domain [4].

Since the method of self-supervised sound feature extraction in DCASE 2021 performs well [5, 6, 7], this report uses the section names classification under different machines as a proxy task to mine the features of normal sounds, to learn valuable information for anomalous sound detection.

In this report, we apply an anomalous sound detection method using feature extractor and anomaly detector. The rest of this article is organized as follows. Section 2 introduces the anomaly scoring system, including preprocessing, feature extraction, anomaly detection, domain generalization techniques and selection. Section 3 presents some additional approaches for the challenge as a complement to the anomaly scoring system. Section 4 shows the results of the methodology proposed in this report on development dataset. Finally, section 5 summarizes the methods and conclusion of this report.

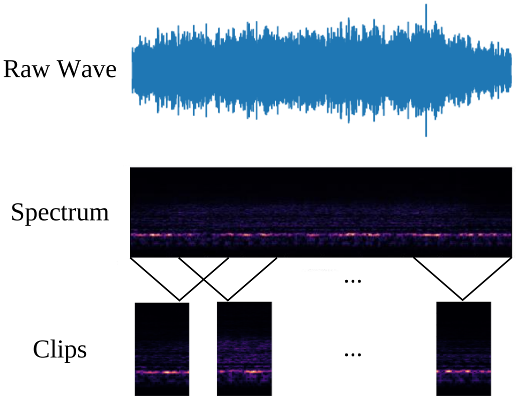

Figure 1: Diagram of the preprocessing stage.

2. ANOMALY SCORING SYSTEM

2.1. Preprocessing

The schematic diagram of preprocessing is shown in Fig. 1, and then the parameter are discussed.

The short-time Fourier transform (STFT) is used to analyze how the frequency content of the sound changes over time. This report slides the 2046 points analysis window over the signal and calculates the discrete Fourier transform of the windowed data. The window hops over the original signal at intervals of 512 samples. Raised cosine window functions taper off at the edges to avoid spectral ringing.

Then, the input of the feature extractor is obtained by concatenating 32 consecutive frames of the spectrogram and shift 16 frames (for source domain) or 2, 4, 8 frames (for target domain and different submissions) at a time.

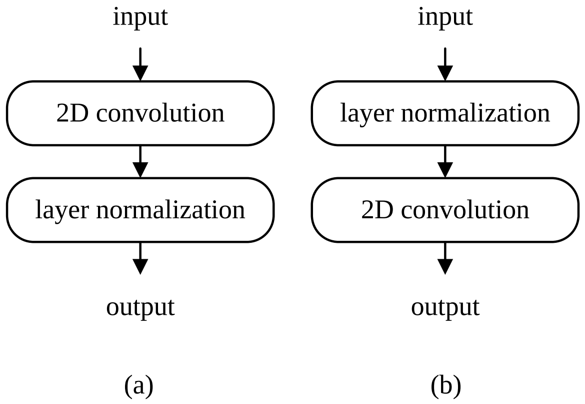

Figure 2: Structure of downsample block of the ConvNeXt. (a) after input; (b) after residual block.

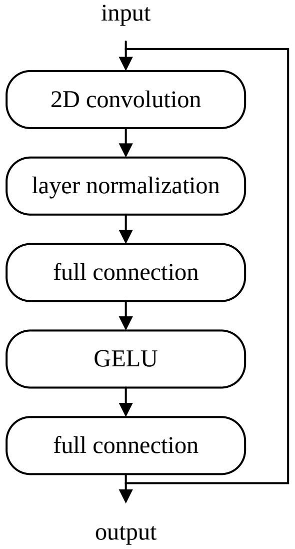

Figure 3: Structure of residual block of the ConvNeXt.

2.2. Feature extractor

A state-of-the-art convolutional network architecture called ConvNeXt [8] is used to extract features that characterize normal sound. The main blocks for ConvNext are shown in Fig. 2 and Fig. 3.

The input of the neural network is the 32 consecutive frames with 1024 frequency points, through a down-sampling layer shown in Fig. 2 named Patchify Stem which is implemented using a 4×4, stride 4 non-overlapping convolutional layer.

Four residual structures with large 7×7 convolution kernels and inverse bottleneck layer are used in the middle stages which is shown in Fig. 3. From Fig. 2, layer normalization and 2×2 convolution are used as the down-sampling layer between residual structures to reduce the size of features.

Finally, layer normalization is used to compress features to only retain channel information, and then a fully connected layer is used to output 1000-dimensional features.

The specific configurations of feature extractor are shown in the Table 1.

Table 1: Configurations of the ConvNeXt feature extractor

Layers	Configurations	Shape
input		$1 \times 1024 \times 32$
downsample1(stem)	4×4, 96, stride 4	$96 \times 256 \times 8$
res-block1	$\begin{bmatrix} \text{d}7 \times 7, 96 \\ 1 \times 1, 384 \\ 1 \times 1, 96 \end{bmatrix} \times 3$	$96 \times 256 \times 8$
downsample2	2×2, 192, stride 2	$192 \times 128 \times 4$
res-block2	$\begin{bmatrix} \text{d}7 \times 7, 192 \\ 1 \times 1, 768 \\ 1 \times 1, 192 \end{bmatrix} \times 3$	$192 \times 128 \times 4$
downsample3	2×2, 384, stride 2	$384 \times 64 \times 2$
res-block3	$\begin{bmatrix} \text{d}7 \times 7, 384 \\ 1 \times 1, 1536 \\ 1 \times 1, 384 \end{bmatrix} \times 9$	$384 \times 64 \times 2$
downsample4	2×2, 768, stride 2	$768 \times 32 \times 1$
res-block4	$\begin{bmatrix} \text{d}7 \times 7, 768 \\ 1 \times 1, 3072 \\ 1 \times 1, 768 \end{bmatrix} \times 3$	$768 \times 32 \times 1$
layer norm		$768 \times 1 \times 1$
dense		1000

To make the features of normal sound more compact, this report tries SphereFace[9], CosFace[10], ArcFace[11], finally using the Additive Angular Margin Loss[12].

2.3. Anomaly detector

After obtaining the deep features of normal sounds, the distributions of all normal sounds should be calculated. Therefore, the Local Outlier Factor [13] algorithm is used as the anomaly detector to describe the intensity of anomaly.

When the test sound is input, the trained feature extractor is used to output the features of the test sound. The features deviating from the normal distribution will be regarded as anomaly.

2.4. Domain generalization techniques

Since there is very little training data available in the target domain, the feature extractor is only trained by the sound in the source domain, and then a new full connection layer is fine-tuned to extract the features of the target domain. When training the anomaly detector, the anomaly scores of normal sound in the source domain and target domain are scaled to the same maximum and mean values, so the model can detect anomalies with the same threshold value regardless of the domain.

Another way to balance source and target domain data is to select different shifts during preprocessing step. In other words, the shift of the target domain is smaller, resulting in an increase in the number of normal sound clips of the target domain used for training.

2.5. Selection

The initial state and hyperparameters in the model will affect the final performance, to make the final performance better, and considering the development and evaluation datasets have different sections, the model that perform best for each machine type is selected.

Table 2: Harmonic mean of the AUC and partial AUC on Development Dataset

		Toy car	Toy train	Bearing	Fan	Gearbox	Slide rail	Valve
Autoencoder-based baseline	AUC (source)	88.53%	76.66%	54.37%	79.25%	68.99%	78.77%	51.93%
	AUC (target)	36.16%	20.61%	58.19%	48.09%	63.15%	46.54%	48.79%
	pAUC	53.07%	48.57%	52.96%	57.96%	57.81%	55.19%	50.47%
MobileNetV2-based baseline	AUC (source)	56.61%	59.41%	**73.85%**	74.07%	65.36%	64.36%	67.50%
	AUC (target)	48.61%	47.50%	**59.68%**	43.74%	53.98%	38.85%	64.00%
	pAUC	50.94%	52.71%	**57.34%**	54.85%	56.46%	54.71%	63.90%
Section-specific AE system (Not submitted)	AUC (source)	88.22%	78.13%	54.04%	77.82%	72.02%	**81.27%**	54.19%
	AUC (target)	54.58%	34.19%	57.13%	47.40%	62.85%	**55.95%**	50.53%
	pAUC	53.52%	51.61%	51.56%	58.15%	57.63%	**57.60%**	50.59%
Proposed system, shift=2 (Submission 1)	AUC (source)	**79.00%**	**67.08%**	59.34%	67.43%	**87.12%**	66.71%	65.90%
	AUC (target)	**62.92%**	**42.99%**	63.08%	58.37%	**76.25%**	67.29%	57.30%
	pAUC	**62.07%**	**53.49%**	53.49%	64.45%	**63.25%**	62.02%	51.96%
Proposed system, shift=4 (Submission 2)	AUC (source)	48.50%	58.31%	53.35%	70.06%	68.13%	72.17%	**73.05%**
	AUC (target)	57.91%	48.56%	61.77%	52.94%	57.59%	64.53%	**77.10%**
	pAUC	52.75%	52.23%	51.80%	64.40%	54.72%	58.09%	**67.91%**
Proposed system, shift=8 (Submission 3)	AUC (source)	69.52%	61.59%	56.40%	**70.38%**	80.22%	67.86%	64.75%
	AUC (target)	60.19%	35.47%	53.85%	**69.96%**	65.37%	58.75%	51.60%
	pAUC	57.15%	50.94%	48.91%	**69.61%**	59.44%	58.38%	55.50%
Selection (Submission 4)	AUC (source)	79.00%	67.08%	73.85%	70.38%	87.12%	81.27%	73.05%
	AUC (target)	62.92%	42.99%	59.68%	69.96%	76.25%	55.95%	77.10%
	pAUC	62.07%	53.49%	57.34%	69.61%	63.25%	57.60%	67.91%

3. OTHER ATTEMPTS

We also tried several methods to supplement the main method. From the perspective of training data, the method proposed in this report divides the data into 7 groups according to machine type. The following attempt is to divide the data into 42 groups according to section, and another attempt is to not distinguish the data.

3.1. Section-specific AE system

The training strategy has been improved for the autoencoder-based baseline system [14, 15], and the autoencoder model is trained for specific sections, increasing the total number of models from 7 to 42. For the scheme of reconstruction normal sound, more detailed data differentiation can make the features of the bottleneck layer more effective.

3.2. One-model system

An attempt is made to detect anomaly sound using only one model, with the proxy task being used to distinguish all sections of all machine types, that is, to distinguish 42 classes at once. As one of the main schemes of DCASE 2020 [16, 17], the advantage of one-model is that all the data can be used for training. A larger amount of data is more likely to improve the accuracy of the model.

4. EXPERIMENTAL RESULTS AND SUBMISSIONS

The method proposed in this report is used to train feature extractors and anomaly detectors respectively for seven machines, including fan, gearbox, bearing, slide rail, toy car, toy train and valve, each of which has six sections on development dataset and additional training dataset.

The proposed system and attempts are compared with the two baseline systems. From the results shown in Table 2, the proposed system performs significantly better in both the source and target domains than the two baseline systems. In the proposed system, the performance of domain generalization that judge the domain of the test sound is mediocre. Section-specific AE system performed well in slide rail, but is mediocre for the rest. However, the one-model system does not outperform the rest of the models in every machine types, so the results are not shown. For the machine type named Bear, the MobileNetV2-based baseline has higher AUC and pAUC harmonic mean performance than all the proposed systems.

In summary, the results obtained by three systems and one selection result have been submitted to the challenge. The first result comes from the process of feature extraction, anomaly detection and domain generalization technology which is proposed in this report, and shift equals to 2 when clipping spectrum in the target domain. The second result comes from changing shift to 4 when clipping spectrum in the target domain during domain generalization. The third result comes from the proposed system whose shift equals to 8. The fourth result is the selection of the above three results and section-specific AE result by machine type.

5. CONCLUSION

In this report, a domain generalization anomalous sound detection system based on feature extraction is proposed. The system is composed of multi-stage residual connection neural network, which is used to extract features from the frequency spectrum of clips and estimate the distribution of features by LOF. These estimated distributions are then used to calculate the intensity of anomaly of the test sound clip and will combine all clips of an sound into the final anomaly score. Experimental evaluations on the dataset of DCASE 2022 task 2 indicate that the proposed system significantly outperforms the challenge's baseline system in AUC and pAUC in both the source and target domains of the development dataset.

6. REFERENCES

[1] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, “Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212–224, 2019.

[2] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, “MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task,” *arXiv preprint arXiv:2205.13879*, 2022.

[3] 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,” in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 1–5.

[4] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, “Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques,” *arXiv preprint arXiv:2206.05876*, 2022.

[5] K. Morita, T. Yano, and K. Tran, “Anomalous sound detection using cnn-based features by self supervised learning,” DCASE2021 Challenge, Tech. Rep., July 2021.

[6] K. Wilkinghoff, “Utilizing sub-cluster adacos for anomalous sound detection under domain shifted conditions,” DCASE2021 Challenge, Tech. Rep., July 2021.

[7] Q. Zhou, “Ensemble of arcface based systems for unsupervised anomalous sound detection under domain shift conditions,” DCASE2021 Challenge, Tech. Rep., July 2021.

[8] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[9] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6738–6746.

[10] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[11] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.

[12] F. Wang, J. Cheng, W. Liu, and H. Liu, “Additive margin softmax for face verification,” *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[13] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “Lof: Identifying density-based local outliers,” *SIGMOD Record*, vol. 29, no. 2, p. 93–104, May 2000.

[14] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, “A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 1996–2000.

[15] Y. Kawachi, Y. Koizumi, and N. Harada, “Complementary set variational autoencoder for supervised anomaly detection,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2366–2370.

[16] 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.

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