A DATA AUGMENTATION-BASED APPROACH FOR FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION

Technical Report

Jiacheng Gou[1], *Chenkun Sun*[2], *Anqi Tu*[3], *Huiyong Li*[4], *Chuang Shi*[5],

University of Electronic Science and Technology of China, Chengdu, China
[1] jcgou@std.uestc.edu.cn
[2] 2020010910022@std.uestc.edu.cn
[3] anqitu@std.uestc.edu.cn
[4] hyli@uestc.edu.cn
[5] shichuang@uestc.edu.cn

ABSTRACT

The detection of abnormal conditions in machinery and equipment through sound diagnosis is of utmost importance in the field of industrial automation. However, acquiring abnormal sound and dealing with machine state transformation can present challenges. In order to address these challenges, a data augmentation combined with unsupervised feature extraction approach has been proposed for abnormal sound detection in machinery and equipment. The method involves the extraction of features from the sound samples using a unsupervised feature extractor, which is constructed using both normal and artificially constructed abnormal log-mel-spectrograms. These features are then fed into a autoencoder for unsupervised abnormal sound recognition. The proposed method has been evaluated using the DCASE 2023 Task 2 Development Dataset, and the results demonstrate that it can adaptively extract sound features of mechanical equipment, achieving an average area under the curve detection result of 56.52%.

Index Terms— Anomalous sound detection, data augmentation, first-shot, deep learning

1. INTRODUCTION

Employing acoustic sensors to monitor machine condition represents a critical subject for the industry, with various applications ranging from factory automation to predictive maintenance [1]. Particularly, the automatic detection of abnormal sounds serves as a pivotal application [2, 3]. Nevertheless, not all potential types of anomalous sounds may be predetermined, and intentionally damaging machinery to capture anomalous sound recordings presents an unfavorable approach. As a result, unsupervised anomalous sound detection has garnered significant research attention in recent years, where only data gathered under normal operating conditions is utilized to train machine learning models [4, 5, 6]. In real-world scenarios, transformations in the operational state of machines or ambient noise can trigger domain shifts. To address frequently occurring or scarcely noticeable domain shifts, domain generalization techniques prove to be valuable [7]. In this task, the system is tasked with applying domain generalisation techniques to effectively manage these domain shifts.

The recent advancements in unsupervised anomalous sound detection owe much to the DCASE challenges centered around this subject. In DCASE 2020, the unsupervised approach was implemented to detect anomalous sounds [8]. DCASE 2021 introduced a domain shifted condition, requiring only a small amount of target domain training data to identify anomalous sounds in both the source and target domains [9]. DCASE 2022, while retaining the features of the previous years, incorporates domain generalization, allowing for the detection of anomalous sounds without prior knowledge of the domain [10]. In DCASE 2023, a first-shot problem has been added to the requirements, necessitating the use of limited machines and sound to train the model, and testing it on entirely new machine types [11].

In recent years, a popular detection framework has emerged for DCASE task2 challenges, involving the use of a classifier that is initially trained to differentiate between machine types or IDs. The final layer of the classifier is then removed, and the remaining layers are utilized as a feature extractor. The feature distribution of normal sound is then modeled and the features of the sound being tested are extracted and compared to this model. The resulting anomaly score is based on the degree of similarity between the extracted features and the modeled feature distribution of normal sound.

This report introduces a data augmentation algorithm for the first-shot problem in DCASE 2023. The algorithm involves generating anomalous sound samples, training a feature extractor, and reconstructing the features using an autoencoder. The resulting reconstruction error is then utilized as the anomaly score. The subsequent sections of the report are structured as follows: Section 2 provides an overview of the anomaly scoring system, which includes preprocessing, data augmentation, feature extraction, and anomaly detection. Section 3 presents the results of the methods outlined in this report on the development dataset. Finally, Section 4 summarizes the methodology and conclusions drawn from this report.

2. ANOMALY SCORING SYSTEM

To detect abnormal sounds in mechanical devices, an anomalous sound is initially produced. Next, a feature extractor is developed by training MobileFaceNet using the log-mel-spectrograms of normal sound and the log-mel-spectrograms of the generated anomalous sound. Subsequently, the feature extractor is employed to extract sample features from the training set, which are then utilized to train an autoencoder to determine the anomaly score. The following steps outline the detection flowchart:

1) The dataset's sound is preprocessed and converted to log-mel-spectrograms;

2) MoblieFaceNet is trained using the log-mel-spectrograms of normal sound from the training set and the log-mel-spectrograms of the generated anomalous sound to construct a feature extractor;

3) The feature extractor is utilized to extract features from the log-mel-spectrograms of normal sound in the training set, and then an autoencoder is trained to serve as an anomaly detector;

4) During testing, the log-mel-spectrograms of the test sound is first fed into the feature extractor for feature extraction, and then the trained autoencoder is used to identify anomalies in the mechanical device.

2.1. Preprocessing

Audio preprocessing commonly involves the extraction of log-mel-spectrograms, which are a spectral representation of audio signals that have been processed to mimic the human auditory system's perception of sound. In this subsection, the librosa library in Python is utilized to compute log-mel-spectrograms, and the different parameters involved in computing log-mel-spectrograms are discussed.

The length of the Fast Fourier Transform window is set to 1024, while the number of samples between adjacent frames in the spectrogram is set to 512. The number of Mel bands to be generated is set to 128, and the resulting spectrogram has 313 frames.

2.2. Data augmentation

The data augmentation techniques of spectral noise addition, spectral masking and spectral warping are used in this subsection to create additional types of abnormal sound samples and expand the classification class, as the first-shot problem requires training a model with a limited amount of data. Using these data augmentation techniques, the algorithm can generate new abnormal sound samples that help improve the model's performance.

1) Spectral noise addition encompasses the addition of various types of noise, including white noise, pink noise, burst moment noise, and burst frequency noise, to the spectrogram;

2) Spectral masking refers to the process of masking certain bands and time intervals of the spectrogram;

3) Spectral warping is a technique that involves bending specific spectral lines in the spectrogram to generate new audio samples.

These techniques can help create more diverse and realistic data for the model to train on, which can lead to better performance when the model is tested on new, unseen machines.

2.3. Feature extractor

In previous DCASE Task 2 challenges, classical models typically used in the field of image classification did not show improved detection performance for log-mel-spectrograms or frequency spectrograms. But MobileFaceNet showed good results in this task [12, 13, 14]. To address this issue, anomalous sounds are initially generated using the approach outlined in the preceding subsection. Subsequently, a model is trained utilizing the MobileFaceNet architecture to differentiate between various machine classes and anomalous sound types [15].

Upon completing the training process, the final layer of the model was omitted and repurposed as a feature extractor. This extractor was leveraged to extract 128-bit feature values from the log-mel-spectrogram of the audio. The architecture of the model is presented in Table 1.

Table 1: Configurations of the MobileFaceNet, where t is the expansion factor, c is the number of channels, n is the number of repeats, and s is the stride.

Input	Operator	t	c	n	s
$128 \times 313 \times 1$	conv2d 3×3	-	64	1	2
$64 \times 157 \times 64$	depthwise conv2d 3×3	-	64	1	1
$64 \times 157 \times 64$	bottleneck	2	64	5	2
$32 \times 79 \times 64$	bottleneck	4	128	1	2
$16 \times 40 \times 128$	bottleneck	2	128	6	2
$8 \times 20 \times 128$	bottleneck	4	128	1	2
$1 \times 1 \times 128$	bottleneck	2	128	2	1
$1 \times 1 \times 128$	conv2d 1×1	-	512	1	1
$1 \times 1 \times 512$	linear GDConv16 $\times$ 1	-	512	1	1
$1 \times 1 \times 512$	linear conv2d 1×1	-	128	1	1

2.4. Anomaly detector

In unsupervised anomaly detection, an autoencoder is trained using the features of normal sound, the structure of which is shown in Table 2. This enables the autoencoder to learn the representation of normal sound features.

During the testing process, the autoencoder can reconstruct normal sound with a small error, whereas reconstructing anomalous sound would result in a larger error. Hence, the reconstruction error can be utilized as a detection criterion for anomaly detection.

Table 2: Configurations of the anomaly detector.

	Input	Operator	c
Encoder	1×1×128	Linear+BN1d+ReLU	128
	1×1×128	Linear+BN1d+ReLU	128
	1×1×128	Linear+BN1d+ReLU	128
	1×1×128	Linear+BN1d+ReLU	128
	1×1×128	Linear+BN1d+ReLU	8
Decoder	1×1×8	Linear+BN1d+ReLU	128
	1×1×128	Linear+BN1d+ReLU	128
	1×1×128	Linear+BN1d+ReLU	128
	1×1×128	Linear+BN1d+ReLU	128
	1×1×128	Linear	128

3. RESULTS AND SUBMISSIONS

The method described in this report utilizes a total of fourteen machines from the development dataset (fan, gearbox, bearing, slide, toy car, toy train, and valve) and the additional training dataset (vacuum, toy tank, toy nscale, toy drone, bandsaw, grinder, and shaker).

When compared with two baseline systems, as presented in Table 3, the proposed system showcased superior performance within the target domain, exhibiting greater resistance to domain shifting. Nonetheless, the source domain was somewhat impacted within the proposed system.

Table 3: Harmonic mean of the AUC and partial AUC on Development Dataset

		Toy car	Toy train	Bearing	Fan	Gearbox	Slide rail	Valve
Baseline (Simple autoencoder)	AUC (source)	70.10%	57.93%	65.92%	80.19%	60.31%	70.31%	55.35%
	AUC (target)	46.89%	57.02%	55.75%	36.18%	60.69%	48.77%	55.35%
	pAUC	52.47%	48.57%	50.42%	59.04%	53.22%	48.77%	51.18%
Baseline (Selective Mahalanobis)	AUC (source)	74.53%	55.98%	65.16%	87.10%	71.88%	84.02%	56.31%
	AUC (target)	43.42%	42.45%	55.28%	45.98%	70.78%	73.29%	51.40%
	pAUC	49.18%	48.13%	51.37%	59.33%	54.34%	54.72%	51.08%
Submission 1	AUC (source)	45.48%	41.54%	67.94%	55.56%	80.96%	86.44%	51.28%
	AUC (target)	52.44%	57.84%	72.40%	32.92%	64.78%	87.40%	60.72%
	pAUC	49.16%	48.05%	55.26%	49.21%	59.44%	75.21%	51.05%
Submission 2	AUC (source)	45.90%	45.40%	72.94%	62.94%	65.00%	96.16%	47.54%
	AUC (target)	50.46%	54.50%	52.52%	45.38%	58.00%	96.48%	53.98%
	pAUC	48.63%	51.37%	54.89%	50.11%	51.79%	93.84%	56.16%
Submission 3	AUC (source)	41.38%	49.20%	61.36%	90.10%	58.10%	73.32%	49.12%
	AUC (target)	60.26%	53.76%	48.52%	34.16%	54.74%	72.20%	45.60%
	pAUC	51.00%	49.68%	51.26%	49.84%	54.11%	75.63%	52.26%
Submission 4	AUC (source)	51.48%	60.24%	50.68%	35.92%	71.72%	64.20%	58.40%
	AUC (target)	53.96%	57.80%	47.88%	39.32%	62.44%	71.36%	48.88%
	pAUC	48.52%	51.63%	49.00%	47.63%	52.00%	59.68%	67.91%

4. CONCLUSION

This report proposes a data augmentation-based approach for the first-shot unsupervised anomalous sound detection problem in DCASE 2023 Task 2. The proposed method involves generating anomalous sound samples, training a feature extractor, and leveraging an autoencoder to reconstruct the features, with the reconstruction error serving as the anomaly score.

The experimental results on the development dataset demonstrate the effectiveness of the proposed approach. The proposed method achieved an AUC of 56.52%. The proposed approach is also shown to be robust to domain shifts and can effectively detect anomalous sounds in both the source and target domains.

5. REFERENCES

[1] L. Erhan, M. Ndubuaku, M. Di Mauro, W. Song, M. Chen, G. Fortino, O. Bagdasar, and A. Liotta, "Smart anomaly detection in sensor systems: A multi-perspective review," *Information Fusion*, vol. 67, pp. 64–79, 2021.

[2] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 1306–1309.

[3] F. Cheng, A. Raghavan, D. Jung, Y. Sasaki, and Y. Tajika, "High-accuracy unsupervised fault detection of industrial robots using current signal analysis," in *Proceedings of the 2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 2019, pp. 1–8.

[4] 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," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[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," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.

[6] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 271–275.

[7] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A fourier-based framework for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 14 383–14 392.

[8] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, November 2020, pp. 81–85.

[9] 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 detection for machine condition monitoring under domain shifted conditions," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 186–190.

[10] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, 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," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022, pp. 1–5.

[11] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," *In arXiv e-prints: 2303.00455*, 2023.

[12] Y. Zeng, H. Liu, L. Xu, Y. Zhou, and L. Gan, "Robust anomaly sound detection framework for machine condition monitoring," DCASE2022 Challenge, Tech. Rep., July 2022.

[13] F. Xiao, Y. Liu, Y. Wei, J. Guan, Q. Zhu, T. Zheng, and J. Han, "The dcase2022 challenge task 2 system: Anomalous sound detection with self-supervised attribute classification and gmm-based clustering," DCASE2022 Challenge, Tech. Rep., July 2022.

[14] S. Venkatesh, G. Wichern, A. Subramanian, and J. Le Roux, "Disentangled surrogate task learning for improved domain generalization in unsupervised anomalous sound detection," DCASE2022 Challenge, Tech. Rep., July 2022.

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