FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION BASED ON MEMORY-AUGMENTED CONVOLUTIONAL AUTOENCODER

Technical Report

Yuren Bian[1],*Jun Li*[2],*Jiayun Chen*[3]

Zhejiang New Rise Digital Technology Co., Ltd.
Hangzhou 311899, China
[1]yurenbian@outlook.com
[2]lijun@newrisedt.com
[3]jiayunchen1@outlook.com

ABSTRACT

This technical report outlines our team's submission to DCASE 2024 Task 2. A novel challenge introduced by this year's DCASE is the concealment of attribute information, such as machine operation conditions, for several types of machines. This approach more closely emulates real-world factory settings. We propose an anomaly detection model based on a memory-augmented convolutional autoencoder that directly operates on spectrograms without attribute information. Experimental results demonstrate that our method outperforms the baseline model for certain types of machines.

***Index Terms*—** DCASE, unsupervised anomalous sound detection, convolutional autoencoder

1. INTRODUCTION

Anomalous sound detection (ASD) involves identifying whether the sounds emitted by a target machine are normal or anomalous. This capability enables the automatic detection of mechanical failures, which is essential for AI-based factory automation in the context of the fourth industrial revolution. Utilizing machine sounds for the prompt detection of anomalies is highly beneficial for monitoring the condition of machines [1].

A major challenge concerning the application of ASD systems is that both the number and variety of anomalous samples can be inadequate in training. Initially, DCASE 2020 Task 2 focused on unsupervised ASD using only normal sound samples for training, thereby addressing the challenge of insufficient anomalous data. In 2021, the DCASE Challenge expanded to incorporate domain adaptation techniques to address domain shifts resulting from variations in machine operating conditions and environmental noise. The 2022 iteration focused on domain generalization, necessitating that models sustain performance across diverse domains without prior domain-specific information. The 2023 challenge introduced "first-shot" ASD tasks, simulating real-world scenarios necessitating rapid deployment of ASD systems without machine-specific hyperparameter tuning. Building upon this foundation, DCASE 2024 incorporates new machine types and conceals attribute information for certain machine types, thereby further enhancing the robustness and applicability of ASD systems in diverse and unforeseen real-world environments [2-5].

For DCASE 2024 Task 2, we propose an anomaly detection model based on a memory-augmented convolutional autoencoder that directly operates on spectrograms without relying on attribute information.

The remainder of this report is organized as follows: Section 2 details the proposed method, including data preprocessing and the neural network model architecture. Section 3 presents the experimental setup and results. Finally, Section 4 provides the conclusion, summarizing the findings and implications of the study.

2. PROPOSED METHOD

2.1 Data preprocessing

In this study, we preprocess time-domain audio signals by initially applying a speed perturbation to increase the number of training samples. The signals are then converted into spectral magnitude plots by applying the Short-Time Fourier Transform (STFT) with an FFT size of 512, a hop size of 128, and a Hann window function to transition the signal into the frequency domain, followed by computing the magnitude of the resulting complex spectrogram.

2.2 Neural network architecture

Memory-augmented Convolutional Autoencoder： Deep autoencoders have been widely utilized for anomaly detection, with the expectation that training on normal data will result in higher reconstruction errors for abnormal inputs compared to normal ones, thereby serving as a criterion for identifying anomalies. However, this expectation does not always align with practical outcomes. It has been observed that autoencoders can sometimes generalize to the extent that they also accurately reconstruct anomalies, leading to a failure in detecting such anomalies. To mitigate this drawback in autoencoder-based anomaly detection, Gong et al. propose augmenting the autoencoder with a memory module, resulting in an improved model called the memory-augmented autoencoder (MemAE). The effectiveness of MemAE has been demonstrated in image and video anomaly detection [6].

We have made several modifications to the structure of MemAE to better suit it for spectrogram analysis, referring to this enhanced version as the memory-augmented convolutional autoencoder (MemCAE).

The core mechanism of MemCAE involves three components: a convolutional encoder, a transposed convolutional decoder, and a memory module. The encoder $f_e(\cdot)$ converts an input $\mathbf{x}$ into an encoded representation $\boldsymbol{z}$:

$$\boldsymbol{z} = f_e(\mathbf{x}; \theta_e) \tag{1}$$

The decoder $f_d(\cdot)$ reconstructs the input $\mathbf{x}$ from the latent representation $\hat{\mathbf{z}}$:

$$\hat{\mathbf{x}} = f_d(\hat{\mathbf{z}}; \theta_d) \tag{2}$$

The memory module consists of a memory matrix $\mathbf{M} \in \mathbb{R}^{N \times C}$ with N memory slots, each of dimension C. Let the row vector $\mathbf{m}_i, \forall i \in [N]$ denote the i-th row of $\mathbf{M}$, where $[N]$ denotes the set of integers from 1 to N. For a given encoding $\mathbf{z}$, the memory module retrieves the most relevant memory items using a soft addressing vector $\mathbf{w}$:

$$\hat{\mathbf{z}} = \mathbf{w}\mathbf{M} = \sum_{i=1}^{N} w_i \mathbf{m}_i \tag{3}$$

Addressing weights $\mathbf{w}$ are computed based on the dot product between $\mathbf{z}$ and the memory items $\mathbf{m}_{\mathrm{i}}$, then normalized using the softmax function:

$$w_i = \frac{\exp(\boldsymbol{z} \cdot \mathbf{m}_{\mathrm{i}})}{\sum_{j=1}^{N} \exp(\mathbf{z} \cdot \mathbf{m}_j)} \tag{4}$$

$\boldsymbol{z} \cdot \mathbf{m}_{\mathrm{i}}$: dot product between the encoding $\mathbf{z}$ and memory item $\mathbf{m}_{\mathrm{i}}$.

During training, the encoder and decoder are optimized to minimize reconstruction error, while the memory is updated to capture prototypical elements of the normal data. During testing, reconstruction is performed using a limited set of normal patterns stored in memory, resulting in small reconstruction errors for normal samples and large errors for anomalies, thereby facilitating anomaly detection.

3. EXPERIMENT

3.1 Dataset

The dataset used for our system consists of MIMII DG [7] and ToyADMOS2 [8], which contain normal and abnormal sounds from seven real/toy machines: Fan, Gearbox, Bearing, Slider, ToyCar, ToyTrain, and Valve. Each piece of audio is a 10-second single-channel recording, including sounds from machines and related equipment as well as ambient sounds. However, in the evaluation dataset, the sets of machine types are completely different from the development dataset. It is worth noting that for the DCASE 2024 Challenge Task 2 Development Dataset, attributes representing operational or environmental conditions are provided in the file names and attribute CSV files for four machine types (Fan, Bearing, Valve, ToyCar). For the remaining three machine types, these attributes are concealed. We also utilized the DCASE 2023 Challenge Task 2 Development Dataset to validate our model, as our model does not use any attribute information.

3.2 Experimental Setup

All experiment implementations are based on PyTorch. During training, the model is trained for 100 epochs with Adam as the optimizer, using a batch size of 8 and a learning rate of 0.001. The hyperparameter C in the memory module is set to 2000.

3.3 Results

The performance of our system is given in Table 1. We employed the area under the receiver operating characteristic curve (AUC) to evaluate the overall detection performance, while the partial AUC (pAUC) was also utilized to measure performance in a low false-positive rate (FPR) range $[0, p]$, where we set $p = 0.1$. We use the mean square error (MSE) to measure the reconstruction quality, which is used as the criterion for anomaly detection. The baseline model used for comparison is the Simple Autoencoder mode[1].

Table 1: Results of the MemCAE method on the development set (%). "AUC-S" and "AUC-T" represent the AUC of the source and target domains, respectively. "2023" and "2024" represent the DCASE 2023 and DCASE 2024 Challenge Task 2 Development Datasets, respectively.

		2023		2024	
		Baseline	Our	Baseline	Our
ToyCar	AUC-S	70.10	42.9	66.98	49.3
	AUC-T	46.89	58.84	33.75	35.6
	pAUC	52.47	49.58	48.77	51.05
ToyTrain	AUC-S	57.93	44.76	76.63	61.34
	AUC-T	57.02	55.18	46.92	60.74
	pAUC	48.57	48.79	47.95	52.63
bearing	AUC-S	65.92	53.16	62.01	61.58
	AUC-T	55.75	54.78	61.4	62.25
	pAUC	50.42	56.63	57.58	56.95
fan	AUC-S	80.19	83.44	67.71	41.36
	AUC-T	36.18	76.7	55.24	67.68
	pAUC	59.04	65.74	57.53	50
gearbox	AUC-S	60.31	56.22	70.4	58.80
	AUC-T	60.69	52.98	69.34	50.84
	pAUC	53.22	52.21	55.65	51.63
slider	AUC-S	70.31	53.84	66.51	50.12
	AUC-T	48.77	46.12	56.01	49.26
	pAUC	56.37	50.31	51.77	50.58
valve	AUC-S	55.35	42.02	51.07	47.28
	AUC-T	50.69	47.2	46.25	48.84
	pAUC	51.18	49.78	52.42	49.26

4. CONCLUSION

In this technical report, we propose an anomaly detection model based on a memory-augmented convolutional autoencoder that directly operates on spectrograms without attribute information. Experimental results demonstrate that our method outperforms the baseline model for certain types of machines.

5. REFERENCES

[1] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo

Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo, and Yohei Kawaguchi. Description and discussion on DCASE 2024 challenge task 2: first-shot unsupervised anomalous sound detection for machine condition monitoring. In arXiv e-prints: 2406.07250, 2024.

[2] 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.

[3] 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.

[4] 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.

[5] 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 eprints: 2303.00455, 2023.

[6] Gong D, Liu L, Le V, et al. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 1705-1714.

[7] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Ya mamoto, Y. Nikaido, and Y. Kawaguchi, “Mimii dg: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task,” in Proceed ings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022). Nancy, France: DCASE, 2022.

[8] 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 un der domain shift conditions,” in Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE). Barcelona, Spain: DCASE, 2021, pp. 1–5.