FINE-TUNING PRE-TRAINED AUDIO MODELS FOR ANOMALOUS SOUND DETECTION

Technical Report

Junjie Wang

Unisound AI Technology Co., Ltd.
Shanghai, China

ABSTRACT

This technical report presents our solution to Task 2 of the DCASE 2025 Challenge, which focuses on unsupervised anomalous sound detection for machine condition monitoring. We developed four subsystems, all of which detect anomalies by extracting embeddings and applying outlier detection algorithms. Among them, three systems utilize fine-tuned audio pre-trained models for embedding extraction, while the remaining one employs a convolutional neural network. Unlike previous approaches that classify machine metadata, our system enhances domain generalization by training models to distinguish between machine sounds and background noise.

***Index Terms*—** Self-supervised, pre-trained, domain generalization

1. INTRODUCTION

This technical report presents our solution to Task 2 of the DCASE 2025 Challenge[1, 2, 3], which addresses the problem of one-class unsupervised anomalous sound detection for machine condition monitoring. Unlike previous editions, this year's challenge provides additional clean machine sounds and background noise samples for some machine types, offering more information to tackle the task.

We adopt a self-supervised approach by constructing an auxiliary classification task to extract embeddings, which are then used to compute anomaly scores via KMeans clustering. Typically, the auxiliary task involves classifying machine metadata. To enrich the discriminative power of the embeddings, our auxiliary task is extended to include classification of clean machine sounds and background noise in addition to metadata. Furthermore, we leverage pre-trained models, BEATs[4], and Dasheng[5], to enhance the robustness of the embeddings.

2. METHODOLOGY

2.1. Dataset

The data set used for this task is derived from the ToyADMOS2 data set[3], consisting of normal and abnormal operating sounds from 14 types of toys/real machines. Each recording is in mono and has a duration of 10 s. For files that are not exactly 10 seconds long, we employ a strategy of trimming or padding to meet the desired duration. These signals are a mixture of machine sounds from several real factories and ambient noise samples. Each type of machine has only one section included in both the development dataset and the additional dataset. In this report, all training data from the development dataset and the additional training dataset are used to train the model. The performance of the model is evaluated on the testing data from the development dataset.

2.2. Using a Non-pretrained CNN

Following the approach described in [6, 7], A dual-branch network is employed. To capture the characteristics of the signal and provide a strong feature initialization, we use both linear magnitude spectrograms and magnitude spectrograms as input features. The linear magnitude spectrograms, with a dimension of 513, are computed using the Short-Time Fourier Transform (STFT), where the window size is set to 1024, the hop size to 512, and the frequency range is limited to 200–8000 Hz. To achieve higher frequency resolution and better capture stationary sounds, we also compute the full-range magnitude spectrum (up to 8000 points) over the entire signal. Before being fed into the network, all STFT spectrograms are normalized by subtracting the time-wise mean and dividing by the time-wise standard deviation computed over the entire training dataset. The model is trained using the Adam optimizer with a default initial learning rate of 0.001.

2.3. Using pre-trained models

In order to improve the robustness of the embeddings, two pretrained audio models were chosen, and classification heads were added to fine-tune BEATs and Dasheng. Specifically, after extracting the features from the last few layers of the pretrained model, we introduce an adaptation module composed of several linear and pooling layers, followed by a task-specific classification head for fine-tuning. To ensure better initialization, we freeze the pretrained layers during the first two epochs and train only the adaptation module. After this warm-up phase, we perform end-to-end fine-tuning of the entire network. To mitigate overfitting, we apply SpecAugment[8] to randomly mask regions on the mel-spectrogram during training.

3. CONCLUSION

Four systems are submitted for the task of anomalous sound detection in machines. Previous challenges have demonstrated that classifying machine identities helps in learning discriminative features for distinguishing between normal and anomalous sounds. In addition, training the model to differentiate between machine sounds with background noise, pure noise, and clean machine sounds enables the extraction of richer and more informative embeddings.

4. REFERENCES

[1] 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.

[2] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2506.10097*, 2025.

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

[4] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[5] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," in *Interspeech 2024*, 2024.

[6] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 276–280.

[7] S. Chen, J. Wang, J. Wang, and Z. Xu, "Mdam: Multi-dimensional attention module for anomalous sound detection," in *International Conference on Neural Information Processing*. Springer, 2023, pp. 48–60.

[8] 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.