PRE-TRAINED MODEL ENHANCED ANOMALOUS SOUND DETECTION SYSTEM FOR DCASE2025 TASK2

Technical Report

Lei Wang

MYPS
Fuyang, China
1022160842@qq.com

ABSTRACT

This study proposes a robust approach to address DCASE2025 Challenge Task2, focusing on First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. The task presents a unique challenge of training models with and without attribute information, necessitating robust performance under both scenarios.

To address this challenge, we utilize advanced pre-trained models as the feature extraction backbone, integrate attribute classification and domain classification networks, and fine-tune them on the DCASE2025 Task2 dataset. Finally, we employ a KNN model as the backend for calculating anomaly scores. Benefiting from the powerful feature extraction capability of the pre-trained model, our system achieves a competitive harmonic mean of AUC and PAUC(p = 0.1) of 60.9% on the development set。

Index Terms— Anomaly detection, pre-trained model, fine-tune, KNN

1. INTRODUCTION

Anomalous sound detection (ASD) is the task of identifying whether the sound emitted from a target machine is normal or anomalous. The DCASE challenge 2025 Task 2[1, 2, 3, 4],″ First-shot unsupervised anomalous sound detection for machine condition monitoring″is the follow-up from DCASE 2020 Task 2 to DCASE 2024 Task 2. The key technical emphases of this challenge include:

- **Unsupervised Learning**: Reliance on normal data for anomaly representation.
- **Domain Generalization**: Algorithm resilience against distribution shifts.
- **Zero-Shot Adaptation**: Model flexibility for novel machine types.
- **Data Heterogeneity**: Handling labeled/unlabeled and noisy/clean data scenarios.

Pretrained models are typically pretrained based on massive audio data, possessing strong generalization capabilities and universal feature extraction abilities. In this work, we leverage the powerful generalization ability of pretrained models to address the generalization challenges between the source domain and the target domain, as well as between machines. We use the pretrained model as a feature extractor, followed by a classification head which are used to distinguish attributes or domains. Finally, KNN is employed as the backend to calculate the anomaly score.

This paper is organized in the following manner: Section 2 describes the proposed approach. In the final of this section we show the experimental results. Section 3 contains the conclusions based on our report.

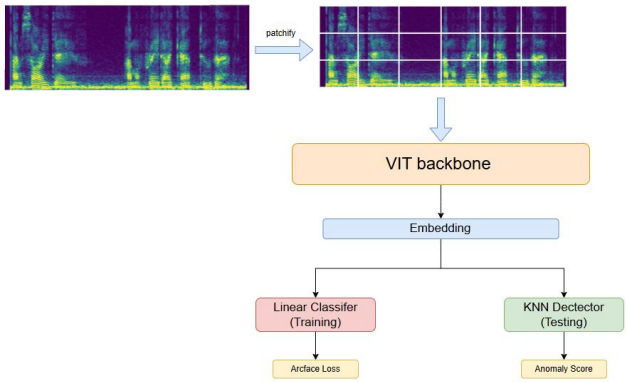

Figure 1: Architecture of proposed ASD system

2. PROPOSED ASD SYSTEM

2.1 Backbone

EAT [5] is a model crafted for self-supervised audio learning, dedicated to efficient representation learning from unlabeled audio data. It proposes a novel objective that fuses global utterance-level and local frame-level learning, thereby enhancing comprehensive audio understanding. Moreover, EAT employs a customized bootstrap self-supervised training strategy tailored to the audio domain. We utilize the EAT base model pretrained on AudioSet-2M[9], which comprises 88 million parameters.

Figure 1 illustrates the architecture of our ASD system. We employ the Encoder of EAT as the backbone, which takes mel - spectrogram features as input. These features are then split into 16×16 patches, and each patch eventually outputs an embedding containing deep representational information. We use the average pooling of all patch embeddings as the input to the classifier.

2.2 Fine-Tuning

The ASD system is fine-tuned on data from all machine types by classifying machine attributes and domains. Specifically, since some machines lack attribute information, we only perform domain classification for such cases. As this year's data provides additional clean machine data and noise data, we have taken further actions: first, for machines with additional clean machine sound data provided, we directly add them to the training set as an expanded training set; for machine types with noise provided, we use the noise for data noise addition to perform data augmentation.

During fine-tuning, we use ArcFace loss[6], and the objective function of the task can be expressed by the following formula:

$$L = \frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j=1,j\neq y_i}^{c} e^{s\cos\theta_j}}$$

where y_i is the label of sample i, and s and m are two hyperparameters. θ_j is the angle between the embedding of sample i and the registered embedding of the j-th class, which is the j-th column of the weight W of the classification head:

$$\theta_j = arccos(\frac{W_j^T x_i}{||w_j||_2 \cdot ||x_j||_2})$$

Where x_i is the final output of the backbone, T represents the transpose operation, and $||\cdot||$ denotes the L2 norm distance.

2.2.1 Full Fine-tune

Full Fine-tune (FFT) is an optimization method that adapts a pre-trained model to specific tasks or domain requirements by adjusting all of its parameters.We fine-tuned all parameters of the classification head and EAT backbone, training for 20,000 steps using the AdamW optimizer [11] with a maximum learning rate of 5e-5 and a batch size of 32. This model is referred to as EAT-FFT.

2.2.2 LoRA Fine-tune

In addition to FFT, we adopt the Low-Rank Adaptation (LoRA)[10] method for model parameter tuning which we refer to as EAT-LoRA. LoRA Freezes the pre-trained model weights and injects trainable rank decomposition matrices into specific layers (such as the attention mechanism). This allows the model to adapt to new tasks while maintaining most of the pre-trained knowledge.

LoRA's core idea is to approximate parameter updates using low-rank matrices, significantly reducing the number of trainable parameters. For the weight matrix W of a pre-trained model, LoRA decomposes its update into the product of two low-rank matrices A and B:

$$W = W_0 + \Delta W = W_0 + B \cdot A \cdot \alpha$$

Where W_0 represents the pre-trained weight matrix, B and A are matrices of dimensions $d \times r$ and $r \times k$ respectively ($r \ll \min(d,k)$), and α is a scaling factor used to adjust the magnitude of the update.

2.3 Backend

We use KNN[7] as the backend of the ASD system. We take the embeddings of normal data as the library, calculate the cosine distance from the embedding of each sample in the eval set to all embeddings of normal data, and take the minimum value as the anomaly score (k = 1).

2.4 Mechanism-based Analysis

All machines are fundamentally driven by electric motors, and machine failures are often attributable to motor anomalies. Therefore, we conducted anomaly detection for machines based on motor mechanism characteristics, including AutoTrash, Polisher, and ScrewFeeder. Specifically, the following mechanism-based analyses were performed on these three machines:

Time-Domain Analysis: Signal energy, impact intensity, and distribution characteristics were quantitatively characterized by calculating the Root Mean Square (RMS), Crest Factor (CF), Kurtosis, and Skewness.For instance, under normal conditions, these metrics remain stable (e.g., kurtosis≈3), whereas anomalies manifest as significantly increased kurtosis (>6) or a sharp rise in the crest factor.

Frequency-Domain Analysis: Focused on spectral structures, this analysis extracted amplitudes at the fundamental frequency (rotational frequency) and fault-characteristic frequencies (e.g., bearing inner/outer race frequencies and sideband frequencies).

Normal spectra exhibit stable fundamental frequencies and harmonics, while anomalies are marked by abrupt increases in characteristic frequency amplitudes or the emergence of sidebands.

Time-Frequency Analysis: Wavelet transforms and short-time Fourier transforms were employed to reveal transient features in non-stationary signals. Local mutations in energy distribution were detected by computing wavelet energy entropy or time-frequency aggregation. Under normal conditions, energy distribution is uniform; anomalies manifest as a sharp drop in energy entropy within specific frequency bands or localized time-frequency aggregation, indicating transient impacts from incipient faults.

Integrating the above analyses, we established multi-domain baseline indicators and achieved anomaly localization through threshold comparison. For systems that use mechanism-based analysis, we denote them with the suffix "-MA".

2.5 Submitted Systems

We trained our ASD system with different learning rates. Additionally, we used checkpoints from different iteration steps. Finally, we employ the ensemble learning strategy [8] to integrate the methods proposed above. To balance the various systems, we apply z-score normalization based on the score distributions of different checkpoints, and then use their weighted sum.

The systems we finally submitted as System-1 and System-3 are EAT-LoRA and EAT-FFT respectively, with the core difference being that the former employs LoRA fine-tuning while the latter adopts a full fine-tuning strategy. System-2 and System-4 are named EAT-LoRA-MA and EAT-FFT-MA respectively. Based on System-1 and System-3, these two systems replace the anomaly scores of three devices—AutoTrash, Polisher and ScrewFeeder—with the output results of the mechanistic analysis model.

2.6 Results

We compare our systems with the baseline systems of DCASE 2025challenge task 2, the AE-MSE and the AE-MAHALA. Our best system outperform the baseline systems, the AUC scores of each machine is shown in Table 1.

Table 1: AUCs and pAUCs per machine type obtained on the development set

		baseline		our system
		mse	mahala	
bearing	AUC(source)	**66.53%**	63.63%	65.32%
	AUC(target)	53.15%	**59.03%**	47.82%
	pAUC	61.12%	**61.86%**	50.95%
fan	AUC(source)	70.96%	**77.99%**	52.80%
	AUC(target)	38.75%	38.56%	**58.68%**
	pAUC	49.46%	50.82%	**53.79%**
gearbox	AUC(source)	64.80%	**73.26%**	69.36%
	AUC(target)	50.49%	51.61%	**72.82%**
	pAUC	52.49%	55.07%	**57.42%**
slider	AUC(source)	70.10%	**73.79%**	72.12%
	AUC(target)	48.77%	50.27%	**57.12%**
	pAUC	52.32%	**53.61%**	52.53%
ToyCar	AUC(source)	71.05%	**73.17%**	62.10%
	AUC(target)	53.52%	50.91%	**68.50%**
	pAUC	**49.70%**	49.05%	47.37%
ToyTrain	AUC(source)	61.76%	50.87%	**72.26%**
	AUC(target)	56.46%	46.15%	**65.94%**
	pAUC	50.19%	48.32%	**53.63%**
valve	AUC(source)	63.53%	56.22%	**78.60%**
	AUC(target)	67.18%	61.00%	**81.24%**
	pAUC	57.35%	52.53%	**72.63%**
All(hmean)	AUC(source)	**66.78%**	65.51%	66.54%
	AUC(target)	51.39%	50.05%	**62.91%**
	pAUC	52.94%	52.72%	**54.60%**

3. CONCLUSION

This paper proposes a pre-trained model-enhanced anomalous sound detection system for DCASE2025 Task2, which focuses on first-shot unsupervised anomaly detection for machine condition monitoring. The system uses EAT as the feature extraction backbone, fine-tunes with attribute and domain classification, employs KNN for anomaly scoring, and integrates models via ensemble learning. It outperforms baseline systems, achieving a harmonic mean of AUC and pAUC(p=0.1) of 60.9% on the development set.

4. 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 2025 challenge task 2: first-shot unsupervised anomalous sound detection for machine condition monitoring. In arXiv e-prints: 2506.10097, 2025..

[2] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda,and S. Saito, ″ToyADMOS2: Another dataset of miniaturemachine 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.

[3] 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.

[4] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, ″ First-shot anomaly detection for machine condition monitoring: A domain generalization baseline,″ Proceedings of 31st European Signal Processing Conference (EUSIPCO), pp. 191–195, 2023.

[5] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, ″Eat: Selfsupervised pre-training with efficient audio transformer,″ in Proceedings of the 33rd International Joint Conference on Artificial Intelligence, 2024.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, ″ Arcface: Additive angular margin loss for deep face recognition,″ in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4690–4699.

[7] S. Ramaswamy, R. Rastogi, and K. Shim, ″ Efficient algorithms for mining outliers from large data sets,″ in Proc. 2000 ACM SIGMOD Int. Conf. Manag. Data, 2000, pp. 427–438.

[8] R. L. Sagi Omer, ″Ensemble learning: A survey,″ Wiley interdisciplinary reviews. Data mining and knowledge discovery, vol. 8, 2018.

[9] Jort Gemmeke, Daniel Ellis, Dylan Freedman, Aren Jansen, et al. Audio set: An ontology and human-labeled dataset for audio events. In Proc. ICASSP. IEEE, 2017.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, ″Lora: Low-rank adaptation of large language models,″ in International Conference on Learning Representations, 2022.

[11] I. oshchilov and F. Hutter, ″ Decoupled weight decay regu larization,″ in International Conference on Learning Repre sentations, 2019.