AITHU SYSTEM FOR FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION

Technical Report

Zhiqiang Lv[1], *Anbai Jiang*[1], *Bing Han*[3], *Yuzhe Liang*[3]
Yanmin Qian[3], *Xie Chen*[3], *Jia Liu*[1], *Pingyi Fan*[2],

[1] Huakong AI Plus Company Limited, Beijing, China
[2] Tsinghua University, Beijing, China
[3] Shanghai Jiao Tong University, Shanghai, China
{lvzhiqiang,liujia}@aithu.com

ABSTRACT

Unsupervised pre-trained models have demonstrated significant promise in anomaly detection with domain shifts. The DCASE 2024 Challenge Task 2 focuses on first-shot unsupervised anomalous sound detection. Compared with last year, this year's challenge omit the attributes for some machine types. To solve this, we leverage large pre-trained models to generate robust representations for the audio. Novel usage of pseudo labeling and Low-Rank Adaptation (LoRA) are explored in the work. Additionally, we introduce SMOTE for domain equalization. Through the fusion of various models and methods, we have achieved a hmean of 68.02% on the development dataset.

***Index Terms*—** Anomaly detection, fine-tune, pseudo labeling, sound, pre-trained model

1. INTRODUCTION

In the realm of industrial automation, the ability to detect unusual sounds is vital for ensuring operational reliability and preventing potential failures. The DCASE 2024 Challenge Task 2 [1, 2, 3, 4], First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring, focuses on identifying anomalies in sounds from specific machine types. The complexity of this task lies in distinguishing normal operational noise from genuine anomalies, requiring sophisticated algorithms capable of learning from diverse acoustic patterns. In practical production environments, the diversity of equipment types, complex surroundings, and challenges with sound data collection make it difficult to develop systems that can accurately identify and classify abnormal sounds across different devices and environments. The main challenges can be summarized as follows:

- The lack of data for training. In real industrial production, anomalies are quite rare to happen, and the normal operating sounds of a certain machine are also limited. A proxy task must be introduced, and it is challenging to train a large-scale model with limited data.
- The existence of domain shifts. The complexity of industrial production environments, varied noises, and differences in recording equipment lead to disparate distributions of collected audio data, resulting in potential domain shift issues that may impact the outcomes of anomaly detection.
- Missing training labels: During the actual data collection process, not all instances may have available attribute labels for training. Models must still achieve good generalization performance even when only a subset of the data has labels.

In line with our previous works [5, 6], we leverage multiple pre-trained models to provide the necessary generalization capability across different machines. Additionally, we apply LoRA finetuning [7] to address overfitting issues. To tackle the problem of missing labels, we utilize pre-trained models to generate pseudo-labels for training. Furthermore, SMOTE [8] is applied to balance the differences in sample quantities between different domains. All submitted systems are ensemble systems where the scores of single models are linearly combined. The best system achieves a general harmonic mean of 68.02% on the development set.

The structure of the paper is organized as follows. Section 2 elaborate all single models adopted in the proposed scheme. Section 3 give an overview of the submitted systems, and Section 4 presents the detection results.

2. MODEL ZOO

2.1. BEATs

BEATs [9], short for Bidirectional Encoder representation from Audio Transformers, is a self-supervised learning (SSL) framework designed for comprehensive audio representation pre-training. The model comprises an acoustic tokenizer and an audio SSL model, both optimized iteratively. This approach enhances the learning of audio representations by generating discrete labels with rich audio semantics, thereby improving performance in audio classification tasks. Specifically, we employ the BEATs-iter3 version, pretrained on the entire training set of the AudioSet dataset and featuring 90M parameters.

BEATs is fine-tuned on the data of all machine types by classifying the attributes. Audio waveforms are first pad or truncate to 10s and converted to log-mel spectrograms with a frame length of 25ms, a frame shift of 10ms and 128 mel bins, which is identical with the original implementation. SpecAug with a maximum mask length of 80 is applied to improve the robustness [10]. To fine-tune BEATs by machine attributes, an attentive statistics pooling layer proposed in ECAPA-TDNN [11] is appended to BEATs to fuse frame embeddings into utterance embeddings, and two dense layers further maps the embedding to the predicted logits. Each unique combination of machine type and attribute is considered as a unique class

Table 1: Performances of single models on the development set

Base	Model	Total	Trainable	bearing	fan	gearbox	slider	ToyCar	ToyTrain	valve	hmean
BEATs	BEATs-full	90M	90M	60.17	62.96	68.92	75.61	55.59	62.18	65.47	63.88
	BEATs-LoRA	90M	3.73M	68.75	61.85	67.22	70.87	55.43	60.24	68.51	64.26
	BEATs-pse	90M	90M	60.16	63.44	66.32	76.74	56.79	56.10	69.64	63.47
EAT	EAT-full	88M	88M	62.56	62.35	66.26	72.62	56.91	60.74	70.72	64.18
	EAT-LoRA	88M	11.25M	65.24	61.34	71.99	76.39	57.02	57.37	72.79	65.23
	EAT-pse	88M	88M	62.87	64.69	66.64	81.87	56.39	58.69	71.68	65.23
Dual Pre-trained		180M	14.98M	63.79	62.01	61.87	73.01	62.77	66.32	72.70	65.77

for classification, and a machine type without attributes is regarded as one class. The model is trained by ArcFace loss [12] for 10000 steps by AdamW [13] with a maximum learning rate of 0.0001, a gradient accumulation step of 8, a warm-up stepof 120 and a batch size of 32. All parameters are fine-tuned, and we refer to this model as **BEATs-full**.

Additionally, we also fine-tune BEATs with low-Rank Adaptation (LoRA) [7], which we refer to as **BEATs-LoRA**. LoRA introduces additional trainable parameters to dense layers within the Transformer, while the original weights are not updated. However, different from the original implementation, we figure out that only injecting additional parameters to the q projection layer, v projection layer and out projection layer of self attention modules works best. Therefore, only these layers are updated by LoRA and the hyperparameter r is set to 64. BEATs-LoRA is also trained by classifying machine attributes, and the pooling layer and dense layers are set to be trainable. The model undergoes 50,000 training steps with an initial learning rate set to 1e-4 and is optimized using AdamW. Additionally, a cosine scheduler is employed, with a maximum learning rate of 5e-3, a minimum learning rate of 1e-5 and 10 warm-up restart steps.

K-nearest neighbor (KNN) detectors are applied to the embeddings to detect anomalies, where the distance metric is cosine and k is selected as 1. For each query embedding, the distance to its closest neighbor is adopted as the anomaly score. To improve the performance on the target domain, SMOTE [8] is employed to oversample the embeddings of the target domain.

2.2. EAT

EAT [14] is a model designed for self-supervised audio learning, focused on efficient representation learning from unlabeled audio data. It introduces a novel objective that integrates global utterance-level and local frame-level learning, enhancing overall audio understanding. Additionally, EAT adopts a tailored bootstrap self-supervised training approach specific to the audio domain. We leverage the EAT base model pretrained on AudioSet-2M, encompassing 88M parameters.

Similar with BEATs, EAT is both full fine-tuned and fine-tuned by LoRA, which we refer to as **EAT-full** and **EAT-LoRA** respectively. Both models pad or truncate raw waveforms to 10s, convert them to log-mel spectrograms with a frame length of 25ms, a frame shift of 10ms and 128 mel bins, which is identical with the original implementation. SpecAug [10] with a maximum length of 80 is also applied. Pooling layers and dense layers are also added, and the models are also trained by attribute classification. EAT-full is optimized by an Adam optimizer [15] with a maximum learning rate of 5e-5 and a warm-up step of 120, while the rest hyperparameters are identical with BEAT-full. EAT-LoRA is trained by an Adam optimizer [15] with a maximum learning rate of 1e-4, while the rest hyperparameters are identical with BEAT-LoRA.

Table 2: Number of classes for both sets

Set	Machine	BEATs-pse	EAT-pse
dev	gearbox	20	20
	slider	20	20
	ToyTrain	25	25
eval	AirCompressor	7	5
	BrushlessMotor	9	8
	HoveringDrone	3	2
	ToothBrush	10	6

Both EAT-full and EAT-LoRA adopt the KNN detector introduced in Section 2.1.

2.3. Dual Pre-trained

The multi-branch model integrates embeddings from both models to improve performance in classification tasks and anomaly detection, which has been proved effective [16, 17]. We improve this scheme by substituting two CNN branches with two powerful pre-trained models, i.e. BEATs and EAT. We refer to this model as **Dual Pre-trained**. During training, a novel multi-backpropagation strategy is employed: initially, the loss is back-propagated through each branch weighted accordingly, followed by a unified backpropagation across the entire model. This method ensures that each model's influence is effectively incorporated and balanced throughout the training process. The settings for training and anomaly detection are the same with BEATs and EAT.

2.4. Pseudo Labeling

Pseudo labeling is a commonly adopted solution for semi-supervised problems, where the model is first trained on the labeled data and assigns pseudo labels to the unlabeled data, and then the model is re-trained on the full dataset using both real labels and pseudo labels. Since attributes are missing for some machine types, we adopt a two-stage pseudo labeling approach to mitigate the problem. In the first stage, the model is trained by classifying all available attributes, where a machine type without attributes is considered as one class. After training, the model extracts the embeddings of audio clips without attributes, and applies agglomerative hierarchical clustering (AHC) on these embeddings to assign pseudo attribute labels. The number of classes is presented in Table 2, which

Table 3: Combination coefficients of four submitted systems

System	BEATs-full	BEATs-LoRA	BEATs-pse	EAT-full	EAT-LoRA	EAT-pse	Dual Pre-trained
System 1	0.0	0.2	0.0	0.0	0.4	0.0	0.4
System 2	0.0	0.5	0.0	0.0	0.1	0.0	0.4
System 3	0.0	0.0	0.0	0.1	0.3	0.2	0.4
System 4	0.1	0.2	0.0	0.1	0.2	0.1	0.3

is manually determined by applying UMAP [18] visualization on these embeddings. In the second stage, the model is re-trained from the initial checkpoint by both the real attribute labels and the pseudo attribute labels.

Both BEATs and EAT are utilized as the backbone of the model, which we refer to as **BEATs-pse** and **EAT-pse**. The network architecture, training hyperparameters and detection processes are identical with Section 2.1 and Section 2.2 respectively.

3. SUBMITTED SYSTEMS

All four submitted systems are model ensembles. Scores of each single model are first normalized to zero mean and unit standard deviation, and scores of different models are linearly combined with the coefficients obtained by grid search on the development set.

Table 3 presents the combination coefficients of four submitted models. For system 1 and 2, only BEATs-LoRA, EAT-LoRA and Dual Pre-trained are incorporated for grid search, where system 1 adopts the optimal coefficients, and system 2 adopts an adjusted version where the coefficient of BEATs-LoRA is manually increased. System 3 and system 4 conduct grid search on all seven single models, where system 3 adopts the optimal coefficients. However, the optimal coefficients overly emphasize EAT models, thus we slightly increase the coefficients for BEATs models, resulting in system 4.

4. EXPERIMENT RESULTS

The detection performance is measured by the Receiver Operating Characteristic (ROC) Curve (AUC) and partial AUC (pAUC). We calculate the source AUC, the target AUC, pAUC and a harmonic mean for each machine type, which is in line with the challenge rule.

Table 1 demonstrates the results of seven single models by the harmonic mean of AUCs and pAUC for each machine type. EAT models generally outperforms BEATs model, and the best performance is achieved by the Dual Pre-trained model with a general harmonic mean of 65.77%.

Table 4 presents the detailed results of four submitted systems, where the best performance is achieved by system 3 with a general harmonic mean of 68.02%.

5. CONCLUSION

This paper described the AITHU system for first-shot unsupervised anomalous sound detection, where we developed seven single models based on two powerful pre-trained models, i.e. BEATs and EAT. We investigated the use of dual branch, pseudo labeling and SMOTE oversampling to improve the detection performances of single models. We also merged these single models into four ensemble systems by the linear combination of anomaly score. As a result, the best system achieved a general harmonic mean of 68.02% on the development set.

6. REFERENCES

[1] 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 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *In arXiv e-prints: 2406.07250*, 2024.

[2] 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.

[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] Z. Lv, B. Han, Z. Chen, Y. Qian, J. Ding, and J. Liu, "Unsupervised anomalous detection based on unsupervised pretrained models," DCASE2023 Challenge, Tech. Rep., June 2023.

[6] B. Han, Z. Lv, A. Jiang, W. Huang, Z. Chen, Y. Deng, J. Ding, C. Lu, W.-Q. Zhang, P. Fan, J. Liu, and Y. Qian, "Exploring large scale pre-trained models for robust machine anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1–5.

[7] 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.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[9] 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.

Table 4: Detection results of four submitted systems on the development set

Machine	Metric	System 1	System 2	System 3	System 4
bearing	AUC_s	72.70	70.58	70.26	70.28
	AUC_t	73.40	76.34	74.08	76.50
	pAUC	57.95	59.05	59.68	62.11
	hmean	67.21	67.87	67.44	69.12
fan	AUC_s	63.14	64.66	64.96	65.26
	AUC_t	65.60	65.36	65.70	65.56
	pAUC	58.37	57.84	61.16	60.05
	hmean	62.22	62.43	63.88	63.52
gearbox	AUC_s	77.18	76.54	76.00	78.36
	AUC_t	74.12	71.16	76.42	75.50
	pAUC	61.79	58.89	60.37	62.63
	hmean	70.37	68.03	70.08	71.47
slider	AUC_s	91.64	90.94	93.90	93.46
	AUC_t	81.60	79.56	85.16	83.54
	pAUC	61.95	59.21	64.42	61.53
	hmean	76.32	74.16	79.12	77.07
ToyCar	AUC_s	61.94	61.02	64.18	61.92
	AUC_t	65.50	66.26	63.58	64.98
	pAUC	52.26	49.58	50.68	48.63
	hmean	59.35	58.08	58.78	57.58
ToyTrain	AUC_s	78.90	78.66	75.44	78.04
	AUC_t	60.14	62.34	63.70	62.70
	pAUC	56.84	56.58	56.74	57.68
	hmean	63.97	64.62	64.41	65.08
valve	AUC_s	78.62	78.88	79.20	79.02
	AUC_t	80.86	81.22	85.22	84.84
	pAUC	66.58	61.74	68.68	64.68
	hmean	74.80	72.84	77.08	75.18
hmean	AUC_s	73.68	73.32	73.76	73.97
	AUC_t	70.80	71.09	72.38	72.41
	pAUC	59.09	57.31	59.77	59.16
	hmean	67.23	66.44	68.02	67.82

AUC_s and AUC_t are the AUC of the source and target domains, respectively.

[10] 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.

[11] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[12] 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.

[13] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[14] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: Self-supervised pre-training with efficient audio transformer," *arXiv preprint arXiv:2401.03497*, 2024.

[15] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[16] J. Jie, "Anomalous sound detection based on self-supervised learning," DCASE2023 Challenge, Tech. Rep., June 2023.

[17] K. Wilkinghoff, "Design choices for learning embeddings from auxiliary tasks for domain generalization in anomalous sound detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[18] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *ArXiv e-prints*, Feb. 2018.