THUEE SYSTEM FOR FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION

Technical Report

Anbai Jiang[1*], *Xinhu Zheng*[1], *Yihong Qiu*[2], *Weijia Zhang*[1], *Boyuan Chen*[1],
Pingyi Fan[1*], *Wei-Qiang Zhang*[1], *Cheng Lu*[2], *Jia Liu*[1],

[1] Tsinghua University, Beijing, China
[2] North China Electric Power University, Beijing, China

ABSTRACT

This report presents our work for DCASE 2024 Task 2: first shot unsupervised anomalous sound detection for machine condition monitoring. This year's challenge is heightened by the introduction of additional machine types and the absence of training labels. To solve these problems, multiple pre-trained models are employed in the submission, along with a Dual Branch CNN model and a flow model. The pre-trained models are fine-tuned with Low-Rank Adaptation (LoRA). Finally, by fusing the systems above, we achieve the best hmean of 68.38% on the development set.

***Index Terms*—** Anomaly detection, LoRA, pre-trained models, normalizing flows, sound

1. INTRODUCTION

Anomalous sound detection (ASD) is a crucial task for machine condition monitoring, which aims to distinguish between normal and abnormal machine sounds without knowing the pattern of anomaly in prior. The task 2 series of the DCASE challenges [1, 2, 3, 4] focus on identifying anomalous sounds from multiple machine types, while featuring the complexity of real-world industrial environments and the domain shift problem.

This year's challenge emphasizes the "first-shot problem under attribute-available and unavailable conditions". In reality, the diversity of machine types makes it challenging to collect operating sounds with trainable attribute labels. Therefore, this year's task features:

- An updated and expanded set of machine types for evaluation.
- Training without attribute labels for some machine types.

In consideration of these challenges, the submitted systems incorporate five distinct single models, where two models are trained from scratch and the other three models are initialized from pre-trained models. First of all, a flow model is trained to learn the distribution of mel features, which has been proved to be robust in previous challenges [5, 6]. Secondly, since the multi-branch architecture has also been demonstrated to be effective [7, 8], an improved dual branch convolutaional neural network (CNN) is trained to extract semantic features of machine audio. Finally, we explore the use of multiple AudioSet pre-trained models, namely BEATs [9] and EAT [10]. These models exhibit remarkable capabilities in audio classification and are better suited for the ASD task than the pre-trained speech models adopted in the previous works [11, 6, 12]. Low-Rank Adaptaion (LoRA) [13] is employed to efficiently tune these pre-trained models. Furthermore, we combine BEATs and EAT into a dual branch scheme in order to leverage the power of both pre-training and multi-branch architecture.

The four submitted systems are all model ensembles, each of which roughly contains 360M parameters. We combine the five single models by model average, score fusion and group fusion, which will be elaborated in Section 3. As a result, the best system achieves an overall harmonic mean of 68.38% on the development set.

The structure of the paper is organized as follows. We will commence by delineating the single models utilized, followed by a concise overview of model ensembles, and conclude with our results on the development set.

2. MODEL ZOO

This section introduces the five single models developed in the proposed scheme.

2.1. NF-CDEE

As a continuation of our previous work [6], a probabilistic model named NF-CDEE is developed to learn the distribution of mel features. Audio waveforms are converted to log-mel spectrograms with a window size of 8192, a hop size of 512 and 256 mel bins. Frequency-wise normalization is independently applied to each mel bin within the batch, which is implemented by batch normalization (BN). Since the performance of the flow model degrades greatly for high dimensional data, we take the mean across the time dimension and employ the model to learn the distribution of the 256-dim mel features. The loss function is the sum of the negative log-likelihoods, which also serves as the anomaly score. The model is implemented by the Pyro library [14]. An Adam optimizer [15] with a learning rate of 1e-3 is utilized to train the model, with the batch size set to 128 and a maximum step of 20,000.

In addition, we find that the NF-CDEE model is more suitable for grouped machine training than training all machines together. Therefore, based on the signal characteristics, we divide the machines into the following four groups across both sets, and a unique model is trained for each group:

1. Stationary: bearing, fan, ToyCar, 3DPrinter, BrushlessMotor, ToothBrush, HairDryer, AirCompressor,
2. Non-stationary: ToyTrain, HoveringDrone, ToyCircuit
3. Periodic: gearbox, slider, RoboticArm
4. Aperiodic with impulse: valve, Scanner

* Contact: jab22@mails.tsinghua.edu.cn, fpy@tsinghua.edu.cn

Table 1: Performances of single models on the development set of the DCASE 2024 dataset

Model	Total	Trainable	bearing	fan	gearbox	slider	ToyCar	ToyTrain	valve	hmean
NF-CDEE	7.28M	6.27M	56.62	56.42	76.79	76.54	50.16	51.34	54.58	58.72
Dual Branch CNN	2.75M	2.75M	54.71	57.64	67.11	85.83	52.47	60.07	72.23	62.65
BEATs-LoRA	90M	3.73M	68.75	61.85	67.22	70.87	55.43	60.24	68.51	64.26
EAT-LoRA	88M	11.25M	65.24	61.34	71.99	76.39	57.02	57.37	72.79	65.23
Dual Branch BEATs and EAT	180M	14.98M	63.79	62.01	61.87	73.01	62.77	66.32	72.70	65.77

2.2. Dual Branch CNN

The Dual Branch CNN model incorporates two independent branches dedicated to analyzing the spectrum and spectrogram of each audio input, enabling examination of both static and dynamic semantic features. Inspired by [16], we improve the dual branch CNN model proposed by Wilkinghoff [7] by integrating self-attention modules [17] that process the feature map along specific dimension into both branches. In the spectrum branch, three self-attention modules are inserted after each convolution layer, which operate on the frequency, frequency and channel dimension respectively. Conversely, in the spectrogram branch, two self-attention modules are inserted following the residual blocks and preceding the max pooling layer, which operate on the frequency and time dimension respectively. The self-attention modules in the spectrum branch incorporate all projection heads but neglect residual connections, while the self-attention modules in the spectrogram branch incorporate residual connections but neglect all projection heads. Figure 1 is the illustration of this neural network model.

Unlike previous approaches that concatenate branch embeddings during or after training, we combine both methods, facilitating three flows of gradient back-propagation in each training iteration. This multi-backpropagation strategy aims to potentially enhance the model's performance and robustness by updating based on both overall compatibility and the independent contributions of each branch embedding.

Identical with the work of Wilkinghoff [7], raw audio inputs are 10 seconds long with a sampling rate of 16kHz. STFT is conducted with a window length of 1024 and a window shift of 512. Hanning window is used in both DFT and STFT, and only the magnitude of spectra and spectrograms is used. Temporal mean normalization is applied to the magnitude spectrograms, but no other data augmentation is applied. Regarding the model, the head numbers of the self-attention modules are [4,4,2] for the spectrum branch and [2,2] for the spectrogram branch. During training, the batch size is set to 64, and the loss function is ArcFace [18], same for general and branch embeddings. The optimizer is AdamW [19] with a learning rate of 5e-3, betas [0.9,0.98] and weight decay 1e-5. The scheduler is a cosine scheduler with a warmup restart step of 10, a cycle step of 1k and a gradient accumulation of 8.

2.3. BEATs-LoRA

BEATs, which stands for Bidirectional Encoder representation from Audio Transformers, is a self-supervised learning (SSL) framework for general audio representation pre-training. The model includes an acoustic tokenizer and an audio SSL model that are optimized iteratively. This approach aims to improve the learning of audio representations by generating discrete labels with rich audio semantics, which in turn enhances the performance of audio classification tasks. We use the BEATs-iter3 version, which is pre-trained on the full training set of the AudioSet dataset and utilizes 90M parameters.

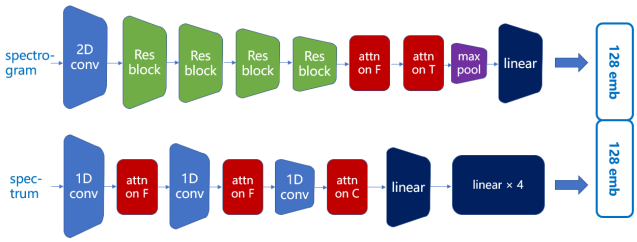

Figure 1: Illustration of Dual Branch CNN. Note that to conduct the auxiliary task of classification, another linear layer is added at the back to project the embedding dimension from 256 to the number of classes.

Rather than fully fine-tuning the model, we apply LoRA for fine-tuning purposes. LoRA introduces parameters to the q and v matrices and the out_projection matrix within the Transformer. The hyperparameter r is configured to 64. For the v matrix in the latter half of the Transformer layers, r is increased to 96. Additionally, throughout training, both the pooling module and the terminal fully-connected layer are rendered trainable to facilitate classification tasks.

For the training process, the input features are log-mel spectrograms with a frame length of 25ms and a frame shift of 10ms. The number of mel bins is set to 128. SpecAugment [20] is employed with a maximum mask length of 80 for both time and frequency axes. The loss function is ArcFcae [18]. Batch size is set to 8, and gradient accumulation occurs every 8 steps. The model is trained for 50,000 steps with an initial learning rate set to 1e-4, optimized using AdamW. Additionally, a CosineAnnealingWarmupRestarts scheduler is implemented, with a max learning rate of 5e-3 and a minimum learning rate of 1e-5 and 10 steps of warm up steps. Prior to anomaly detection, SMOTE is also employed to balance distributions between different domains, and the same technique is also used in the following two pre-trained models.

2.4. EAT-LoRA

EAT is a model designed for audio self-supervised learning, aiming to learn representations from unlabeled audio efficiently. It introduces a unique objective that incorporates global utterance-level and local frame-level learning, thereby enriching audio comprehension. Furthermore, EAT adopts a bootstrap self-supervised training paradigm tailored to the audio domain. We utilize the EAT-base model pre-trained on AudioSet-2M, which comprises 88M parameters.

Similar to the BEATs model, we employ LoRA instead of full

Table 2: Combination coefficients of four submitted systems

Model	Group	NF-CDEE	Dual Branch CNN	BEATs-LoRA	EAT-LoRA	Dual Branch BEATs and EAT
System 1	Scanner	0.0	0.0	0.5	0.4	0.1
	Others	0.0	0.2	0.1	0.3	0.4
System 2	Periodic	0.5	0.0	0.0	0.5	0.0
	Others	0.0	0.0	0.5	0.4	0.1
System 3	All	0.0	0.2	0.3	0.2	0.3
System 4	Periodic	0.2	0.1	0.3	0.4	0.0
	Others	0.0	0.2	0.3	0.2	0.3

Table 3: Performances of submitted systems on the development set of the DCASE 2024 dataset

Model	bearing	fan	gearbox	slider	ToyCar	ToyTrain	valve	hmean
System 1	66.09	61.60	71.08	79.75	58.25	65.71	76.43	67.67
System 2	67.87	62.43	77.94	80.92	58.08	64.62	72.84	68.38
System 3	67.94	61.65	71.34	78.51	56.80	65.01	75.78	67.39
System 4	67.94	61.65	76.89	81.25	56.80	65.01	75.78	68.33

fine-tuning during the training process. LoRA parameters are introduced to the q and v matrices of the Transformer layers, as well as the fully connected layers. The training settings are consistent with the previous model.

2.5. Dual Branch BEATs and EAT

The Multibranch model is a combination of the two models introduced above. It combines the embeddings extracted from both models for enhanced performance in classification tasks and anomaly detection. Throughout the training process, we employ the multi-backpropagation strategy to balance the influence of each model's contributions, where the loss is first back-propagated through each branch according to their respective weights, and then propagated through the entire model. This approach ensures that each model's contributions are appropriately balanced and integrated into the overall training procedure. The training settings are consistent with those used previously.

3. ENSEMBLE

Model ensemble has been proved effective for improving ASD performance and robustness in previous works [21, 11]. We also adopt ensemble techniques in the submitted systems, namely model average, score fusion and group fusion, which are applied progressively.

3.1. Model Average

Averaging the parameters of multiple checkpoints is a vital technique for classification task. In the proposed scheme, we apply model average to the Dual Branch CNN model introduced in Section 2.2, where we average the top 3 checkpoints from a single run. Meanwhile, model average is not applicable for the NF-CDEE model and LoRA models, since applying it deprecates the performances of these models.

3.2. Score Fusion

Score fusion is a commonly adopted technique in previous challenges [11, 22] for combining heterogeneous single models into a powerful ensemble, which is also adopted in the submitted systems. We first calibrate each single model by normalizing the scores via mean and standard deviation. The calibrated scores are then linearly combined into a general score, and the combination coefficients are obtained by grid search on the development set. However, since the greedy grid search tends to overfit on the development set resulting in an imbalance distribution of coefficients, we manually adjust the coefficients for some systems to improve the generalizability.

3.3. Group Fusion

To further improve the robustness, we introduce group fusion which applies score fusion in groups, since the trained-from-scratch models may not scale well on particular machine types. On the one hand, the NF-CDEE model only showcases consistent excellence on the periodic group, while the performances on other groups are severely lagged behind pre-trained models. Therefore, the NF-CDEE model only contributes to the scores of the periodic group, while the combination coefficient of the NF-CDEE model on other groups is set to 0.0. On the other hand, the Dual Branch CNN model yields a naive score distribution for Scanner, thus it is not applied on this machine type.

3.4. Submitted Systems

Table 2 presents the combination coefficients of five single models in four submitted systems. Since three pre-trained models generalize well to all machine types, they are considered as base models and are employed in all submitted models. System 1 combines base models with the Dual Branch CNN model on all machine types except Scanner. System 2 combines base models with NF-CDEE on the periodic group. System 3 combines all available models with the optimal coefficients. System 4 also utilizes all available models, but only applies NF-CDEE on the periodic group and slightly increases the coefficient of BEATs-LoRA.

4. EXPERIMENT

The Receiver Operating Characteristic (ROC) Curve (AUC) and partial AUC (pAUC) are calculated in accordance with the challenge rule, and we report the harmonic mean of AUCs and pAUC for each machine type and the overall harmonic mean across machine types.

Table 1 presents the performances of five single models. Three pre-trained models demonstrate superior performances than two train-from-scratch models, where the Dual Branch BEATs and EAT model showcase the best detection results. The NF-CDEE model only excels on gearbox and slider, thus it is only applied on the periodic group in group fusion.

Table 3 presents the performances of four submitted systems, where system 2 achieves the best detection result of 68.38%.

5. CONCLUSION

This paper depicted the THUEE system for the ASD task, where we employed two train-from-scratch models and three pre-trained models, and ensembled them via model average, score fusion and group fusion. The best system achieved 68.38% on the development set of the DCASE 2024 dataset.

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] J. Lopez, G. Stemmer, and P. Lopez-Meyer, "Ensemble of complementary anomaly detectors under domain shifted conditions," DCASE2021 Challenge, Tech. Rep., July 2021.

[6] A. Jiang, Q. Hou, J. Liu, P. Fan, J. Ma, C. Lu, Y. Zhai, Y. Deng, and W.-Q. Zhang, "Thuee system for first-shot unsupervised anomalous sound detection for machine condition monitoring," DCASE2023 Challenge, Tech. Rep., June 2023.

[7] 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.

[8] Y. Zhang, J. Liu, Y. Tian, H. Liu, and M. Li, "A dual-path framework with frequency-and-time excited network for anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1266–1270.

[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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman, "Pyro: Deep universal probabilistic programming," *J. Mach. Learn. Res.*, vol. 20, pp. 28:1–28:6, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-403.html

[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] W. Junjie, W. Jiajun, C. Shengbing, S. Yong, and L. Mengyuan, "Anomaly sound detection system based on multi-dimensional attention module," DCASE2023 Challenge, Tech. Rep., June 2023.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[18] 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.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[20] 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.

[21] K. Wilkinghoff, "Sub-cluster adacos: Learning representations for anomalous sound detection," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[22] Y. Deng, J. Liu, and W.-Q. Zhang, "Aithu system for unsupervised anomalous detection of machine working status via sounding," DCASE2022 Challenge, Tech. Rep., July 2022.