THE SMALL RICE SUBMISSION FOR THE DCASE 2024 TASK 2: ANOMALOUS SOUND DETECTION USING SUB-CLUSTER NOISY-ARCMIX

Technical Report

Chenyu Liu, Gang Li, Junbo Zhang, Jizhong Liu, Heinrich Dinkel†, Yongqing Wang†, Zhiyong Yan†, Yujun Wang, Bin Wang*

AI Lab, Xiaomi Corporation, China
{liuchenyu3, ligang5, zhangjunbo1, liujizhong1}@xiaomi.com

ABSTRACT

This paper describes our submission for the DCASE 2024 task 2. The objective is identifying whether the sound emitted from a machine is normal or anomalous without having access to anomalous samples. The ASD model we designed to calculate the anomaly scores is a CED based supervised model. To alleviate the problem of domain shifts, we use sub-cluster noisy-arcmix combined with asymmetric focal loss to balance the data weights while learn more compact intra-class representations for normal samples. In addition, we explore data augmentation methods such as manifold mixup and FeatEx to further improve the model perfomance. Our best single model achieves a pAUC of 55.81%, a source domain AUC of 67.79%, and a target domain AUC of 65.88% on the development dataset.

***Index Terms*—** Anomalous Sound Detection, Domain Generalization, First-Shot, Audio Pretrained Model

1. INTRODUCTION

Anomalous sound detection (ASD) has a wide range of applications in the field of machine condition monitoring, enabling the assessment of machine functionality based on sound analysis. This technology can be applied to industrial monitoring, helping to reduce labor costs in factories. It can also be applied to the field of autonomous driving, providing timely warnings of vehicle safety hazards and preventing loss of life and property.

In recent years, the DCASE task 2 has focused on the detection of machine anomaly sounds in first shot scenarios. This task involves the use of development and evaluation datasets with machine classes that are completely different. In addition, there is a domain shifts problem in the training data caused by machine state or environmental changes. Currently, there are two main categories of mainstream ASD methods: supervised learning methods based on a classification pretext task and unsupervised learning methods that utilize AutoEncoders for a reconstruction pretext task [1]. Supervised methods often demonstrate exceptional performance when there is an abundance of labeled data, whereas unsupervised methods can maintain relatively stable performance regardless of the availability of labeled data. However, unsupervised methods generally have inferior upper bounds on performance compared to supervised methods. Due to the fact that annotated attributes information is not always available in reality, the DCASE 2024 task 2 [2] does not provide attributes information for some of the machine types. As a result, the ASD system is required to work well regardless of whether or not machine attributes are provided.

Inspired by recent advancements in large-scale audio pretraining models [3], we propose a supervised ASD method that utilizes CED-base as the backbone network for feature extraction. To address the challenge of data imbalance, particularly domain shifts, we introduce a sub-cluster noisy-arcmix combined with asymmetric focal loss. This approach aims to balance sample weights while learning more compact intra-class representations for normal data and pushing away representations of anomalous data. Furthermore, we introduce time-shift, FeatEx [4], and two mixup [5] strategies for data augmentation to enhance the model's generalization ability by increasing the difficulty of the classification task.

The paper is structured as follows: In Section 2 we introduce our ASD method. In Section 3 details regarding the experimental setup and results are provided. The conclusion is given in Section 4.

2. PROPOSED METHOD

2.1. CED-based Supervised Classification

Our ASD model is trained using a classification-based supervised learning approach, where the joint labeling of machine type and attributes is used as the classification target. For machine types that do not have attributes, it is assumed that the machine type has only one attribute with the value 'None'. In the training phase, the model learns the characteristic patterns of normal machine audios by performing the classification task. In the testing phase, the model calculates the anomaly score by measuring the cosine distance between samples and the normal samples in the learned feature space.

We expect to exploit the generalization ability of models pretrained on large-scale audio data. Therefore, we use the pre-trained CED-base [6] model as our audio encdoer and fine-tune it with the training data. For encoder outputs, we utilize a Sub-Cluster Noisy-arcmix combined with Asymmetric focal Loss (SC-NAL), which helps to improve the compactness of intra-class representations and alleviate the issue of domain shifts.

2.2. Data Augmentation

To further enhance the generalization ability of the model, we apply time shift, mixup and FeatEx for data augmentation. In addition to the standard mixup, we also utilize manifold mixup [7].This method

*Corresponding author.
†Equal contribution.

	Metric	baseline (MSE)	baseline (MAHALA)	Our (32sc)	Our (32sc+FeatEx)	Our ensemble
ToyCar	AUC(source)	**66.98%**	63.01%	48.60%	53.60%	55.28%
	AUC(target)	33.75%	37.35%	55.36%	**60.84%**	55.48%
	pAUC	48.77%	**51.04%**	49.53%	49.26%	50.84%
ToyTrain	AUC(source)	**76.63%**	61.99%	65.20%	68.60%	70.24%
	AUC(target)	46.92%	39.99%	57.92%	**62.24%**	59.08%
	pAUC	47.95%	48.21%	53.68%	53.79%	**55.11%**
bearing	AUC(source)	62.01%	54.43%	51.32%	60.32%	**64.08%**
	AUC(target)	61.40%	51.58%	66.96%	67.88%	**70.52%**
	pAUC	57.58%	58.82%	56.53%	53.84%	**60.05%**
fan	AUC(source)	67.71%	**79.37%**	59.88%	60.84%	59.84%
	AUC(target)	55.24%	42.70%	69.04%	68.96%	**70.60%**
	pAUC	57.53%	53.44%	57.37%	58.21%	**58.79%**
gearbox	AUC(source)	70.4%	**81.82%**	74.28%	77.48%	78.12%
	AUC(target)	69.34%	74.35%	70.64%	75.32%	**75.96%**
	pAUC	55.65%	55.74%	56.42%	**60.16%**	56.16%
slider	AUC(source)	66.51%	75.35%	81.92%	76.08%	**86.96%**
	AUC(target)	56.01%	68.11%	64.12%	60.40%	**67.12%**
	pAUC	51.77%	49.05%	50.95%	51.47%	**52.63%**
valve	AUC(source)	51.07%	55.69%	89.96%	91.28%	**91.32%**
	AUC(target)	46.25%	53.61%	**68.28%**	68.12%	63.44%
	pAUC	52.42%	51.26%	63.89%	**67.89%**	64.95%
All(hmean)	AUC(source)	65.00%	65.77%	64.29%	67.79%	**70.07%**
	AUC(target)	50.28%	49.51%	64.13%	**65.88%**	65.35%
	pAUC	52.84%	52.28%	55.14%	55.81%	**56.60%**

Table 1: Main results proposed in our work for the DCASE 2024 Task 2 challenge on the development dataset

involves interpolating at the embedding level, aiming to further refine the decision boundary. The mixed embedding is calculated as follows:

$$\begin{aligned} \tilde{e} &= \lambda e_i + (1-\lambda) e_j \\ \tilde{y} &= \lambda y_i + (1-\lambda) y_j \end{aligned} \tag{1}$$

Where e_i and e_j represent the embedding of two random training samples, and y_i and y_j are the corresponding labels. During training, we apply standard mixup with a probability of $p_1 = 0.25$, and Manifold mixup with a probability of $p_2 = 0.25$.

Furthermore, in order to enhance the complexity of the classification task, we apply a simplified FeatEx. FeatEx was originally designed for CNN with two sub-networks to generate a new embedding by exchanging the outputs of different samples in each branch. It also incorporates label expansion to simulate anomaly classes. However, our experiments revealed that label expansion has limited effects on the model. So as to utilize the noisy-arcmix [8] loss, we choose not to perform label expansion. Additionally, as CED does not have a two-branch structure, we replicated the model output twice to simulate two branches. The simplified FeatEx calculates a new augmented embedding as follows:

$$\begin{aligned} e_{\text{new}} &= (e_i, e_j) \in \mathbb{R}^{2D} \\ y_{\text{new}} &= (0.5 \cdot y_i + 0.5 \cdot y_j) \in [0,1]^N \end{aligned} \tag{2}$$

Here, e_i and e_j denote the embedding of two random training samples, and y_i and y_j are the corresponding labels. For the samples that are not enhanced by simplified FeatEx, the new embedding is defined as $e_{\text{new}} = (e, e)$, while the label y remains the same.

2.3. Sub-Cluster Noisy-Arcmix with Asymmetric Focal Loss

In the DCASE task 2, the training data only consists of normal machine audios. To improve the model's ability for distinguish abnormal audio during the testing phase, it is necessary for the model to learn compact representations. To this end, we apply noisy-arcmix, which combines the benefits of mixup and ArcFace [9]. Choi et al. demonstrated that noisy-arcmix can effectively compact the distance between intra-class normal samples without significantly affecting the proximity of abnormal samples. In addition, inspired by SCAdaCos [10], we model multiple sub-clusters for each class. After applying noisy-arcmix, we sum the softmax scores of all sub-clusters for each class to encourage our model learning more complex distributions. The calculation of the sub-cluster noisy-arcmix loss for sample x is as follows:

$$L_{SC\text{-}NAL}(x,y) = -y^{\mathsf{T}} log \frac{e^{cos(\theta + my)}}{\sum_{k=1}^{NS} e^{cos(\theta_k + m y_k)}} \tag{3}$$

Here, N is the number of classes and S is the number of sub-clusters.

To alleviate the issue of data imbalance, where there are only 10 target domain samples out of 1000 training samples for each machine type, we use asymmetric focal loss [11] instead of the standard cross-entropy loss, which can flexibly adjust the loss weights according to the training difficulty, assigning higher weights to samples that are harder to classify.

		SC(FeatEx)			
		1	16	32	64
SC	1	61.37%	61.41%	60.76%	60.66%
	16	60.34%	61.67%	61.84%	60.74%
	32	59.48%	61.52%	61.81%	61.92%
	64	60.42%	60.34%	60.01%	60.15%

Table 2: mean AUCs for different sub-cluster settings

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

The data used for this task including two recent machine audio datasets, ToyADMOS2 [12] and MIMII DG [13]. We train our model on both the development dataset and the additional training dataset. The development dataset contains audio recordings from 7 machine types that are different from the evaluation dataset, while the additional training dataset includes audio recordings from 9 machine types that are the same as the evaluation dataset. All training audios have a sample rate of 16kHz and durations ranging from 6s to 12s. We normalize all audios to a duration of 10s. For audios shorter than 10s, we perform copy padding, while for audios longer than 10s, we randomly select a 10s segment.

During the training phase, we employ the AdamW optimizer to train the model for a total of 20 epochs, with a batch size of 64. For fine-tuning the CED-base, we follow the pre-training settings and extract the mel-spectrogram of the audio using a mel bin size of 64, an fft size of 512, and a hop size of 160. In the testing phase, we determine the anomaly score by calculating the minimum cosine distance between the test sample and all centroids of the source domain samples, as well as all target samples.

3.2. Results

Table 1 presents a comparison between our ASD model and the baseline [14] on the development dataset. The results demonstrate that our method achieves higher average scores across all metrics, with the most significant improvement on the AUC target. SC-NAL can effectively enhance the model's performance on the target domain, even without balanced sampling. Additionally, the simplified FeatEx contributes to a slight improvement across most metrics.

Table 2 presents the mean values of all the AUC metrics under different sub-cluster settings, including AUC, pAUC, AUC source, pAUC source, AUC target and pAUC target. It is can be seen that the performance of the model shows variability when different numbers of sub-clusters are employed. A favorable performance can be achieved by appropriately increasing the sub-cluster number. However, setting excessively large sub-cluster number can also negatively affect the results.

Observing that there are variations in the model's performance across machine types with different sub-cluster settings, we choose several different single models used for ensemble. Our submissions system S1, S2 and S4 adopt different model fusion methods, while S3 is the optimal single model.

4. CONCLUSIONS

In this paper, we introduce our submissions for the DCASE 2024 task 2. We focus on developing an effective supervised ASD model by utilizing a large-scale pre-trained CED-based encoder. We also incorporate sub-cluster noisy-arcmix with asymmetric focal loss to mitigate the domain shifts problem. The experimental results show that our best single model improves against the baseline(MSE) by an average AUCs of 7.12% on the development dataset.

5. REFERENCES

[1] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on dcase 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2305.07828*, 2023.

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

[3] B. Han, Z. Lv, A. Jiang, W. Huang, Z. Chen, Y. Deng, J. Ding, C. Lu, W.-Q. Zhang, P. Fan, *et al.*, "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. 1326–1330.

[4] 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.

[5] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[6] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "Ced: Consistent ensemble distillation for audio tagging," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 291–295.

[7] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International conference on machine learning*. PMLR, 2019, pp. 6438–6447.

[8] S. Choi and J.-W. Choi, "Noisy-arcmix: Additive noisy angular margin loss combined with mixup for anomalous sound detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 516–520.

[9] 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.

[10] 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.

[11] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 82–91.

[12] 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.

[13] 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.

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