ANOMALOUS SOUND DETECTION BASED ON PSEUDO LABELS FROM GUIDED CLUSTERING

Technical Report

Yaocong Wang[1]*, Xinlong Deng*[1]*, Jie Jiang*[1*]*, Qiuqiang Kong*[2]

[1] China University of Petroleum (Beijing), College of Artificial Intelligence, Beijing, China
{2023216511, 2022012257}@student.cup.edu.cn, jiangjie@cup.edu.cn
[2] The Chinese University of Hong Kong, Department of Electronic Engineering, Hong Kong, China
qqkong@ee.cuhk.edu.hk

ABSTRACT

This technical report presents a description of the CUP submission for Task 2 "first-shot unsupervised anomalous sound detection for machine condition monitoring" of the DCASE 2024 Challenge. The submitted system is an adaptation of a previously proposed model which utilizes static and dynamic frequency information and is trained through an auxiliary classification task with sub-cluster AdaCos loss. In this work, we focus on the clustering of machine sound clips under attribute-unavailable conditions such that attribute classification based methods can be extended to machine sound clips without attribute information for detecting anomalous sounds.

***Index Terms*—** anomalous sound detection, clustering, first-shot classification

1. INTRODUCTION

Anomalous sound detection (ASD) has been a task at the DCASE challenge since 2020 [1, 2, 3, 4]. It aims at determining whether the sound emitted from a target machine is normal or anomalous. For the ASD tasks at DCASE, only normal data are provided for model training and different requirements were introduced including domain shifts between a source domain and target domain, mutually exclusive machine types in development and evaluation set [2, 3, 4, 5, 6]. The focus of this year's ASD task is the same as that of 2023 with one modification, i.e., the additional attribute information for some machine types is not provided [7].

In this work, we focus on how subcategories of machine sound clips can be obtained when attribute information is not available for some machine types. That is, given a set of machine sound clips, we try to cluster these machine sound clips into subcategories. In this way, the existing attribute classification based ASD methods can also be applied to the machine types without attribute information, such as the ASD method proposed by Wilkinghoff in the DCASE 2023 challenge [8]. Concretely, we first leverage a large-scale dataset such as AudioSet [9] to pre-train a model to obtain audio features by alternating between clustering of the audio samples by their embeddings and updating the weights of the the model by predicting the cluster assignments. Secondly, we fine-tune the model using the sound clips from the machines with their attribute information as the classification labels. Moreover, we use a modified loss that considers both classification accuracy and distances of contrastive samples. Thirdly, we use the fine-tuned model to cluster the sound clips from the machines without attribute information to generate pseudo-subcategories of these machines. Finally, the subcategories of the sound clips of all the machine types can be used to train an attribute classification based ASD model.

*Thanks to NSFC for funding.

2. BASELINE METHODS

The DCASE 2024 challenge Task 2 organizers provide a baseline system containing two different methods [10]. The first method is based on a simple autoencoder which calculates the anomaly score in terms of the reconstruction error of a machine sound clip. The second method is based on a selective Mahalanobis which calculates the anomaly score in terms of the reconstruction error of a machine sound clip in the Mahalanobis metric. In addition, we use the ASD method proposed by Wilkinghoff in the DCASE 2023 challenge [8] as another baseline.

3. PROPOSED METHOD

The proposed ASD method consists of four steps. The framework of the first two steps is shown in Fig. 1. First, we adapt the DeepCluster method [11] from the image clustering field to pre-train an embedding network based on AlexNet [12] to learn useful audio features using the AudioSet dataset [9]. The pre-training iteratively groups the audio features with a clustering algorithm, k-means, and uses the subsequent assignments as pseudo-labels to update the weights of the network in a supervised manner. Then, we fine-tune the pre-trained network with two sub-tasks using the sound clips from machines with attribute information. The first sub-task is predicting the sub-category to which an audio clip belongs. The true label is obtained by grouping sound clips into sub-categories based on their attribute information. The loss of the first sub-task is:

$$\mathcal{L}_a = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp(x_{n,y_n})}{\sum_{c=1}^{C}\exp(x_{n,c})} \tag{1}$$

where C is the number of sub-categories, N is batch size, y_n is the true label of x_n.

The second sub-task measures the quality of the embedding network by the distances between pairs of machine sound clips with the following contrastive loss:

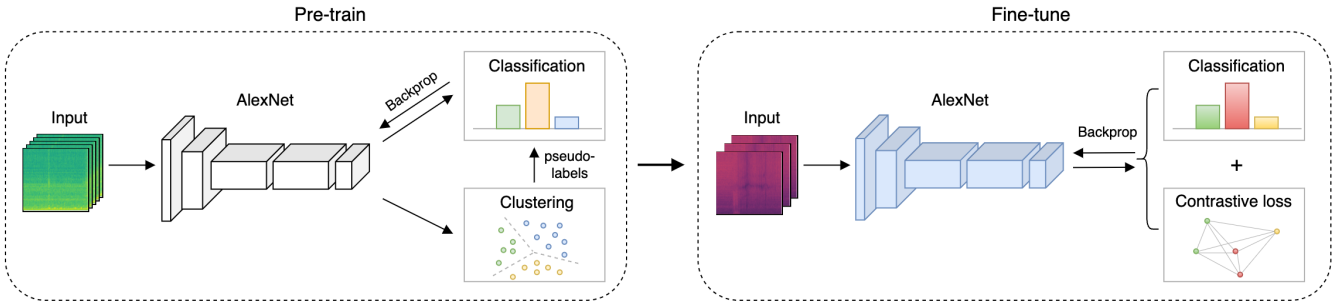

Figure 1: The first steps of the proposed ASD method.

$$\mathcal{L}_b = \frac{1}{N} \sum_{1 \le n \le N} \mathbb{1}_{c_n = c_{n'}} d(x_n, x_{n'})^2 + \mathbb{1}_{c_n \neq c_{n'}} \frac{1}{d(x_n, x_{n'})^2} \quad (2)$$

where d is a function that returns the euclidean distance between the embeddings of two machine sound clips, n' is an index sampled from 1 to N. The first part of $\mathcal{L}_b$ is applied when the two sound clips belong to the same sub-category. The second part of $\mathcal{L}_b$ is applied when the two sound clips come from different sub-categories. In this way, the embedding network will minimize the distance between pairs of samples from the same sub-category while maximize the distance between pairs of samples from different sub-categories.

The final loss is a combination of $\mathcal{L}_1$ and $\mathcal{L}_2$ as follows:

$$\mathcal{L} = \mathcal{L}_a + \lambda \mathcal{L}_b \quad (3)$$

where λ is a hyper-parameter to balance the two loss terms.

At the third step, we take the sound clips from the machines without attribute information and obtain their embedddings from the fine-tuned embeddding network. Then we carry out k-means clustering with the embedddings to obtain pseudo sub-categories for the machine sound clips without attribute information. Finally, we use the sound clips of all machine types with their true or pseudo sub-category labels to train an attribute classification based model to detect anomalous sound. In this work, we choose the attribute classification based ASD model proposed by Wilkinghoff in the DCASE 2023 challenge [8].

For all the four steps aforementioned, the model inputs are spectrograms of the raw waveform data. The data pre-processing follows the procedure in [8].

4. RESULTS

Table 1 shows the experiment results of the proposed method together with the three baseline methods on the development set. It can be seen that the proposed method achieves the best average AUC and pAUC over all the 7 machine types. In specific, the proposed method and the baseline [8] output the other two baselines with a large margin for Slider and Valve, with the proposed method achieves the best AUC for both source and target domains as well as the best pAUC. As for ToyCar and Fan, the proposed method achieves higher AUC for the target domain as well as higher pAUC. In the case of Bearing, the proposed method achieves similar results as the baseline MSE and outperforms the baseline [8]. For Gearbox, the baseline MAHALA achieves the best performance. For ToyTrain, the baseline MSE achieves the best AUC for the source domain, while the baseline [8] and the proposed method achieves similar target domain AUC and pAUC which outperform the other two baselines. In particular, the proposed method improves the AUC and pAUC by a large margin for Slider which has no attribute information.

Table 1: AUCs and pAUCs per machine type on the development set obtained by different methods. The last row shows the harmonic mean over all machine types. Highest AUCs and pAUCs in each row are highlighted in bold letters.

	Method	Baseline MSE	Baseline MAHALA	Baseline [8]	Proposed method
ToyCar	AUC(source)	**66.98%**	63.01%	46.44%	51.04%
	AUC(target)	33.75%	37.35%	44.2%	**46.08%**
	pAUC	48.77%	51.04%	49.05%	**49.31%**
ToyTrain	AUC(source)	**76.63%**	61.99%	36.56%	56.48%
	AUC(target)	46.92%	39.99%	**57.96%**	55.72%
	pAUC	47.95%	48.21%	50.47%	**51.47%**
Bearing	AUC(source)	62.01%	54.43%	58.16%	**62.56%**
	AUC(target)	61.4%	51.58%	56.52%	**63.28%**
	pAUC	57.58%	**58.82%**	54.37%	55.10%
Fan	AUC(source)	67.71%	**79.37%**	51.24%	57.28%
	AUC(target)	55.24%	42.7%	61.04%	**68.04%**
	pAUC	57.53%	53.44%	51.31%	**59.89%**
Gearbox	AUC(source)	70.4%	**81.82%**	61.96%	76.67%
	AUC(target)	69.34%	**74.35%**	71.0%	62.56%
	pAUC	55.65%	**55.74%**	51.16%	50.79%
Slider	AUC(source)	66.51%	75.35%	83.16%	**93.6%**
	AUC(target)	56.01%	68.11%	80.76%	**93.16%**
	pAUC	51.77%	49.05%	56.79%	**71.57%**
Valve	AUC(source)	51.07%	55.69%	92.52%	**96.16%**
	AUC(target)	46.25%	53.61%	67.44%	**71.12%**
	pAUC	52.42%	51.26%	64.26%	**69.58%**
All (hmean)	AUC(source)	65.00%	65.77%	56.19%	**66.75%**
	AUC(target)	50.28%	49.51%	60.74%	**63.10%**
	pAUC	52.84%	52.28%	53.52%	**57.10%**

5. REFERENCES

[1] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo,

M. Yasuda, and N. Harada, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 81–85.

[2] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on dcase 2021 challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 186–190.

[3] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[4] 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," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, Tampere, Finland, September 2023, pp. 31–35.

[5] 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.

[6] 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.

[7] 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:2406.07250*, 2024.

[8] K. Wilkinghoff, "Fraunhofer fkie submission for task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," DCASE2023 Challenge, Tech. Rep, Tech. Rep., 2023.

[9] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[10] 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.

[11] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *European Conference on Computer Vision*, 2018.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.