ANOMALOUS SOUND DETECTION BASED ON UNSUPERVISED LEARNING AND ENSEMBLE METHOD

Technical Report

Kai Guo, Fengrun Zhang

Beijing Institute of Technology
School of Information and Electronics
Beijing, China
guokai001123@126.com, 3120220766@bit.edu.cn

ABSTRACT

Over the past few years, self-supervised methods have significantly advanced the field of Anomalous sound detection. However, the real-world application of these methods is often hampered by the lack of available prior knowledge and auxiliary labels, which restricts the models' ability to generalize. In the DCASE 2024 Challenge Task 2, attribute information for several machine types is absent from both the development and evaluation datasets, reflecting the realities of practical scenarios. To solve the above problems, we design two unsupervised machine anomalous sound detection models and an ensemble system that achieves better performance than the baseline model.

Index Terms— Anomalous Sound Detection, Unsupervised Learning, Diffusion, Audio Pretrained Model, GMM

1. INTRODUCTION

With the development of industrial technology, it is necessary to ensure the continuous and stable operation of industrial machines. In recent years, with the development of machine learning and deep learning technology, the method of using sound features to determine whether an industrial machine is functioning properly has become popular, and this method is often referred to as Anomalous Sound Detection (ASD).

The goal of Detection and Classification of Acoustic Scenes and Events (DCASE) task 2 [1] is to train a model using the sound of the machine when it is working normally, and finally the model has the ability to judge abnormal sounds. As of 2023, the competition hopes that the methods proposed by the contestants can have a certain domain generalization ability, and at the same time, they can have a certain ability to deal with first-shot problems [2]. As a result, in 2023, the development dataset and evaluation dataset of the task contain completely different machine types, which makes it impossible for entrants to set hyperparameters for each machine type for better overall performance.

Nonetheless, there is one category of approaches that have remained popular over the past few years, and that is the self-supervised learning method based on machine attributes [3]. People usually use the attributes of the machine as a label to train a sound feature extractor and classifier, and calculate the anomaly score of the sound by clustering latent features [4] or indirectly using the output results of the classifier [5]. However, attribute information for real-world machine sounds are not always easy to obtain, for example, some machines always operate under the same conditions, or the operating conditions of the machines cannot be measured by a fixed value. In order to simulate the difficulty of obtaining attribute information for machines in reality, the attribute information of some machine types has been removed for the first time in the 2024 competition [6].

In this case, we design machine abnormal sound detection methods based on the following assumptions: First, the attribute information of the machine is always difficult to obtain, so we do not use any attribute information in the development set. Second, the model needs to be able to handle first-shot problems, so we set the same hyperparameters for all machine types, both for the development set and the evaluation set.

Fine-tuning the pre-trained model is a self-supervised abnormal sound detection method with certain generality [7]. In the past, people often used the machine's attribute information to fine-tune the pre-trained model like HuBERT [8]. In this competition, we use machine type classification as an auxiliary task to fine-tune the pre-trained audio tagging model CED [9], and achieve better performance than the baseline model [10]. Diffusion models [11] are widely used in computer vision fields such as image anomaly detection. We explore the use of a diffusion model for anomalous sound detection, and improve the performance of the model through masking. The popular deep learning model has large parameters and slow training speed, TWFR-GMM [12] has proven to be an efficient method for anomalous sound detection based on unsupervised machine learning.

The paper is organized as follows: Section 2 describes our submitted four systems. The first three systems are named AnoCED, CMDM, and TWFR-GMM [12], respectively, while the fourth system is an ensemble system of the first three systems. Section 3 describes the results of four systems on the development dataset and our discussions.

2. SYSTEM SUBMISSION

The development dataset consists of two datasets: MIMII DG [13] and ToyADMOS2 [14]. We all use mel-spectrogram as the input feature of the model, but the parameters are different for different models. For AnoCED we follow the default parameters of the CED [9], setting the window length to 512, the hop length to 160, and the number of mel filterbanks to 64. For CMDM, we set the window length to 1024, the hop length to 512, and the number of mel filterbanks to 128 to obtain the feature size suitable for training the

diffusion model. For TWFR-GMM, we use the same parameters as CMDM.

2.1. AnoCED

CED [9] is a straightforward training framework that refines student models through large teacher ensembles using consistent instruction. The model was pre-trained in the AudioSet [15]. We choose the CED-Mini for fine-tuning because of its combination of performance and small size.

We simply replace the output layer of the original model with a new linear layer, and we fine-tune the model to have the ability to distinguish the machine type of input audio machine. For example, the output size of the model is 7 for the development set and 9 for the evaluation set. We use the cross-entropy loss to train the model. We use the Adam optimizer [16] and the learning rate is 1e-4. Our model was trained for 200 epochs.

After the model are trained, we use ensemble method to calculate the anomaly score. We use a combination of PCA, KNN and model linear layer outputs to evaluate whether there are anomalies in the sound. We use the embedding of all the training samples extracted from the CED model to train the PCA and KNN models respectively. Let S_{KNN} and S_{PCA} be the anomaly score calculated by the KNN model and the PCA model, respectively, and $S_{Classifier}$ is the negative log-softmax of the model output, we calculate the anomaly score via:

$$S_{AnoCED} = S_{KNN} + S_{PCA} + 1000S_{Classifier} \tag{1}$$

2.2. CMDM

We introduce the Context-Masking Diffusion Model (CMDM) for unsupervised anomalous sound detection. We conceptualize the diffusion process as masks that introduce noise into specific areas of the audio features. The denoising phase of the Denoising Diffusion Probabilistic Models (DDPM) [11] then leverages both the contextual information from the unmasked areas and the noise information from the masked areas to recover the original signal. Furthermore, recognizing the varied nature of potential anomalies across different types of machine, we have developed versatile masking strategies that span across different dimensions—time frames (T) and patches (P) to enhance the robustness and efficacy of the model in detecting anomalies in diverse machine sounds.

In our method, each input data x is partitioned into L positions with two strategies: partitioning from the dimensions of T and P. We uniformly sample positions with an interval l from x with a rate of λ at a fixed grid to obtain the masked positions m_l. During the forward diffusion process, noise is only added to x^{ij} within m_l, where x^{ij} is the element of x. Consider $M_L \in R^{T,F}$ a binary masking matrix where the elements overlap with m_l are set to one while others are set to zero, representing whether to add noise. A context-masking process to obtain the sample $\tilde{x}_t$ is introduced after every forward and reverse step via:

$$\tilde{x}_t = x_t \odot M_L + x_0 \odot \neg M_L \tag{2}$$

where $\odot$ denotes element-wise multiplication and $\neg$ denotes element-wise negation. In the training phase, $\tilde{x}_t$ is fed into $\epsilon_\theta(\tilde{x}_t, t)$ to estimate the given noise of both noisy positions M_L and original positions $\neg M_L$ via:

$$\mathcal{L}_{denoising} = \begin{cases} \left\| \epsilon_t - \epsilon_\theta(x_t, t) \right\|^2, x^{ij} \in M_L \\ \left\| \epsilon_\theta(x_t, t) \right\|^2, x^{ij} \in \neg M_L \end{cases} \tag{3}$$

where the prediction error of both positions is minimized. The architecture of denoising U-Net we used is same as [17]. We divide the Mel spectrogram features into 128×128 fragments as model inputs. The Adam optimizer with a learning rate of 1e-4 is used for optimization. For each machine type, the model is trained with 64000 steps. For more details, you can refer to the original paper [11] or our technical report from last year [18]. We used the mean of the top 64 absolute errors in the input and output of the model as the sound anomaly score. Let S_T and S_P be the anomaly score calculated by two models trained using T and P mask strategies, respectively. We calculate the final anomaly score via:

$$S_{CMDM} = 0.75S_T + S_P \tag{4}$$

2.3. TWFR-GMM

We simply reproduce the TWFR-GMM proposed by Guan *et al.* [12]. We believe TWFR-GMM is a highly efficient machine learning method for detecting anomalous sounds. We have conducted a number of experiments using the Time-Weighted Frequency Domain Representation proposed in the original paper, including training an autoencoder or a self-supervised neural network using TWFR. But in fact, simply using GMM for anoamlous sound detection can achieve the best performance.

We don't use any hyperparameter search techniques. For all machine types, we set the number of mixture components of GMM to 2 and the decay weighting parameter of GWRP to 1. SMOTE [19] is employed to mitigate sample insufficiency by over-sampling the samples across all machine types in the target domain.

2.4. Ensemble system

We use a model ensemble technique based on weighted anomaly scores to achieve better model robustness. Let S_{CED} , S_{GMM} and S_{CMDM} be the anomaly score calculated by the AnoCED model, TWFR-GMM model and CMDM respectively, we calculate the anomaly score for the ensemble system via:

$$S_{ensemble} = S_{CED} + S_{GMM} + 1000S_{CMDM} \tag{5}$$

3. RESULTS AND DISCUSSIONS

The results of our anomalous sound detection experiments on the development dataset are shown in table 1. In general, the proposed CMDM system achieves the best harmonic average performance. Especially for the target domain, the CMDM achieves the best AUC on four machine types, indicating that the proposed CMDM has a certain domain generalization ability. Although the proposed AnoCED method does not achieve the best hmean performance, the overall performance is more stable, and the model complexity and training cost are much smaller than those of CMDM. The TWFR-GMM method appears to have achieved the highest number of best results, however, it has the worst hmean performance. This is due to its low tAUC on some machine types, which proves that its domain generalization ability needs to be improved. The ensemble model achieves the best sAUC on some machine types, but the overall hmean performance is not the best, because on some machine types, the model with poor domain generalization ability drags down the model with better performance. Therefore, in the future, we will focus on how to use model ensemble methods to achieve better domain generalization capabilities.

4. CONCLUSION

In DCASE 2024 task 2, we propose two unsupervised machine anomaly sound detection algorithms that do not make use of machine attribute information: AnoCED and CMDM, and introduce TWFR-GMM [12] to design ensemble system. The proposed method has certain performance, and we will further improve the domain generalization ability of the model in the future.

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)*, November 2020, pp. 81–85. [Online]. Available: http://dcase.community/documents/workshop2020/proceedings/DCASE2020Workshop_Koizumi_3.pdf

[2] 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 arXiv e-prints: 2305.07828*, 2023.

[3] R. Giri, S. V. Tenneti, K. Helwani, F. Cheng, U. Isik, and A. Krishnaswamy, "Unsupervised anomalous sound detection using self-supervised classification and group masked autoencoder for density estimation," DCASE2020 Challenge, Tech. Rep., July 2020.

[4] K. Wilkinghoff, "An outlier exposed anomalous sound detection system for domain generalization in machine condition monitoring," DCASE2022 Challenge, Tech. Rep., July 2022.

[5] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 816–820.

[6] 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.

[7] 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)*, 2024, pp. 1326–1330.

[8] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[9] 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)*, 2024, pp. 291–295.

[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] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[12] J. Guan, Y. Liu, Q. Zhu, T. Zheng, J. Han, and W. Wang, "Time-weighted frequency domain audio representation with gmm estimator for anomalous sound detection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 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, 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.

[15] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6628106

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[18] F. Zhang, C. Hu, and K. Guo, "Unsupervised learning for anomalous sound detection based on prediction and reconstruction tasks," DCASE2023 Challenge, Tech. Rep., June 2023.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, no. 1, p. 321–357, jun 2002.

Table 1: Results of submitted systems on DCASE 2024 task 2 development set. Best in bold.

Machine	AnoCED			CMDM			TWFR-GMM			Ensemble		
	sAUC↑	tAUC↑	pAUC↑	sAUC↑	tAUC↑	pAUC↑	sAUC↑	tAUC↑	pAUC↑	sAUC↑	tAUC↑	pAUC↑
bearing	**60.94**	63.98	54.63	60.65	**65.84**	57.00	53.24	58.60	**59.11**	60.16	63.70	56.37
fan	68.88	43.12	54.11	57.62	**72.54**	56.68	**79.44**	32.40	50.84	69.58	42.62	**57.11**
gearbox	71.88	68.32	**58.42**	57.02	60.12	52.74	**83.50**	**79.12**	57.79	74.66	72.42	57.95
slider	81.82	54.24	49.16	85.74	55.82	55.05	80.90	**75.12**	**58.32**	**92.22**	48.26	48.89
ToyCar	50.10	43.84	47.58	53.84	**48.62**	48.84	**62.62**	33.70	**51.58**	55.64	47.42	48.84
ToyTrain	63.04	50.42	49.05	76.34	**55.94**	**51.84**	73.30	45.14	49.16	**77.38**	50.54	49.74
valve	59.22	**68.00**	**56.95**	59.46	59.67	50.53	52.12	46.44	51.11	**64.62**	65.32	54.68
hmean		56.42			**57.99**			54.96			57.80	