FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION USING ATTRIBUTE CLASSIFICATION AND CONDITIONAL AUTOENCODER

Technical Report

Lei Wang[1], Fan Chu[1], Yuxuan Zhou[1], Shuxian Wang[2], Zulong Yan[3], Shifan Xu[3], Qing Wu[1], Mingqi Cai[4], Jia Pan[4], Qing Wang[2], Jun Du[2], Tian Gao[4], Xin Fang[4], Liang Zou[3]

[1] National Intelligent Voice Innovation Center, Hefei, China
{leiwang32, fanchu, yxzhou15, qingwu6}@nivic.cn
[2] University of Science and Technology of China, Hefei, China
{sxwang21}@mail.ustc.edu.cn, {qingwang2, jundu}@ustc.edu.cn
[3] China University of Mining and Technology, Xuzhou, China
{zulongyan, sfxublues, liangzou}@cumt.edu.cn
[4] iFLYTEK, Hefei, China, {mqcai, jiapan, tiangao5, xinfang}@iflytek.com

ABSTRACT

This technical report outlines our solution to DCASE 2023 Challenge Task 2, First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. This year's task focuses on the first-shot problem: the development dataset and the evaluation dataset have completely different sets of machine types, and each machine type contains only one section. We propose an anomaly detection method based on attribute classification and conditional autoencoder. The attribute classification method includes model pre-training, embedding extraction and inlier modeling, and the conditional autoencoder uses attribute information as conditions. The proposed system achieves 78.35% in the harmonic mean of all machine types, sections, and domains for the area under the curve (AUC) and partial AUC (p = 0.1) on the development set.

Index Terms— DCASE, unsupervised anomalous sound detection, first-shot, attribute classification, autoencoder

1. INTRODUCTION

In DCASE challenge 2023 Task 2 "*First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring*" [1, 2], it is required to detect anomalous sounds of machines. In real-world conditions, it is often easier for us to obtain the sound of the machine working normally, while the anomalies are rare and highly diverse. Therefore, we need to use the normal sounds in the training data to detect anomalous sounds in the test data. Furthermore, the operational states of a machine or the environmental noise can change to cause domain shifts. The system needs to use domain generalization techniques to handle frequent or hard-to-notice domain shifts. In the DCASE 2023 task, first-shot problem is introduced, that is, we need to train a model for a completely new machine type, and can only use a limited number of machines from its machine type.

Our submission includes two major approaches for anomalous sound detection. The first method is based on machine attribute classification. The second approach is to detect anomalies with the conditional autoencoder using machine attribute information as conditions, and the Mahalanobis distance is used to calculate the anomaly score.

In the following, we describe each approach and our experimental results in detail. Each recording used in this challenge is a single-channel and 10-second long audio. The development set includes seven machines: ToyCar, ToyTrain, Fan, Gearbox, Bearing, Slide rail and Valve, and the evaluation set includes seven new machines: Vacuum, ToyTank, ToyNscale, ToyDrone, Bandsaw, Grinder, and Shaker [3, 4].

2. PROPOSED APPROACH

2.1. Attribute Classification

According to the results of previous challenges [5, 6, 7], methods based on self-supervised classification usually achieve better performance [8, 9, 10]. For this year's task, although there is only one section for each machine, we can train a classifier with machine attribute information. Specifically, in order to get a more robust anomaly detector, first, we use the training data of all machines in the development set to train a 14-category domain classifier, and then we fine-tune the model parameters to get the attribute classifier for each machine. Each attribute of the machine has a classification head, and there is also a classification head for distinguishing positive machines from negative machines (the other six types of machines). Then, based on the attribute classifier, we extract embeddings of training data to train inlier models (IM) to model the probability distribution of normal data. In the inference stage, after each test data embedding is extracted, data that deviates from the probability distribution of normal data is detected as abnormal.

2.1.1. Acoustic Features

We transformed all audio clip into spectrograms with a Mel transformation. At the same time, Liu et al. [11] proposed the STgram structure, which can use time-domain information to complement the spectrogram. Therefore, we extract temporal features from the raw wave and then concatenate it with the mel spectrogram. In addition, in order to improve the generalization of the model

Table 1: Results of the attribute classification method on the development set (%). "AUC-S" and "AUC-T" represent the AUC of the source and target domains, respectively.

		ToyCar	ToyTrain	bearing	fan	gearbox	slider
Model 1	AUC-S	67.71	73.40	84.62	82.94	76.56	99.46
	AUC-T	66.54	63.20	75.08	83.94	76.00	90.66
	pAUC	53.89	54.26	64.59	73.26	68.89	82.05
Model 2	AUC-S	66.56	68.86	88.36	95.74	78.39	99.66
	AUC-T	67.61	61.38	77.86	83.44	74.83	93.86
	pAUC	58.11	50.95	74.42	80.37	62.37	88.89
Model 3	AUC-S	68.48	75.78	78.64	84.96	79.08	99.52
	AUC-T	62.70	65.86	74.94	69.48	79.28	92.56
	pAUC	59.21	54.79	67.00	58.89	59.42	82.74

and the representation of the feature, we use the pre-trained model wav2vec [12] to extract feature vectors to connect with the above features.

2.1.2. Training and Results

We choose EfficientNet-B0 [13] as the network structure for domain classification and attribute classification, and mixup [14] is used for data augmentation. Further, a domain generalization strategy is applied, that is, when creating a mini-batch, we sample normal data in the target domain to ensure that there are two target domain samples in the mini-batch. AdamW [15] optimizer is used with the OneCycleLR scheduler for 300 epochs, and the initial learning rate is 0.001. The batch size is set to 128, and BCEWithLogitsLoss is adopted. LOF, KNN, and GMM are used as IM to model the probability distribution of normal data and then calculate anomaly scores. On this basis, we employ manifold mixup [16] to improve the generalization of the model, and then add batch hard triplet loss [17] to improve the anomaly detection performance, and try to use 3-channel features (Mel spectrogram, Tgram features and wav2vec pre-trained features) to replace the mel spectrogram. The three models we used are summarized as follows:

- **Model 1**: Mel spectrogram, BCEWithLogitsLoss
- **Model 2**: Mel spectrogram, Manifold mixup, BCEWithLogitsLoss+Batch hard tripet loss
- **Model 3**: Mel spectrogram+Tgram features+Wav2vec pre-trained features, Manifold mixup, BCEWithLogitsLoss

The results of the three models on the development set are shown in Table 1. Among them, for the valve, on the basis of Model 1, the performance is better when the 14-category domain classification model is not loaded and the machine classification head is not used, and the attribute classification is directly used. Its AUC (source), AUC (target) and pAUC are 91.50%, 93.82%, 85.37%, respectively. We keep this fixed experimental configuration for the valve, so the results for the valve are no longer shown in Table 1.

2.2. Conditional Autoencoder

The autoencoder (AE) is based on the reconstruction error to realize the detection of anomalous sound. That is, the input feature vector is first mapped to a hidden representation with a lower dimensional space by the encoder component, and then, the decoder component attempts to reconstruct the inverse transformation from the hidden representation to the original input signal. The difference between the feature vector of the original input and the output vector of the autoencoder is the reconstruction error. In the training phase, we use the attribute information of the machine as the condition, and the attribute labels are encoded and input into AE for training along with the audio features. In the test phase, the test data uses the AE model of the corresponding machine, and we calculate the Mahalanobis distance according to different attributes, and take the minimum value as the anomaly score. In addition, we perform score normalization by source and target domains. For the seven machines in the evaluation set, since the attribute labels and domain labels of the test set are unknown, we train an attribute classifier and domain classifier, and then use the predicted labels to get anomaly scores.

The network architecture we use is convolutional AE [18]. 128-dimensional log-Mel spectrogram features are used as input to the network. The batch size of training is set as 256 and Adam optimizer is used to train the model with the learning rate of 0.0005. The results of conditional AE are shown in Table 2.

Table 2: Results of the conditional AE on the development set (%). "AUC-S" and "AUC-T" represent the AUC of the source and target domains, respectively.

	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve
AUC-S	68.04	62.44	71.02	69.46	75.32	89.70	57.95
AUC-T	55.38	56.72	72.04	62.60	76.06	76.02	54.08
pAUC	49.84	49.92	53.68	59.74	57.53	60.00	52.47

Table 3: DCASE 2023 Task 2 experimental results on development dataset (%). The value in the row "Total Score" represents the harmonic mean of the AUC and pAUC scores over all the machine types, sections, and domains.

		Baseline (AE-MSE)	Baseline (AE-MAHALA)	Our system
ToyCar	AUC (source)	70.10	**74.53**	72.58
	AUC (target)	46.89	43.42	**68.04**
	pAUC	52.47	49.18	**58.95**
ToyTrain	AUC (source)	57.93	55.98	**73.80**
	AUC (target)	57.02	42.45	**66.62**
	pAUC	48.57	48.13	**53.32**
bearing	AUC (source)	65.92	65.16	**88.94**
	AUC (target)	55.75	55.28	**80.62**
	pAUC	50.42	51.37	**77.11**
fan	AUC (source)	80.19	87.10	**91.90**
	AUC (target)	36.18	45.98	**86.18**
	pAUC	59.04	59.33	**76.11**
gearbox	AUC (source)	60.31	71.88	**86.18**
	AUC (target)	60.69	70.78	**83.78**
	pAUC	53.22	54.34	**67.89**
slider	AUC (source)	70.31	84.02	**99.94**
	AUC (target)	48.77	73.29	**95.68**
	pAUC	56.37	54.72	**90.89**
valve	AUC (source)	55.35	56.31	**91.50**
	AUC (target)	50.69	51.40	**93.82**
	pAUC	51.18	51.08	**85.37**
Total Score		55.02	56.91	**78.35**

2.3. Ensemble

Considering that the results of the above several anomaly detection methods are complementary, so we can ensemble them. We combined these models by grid search [19]. We explored four different sets of weights as the final four systems submitted, and Table 3 shows our best results on the development set through system ensembles.

3. CONCLUSIONS

In this paper, we propose a method for anomalous sound detection based on attribute classification and conditional AE. Experimental results show that by integrating our different methods, we can achieve better results than the baseline. In the future, we will develop more effective generative-based anomalous sound detection methods to deal with domain generalization and first-shot problems.

4. 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," *In arXiv e-prints: 2305.07828*, 2023.

[2] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "First-shot anomaly detection for machine condition monitoring: A domain generalization baseline," *In arXiv e-prints: 2303.00455*, 2023.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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, pp. 1–5.

[8] 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.

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

[10] Y. Zeng, H. Liu, L. Xu, Y. Zhou, and L. Gan, "Robust anomaly sound detection framework for machine condition monitoring," DCASE2022 Challenge, Tech. Rep., July 2022.

[11] 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)*. IEEE, 2022, pp. 816–820.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[13] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 687–10 698.

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

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

[16] 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.

[17] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[18] A. Ribeiro, L. M. Matos, P. J. Pereira, E. C. Nunes, A. L. Ferreira, P. Cortez, and A. Pilastri, "Deep dense and convolutional autoencoders for unsupervised anomaly detection in machine condition sounds," *arXiv preprint arXiv:2006.10417*, 2020.

[19] P. Daniluk, M. Góździewski, S. Kapka, and M. Kośmider, "Ensemble of auto-encoder based and wavenet like systems for unsupervised anomaly detection," *Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2020 Challenge), Tech. Rep*, 2020.