ANOMALOUS SOUND DETECTION VIA MULTITASK LEARNING AND ADVERSARIAL LEARNING

Technical Report

Yucong Zhang, Ming Li

Data Science Research Center, Duke Kunshan University, Kunshan, China

ABSTRACT

This technical report describes our submitted systems to DCASE 2023 Challenge Task 2. We propose two different methods. The first one is a multitask learning method, which incorporates a self-supervised attribute classification and a GMM-based scoring. The second one is to directly train an anomaly evaluator via adversarial learning, which achieves domain generalization by learning inherit properties other than the attributes. Experimental results on the development dataset show that both our methods outperform the baseline methods. The ensemble system has an average improvement of 8% based on the baseline results.

Index Terms— Anomalous Sound Detection, Domain Generalization, Multitask Learning, Gaussian Mixture Model, Adversarial learning

1. INTRODUCTION

DCASE 2023 Challenge Task 2 [1] focuses on machinery monitoring, requiring teams to detect anomalous data using only normal data. The competition uses datasets from[2] and[3]. Compared to the previous challenges [4, 5], the challenge this year uses completely different machine types for the evaluation, which prevents teams to tune hyper-parameters for each machine type in the training dataset. In the last challenge, top-ranked teams [6, 7] effectively learned the properties of the normal samples by training a classifier according to different kinds of domain shifts as an auxiliary task. However, no such information are provided this time. Only machine attributes are provided for training. Hence, we propose two Anomalous Sound Detection (ASD) methods that try to take the machine attributes into consideration.

The first method is to build a machine encoder for each machine type by adopting multitask learning. Since one machine might have various kinds of attributes, multiple attribute classifiers are added after the machine encoder to force the learned embedding to learn useful properties. The second method is to directly train an anomaly evaluator by adversarial training. The intuition behind this method is that if we can generate fake machine samples using some attributes, then by distinguishing the fake and the real samples, the model can focus on the characteristics other than the attributes, which might be the key to detect anomalies.

2. PROPOSED METHOD

In this challenge, we submit four systems. Zhang_DKU_task2_1 and Zhang_DKU_task2_2 are trained by multitask learning, which are introduced in Section 2.1. Zhang_DKU_task2_3, described in Section 2.2, is the system trained by adversarial learning. Zhang_DKU_task2_4 is the fused system using the scores of all the systems.

2.1. Multitask learning

2.1.1. Audio Encoder

We use two different ways of feature extracting techniques. One way is to use the naive log Mel-spectrogram, the other way is to use a neural network to model the process of Short-Time Fourier Transform (STFT). In the past few years, researchers have proposed to directly build the model from the raw signal, instead of the spectrogram or filter bank. The authors in [8] proposed a large-scaled pretrained audio neural network, using a deep neural network to directly model the signal. In [9, 10], the authors uses a simple CNN-based network to directly extract features from raw signal and showed its effectiveness in the ASD task.

In this challenge, we implement the audio encoder using the same architecture described in [9]. After the features are extracted by the audio encoder, they are concatenated with the log Mel-spectrogram and sent to the backend as the input for the classifier.

2.1.2. Classifier

In our experiment, we use a conformer-based network [11] to map the audio features into the latent space. Then, we add multiple classifiers in the backend and train the model using a multitask learning. Specifically, we imposes three tasks: a) attribute classification b) binary machine type classification c) auxiliary augmentation classification.

The overall pipeline is shown in Figure 1, where e in the figure stands for the machine embedding in the latent space, $a_1, \ldots, a_n$ are the posterior probability vector for the $1^{st}, \ldots, n^{th}$ attribute respectively, p_{aug} tells the likelihood of different augmentation techniques and p describe the likelihood that the given embedding e belongs to the target machine type.

Inspired by the systems in the previous challenges [4, 5, 12, 13], machines with types other than the target type can be seen as the psuedo-anomalous data of the target machine type. Thus, the likelihood p can also be seen as the anomaly score measurement. In this challenge, we submit the results using log Mel-spectrogram and concatenated audio features as Zhang_DKU_task2_1 and Zhang_DKU_task2_2 respectively.

2.1.3. Domain generalization

We did not include any specific domain generalization modules in our model design. However, in order to mitigate the negative effect brought by domain shifts, we carefully design the training batches to make sure each batch to contain at least one sample from the target domain.

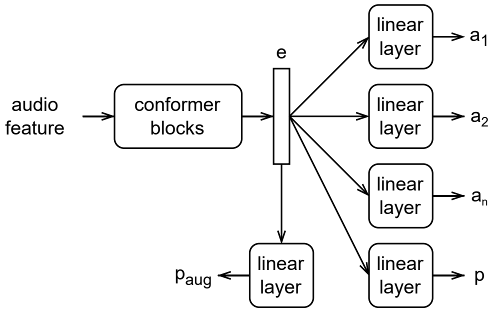

Figure 1: Multitask learning workflow.

2.1.4. Gaussian Mixture Model

Instead of using only the likelihood p, we introduce the negative log-likelihood (NLL) from the Gaussian Mixture Model (GMM) when calculating the anomaly scores. For each machine type, we train a GMM on the learned embedding of the normal samples, and calculate the NLL for all the test samples. In order to combine both scores, we first transform each of them into a standardized scale, and then we calculate the weighted sum of them. In this way, the final result of each system is a score fusion of the likelihood p and the NLL from the GMM.

2.2. Adversarial learning

Since the aforementioned method tries to learn inherit characteristics by distinguishing the machine attributes with various augmentation techniques, and the machine attributes are used to describe different domains, it lacks domain generalization ability. Hence, we design another system which directly train an anomaly evaluator which can learn properties other than the attributes in an adversarial learning scheme. The overview of this system is shown in Figure 2.

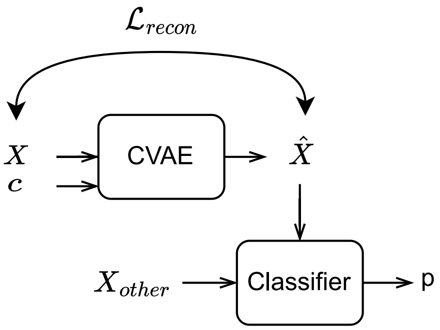

Figure 2: Adversarial learning workflow. X and $\hat{X}$ are the original spectrogram and the reconstructed spectrogram respectively. c represents the conditional embedding. p is the likelihood that describes the anomaly score.

First, we train a CVAE that can generate fake samples according to the given attributes. The reason we choose CVAE is that it can bring variations to the generated samples that follow given conditions. Then, we build a simple CNN-based network to distinguish the real target samples from the fake target samples and samples of other types. As the fake samples are generated according to the given attributes, we want the classifier to learn characteristics other than the given attributes.

We train the classifier in an adversarial training manner. For each batch of data, we have two steps. In the first step, we only update the parameters in the classifier. We first generate fake samples using the trained CVAE, then we feed them to the classifier to distinguish the real samples from the fake samples and samples of other machine types. In the second step, we fool the classifier to judge the fake samples as real samples, and only update the parameters of the generator. In this way, we aim to derive a better classifier as well as a better generator.

Similar to the type classifier described in Section 2.1.2. We submit the output of the classifier p as Zhang_DKU_task2_3.

3. EXPERIMENT

3.1. Settings

3.1.1. Data Processing

In our systems, we employ a log-Mel spectrogram with 128 Mel filters as input, with the number of FFT points and hop length set to 1024 and 160 respectively. Instead of using the whole spectrogram, we choose to use a window size of 256. The augmentation techniques used in this challenge is the same as in [14].

3.1.2. Multitask learning

In both Zhang_DKU_task2_1 and Zhang_DKU_task2_2, the system contains three conformer blocks without positional encoding, with 512 linear units for FFN modules in each block. We adopt four heads in the MHSA module, with an output dimension of 256. To extract deep features, we utilize an attentive statistical pooling layer after the conformer blocks to get 128d features. All the classifiers contain only one linear layer. The audio encoder used in Zhang_DKU_task2_2 is the CNN-based network same as it is in [9].

3.1.3. Adversarial learning

Table 1: The network structure of the encoder in CVAE.

Operator	c	k	s
Conv1d	32	1	1
Conv1d	16	3	2
Conv1d	8	3	2
Linear	256	-	-
Linear	16	-	-

In Zhang_DKU_task2_3, the architecture of the encoder part in CVAE is shown in Table 1. The dimension in the latent space is 16. Before decoding, the attributes are encoded as one-hot vectors, and then concatenated with the embedding sampled from the latent space.

4. RESULTS AND CONCLUSION

The results of our systems on the development dataset are shown in Table 2, including the AUC for both source and target domain, the pAUC, and the harmonic mean of both AUC and pAUC. It is shown in the table that all of our proposed systems outperform the baseline systems [15].

Table 2: Anomaly detection results [%] for different machine types.

	Criteria	Baseline AE	Baseline AE	Multitask+GMM	Multitask+GMM	Adversarial	Ensemble
		with MSE [15]	with MAHA [15]	with MelSpec.	with Combined	learning	
ToyCar	AUC (source)	70.10	**74.53**	52.00	48.76	43.12	48.32
	AUC (target)	46.89	43.42	45.52	**56.80**	55.12	54.68
	pAUC	**52.47**	49.18	49.05	51.79	49.68	52.21
ToyTrain	AUC (source)	57.93	55.98	43.66	**61.84**	42.56	58.68
	AUC (target)	**57.02**	42.45	55.64	49.08	47.40	49.16
	pAUC	48.57	48.13	47.79	**54.74**	47.37	51.16
Bearing	AUC (source)	65.92	**65.16**	62.08	66.44	62.80	66.44
	AUC (target)	55.75	55.28	**69.08**	56.84	57.84	65.00
	pAUC	50.42	51.37	56.84	56.84	60.00	**57.26**
Fan	AUC (source)	80.19	**87.10**	59.88	82.20	73.20	65.92
	AUC (target)	36.18	45.98	61.52	55.44	**65.00**	61.40
	pAUC	59.04	59.33	57.26	58.11	**61.68**	57.89
Gearbox	AUC (source)	60.31	71.88	67.76	**79.08**	69.92	76.88
	AUC (target)	60.69	70.78	68.24	70.32	71.12	**73.12**
	pAUC	53.22	54.34	54.53	65.89	**69.47**	56.42
Slider	AUC (source)	70.31	84.02	**97.80**	95.76	70.40	97.24
	AUC (target)	48.77	73.29	**95.68**	89.52	77.08	91.76
	pAUC	56.37	54.72	**95.58**	69.89	66.11	82.11
Valve	AUC (source)	55.35	56.31	**80.72**	62.08	63.72	59.76
	AUC (target)	50.69	51.40	50.84	**72.88**	72.76	72.28
	pAUC	51.18	51.08	53.68	49.26	**63.37**	48.84
Average	AUC (source)	55.02	56.91	59.71	**61.55**	58.21	61.48

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] 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] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, 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," *arXiv preprint arXiv:2206.05876*, 2022.

[5] 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 sound detection for machine condition monitoring under domain shifted conditions," *arXiv preprint arXiv:2106.04492*, 2021.

[6] I. Kuroyanagi, T. Hayashi, K. Takeda, and T. Toda, "Two-stage anomalous sound detection systems using domain generalization and specialization techniques," DCASE2022 Challenge, Tech. Rep., July 2022.

[7] 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.

[8] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[9] 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.

[10] F. Xiao, Y. Liu, Y. Wei, J. Guan, Q. Zhu, T. Zheng, and J. Han, "The dcase2022 challenge task 2 system: Anomalous sound detection with self-supervised attribute classification and gmm-based clustering," DCASE2022 Challenge, Tech. Rep., July 2022.

[11] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

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

M. Yasuda, *et al.*, “Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring,” *arXiv preprint arXiv:2006.05822*, 2020.

[13] P. Primus, “Reframing unsupervised machine condition monitoring as a supervised classification task with outlier-exposed classifiers,” DCASE2020 Challenge, Tech. Rep., July 2020.

[14] H. Hojjati and N. Armanfard, “Self-supervised acoustic anomaly detection via contrastive learning,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3253–3257.

[15] 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.