TWO-STAGE CONTRASTIVE LEARNING FOR ANOMALOUS SOUND DETECTION

Technical Report

Seunghyeon Shin[1], *Seokjin Lee*[1,2],

[1] School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, Republic of Korea, {sh.shin, sjlee6}@knu.ac.kr
[2] School of Electronics Engineering, Kyungpook National University, Daegu, Republic of Korea

ABSTRACT

This technical report describes our anomalous sound detection system submission for DCASE 2023 Task 2. Our system is composed of two stages: a self-supervised contrastive learning network as a feature extractor and a covariance estimator for anomalous scoring. The feature extractor network is trained only once and used across all classes, while the anomalous score is calculated using the mahalanobis distance with the covariance estimator. Our system tested with two kinds of covariance estimation method. Our system with maximum likelihood covariance estimation method achieved a performance improvement of 7.39% and 5.5% over the baseline system which uses mean square error loss and mahalanobis distance loss, based on the official scoring metric of DCASE 2023 Task 2

Index Terms— Anomalous sound detection, contrastive learning, data augmentation

1. INTRODUCTION

In DCASE 2023 Task 2 [1], the goal is to analyze machine operating sound and determine whether machine operating is anomaly or normal. During the task, the system is provided with only normal state sound clips for training because anomalous sounds can arise from various reasons and are difficult to collect. In previous year, domain generalization was required for the anomalous sound detection system to perform generally in different working environments and conditions. This year, the system is required to operate under the first-shot condition, where the class of training data differs from that of the evaluation data. At first-shot condition, anomalous sound detection system shall respond to operating sounds of various characteristics and which means a general system between various classes. To satisfy first-shot condition, We propose a two-stage anomalous sound detection system that consists of a general feature extracting network and a anomalous score calculator. First stage of our system is a general feature extracting network that is commonly used across all classes.Our general feature extracting network is trained using contrastive learning strategies. In the second stage, the system with a covariance estimator to calculate an anomalous score from the output of general feature extracting network.

2. PROPOSED METHOD

2.1. Self-supervised contrastive learning network

First stage of our model is self-supervised feature extracting network trained with a contrastive learning strategy. Contrastive learning commonly involves using many data augmentation methods during the training process, but the available augmentation techniques for the acoustic domain are more limited than those for the image domain. We utilized three types of data augmentation techniques, pitch shifting, adding white Gaussian noise, and time-domain masking. By applying three kinds of data augmentation, resulting in an augmented dataset six times larger than the original. Both the original and augmented datasets were converted to mel-spectrograms and used as input for our self-supervised contrastive learning network. Our training strategy is a modified version of SimCLR[2]. In SimCLR, network is trained to maximize agreement between the same samples that adopt different data augmentation methods. Our system uses augmented samples in combination with the original samples, unlike in SimCLR and trained to maximize agreement by using normalized temperature-scaled cross entropy(NT-Xent) loss function[2]. NT-Xent loss function are defined as

$$l_{i,j} = -log\frac{exp(sim(z_i, z_j/\tau))}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} exp(sim(z_i, z_k/\tau))}, \quad (1)$$

where $\mathbb{1}_{[k\neq i]}$ is indicator function which is 1 when $k \neq i$, τ is temperature hyper-parameter, (z_i, z_j) are augmented or original data from sample i from batch N and $sim(z_i, z_j)$ is dot product between l_2 normalized (z_i, z_j). In SimCLR, a ResNet-50[3] and non-linear projection head were used, but due to the limited quantity of available data, we used a ResNet-18 with a non-linear projection head consisting of a combination of a linear layer, ReLU non-linear activation function, and another linear layer. Despite this difference, ResNet-50 and ResNet-18 showed similar performance, but ResNet-50 required higher computational power. The output of the network adopting the projection head is 128 latent dimensions.

2.2. Anomaly score calculator

We calculated anomalous scores by using the Mahalanobis distance from the normal state training sample. To compute the Mahalanobis distance, we needed to estimate the covariance from the network output. So we used two kinds of covariance estimator. One is maximum likelihood covariance estimator and the other is elliptic envelope algorithm. We utilize both covariance estimators from the scikit-learn library[4]. The covariance estimator was trained using normal state samples of the target class, which were obtained as outputs of the feature extraction network. Overall structure of proposed anomalous sound detection illustrated in Figure 1.

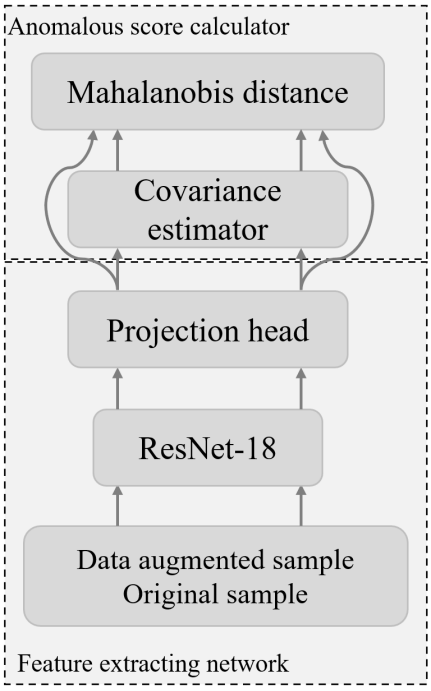

Figure 1: Structure of the proposed anomalous sound detection system.

2.3. Experiment and training settings

The dataset of DCASE 2023 Task 2 includes 14 different types of machines. [5, 6]During the development process, validation data containing normal and anomalous sounds was provided for seven machine types. The additional and evaluation dataset consisted of different machine types from development dataset. As input for the network, we converted each audio clip to a mel-spectrogram. We converted audio file to spectrogram using STFT with 2048 filter length and 512 hop size. And each spectrum was compressed through a Mel filter with a number of bins of 128. For contrastive learning, we applied three types of data augmentation methods. First, we shifted the pitch of the audio samples, with half of the augmented samples lowered and the other half raised. The pitch shift range was lower than plus one octave and higher than minus one octave. Second, we added white Gaussian noise with a range of -6 to -24 dB. Third, we applied time-domain masking to the mel-spectrogram, masking approximately 500ms and 1000ms of the audio. We trained our self-supervised contrastive learning network with the Adam optimizer[7] and cosine annealing with warm restarts learning rate scheduler[8]. Hyper-parameters used in our model are a temperature τ of 0.05, a batch size of 48, an initial learning rate of 0.5, and three restarts of the learning rate using the cosine annealing warm restarts scheduler.

3. RESULTS

The performance metric for evaluation in DCASE 2023 Task2 is the harmonic mean of the area under the curve (AUC) and the partial area under the curve (pAUC). We compared our proposed system with the DCASE 2023 Task2 baseline system[9] in Table 1. In the table, we use "ML" to refer to the maximum likelihood covariance estimator, "elliptic" to refer to the elliptic envelope covariance estimator, and "MSE" to refer to the mean square error. The harmonic mean of our proposed system using maximum likelihood covariance estimator scored 62.41% and elliptic envelope covariance estimator scored 61.31% compared to the baseline system using mean sqaure error(MSE), which scored 55.02%, and baseline system using the mahalanobis distance, which scored 56.58% Our system exhibited a more generalized performance across domains, with a lower decrease in performance when moving from the source domain to the target domain compared to the baseline system.

4. CONCLUSION

In conclusion, we proposed self-supervised contrastive learning network based two-stage system for anomalous sound detection in DCASE 2023 Task 2. In the first stage, we employed a self-supervised contrastive learning network as a feature extractor, and we used three types of data augmentation methods, including pitch shifting, gaussian noise, time domain masking, during the contrastive learning process. In the second stage, we used a covariance estimator and the mahalanobis distance to score whether sound is anomalous. We tested two covariance estimation method, maximum likelihood estimator and elliptic envelop estimator, and found that our system with the maximum likelihood estimator achieved a performance improvement of 7.39% and 5.58% over the baseline system which uses mean square error loss and mahalanobis distance loss, respectively, based on the official scoring metric of DCASE 2023 Task 2.

Table 1: AUC and pAUC result of proposed model

Class		Ours ML	Ours Elliptic	Baseline MSE	Baseline Mahalanobis
ToyCar	AUC(target)	39.36%	40.72%	46.89%	43.42%
	AUC(source)	51.64%	52.04%	70.10%	74.53%
	pAUC	54.57%	51.47%	52.47%	49.18%
ToyTrain	AUC(target)	44.36%	45.68%	57.02%	42.45%
	AUC(source)	47.63%	49.80%	57.93%	55.98%
	pAUC	49.37%	48.27%	48.57%	48.13%
Bearing	AUC(target)	73.40%	70.64%	55.75%	55.28%
	AUC(source)	73.40%	71.04%	65.92%	65.16%
	pAUC	59.73%	50.58%	50.42%	51.37%
Fan	AUC(target)	65.44%	65.52%	36.18%	45.98%
	AUC(source)	65.60%	65.10%	80.19%	87.10%
	pAUC	48.16%	52.16%	59.04%	59.33%
Gearbox	AUC(target)	74.32%	77.68%	60.69%	70.78%
	AUC(source)	80.80%	78.72%	60.31%	71.88%
	pAUC	54.53%	51.58%	53.22%	54.34%
Slider	AUC(target)	83.84%	82.52%	48.77%	63.29%
	AUC(source)	82.92%	81.32%	70.31%	84.02%
	pAUC	62.89%	51.90%	56.37%	54.73%
Valve	AUC(target)	99.80%	99.56%	50.69%	51.4%
	AUC(source)	98.73%	99.24%	55.35%	56.31%
	pAUC	98.74%	99.24%	51.18%	51.08%
All	harmonic mean	62.41%	61.31%	55.02%	56.58%

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

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[5] 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 ggeneralization task," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[6] 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.

[7] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[8] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," 2016. [Online]. Available: https://arxiv.org/abs/1608.03983

[9] 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.