Anomaly Sound Detection System Based on Multi-Dimensional Attention Module

Technical Report

Wang Junjie[1], Wang Jiajun[2], Chen Shengbing[2], Sun Yong[1], Liu Mengyuan[1]

[1]Industrial Equipment States Evaluation and Fault Prediction Technology R&D Lab
[2]Multimodal Information Processing and Modeling for Industrial Equipment R&D Lab
School of Artificial Intelligence and Big Data
Hefei University, Hefei, China
15656793081@163.com, shbchen@hfuu.edu.cn
{2350846451, 2032336570, 2863845001}@qq.com

ABSTRACT

This technical report presents our approach for Task 2 of the DCASE 2023 Challenge, which focuses on unsupervised anomaly sound detection for machine condition monitoring. We constructed four subsystems, where the first two are based on self-supervised learning methods that utilize feature vectors extracted from convolutional neural networks and employ outlier detection algorithms to identify abnormal sounds. The third subsystem incorporates a modification of the Mahalanobis distance autoencoder (AE) to better adapt to domain shift. The fourth subsystem integrates the previous three systems. The experimental results demonstrate that the proposed system outperforms the baseline significantly on the development set.

Index Terms— self-supervised learning, domain shift, anomaly sound detection, convolutional neural networks

1. INTRODUCTION

The DCASE 2023 Challenge Task 2 [1] aims at First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring. This year, the task focuses on the first-shot problem. In practical applications, the machine types may be novel or have limited testing data, making it impractical to adjust hyperparameters for each machine type using the testing data from the development set. Additionally, the source and target domain data are severely imbalanced, and when a model trained on the source domain data is applied to the target domain data, its performance deteriorates due to domain shift caused by factors other than anomalies.

For Task 2, [2] provides two baseline methods. The first method utilizes a standard autoencoder, which performs well in unsupervised anomaly detection but faces challenges in domain generalization. The second method is based on a Mahalanobis distance autoencoder, which performs well on the source domain but has subpar performance on the target domain.

We propose a self-supervised learning approach, where we first train a neural network to extract embeddings by classifying the labels extracted from the metadata. Then, we use an outlier detection algorithm to score the abnormality level of the embeddings. However, embeddings extracted by conventional networks sometimes fail to capture important frequency bands and critical time segments. To address this, we employ an attention mechanism to enable the model to focus on embeddings that are more relevant to the task.

In the field of image recognition, channel attention mechanisms such as SENet [3] and ECANet [4] have achieved excellent results. We attempted to apply these methods to Anomalous Sound Detection (ASD), but they yielded unsatisfactory performance due to the differences between spectrograms and regular images. Therefore, we design a Multi-Dimensional Attention Module (MDAM) tailored for anomaly sound detection to enhance the detection performance.

2. METHODOLOGY

2.1. Dataset

The dataset used for this task is derived from the MIMII DG [5] and ToyADMOS2 [6] datasets, consisting of normal and abnormal operation sounds from 14 types of toys/real machines. Each recording is in mono and has a duration of 10 seconds. For files that are not exactly 10 seconds long, we employ a strategy of trimming or padding to meet the desired duration. These signals are a mixture of machine sounds from several real factories and ambient noise samples. Each machine type has only one section included in both the development dataset and the additional dataset. In this report, all training data from the development dataset and the additional training dataset are used to train the model. The performance of the model is evaluated on the testing data from the development dataset.

2.2. Classification-Based Model

We utilize Wilkinghoff [7] as our backbone network, which consists of an enhanced ResNet[8] and a dual-branch network composed of three 1D convolutions and five dense layers. To capture the signal dimension and obtain a good initialization feature, we use linear magnitude spectrograms and magnitude spectrograms as input features. Linear magnitude spectrograms with a dimension of 513 are obtained through Short-Time Fourier Transform (STFT), where the sampling window size is set to 1024, the hop size is 512, the maximum frequency is set to 8000

Table 1: Anomaly detection results for different machine types

	Method	Baseline MHLAE	Our MHLAE	MDAM + cos	MDAM + knn
ToyCar	AUC(source)	**74.53 %**	72.00%	51.36%	51.72%
	AUC(target)	43.42 %	50.90%	64.0%	**68.06%**
	pAUC	**49.18 %**	49.00%	42.73%	47.76%
ToyTrain	AUC(source)	**55.98 %**	53.3%	50.18%	51.28%
	AUC(target)	42.45 %	43.0%	62.95%	**63.97%**
	pAUC	48.13 %	47.0%	47.47%	**49.47%**
Bearing	AUC(source)	65.16 %	67.0%	74.56%	**79.55%**
	AUC(target)	55.28 %	53.24%	**73.12%**	70.21%
	pAUC	**51.37 %**	50.23%	51.31%	50.23%
Fan	AUC(source)	87.1 %	**95.00%**	91.60%	92.67%
	AUC(target)	45.98 %	51.00%	82.32%	**82.62%**
	pAUC	59.33 %	60.26%	**61.22%**	60.23%
Gearbox	AUC(source)	71.88 %	73.0%	82.44%	**88.45%**
	AUC(target)	70.78 %	73.2%	81.83%	**83.85%**
	pAUC	54.34 %	56.42%	**63.43%**	63.42%
Slider	AUC(source)	84.02 %	82%	97.72%	**98.71%**
	AUC(target)	73.29 %	74%	**96.52%**	95.22%
	pAUC	54.72 %	54%	82.81%	**83.66%**
Valve	AUC(source)	56.31 %	56%	95.33%	**95.74%**
	AUC(target)	51.4 %	51%	91.42%	**91.93%**
	pAUC	51.08 %	51%	**73.02%**	73.01%
All(hmean)	AUC(source)	68.84%	68.72%	72.45%	**74.40%**
	AUC(target)	52.37%	54.65%	77.05%	**77.80%**
	pAUC	52.36%	52.22%	57.46%	**58.78%**

Hz, and the minimum frequency is 200 Hz. The magnitude spectrum of the entire signal (8000) is used to achieve higher frequency resolution for better capturing stationary sounds. Before feeding them into the neural network as features, all STFT spectrograms are normalized by subtracting the time average and dividing by the time standard deviation of all files belonging to the training dataset. We use the Adam optimizer with a default initial learning rate of 0.001 to train our model. The SCAdaCos [9] loss function is employed, with the number of classes set as the joint category of machine ID and attributes. The alpha parameter of the Elu activation function is set to 1.0.

By integrating the MDAM into the ResNet branch, the model can effectively focus on capturing features that are more relevant to anomalies. This attention mechanism helps the model prioritize and highlight the discriminative information in the spectrograms, thereby enhancing the detection performance.

2.3. Autoencoder-Based Model

We compared the standard AE with the Mahalanobis distance-based AE (MHLAE) and observed that MHLAE performs better overall on the development set but exhibits poorer performance on the target domain. This could be attributed to the limited number of samples in the target domain. To address this issue, we employed the SMOTE (Synthetic Minority Over-sampling Technique) [10] oversampling technique to alleviate the sample imbalance between the source and target domains. Additionally, we utilized Mixup [11], which involves interpolating between samples from the target and source domains to generate new samples that simulate a new domain. This approach enhanced the model's ability to generalize to unseen data.

3. RESULTS AND DISCUSSIONS

Table 1 presents the results of all our models. Compared to the baseline, our modified Mahalanobis distance AE performs similarly on the source domain and shows better performance on the target domain. Our anomaly sound detection model based on the multi-dimensional attention module demonstrates significant overall improvement over the baseline, enhancing the detection performance. It is important to note that the results presented here do not include model ensemble.

4. REFERENCES

[1] Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, and Yohei 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] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, and Masahiro Yasuda. First-shot anomaly detection for machine condition monitoring: a domain generalization baseline. In arXiv e-prints: 2303.00455, 2023.

[3] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141

[4] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11534-11542.

[5] Kota Dohi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, Masaaki Yamamoto, Yuki Nikaido, and Yohei 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.

[6] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Masahiro Yasuda, and Shoichiro 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), 1–5. Barcelona, Spain, November 2021.

[7] Wilkinghoff K. Design Choices for Learning Embeddings from Auxiliary Tasks for Domain Generalization in Anomalous Sound Detection[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[9] Wilkinghoff K. Sub-cluster AdaCos: Learning representations for anomalous sound detection[C]//2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.

[10] Fernández A, Garcia S, Herrera F, et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary[J]. Journal of artificial intelligence research, 2018, 61: 863-905.

[11] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization[J]. arXiv preprint arXiv:1710.09412, 2017.