SELF-SUPERVISED REPRESENTATION LEARNING FOR FIRST-SHOT UNSUPERVISED ANOMALOUS SOUND DETECTION

Technical Report

Wang Jiajun[1]*, Wang Junjie*[2]*, Chen Shengbing*[1]*, Xu Zhiqi*[1]*, Wan Mengyuan*[1]*,*

[1] Multimodal Information Processing and Modeling for Industrial Equipment R&D Lab
[2] Industrial Equipment States Evaluation and Fault Prediction Technology R&D Lab
HeFei University, Hefei, China,
Shcool of Artificial Intelligence and Big Data,
2350846451@qq.com,15656793081@163.com,
shbchen@hfuu.edu.cn,{974404857,1284771501}@qq.com

ABSTRACT

This paper describes a self-supervised representation learning system for the DCASE 2023 Challenge Task 2: "First-shot compliant unsupervised anomaly detection (ASD) for machine condition monitoring".First-shot ASD does not allow systems to do machine-type dependent hyperparameter tuning or tool ensembling based on the performance metric calculated with the grand truth. Due to the challenges in extracting meaningful features from exposure methods of outlier values in anomaly detection, a novel approach of self-supervised representation learning is introduced. The proposed method involves initial classification based on sound metadata, and subsequent feature extraction, and ultimately, anomaly scores are obtained through an anomaly detection algorithm. Our final system is a result of integrating multiple systems together. The proposed system achieves a 63.16% area under the curve (AUC) and partial AUC (p = 0.1) in the harmonized average across all machine types, subsets, and domains on the development dataset.

***Index Terms*—** self-supervised representation learning;First-shot;Anomalous sound detection

1. INTRODUCTION

Automatic machine condition monitoring with deep learning techniques for predictive maintenance is a crucial application in Industry 2.0/4.0. Unsupervised anomalous sound detection tasks are held in the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge. However, most winning systems have utilized techniques specific to the challenge task setting[1]. In the task setting, many sound samples of similar but different machine instances are available as different data sections for training. Some systems have used these different sections of data as pseudo anomaly data samples. Besides, most of the winning systems have relied on machine-type dependent hyperparameter tuning and tool ensembling based on the performance observation with the given anomaly samples for performance assessment. These solutions are not always applicable to the industry's realistic application scenarios[2].

We conduct an experimental evaluation of the developed system using the DCASE 2023 Task 2 Challenge development datasets. Here, the datasets[3, 4] contain fourteen machine types. Each machine type has one section ID, reflecting a domain shift scenario type. The training data contains domain data for both the source and target domains, but only a few samples of target domain data. Experiments on the datasets show that all of the created systems significantly outperformed the official baseline system in the evaluation metric, the harmonic mean of the area under the curve (AUC), and partial AUC (p = 0.1) for all machines types, and section IDs (all / har-mean). The domain generalization approach achieved 63.16 %

2. METHOD

2.1. self-supervised representation learning

The proposed framework can be divided into two main components: the upstream model and the downstream model. In the upstream model, a multi-task approach is employed. The first task involves the classification of fourteen different machine types, while the second task focuses on classifying machine attributes. The upstream model is responsible for learning representations that capture the relevant information for these tasks.

Moving on to the downstream model, the representations learned by the upstream model are utilized. In this stage, commonly used anomaly detection algorithms such as K-Nearest Neighbors (KNN), Local Outlier Factor (LOF), and Gaussian Mixture Models (GMM) are employed to evaluate the anomaly scores. These algorithms utilize the learned representations to assess the abnormality of the input data.

By employing this approach, the framework effectively combines the benefits of multi-task learning in the upstream model and the power of established anomaly detection algorithms in the downstream model. This enables the system to achieve robust anomaly detection performance, leveraging the discriminative features learned from the upstream model and the evaluative capabilities of the downstream algorithms.

2.2. Improvement of an upstream model

The method is to use not only EfficientNet-B0 [5] but also Transformer[6] for the models used in the feature extractor. We obtain feature extractors that focus on different features using convolutional neural network-based and self-attention-based models

2.3. Domain generalization approach

The key to domain generalization[7] is treating data from the source and target domains as the same. We employ two techniques for domain generalization. The first is to sample the normal data in the target domain in creating a mini-batch so that at least one sample in the target domain is in the mini-batch when training the feature extractor by OE. It reduces the problem of data imbalance between the source and target domains.

The second is to use a Mixup[8] of source and target domain data to generate 50 samples of pseudo-normal data when training the anomalous detector by IM. It is effective since it models the intermediate data representation in the source and target domains as normal data.

2.4. Ensemble

For both approaches, ensembles are effective in improving performance[7]. When ensembling, the anomaly scores are standardized by each section ID before being used since the output scales differently depending on the anomalous detectors h. The approach obtains anomaly scores by selecting multiple models and averaging them.

Table 1: Scores of Different Models

Machines	AUC	Scores 1	Scores 2	Scores 3	Scores 4
ToyCar	source_auc	52.16	49.64	48.76	49.2
	target_auc	47.36	52.08	49.04	49.28
	pauc	48.63	48.68	48.68	48.84
ToyTrain	source_auc	43.42	43.36	44.32	46.03
	target_auc	54.48	56.44	58.24	58.8
	pauc	49.21	49.15	49.10	48.84
fan	source_auc	81.68	81.8	80.36	80.56
	target_auc	81.92	81.88	80.8	80.28
	pauc	66.84	66	61.89	61.73
gearbox	source_auc	87.44	87.68	88.4	87.2
	target_auc	81.52	82.48	81.44	81.76
	pauc	66.84	66.84	65.31	65
bearing	source_auc	73.24	74.16	74.68	75
	target_auc	62.72	62.36	61.68	62.04
	pauc	51.63	52.15	52	52.41
slider	source_auc	92.52	92.84	93.12	92.68
	target_auc	96.36	97	97.08	97.2
	pauc	87.42	89.21	89.57	88.73
valve	source_auc	88.08	83.08	79.48	77
	target_auc	84.16	85.36	84.16	82.16
	pauc	65.21	58.47	55.84	55.68

3. EXPERIMENT

The amplitude of the audio input sequence was standardized to have a mean of 0 and a variance of 1. The audio input sequence was extracted as Mel-spectrogram with a window size of 128 ms, a hop size of 16 ms, and 224 Mel-spaced frequency bins in the range of 50–7800 Hz in 2.0 sec. The feature was passed to the encoder f using EfficientNet-B0 and Transformer. The scheduler was OneCycleLR, and the optimizer was AdamW with a learning rate of 0.001. The batch size was set to 128. It was a hyperparameter that whether or not using Mixup to obtain intermediate features between normal and pseudo-anomalous data during training for the feature extractor. GMM, LOF, or KNN were used for the anomalous detector h. The hyperparameter of the anomalous detector h was the number of components for GMM or the number of neighbors for LOF or KNN, where it was one of 1, 2, 4, 16, 32. During inference, we divided 10.0 sec. clips into S = 10 segments with overlapping. The results are shown in Table 1

4. CONCLUSION

Therefore, this study proposes a self-supervised representation learning approach. Firstly, sound metadata is utilized for classification. Then, representations are extracted, and finally, anomaly scores are obtained through an anomaly detection algorithm. Our final system is obtained by integrating multiple systems. The proposed system achieves an AUC (Area Under the Curve) of 63.16% and a partial AUC (p = 0.1) in the harmonized average across all machine types, subsets, and domains on the development set.

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] 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] 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.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] 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.

[8] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization."