ANOMALOUS SOUND DETECTION BASED ON ENSEMBLE OF AUTOENCODERS

Technical Report

Ee-Leng Tan

Nanyang Technological University
EEE, 50 Nanyang Avenue,
Singapore 639798
etanel@ntu.edu.sg

Santi Peksi

Nanyang Technological University
EEE, 50 Nanyang Avenue,
Singapore 639798
speksi@ntu.edu.sg

Nguyen Duy Hai

Nanyang Technological University
EEE, 50 Nanyang Avenue,
Singapore 639798
ndhai@ntu.edu.sg

ABSTRACT

This technical report outlines our solution to task 2 of the detection and classification of acoustic scenes and events (DCASE) 2021 challenge. The objective of this task is to identify anomalous sounds using an anomaly detector trained with normal sound only and to avoid identifying normal sounds that deviate from the operating condition of the normal sounds in the training dataset as anomalous sounds. Our approach is based on an assemble of autoencoders with different network architectures targeted to different machine types.

Index Terms— Autoencoders, semi-supervised anomalous sound detection

1. INTRODUCTION

In recent years, anomalous sound detection (ASD) has received significant attention from the machine learning community [1]. An anomaly in sound might indicate a fault in a machine, and early detection of anomalies can be applied to major fault avoidance and plays an important role in predictive maintenance [2]. An anomaly detection system can be manually constructed using thresholds on data recommended by domain experts or automatically constructed from the available data using machine learning. Since extensive domain knowledge and expert level foresight may not be readily accessible, an unsupervised system that is trained by available data (mostly sound from healthy machines) would be very attractive.

Task 2 of the detection and classification of acoustic scenes and events (DCASE) challenge involves analyzing sounds emitted by machines and deciding if the operation of the machine is normal or anomalous. An exhaustive collection of anomalous sounds from faulty machines is not practical, therefore the training set of such an anomalous sound detector is often contained sound recorded mainly or solely from healthy machines. Furthermore, the operating conditions of the healthy machines in the training set may differ from the test dataset, and the differences in operating conditions are not limited to operating speed, machine load, and environmental noise.

The challenge dataset was jointly created by Hitachi Ltd. and NTT Corporation [3, 4] and recordings of seven machine types, namely, fan, gearbox, pump, slide rail, toy car, toy train, and valve are provided in the dataset. The proposed solution extends the DCASE 2021 baseline [5], which is based on autoencoder (AE). The anomaly score is calculated as the reconstructed error of the observed sound. Since the AE is trained with the recording of operations of healthy machines, AE would reconstruct anomalous sounds without high reconstruction loss.

The 2021 challenge also highlighted the problem of normal sounds being wrongly classified as anomalies due to the changes in the machine's operating conditions. Often, machines are operated at different conditions based on production demand. For instance, the training data may only contain healthy motor sounds captured at 200-300 rpm, and the testing data may consist of recordings of the motor operating between 100-400 rpm.

The reconstructed loss produced by AEs with various network configurations and the frequency spectrums of the sounds recorded from the seven machine types are analyzed. Using the provided information of the machine type, the proposed solution becomes semi-supervised. Instead of computing an anomaly score for the seven machine types using seven AEs, we have grouped the machine types having similar frequency spectrum into three groups, and the sound recordings of these three groups are then used to train three AEs.

2. PROPOSED APPROACH

The proposed approach adopts the AE-based baseline to compute anomaly scores for the sound recordings. The AE is an unsupervised neural network having two key components, namely, encoder and decoder. The encoder transforms the input data of a higher dimension to a new representation of a lower dimension, and the decoder approximates the input from the low dimensional representation. Since the AE is trained to minimize reconstruction error with recorded sounds of healthy machines, high reconstruction errors are expected with anomalous sounds.

The input of the AE is the log-mel spectrogram of the input $X = \left\{X_t\right\}_{t=1}^{T}$, where F and T are the number of mel-filters and timeframes, respectively. The acoustic feature at t is constructed by concatenating P consecutive frames of the log-mel spectrogram and is denoted as $\psi_t = \left(X_t, \cdots, X_{t+P-1}\right)$. Let r_ϑ denote the vector reconstructed by the AE, the anomaly score A_θ for X is

$$A_\vartheta\left(X\right) = \frac{1}{PFT}\sum_{t=1}^{T}\left\|\psi_t - r_\theta\left(\psi_t\right)\right\|_2^2, \qquad (1)$$

TABLE I NETWORK ARCHITECTURE

Machine Type	Toy car, pump, slider, valve	Toy train, fan	Gearbox
Input shape	1,280 = 256 × 5	1,280 = 256 × 5	640 = 128 × 5
Mel Bands	256	256	128
Architecture	Dense Layer #1 (Units: 256)	Dense Layer #1 (Units: 256)	Dense Layer #1 (Units: 128)
	Dense Layer #2 (Units: 256)	Dense Layer #2 (Units: 256)	Dense Layer #2 (Units: 128)
	Dense Layer #3 (Units: 256)	Dense Layer #3 (Units: 256)	Dense Layer #3 (Units: 128)
	Dense Layer #4 (Units: 256)	Dense Layer #4 (Units: 256)	Dense Layer #4 (Units: 128)
	Bottleneck Layer (Units: 32)	Bottleneck Layer (Units: 4)	Bottleneck Layer (Units: 32)
	Dense Layer #5 (Units: 256)	Dense Layer #5 (Units: 256)	Dense Layer #5 (Units: 128)
	Dense Layer #6 (Units: 256)	Dense Layer #6 (Units: 256)	Dense Layer #6 (Units: 128)
	Dense Layer #7 (Units: 256)	Dense Layer #7 (Units: 256)	Dense Layer #7 (Units: 128)
	Output Layer (Units: 1,280)	Output Layer (Units: 1,280)	Output Layer (Units: 640)

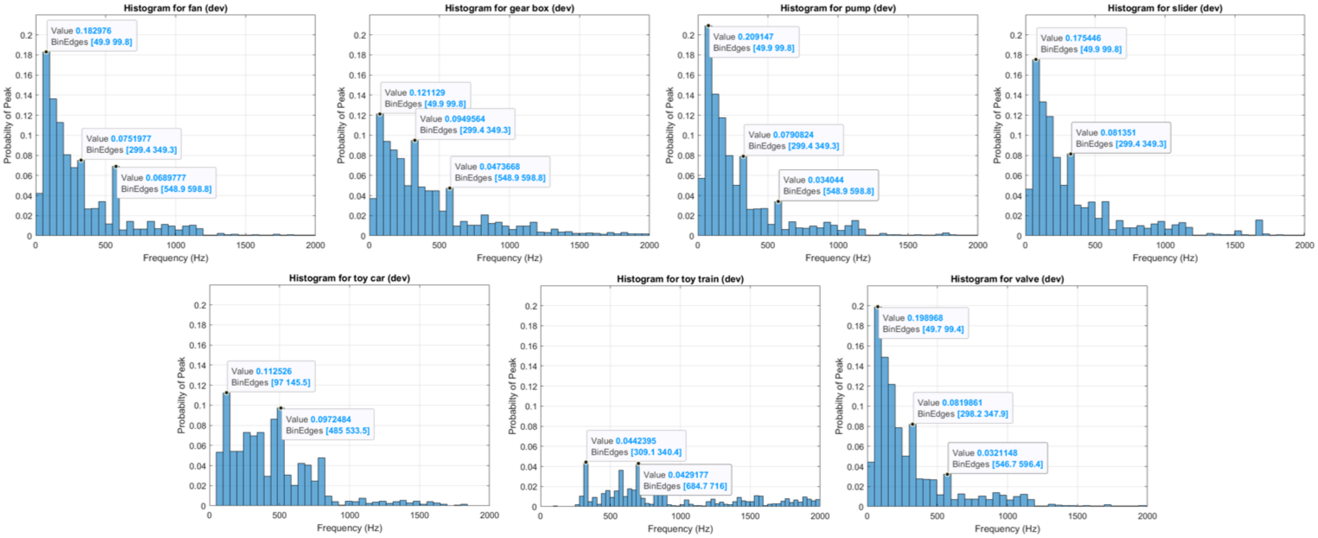

Figure 1. Histograms of peak locations in FFT spectrums computed from sound recordings of seven machine types (Dev).

where $\|\cdot\|_2$ is ℓ_2 norm.

The input to the AE is constructed by five consecutive frames of the log-mel spectrogram, where each frame consists of *F* mel bands. The grouping of the machine type is based on the characteristic of the fast Fourier transform (FFT) of the recorded sounds from the seven machine types (see Figure 1). An ensemble of three AEs based on static classifier selection is used in our proposed method (see Figure 2), and the network architectures of three AEs are summarized in Table I.

3. DATASET

The dataset provided for the DCASE 2021 Challenge 2 consists of normal and anomalous sounds recorded from seven types of machines. Each recording is a single-channel 10-second audio track that captures both a machine's operating as well as environmental noise. In real-world cases, the operating conditions of the machine and environmental noise conditions would fluctuate depending on production demand.

Based on the possible shift in the operating conditions of the machine and environmental noise, a combination of the following data augmentation techniques is used:

1. Frequency shifting augmentation: Features in the log-mel spectrograms are shifted to simulate different operating conditions.
2. Additive noise: Additive noise is added to the recordings to simulate different types and sound levels of noise.

The recordings of all sections from each machine are used for data augmentation and training of the AEs.

4. RESULTS

The classification results for the development dataset are summarized in Table II. Different numbers of epochs (10 epochs for training with the development set is only for illustration in this document) have been used in the training of three AEs in the proposed system to avoid overfitting.

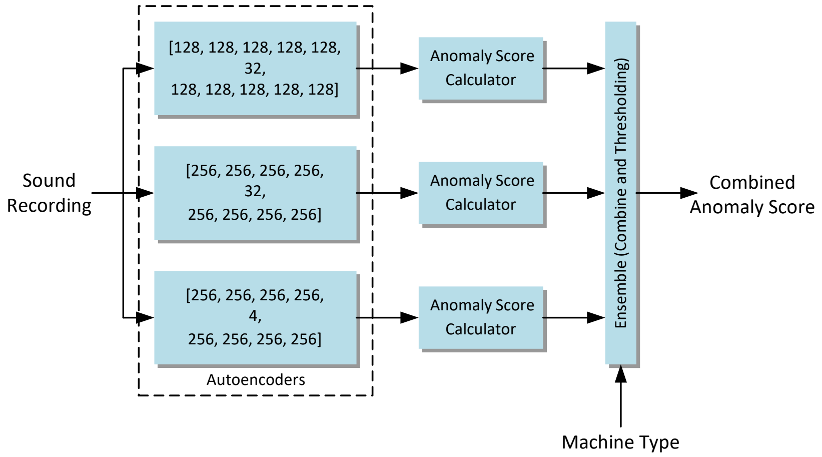

Figure 2. Block diagram of proposed solution.

TABLE II PERFORMANCE SUMMARY

Algorithm	Mean	Fan		Gearbox		Pump		Slide Rail		Toy Car		Toy Train		Valve	
		AUC (%)	PAUC (%)	AUC (%)	PAUC (%)	AUC (%)	PAUC (%)	AUC (%)	PAUC (%)	AUC (%)	PAUC (%)	AUC (%)	PAUC (%)	AUC (%)	PAUC (%)
Proposed (10 epochs)	Arithmetic	64.58	52.84	68.37	52.75	64.88	55.23	68.45	56.04	60.76	53.20	60.56	54.35	54.64	50.82
	Harmonic	64.28	52.70	67.33	52.71	63.08	54.81	66.05	55.51	59.78	53.12	59.64	53.72	54.38	50.74

5. CONCLUSION

Autoencoders have shown great potential in the field of anomalous sound detection. An ensemble of three AEs was proposed to perform anomaly detection for seven machine types. These AEs are configured based on the spectrums of the sound recordings from healthy machines and are then trained by these recordings. From our analysis, we found that the seven machine types can be effectively analyzed using three AEs. An ensemble using static classifier selection is then employed to obtain the anomaly decision of the sound recordings in the test dataset.

6. REFERENCES

[1] Y. Kawaguchi, and T. Endo, "How can we detect anomalies from subsampled audio signals?," *in IEEE 27th International Workshop on Machine Learning for Signal Processing*, Sept. 2017.

[2] T. P. Carvalho, F. A.A. M. N. Saores, R. Vita, R. D. P. Francisco, J. P. Basto, S. G. S. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," *Computer & Industrial Engineering*, vol. 137, Nov. 2019.

[3] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, "MIMII DUE: sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," *In arXiv e-prints: 2006.05822*, pp. 1–4, 2021.

[4] 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 arXiv preprint arXiv:2106.02369*, 2021.

[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," *In arXiv e-prints: 2106.04492*, pp. 1–5, 2021.