ANOMALOUS SOUND DETECTION WITH ENSEMBLE OF CNN-BASED FEATURES AND AUTOENCODER APPROACHES

Technical Report

Xiaoyu Li
Department of Big Data and Artificial Intelligence
China Telecom Corporation Research Institute
Beijing, China
lixy01@chinatelecom.cn

Jie Yang
Department of Big Data and Artificial Intelligence
China Telecom Corporation Research Institute
Beijing, China
yangj72@chinatelecom.cn

Hao Shen
Department of Big Data and Artificial Intelligence
China Telecom Corporation Research Institute
Beijing, China
shenh4@chinatelecom.cn

ABSTRACT

This paper introduces a solution with the ensemble of three anomalous sound detection (ASD) methods for the DCASE2022 Challenge Task 2[1] [2 [3]. This task is required to detect unknown anomalous sound basing on normal sound data. The first ASD method is using the audio clip of the machine which is normal, and the section index of audio clip to train the Convolutional Neural Network (CNN). Then, anomalous sound is detected by using feature vectors extracted from CNN. The second ASD method is an OE-based detector that uses MobileNetV2. The third ASD method is an IM-based detector that uses autoencoder (AE). As a result, our method achieves a harmonic mean of 72.70% over of area under the curve (AUC), and 60.35% in partial AUC (pAUC)

Index Terms— Anomalous Sound Detection, Convolutional Neural Network, Autoencoer, MobileNetV2.

1. INTRODUCTION

In DCASE2022 task2 “Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques” [1] [2 [3], it is required to detect the anomaly sound of the machine. Since only the normal sound of a machine can be obtained, the detection of anomaly sound is an unsupervised problem. We train three models including MobileNetV2[4], Autoencoder and ResNet38[5] in this task. As a result, we find that ResNet38 has best performance on four datasets including Gearbox, ToyCar and ToyTrain. Autoencoder model has best result on Fan dataset and MobileNetV2 works best on Slider, Bearing and Valve.

This paper is organized as follows. In chapter 2, we describe three anomalous sound detection methods including audio preprocessing, neural network and anomaly detector. In chapter 3, we introduce the experiments on evaluation datasets and the results. In chapter 4, we summarize the result of experiments. In chapter 5, we show the submitting model.

2. ANOMALOUS SOUND DETECTION METHODS

2.1. Feature extractor using ResNet38 approach

Since it is easier to extract feature from frequency domain rather than time domain, we transform all audio to spectrograms. The ResNet38 is a pretrained model [6], we set STFT window size as 1024 and hop size as 160 to fix the input shape of the model.

The ResNet38 is a pretrained model [6] from AudioSet introduced in the paper titled "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition" [7].

There are seven types of machine including Gearbox, Slider, ToyCar, ToyTrain, Fan, Bearing and Valve. There are 6 sections (3 sections development datasets and 3 section additionally datasets) in each machine type. For every machine, we fine-tune on ResNet38 pretrained model with spectrograms as input and 6 classes of section as output. In order to get more audio feature information, we concatenate the mean output of conv_block, resnet layer1, resnet layer2, resnet layer3, resnet layer4 and output of fc1 as a 3072-dimension feature embedding, and an anomaly detector is used on embedding vectors.

We apply k-Nearest Neighbors(k-NN) [8] on feature embedding of audio as an anomaly detector. We set the number of neighbors as 1, as a result, the distance to the nearest neighbor reflects the possibility that audio is normal. The larger of the distance, the more deviated from normal.

2.2. MobileNetV2 approach

First, we transform all audio clips into spectrograms using Librosa python package. Then we calculate the log-mel-spectrogram with the spectrogram. We use the parameters as follows: n_mels=128, n_frames=64, n_hop_frames=8, n_fft= 1024, hop_length=512 and power=2.0.

By using log-mel-spectrogram and section indices, we train a MobileNetV2 model with development dataset and additional training dataset. The input shape is 64×128×3, which is the triplication of 64×128 to each color channel, and the output is softmax for 3 sections.

2.3. Autoencoder approach

The input features are log Mel-spectrogram extracted from audio and one-dimensional section ID determined from the filename. The chosen auto-encoder network architecture is as follows::

- Input (640)
- Dense (128) + BN+ ReLU
- Dense (128) + BN+ ReLU
- Dense (128) + BN+ ReLU
- Dense (128) + BN+ ReLU
- Dense (8) + BN+ ReLU
- Dense (128) + BN+ ReLU
- Dense (128) + BN+ ReLU
- Dense (128) + BN+ ReLU
- Dense (128) + BN+ ReLU
- Output (640)

The anomaly score is then calculated as the average reconstruction error over all aggregated frames of the audio sample.

3. RESULTS

The results are shown in Table 1, Table 2 and Table 3, where Table 1 and Table 2 show the source domain and target domain AUC results and Table 3 represents the pAUC in both domain results.

4. CONCLUSION

In this paper, we used three different models and normal sound of the machine to detect anomaly sound. The performance of 72.70% for AUC and 60.35% for pAUC was shown for the development dataset

5. EVALUATION SUBMISSIONS

In this report, we submit three anomaly sound detection systems for the evaluation dataset. Table 4 shows the models we used

6. REFERENCES

[1] 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 arXiv e-prints: 2205.13879, 2022

[2] 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 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021), 1–5. Barcelona, Spain, November 2021.

[3] Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Takashi Endo, Masaaki Yamamoto, and Yohei Kawaguchi. Description and discussion on DCASE 2022 challenge task 2: unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. In arXiv e-prints: 2206.05876, 2022.

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks,” pp. 4510–4520, 2018.E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*, London, UK: Academic Press, 1999.

[5] Wu Z , Shen C , Hengel A . Wider or Deeper: Revisiting the ResNet Model for Visual Recognition[J]. Pattern Recognition, 2016.

[6] https://zenodo.org/record/3576403/#.YqFbGqhBxPY

[7] Kong Q , Cao Y , Iqbal T , et al. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition[J]. 2019.

[8] S. Ramaswamy, R. Rastogi, and K. Shim, “Efficient algorithms for mining outliers from large data sets,” in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. ACM, 2000, pp. 427–438

Table 1: Harmonic Mean of AUC in the source domain of Development Dataset (%)

	Bearing	Fan	Gearbox	Slider	ToyCar	ToyTrain	Valve	Total
Baseline_AE	56.95	78.97	69.22	78.81	91.7	76.98	52.09	69.69
Baseline_MobileNetV2	63.07	71.54	70.37	69.84	61.21	60.4	68.77	66.16
ResNet38	61.42	56.36	80.8	92.46	87.8	75.81	53.53	69.67
AE	55.74	76.86	70.57	77.03	92.52	77	51.42	69.07
MobileNetV2	65.75	62.93	74.25	92.3	48.2	55.4	74.26	65.02
Our Best	65.75	76.86	80.8	92.46	92.52	77	74.26	78.94

Table 2: Harmonic Mean of AUC in the target domain of Development Dataset (%)

	Bearing	Fan	Gearbox	Slider	ToyCar	ToyTrain	Valve	Total
Baseline_AE	59.01	49.19	62.83	49.04	36.64	26.36	49.86	44.06
Baseline_MobileNetV2	61.79	51.76	59.04	48.59	52.81	46.3	60.92	53.86
ResNet38	63.91	54.12	75.12	76.5	68.18	67.75	42.96	61.83
AE	56.77	51.37	64.2	49.55	40.9	32.66	48.94	47.19
MobileNetV2	68.01	44.62	60.01	72.13	66.23	54.47	67.08	60.34
Our Best	68.01	54.12	75.12	76.5	68.18	67.75	67.08	67.37

Table 3: Harmonic Mean of partial AUC of Development Dataset (%)

	Bearing	Fan	Gearbox	Slider	ToyCar	ToyTrain	Valve	Total
Baseline_AE	52.18	57.98	58.72	56.05	52.79	50.56	50.39	53.91
Baseline_MobileNetV2	57.89	57.56	56.53	56.49	52.46	51.59	65.27	56.54
ResNet38	51.7	48.7	64.19	62.64	52.05	52.78	50.12	54.04
AE	53.19	58.85	60.52	56.12	53.77	51.61	50.42	54.71
MobileNetV2	56.04	61.14	61.91	69.36	57.21	52.28	65.09	59.96
Our Best	56.04	61.14	64.19	69.36	57.21	52.78	65.09	60.35

Table 4: Submission of System

Model name	Bearing	Fan	Gearbox	Slider	ToyCar	ToyTrain	Valve
Li_CTRI_task2_1	MobileNetV2	Autoencoder	RestNet38 with k-NN	MobileNetV2	RestNet38 with k-NN	RestNet38 with k-NN	MobileNetV2