ANOMALOUS SOUND DETECTION BASED ON A NOVEL AUTOENCODER

Technical Report

Yaoguang Wang, Xianwei Zhang, Liang He

Department of Electronic Engineering, Tsinghua University, Beijing, China
yaoguang18@mails.tsinghua.edu.cn, zhangxw019@163.com, heliang@mail.tsinghua.edu.cn

ABSTRACT

The DCASE2020 Task2 is Unsupervised Detection of Anomalous Sounds for Machine Condition Monitoring [1]. This technical report describes the approach we used to participate in this task. We utilize the interpolation deep neural network (IDNN) [2] based on the autoencoder (AE). For 5 frames of a spectrogram from sounds in development dataset, we remove the middle frame and send the rest into AE and get the output with the same shape of middle frame. The reconstruction error between the output and the original middle frame is used as anomaly score. Compared with baseline, the AUC score is improved on validation dataset of valve.

Index Terms— Anomalous Sounds Detection, IDNN, autoencoder

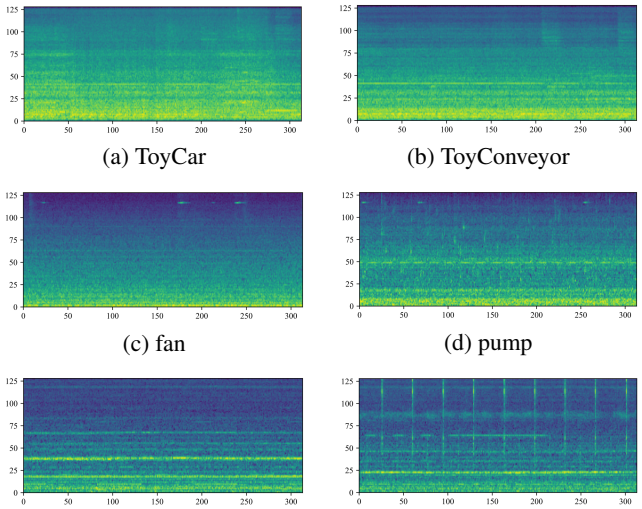
1. INTRODUCTION

The topic of Detection and Classification of Acoustic Scenes and Events (DCASE) Task 2 is anomalous sound detection (ASD), which to identify whether the sound emitted from a target machine is normal or anomalous. This technology can be used for monitoring and maintenance of large-scale equipment.

In this task, six types of machine audio data has been provided: toy-car, toy-conveyor, valve, pump, fan, and slide rail. And each type includes three or four machines that have different IDs. In development dataset, each machine ID's dataset consists of around 1,000 samples of normal sounds for training and 100-200 samples each of normal and anomalous sounds to validation. With the additional training dataset released on April 1st, each machine ID added around 1,000 pieces of normal data that could be trained. The evaluation dataset was released on June 1st. Our submission includes the anomaly scores on these validation samples.

Since anomalous sounds cannot be used for training, this task is unsupervised learning, which is quite different from the previous DCASE tasks. The traditional unsupervised learning methods include hierarchical clustering, Gaussian mixture model, matrix factorization and so on. With the popularity of deep learning, various deep neural networks are also used for anomaly detection [3]. Anomalous sounds are rare, thus detecting such sounds can be formulated as an outlier detection problem. Based on the baseline, we refer to interpolation deep neural network (IDNN) and conduct the experimental evaluation using the development dataset. The AUC score has improved compared to the baseline on the test set.

The rest of this report is organized as follows. The method are introduced in Section 2. Experimental results on test set are shown in Section 3.



(a) ToyCar (b) ToyConveyor

(c) fan (d) pump

(e) slider (f) valve

Figure 1: log-Mel spectrograms of the normal sound

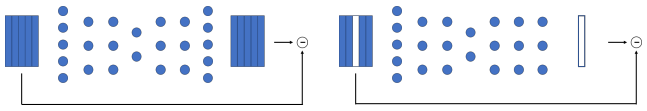
2. METHOD

Autoencoders are employed for unsupervised anomaly detection based on reconstruction errors. To detect anomalies, multiple frames of a spectrogram are used as an input, and the same number of frames are generated by the model as an output, the reconstruction errors are calculated from the mean square error between the input and output. The model is trained with normal data and learns to minimize reconstruction errors. Since the autoencoder is trained with normal data, the reconstruction error of normal data is small while the abnormal data is excepted to be large. So, the reconstruction errors are often used as anomaly scores. Autoencoders are trained to minimize reconstruction errors given as follows:

$$L_{AE} = \|x - D(E(x))\|_2^2$$

where x denotes an input, E represents an encoder and D represents an decoder.

The conventional autocoder remains some issues, for example, the reconstruction error tends to be large due to the difficulty of predicting the edge frame, especially for non-stationary sound. As Figure 1 shows, non-stationarity can be seen in the valve sound.We propose a novel network to solve the issue. The model utilizes multiple frames of a spectrogram whose center frame is removed as an input, and it predicts the removed frame as an output, which it can



(a) The conventional AE (b) The novel AE

Figure 2: Architecture of the conventional AE and the novel AE

be considered an interpolation of the removed frame. So anomalies can be detected based on an interpolation error that is the mean square error between the predicted frame and the true frame. The loss function is given as follows:

$$L_{IDNN} = \left\| x_{\frac{n+1}{2}} - D(E(x_{1,\ldots,\frac{n+1}{2}-1,\frac{n+1}{2}+1,\ldots,n})) \right\|_2^2$$

where n is the sum of the number of the input frames and the output frame.

3. EXPERIMENT

The data used for this task comprises parts of ToyADMOS [4] and the MIMII Dataset [5] consisting of the normal/anomalous operating sounds of six types of toy/real machines. Each recording is a 10-sec length audio that includes both a target machine's operating sound and environmental noise. Each machine type consists of several individual machines. We conducted an experiment to evaluate the performance of the method.

A logMel spectrogram was used as an input feature, we produce 128 log mel-bank magnitudes in which the frame size was set to 1024, the hop size was set to 512. Different from the traditional method which concatenating five frames as an input feature and generating the same number of frames as an ouput, the novel model attempts to only predict the center frame that is removed from the consecutive frames as the input, which it can be considered an interpolation of the removed frame. In that case, the reconstruction error tends to be small without predicting the edge frame, especially for non-stationary sound.

The novel autoencoder network comprises FC(Input, 128), FC(128, 128), FC(128, 128), FC(128, 128), FC(128, 8), FC(8, 128), FC(128, 128), FC(128, 128), FC(128, 128), FC(128, Output), where FC(a, b) represents a fully-connected layer with input neurons a, an output layer b, and each FC is followed by BatchNormalization and ReLU activation functions. The network was trained with an Adam optimization technique and the performance was evaluated based on the area under the curve (AUC) of the receiver operating characteristic. It is worth noting that, the number of parameters of the novel model is relatively small without reconstructing the whole input feature. The number of conventional auoencoder 270.0k while the novel model is 187.6k. Figure 1 shows the results of averaged AUC with the conventional autoconder and the novel autoencoder.

Table 1: Average AUC scores of six sounds

Method	Car	Conv	fan	pump	slider	valve
AE	0.80	0.73	0.65	0.73	0.85	0.66
IDNN	0.81	0.74	0.68	0.73	0.82	0.84

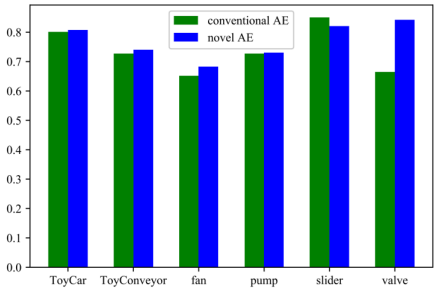


Figure 3: Averaged AUC of the conventional AE and novel AE

4. REFERENCES

[1] http://dcase.community/challenge2020/.

[2] Kaori Suefusa, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, and Yohei Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. IEEE ICASSP*, 2020, pp. 271–275.

[3] Raghavendra Chalapathy, Sanjay Chawla, " Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407v2,* 2019

[4] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, Noboru Harada, and Keisuke Imoto. ToyADMOS: a dataset of miniature-machine operating sounds for anomalous sound detection. In Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 308C312. November 2019.

[5] Harsh Purohit, Ryo Tanabe, Takeshi Ichige, Takashi Endo, Yuki Nikaido, Kaori Suefusa, and Yohei Kawaguchi. MIMII Dataset: sound dataset for malfunctioning industrial machine investigation and inspection. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019), 209C213. November 2019.