UNSUPERVISED ANOMALOUS SOUND DETECTION VIA SEMI-SUPERVISED GANOMALY ADVERSARIAL TRAINING

Technical Report

Wenbin Zhu[1], *Jie Ou*[1], *Ying Zeng*[1], *Yi Zhou*[1], *Hongqing Liu*[2*]

[1] School of Communication and Information Engineering
Chongqing University of Posts and Telecommunications, Chongqing, China
[2] Chongqing Key Lab of Mobile Communications Technology
Chongqing University of Posts and Telecommunications, Chongqing, China
hongqingliu@outlook.com

ABSTRACT

This technical report describes the submission from our team for Task 2 of the DCASE2021 challenge Unsupervised Anomalous Sound Detection for Machine Condition Monitoring under Domain Shifted Conditions. In this work, we adopt a GANomaly semi-supervised anomaly detection method via adversarial training to perform anomalous sound detection. By using the conditional generation of the confrontation network, the generator network effectively fits the data distribution of the normal samples during training, and calculates the reconstruction error of the anomaly score of the test samples.

Index Terms**—** Anomalous Sound Detection, Unsupervised, Fault Detection

1. INTRODUCTION

This task is the follow-up of DCASE 2020 Task 2, and the IEEE Audio and Acoustic Signal Processing Association's 2021 ”Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge” (DCASE) still regards the detection of abnormal sounds in machines as one of the tasks. The purpose of the task is detecting unknown anomalous sounds under the condition that only normal sound clips have been provided as training data, the same as in 2020. This task[1] is also performed under the conditions that the acoustic characteristics of the training data and the test data are different.

The two datasets (MIMII DUE[2] and ToyADMOS2[3]) consist of the normal/anomalous operating sounds of seven types of real/toy machines. Each recording is a single-channel 10 s long audio that includes both machine's operating sound and environmental noise. The source domain means the condition under which most of the training data was recorded, whereas the target domain means a different condition under which some of the test data was recorded. The source and target domains differ in terms of operating speed, machine load, viscosity, heating temperature, environmental noise, signal-to-noise ratio (SNR), etc.

To solve this task, we adopt a GANomaly[4] semi-supervised anomaly detection method via adversarial training. By using the conditional generation of the confrontation network[5], the network can effectively learn the data distribution of normal samples during training. Employing encoder-decoder-encoder sub-networks in the generator network enables the model to map the input data to a lower dimension vector, which is then used to reconstruct the generated output data. The reconstruction error is served as the anomaly score of the samples for anomaly detection.

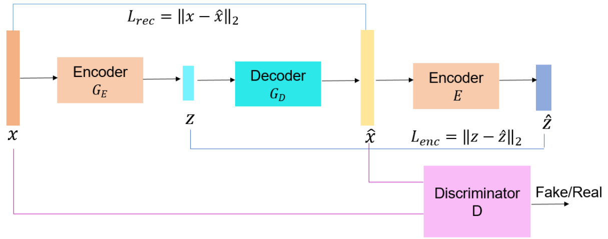

Figure 1: The structure of Ganomaly network

2. PROPOSED APPROACH

2.1. Data preprocessing

Following the baseline model, each input 10 s file is split into frames of length 64 ms, with a hop length of 32 ms between frames. 1024-FFT and 128 Mel bins are used to featurize each frame. In training, 10 frames are concatenated, resulting in 10 × 64 = 640 dimensional input, which are the log-Mel spectrograms computed using the above parameters.

2.2. frequency-based attention

As we know, different sound events have different spectral characteristics, and we should treat these frequency components differently according to the characteristics of each frame to better identify sound events. For some categories in the data set, the difference between normal sounds and abnormal sounds is particularly obvious in frequency components such as ToyCar. We adopt the method of frequency-based attention mechanism, which is similar to the method proposed by He et al.[6]. The input feature will go through a fully-connected layer with 64 hidden units, followed by an sigmoid. We then normalize the weights obtained along the frequency axis function. Finally, the frequency attention weight and the input feature are multiplied element-wise. The weighted feature

[*] This work was partly supported by National Natural Science Foundation of China under Grant 61971075.

Table 1: DCASE 2021 Task 2 Results over Dev Data.

	toycar	toytrain	fan	gearbox	pump	slider	valve
Baseline	63.19(52.42)	63.0(54.9)	64.03(53.58)	66.76(52.8)	63.66(54.74)	69.16(56.4)	53.74(50.61)
Proposed	64.5(55.5)	62.1(54.8)	61.4(53.2)	67.1(53.7)	62.6(55.5)	66.6(54.9)	51.3(50.2)

is computed as follows

$$\hat{\mathrm{W}}_{n,t} = sigmoid(\mathrm{V}_n\mathrm{X} + \mathrm{b}_n), \quad (1)$$

$$\mathrm{W}_{n,t} = N_f \frac{\hat{\mathrm{W}}_{n,t}}{\sum_n \hat{\mathrm{W}}_{n,t}}, \quad (2)$$

$$\bar{X} = \mathrm{W}_{n,t} \otimes \mathrm{X}, \quad (3)$$

where X is the input acoustic feature, V_n and b_n represent the weights and bias for the n-th hidden unit respectively, $\hat{\mathrm{W}}_{n,t}$ is the frequency attention weight without normalization, $\mathrm{W}_{n,t}$ is the normalized result, $\otimes$ represents element-wise multiplication, $\bar{X}$ is the weighted feature and N_f is the number of frequency points in the mel space.

2.3. GANomaly

For this task, since only the normal sound sample are provided for training, which indicates a unsupervised learning. To that end, we adopt a semi-supervised learning method GANomaly, which only trains normal samples and learns the high-dimensional features and latent spatial features of normal audio. During testing phase, if the test result is larger than distance metric from the normal sample distribution, it indicates that the test sample distribution has outliers, i.e., anomalous sound.

Figure illustrates the overview of this network that contains two encoders, one decoder, and discriminator networks, and they are employed within three sub-networks. First sub-network is an autoencoder network behaving as the generator part of the model.

The generator learns the representation of the input data and reconstructs the input log-mel spectrum via the use of autoencoder network, respectively. The structure of the first sub-network is as follows. The generator G first obtains the input log-mel spectrum x, and forward-passes it to the encoder network G_E. The encoder network G_E contains five convolutional layers, and each layers followed by a batch-norm[7] and Swish[8] activation. The Swish activation is defined by

$$f(x) = x \cdot sigmoid(\beta x), \quad (4)$$

where β is a constant or trainable parameter. When $\beta = 0$, Swish is a linear activation function, and when $\beta \to \infty$, Swish becomes a ReLU function. In this sense, the Swish function can be regarded as a smooth function between the linear function and the ReLU function. The studies show that this activation on deep models is better than ReLU function.After encoder, G_E compresses x into a high-dimensional feature vector z that is the bottleneck features of G, where $z = G_E(x)$.The decoder part G_D is the inverse process of the encoder, which uses the convolutional transpose layers, Swish activation, and batch-norm. The decoder upscales the vector z to reconstruct the log-mel spectrum x as $\hat{x}$, where $\hat{x} = G_D(z)$.

The second sub-network is an encoding network E, and this network is used to compress the log-mel spectrum $\hat{x}$ reconstructed by generator G. Although E uses the same structure as G_E, their parameters are obviously different, where E downscales $\hat{x}$ to its feature representation $\hat{z} = E(\hat{x})$. This structure is the core of the network. It abandons most of the autoencoder-based anomaly detection methods that are commonly used to infer anomalies by comparing the difference between the original data and the reconstructed data, and adopts a new method of comparing the original data. The difference between the data and the reconstructed data in the higher level of abstraction space is the way to infer anomalies, and this additional level of abstraction greatly improves the ability to resist noise interference and learn a more robust anomaly detection model.

The first sub-net work and the second sub-network forms the generative network (G-Net) in the generative discriminant network. G-net effectively learns the feature distribution of normal audio data, and when the faulty data is the input to the network, the generator G cannot effectively reconstruct the input x, the vector $\hat{z}$ obtained by E will be more dissimilarity to vector z.

The third sub-network is a discriminant network (D-Net), which is used to distinguish the original data x as true and the reconstructed data $\hat{x}$ as fake. Its structure is the same as G_D. It is of interest to point out that using the idea of generative confrontation can train a better G-Net.

3. RESULTS

In this task, we only report results using the development set. In Table 1, we present AUC results and pAUC in parentheses for both the challenge baseline autoencoder model and our submissions for this task.

4. REFERENCES

[1] 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, 1–5*, 2021.

[2] 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, 1–4*, 2021.

[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," *arXiv preprint arXiv:2106.02369*, 2021.

[4] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian conference on computer vision*. Springer, 2018, pp. 622–637.

[5] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient gan-based anomaly detection," *arXiv preprint arXiv:1802.06222*, 2018.

[6] Y.-H. Shen, K.-X. He, and W.-Q. Zhang, “Learning how to listen: A temporal-frequential attention model for sound event detection,” *arXiv: Sound*, pp. 2563–2567, 2019.

[7] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[8] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for activation functions,” *arXiv preprint arXiv:1710.05941*, 2017.