UNSUPERVISED LEARNING FOR ANOMALOUS SOUND DETECTION BASED ON PREDICTION AND RECONSTRUCTION TASKS

Technical Report

Fengrun Zhang, Chenguang Hu, Kai Guo

Beijing Institute of Technology, School of Information and Electronics, Beijing, China
{3120220766,3220200551}@bit.edu.cn, guokai001123@126.com

ABSTRACT

The purpose of anomalous sound detection is to detect whether the sound emitted by the machine is normal or anomalous. Due to the scarcity and diversity of anomalous data, only normal audio data is used to detect anomalies. The DCASE 2023 challenge is dedicated to developing a general-purpose anomalous detection algorithm that has good anomalous detection results on different machine types. For the problem scenario of DCASE 2023, we have developed four systems for anomalous sound detection, which are called VIDNN, CPC-VAE, VAE-GMM, DDPM.

Index Terms— anomalous sound detection, unsupervised learning, VAE, diffusion

1. INTRODUCTION

With the development of industrial modernization, artificial-intelligence-based factory automation, sound, as an important information modality, is very useful for monitoring the working status of machines. The task 2 of the Detection and Classification of Acoustic Scenes and Events (DCASE) is set to connect academic tasks and real-world problems in anomalous sound detection (ASD)[1].

DCASE 2023 challenge task 2 focuses on the following practical issues. Firstly, since anomalies rarely occur and highly diverse in real-world factories, only normal sounds are provided in the training data to detect unknown anomalous sounds. Secondly, in actual scenarios, a machine may have different working states, such as working voltage, running speed, but the difference between different working states is not the difference between normal and anomalous sounds. The difference between different work states is defined as domain shifts. In this task, the system is required to use domain-generalization techniques for handling these domain shifts. Lastly, this task hopes to develop a generalized anomalous sound detection system[2–5].

In the DCASE task 2 challenge of the past few years, discriminative methods based on auxiliary tasks using the machine type and machine identity (ID) tags attached to given dataset have emerged, which transform the binary classification task of anomalous detection into a machine id recognition problem[6]. However, its generalization is limited. Classification methods suffer from performance instability: performance varies even for machines of the same type. In DCASE 2023 task2, only a limited number of machines from its machine type is provided, which means that outlier exposure (OE) methods like the classification of machine IDs used as an auxiliary task in the past few years cannot be used. A common inlier modeling (IM) approach using unsupervised learning is to model the distribution of normal data through unsupervised learning[7]. According to whether the feature of a piece of audio belongs to this probability distribution, as a criterion for distinguishing, anomalies can be detected from a large amount of normal data.

In the IM approach field of anomalous detection, autoencoder is often adopted as a baseline. An autoencoder is mainly composed of an encoder and a decoder. Its main purpose is to convert the input into a low-dimensional intermediate variable, and then obtain an output from the intermediate variable, finally compare the input and output to make them as close as possible. There are many improved models based on autoencoder that have been applied to the field of anomalous sound detection, such as variational autoencoder (VAE)[8], interpolation deep neural network (IDNN)[9]. We assume that autoencoders trained with normal data can learn the distribution of normal data, so as to reconstruct normal data, and because the distribution of anomalous data is different from normal data, there will be a larger reconstruction error. We added Contrastive Predictive Coding (CPC)[10] to VAE, so that the model can not only be reconstructed, but also use the reconstructed features to predict features adjacent to the input features on the spectrum, which is called CPC-VAE.

Anomalous detection using generative models and reconstruction errors has proven to be a viable approach. Denoising Diffusion Probability Model (DDPM)[11] is an efficient generative model that has achieved leading performance in the fields of image generation, image anomalous detection, and speech synthesis, while there is no published research in the field of anomalous sound detection. We propose a new anomalous detection method based on DDPM, adding Gaussian noise to perturb the input spectrum, estimating the noise with DDPM and obtaining a high-quality approximation of the input. When the test sample is anomalous, the noise-perturbed spectrum will be reconstructed as an approximately normal sample, and the anomalous is detected by the inconsistency between the input spectrum and the generated sample.

The paper is organized as follows: Section 2 describes our submitted four systems, namely VIDNN, CPC-VAE, VAE-GMM[12], and DDPM. Section 3 describes experimental results of four systems on the development dataset and our discussions.

2. SYSTEM SUBMISSION

We introduce four models we used in the competition: VIDNN, CPC-VAE, VAE-GMM, and DDPM. The input characteristics of the four systems are different.Firstly we apply the short-time Fourier transform (STFT) with the Hann window of size 1024 to extract STFT feature. For VIDNN and CPC-VAE, the input is STFT

feature. Then pass the STFT features through the Mel filter and take the logarithm to get the Fbank features, with the number of Mel filter banks is set to 128. For VAE-GMM and DDPM, the input feature is Fbank. The hop size of frame shift is 160 for DDPM, and hop size of other systems is 512.

2.1. SYSTEM1:VIDNN

For VIDNN, continuously five frames are concatenated and used as a sample, and four frames in a sample are used as an input, and one frame was predicted as an output. The loss function of VIDNN is given as follows:

$$\mathcal{L}_{\text{VIDNN}} = \left\| \mathbf{x}_{\frac{n+1}{2}} - D\left(E\left(\mathbf{x}_{1,\ldots,\frac{n+1}{2}-1,\frac{n+1}{2}+1,\ldots,n}\right)\right)\right\|_2^2 \quad (1)$$

The specific structure of the model is shown in Table 1.

Table 1: VIDNN model architecture

Blocks	input size	output size
FC + BN + ReLU	B×2052	B×400
FC + BN + ReLU	B×400	B×128
FC + BN + ReLU	B×128	B×128
2× (FC + BN + ReLU)	B×128	B×24
Reparameterize	B×24×2	B×24
FC + BN + ReLU	B×24	B×128
FC + BN + ReLU	B×128	B×128
FC + BN + ReLU	B×128	B×400
FC + BN + ReLU	B×400	B×513

When calculating anomalous scores, we used Euclidean distance calculated by MSE loss as a reconstruction error.

2.2. SYSTEM2:CPC-VAE

Same as VIDNN, continuously five frames are concatenated and used as a sample. The first four frames are fed into the VAE for reconstruction. Then the reconstructed four frames predict the fifth frame through a layer of CPC network, and then concatenate the first four reconstructed frames with the predicted fifth frame and calculate the reconstruction error with the five frames of the input sample. CPC Network uses cosine similarity to evaluate the reliability of prediction, and uses Contrastive loss to enhance the network's ability. In formula 2, W_k is a linear predictor using GRU output c_t to predict the future information vector after the k-th step of future feature z, and c_t is the t-th contextual feature.

$$\mathrm{s}\left(\mathrm{c_t}, \mathrm{z}\right) = \frac{\langle \mathrm{W_k c_t}, \mathrm{z} \rangle}{\|\mathrm{W_k c_t}\| \, \|\mathrm{z}\|} \quad (2)$$

In order for the model to learn to distinguish the real vector $z_t + k$ from the N negative samples , the contrastive prediction loss of each step (k) prediction is used. The loss function of CPC network is given as follows:

$$\mathcal{L}_{k,t} = -\log \frac{\exp\left(s\left(c_t, z_{t+k}\right)/\tau\right)}{\exp\left(s\left(c_t, z_{t+k}\right)/\tau\right) + \sum_{i=1}^{N} \exp\left(s\left(c_t, z_{ki}^{-}\right)/\tau\right)} \quad (3)$$

where τ is the temperature parameter.

In our experiment, the step parameter k of future feature z is set to 1, the temperature parameter τ is set to 1. Using random shuffle while training, we consider that other samples in the same batch are randomly obtained negative samples in formula 4.

$$D_M(x, y) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)} \quad (4)$$

In the last epoch of the training process, we calculated the covariance matrix Σ_{source} and Σ_{target} for the source and target samples respectively. When calculating the Mahalanobis distance using formula 4, we calculated the Mahalanobis distance D_{source} and D_{target} of the samples in the source domain and the target domain respectively, and the sum of the two serves as the final anomalous score. Compared with calculating the Mahalanobis distance of the source domain and the target domain separately and taking a smaller one, the addition of the two has better domain generalization performance.

The specific structure of the model is shown in Table 2.

Table 2: CPC-VAE model architecture

Blocks	input size	output size
FC + BN + ReLU	B×2052	B×400
FC + BN + ReLU	B×400	B×128
FC + BN + ReLU	B×128	B×128
2× (FC + BN + ReLU)	B×128	B×24
Reparameterize	B×24×2	B×24
FC + BN + ReLU	B×24	B×128
FC + BN + ReLU	B×128	B×128
FC + BN + ReLU	B×128	B×400
FC + BN + ReLU	B×400	B×513
FC + BN + ReLU	B×400	B×2052
GRU	B×4×513	B×1024
FC	B×1024	B×513

2.3. SYSTEM3:VAE-GMM

For VAE-GMM, the 128-dimensional Fbank feature of every 5 frames is concatenate as a sample. We use MSE loss and KL divergence loss to train VAE, and save the latent features of the reparameterize layer as the fitting target data of GMM.

We save the latent features of all training samples to train the GMM model, and use the fitted GMM model to calculate the log likelihood of all training samples, and take the average as the $score_{train}$ of the normal sample. Then calculate the latent features for all samples of each audio i, use the fitted GMM model to calculate the log likelihood, and average all the scores to get the $score_i$ of the audio i in formula 5.

$$score_i = \mathcal{A}\left(h\left(f\left(\mathcal{X}_k\right)\right)\right) \quad (5)$$

X_k is the sample of all audio i, f is the encoder and reparameterization network of VAE, h is the GMM model, A is the function of averaging, and the scoring of all samples belonging to audio i is averaged.

Then the anomalous score of audio i is calculated by subtract the $score_{train}$ and $score_i$, representing the log-likelihood difference between the test audio and the training audio on the latent features.

The specific structure of the VAE model is shown in Table 3.

Table 3: VAE model architecture

Blocks	input size	output size
FC + BN + ReLU	B×640	B×128
FC + BN + ReLU	B×128	B×128
FC + BN + ReLU	B×128	B×64
FC + BN + ReLU	B×64	B×64
2× (FC + BN + ReLU)	B×64	B×30
Reparameterize	B×30×2	B×30
FC + BN + ReLU	B×30	B×64
FC + BN + ReLU	B×64	B×64
FC + BN + ReLU	B×64	B×128
FC + BN + ReLU	B×128	B×128
FC + BN + ReLU	B×128	B×640

2.4. SYSTEM4:DDPM

DDPM can be described as the asymptotic addition of gaussian noise N with standard deviation σ to input data points x sampled from distribution X of the training data, controlled by a time step parameter T. Sampling the sample $x_0 \sim$ X, as T increases, x_0 becomes an isotropic Gaussian noise distribution $p(x, \sigma)$. Sampling the sample $x_t \sim$ N, as T decreases, this point is gradually denoised is a new sample that obeys the distribution of the data set. The detailed derivation process can be found in [11].

In this experiment, DDPM is used to realize anomalous detection. This method first destroys the input anomalous spectrogram x_0, so that it is disturbed by noise within t time steps x_t, and then denoises the time step and restores it to x_0. The model tends to generate noisy samples as conforming to the distribution of training set. To detect anomalies of different sizes, we parameterize x_t as x_λ, where larger values of λ remove larger anomalies.

Same as [11], the denoising network is a U-Net structure with an encoder-decoder. We set the base channel to 64, attention resolution to 32, 16, 8, and head to 4. The encoder consists of 3 layers with 64, 128, 256 channels, and correspondingly, the decoder has 256, 128, 64 channels.

For each audio's FBank, we slice it into 128×128 images along the time dimension T, with no overlap between images. At inference time, for an outlier sample x_0 in the test set, we add noise to it to a parameterized time step λ , and then denoise to get $\bar{X}_0$. In order to detect anomalous samples, the reconstructed spectrogram is compared with the original spectrum, that is, the MSE between $\bar{X}_0$ of the reconstructed spectrum and the original spectrum x_0 is calculated as the anomalous score map, and the audio-level anomalous score $S_{anomalous}$ is calculated by formula 6.

$$\mathrm{S}_{anomalous} = \frac{1}{\mathrm{FT}} \sum_{i=1,j=1}^{i<F,j<T} (\mathrm{x_{ij}} - \hat{x}_{\mathrm{ij}})^2 \tag{6}$$

3. RESULTS AND DISCUSSIONS

The results of our submitted systems on the development dataset are demonstrated in appendix. Table4, Table5, Table6, and Table7 respectively illustrate the AUC and PAUC (p=0.1) of different systems in the source domain and target domain.

Same as the baseline, VIDNN, CPC-VAE after training for 100 epochs. The number of DDPM training steps is set to 64000. We found that the results of VAE-GMM will vary greatly on different epochs. We believe this is due to the fact that although the model loss function converges with increasing epochs, the latent variables in the intermediate layers are still changing. We show the best results in 100 epochs, which is different on different machine types. Since the evaluation set in 2023 is a completely new machine type, it is impossible to choose the best epoch for each machine, so we submit the results of training for 100 epochs.

Through experiments, we found that different machine types have different applicable systems. Some machines are suitable for detecting anomalies based on reconstruction tasks, such as fan and bearing, and some machines are suitable for detecting anomalies based on prediction tasks, such as valve. This means developing a general approach is very difficult. From the results, we can see that our four systems have a certain improvement compared with the baseline, especially DDPM, which is better than other kinds of anomaly detection in the target domain method.

References

[1] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on dcase 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," 2022.

[2] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," 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, 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.

[5] 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 ggeneralization task," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[6] Y. Zeng, H. Liu, L. Xu, Y. Zhou, and L. Gan, "Robust anomaly sound detection framework for machine condition monitoring," DCASE2022 Challenge, Tech. Rep., July 2022.

[7] P. Daniluk, M. Gozdziewski, S. Kapka, and M. Kosmider, "Ensemble of auto-encoder based systems for anomaly detection," DCASE2020 Challenge, Tech. Rep., July 2020.

[8] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013.

[9] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 271–275.

[10] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding." vol. abs/1807.03748, 2018.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[12] K. Tian, G. Fu, S. Li, G. Tang, and X. Shao, "Anomaly machine detection algorithm based on semi variational autoencoder of mel spectrogram," DCASE2020 Challenge, Tech. Rep., July 2020.

Table 4: AUC[%] for source data in baseline, proposed systems

machine	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve
AE+MSE	66.8	60.3	63.0	79.0	59.3	68.9	55.2
AE+MAHALA	71.7	58.5	63.0	81.8	71.9	**83.6**	55.5
VIDNN	65.7	59.6	65.3	76.6	56.2	74.5	**76.9**
CPC-VAE	68.7	48.8	60.9	**94.3**	**72.7**	79.8	55.5
VAE-GMM	**71.8**	**76.1**	69.9	84.7	43.9	53.6	69.3
DDPM	63.0	66.0	**76.9**	73.8	64.4	77.1	49.6

Table 5: AUC[%] for target data in baseline, proposed systems

machine	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve
AE+MSE	39.4	61.0	52.6	37.9	64.9	54.3	54.1
AE+MAHALA	39.8	43.2	55.7	46.1	68.5	70.9	52.3
VIDNN	38.5	**67.9**	54.0	35.0	60.8	52.5	**68.7**
CPC-VAE	38.8	40.1	62.1	55.2	**69.5**	71.1	51.7
VAE-GMM	59.3	33.8	**72.6**	56.6	64.9	**73.1**	57.9
DDPM	**62.9**	56.0	70.6	**81.3**	67.1	64.7	43.6

Table 6: pAUC[%] for source data in baseline, proposed systems

machine	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve
AE+MSE	49.3	47.6	60.0	57.5	52.8	59.6	56.4
AE+MAHALA	47.8	47.4	59.4	56.6	**56.2**	60.8	52.2
VIDNN	50.1	47.4	63.4	60.4	48.4	62.9	**61.9**
CPC-VAE	48.2	47.4	62.5	**76.2**	**56.2**	62.3	50.1
VAE-GMM	51.6	52.3	55.8	54.5	53.2	49.9	53.8
DDPM	**52.6**	**54.5**	**67.6**	67.2	53.7	**64.8**	52.6

Table 7: pAUC[%] for target data in baseline, proposed systems

machine	ToyCar	ToyTrain	bearing	fan	gearbox	slider	valve
AE+MSE	54.1	49.0	48.6	60.8	51.4	52.4	51.2
AE+MAHALA	49.9	49.7	**49.7**	**63.8**	49.9	50.1	50.7
VIDNN	49.5	**57.5**	49.1	64.6	49.5	52.2	**55.8**
CPC-VAE	49.9	49.3	49.3	62.1	52.4	49.3	50.7
VAE-GMM	50.3	46.7	49.5	56.4	53.5	54.3	50.7
DDPM	**54.3**	51.8	48.8	60.0	**62.1**	**56.0**	52.2