VAE-BASED ANOMALY DETECTION WITH DOMAIN ADAPTATION

Technical Report

Jun'ya Yamashita, Hayato Mori, Satoshi Tamura, Satoru Hayamizu

Gifu University
Faculty of Engineering Yanagido 1-1, Gifu, Gifu 5011193, Japan
junya@asr.info.gifu-u.ac.jp

ABSTRACT

This paper presents our anomaly detection scheme for DCASE 2021 Challenge, using a Variational AutoEncoder (VAE) with a framework of Interpolation Deep Neural Network (IDNN) and fine tuning as an adaptation method. VAE is built using normal training data for each machine, to predict a frame from its neighbor frames just like IDNN. In addition, we involve a kind of high-pass filter and a scheme to preserve particular frames or frequencies having larger errors. Finally an anomaly score is calculated based on reconstruction error in VAE. We further apply fine tuning to target data recorded in different settings, to adapt a model.

***Index Terms*—** Variational autoencoder, interpolation deep neural network, domain adaptation.

1. INTRODUCTION

Anomaly detection is a technique to detect abnormal data, using statistics, machine learning and deep learning technology. Since there are high demands to predict or detect any failure in industrial fields, many researchers have devoted their efforts to accomplish a high-performance anomaly detection technique. This paper reports our activities to DCASE 2021 Challenge Task 2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring under Domain Shifted Conditions [1].

In the anomaly detection field, an autoencoder is often adopted. An autoencoder, that employs DL architecture, converts given data into low-dimensional representations in an encoder part, followed by reconstructing the original data from the vectors in a decoder part. We build a model only using normal data, and compute an error between given and reconstructed data as an anomaly score. Since the model cannot well reconstruct anomaly data which are not used for model training, higher error scores are observed for the anomaly data.

In this work we focus on a Variational AutoEncoder (VAE) [2] to model normal data. VAE is one of the autoencoders, in which latent vectors obtained from an encoder should be observed based on a Gaussian distribution. It is confirmed that this model works well for machine sound data, on the other hand, it is not sure that this approach can be still effective for the same machine data in different settings or environments. In other words, we may need domain adaptation to a VAE model.

We also employ an Interpolation Deep Neural Network (IDNN) framework [3] in our scheme. The technique has been successfully chosen in several anomaly detection tasks, including the last DCASE Challenge. IDNN tries to predict a frame from its neighbor frames. Once the model is built using normal data, abnormal data cannot be well predicted, resulting higher reconstruction error.

This paper explorers how to incorporate VAE and IDNN for the anomaly detection challenge with high performance. We at first measure an anomaly score by reconstruction error. Next, we try to improve the score based on our preliminary investigation, by employing a kind of high-pass filter and detecting periodic and frequency-dependent abnormal sounds. In addition, we carry our fine tuning as an adaptation method to improve detection accuracy for anomaly data.

2. METHODOLOGY

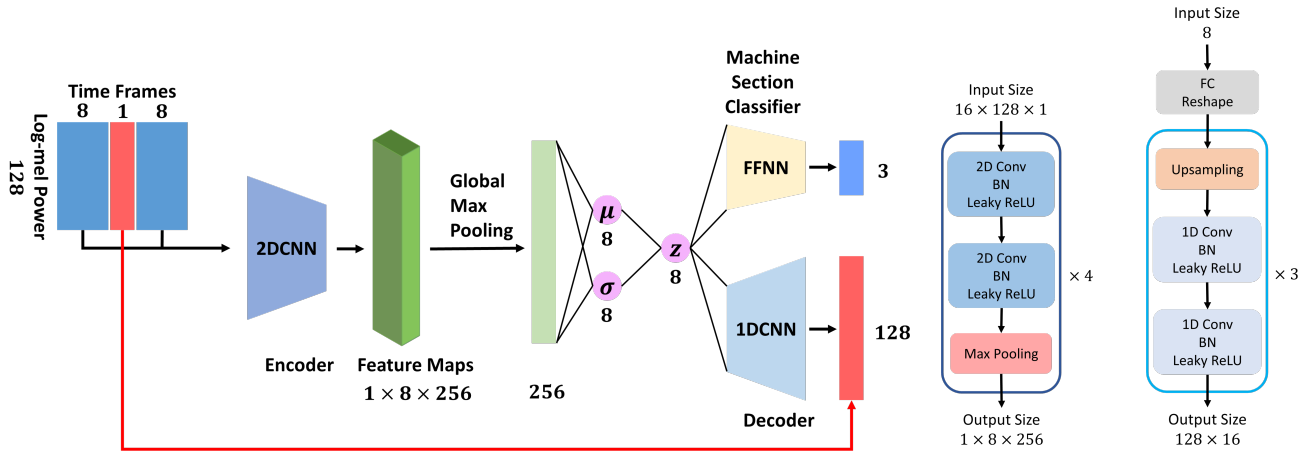
2.1. Preprocessing

First, we split a 10-second audio clip into 313 frames, with a frame length of 64ms and a frame shift of 32ms. In each frame, we secondly compute 128-dimensional log-scale mel-frequency power coefficients. For a particular frame, previous eight frames and following eight frames are collected, to obtain a 16x128 matrix as an input data to VAE.

2.2. VAE and IDNN

As mentioned, we chose VAE to obtain compact representations from original data, or estimate anomaly scores by reconstructing from the latent vectors. Here we also employ a strategy inspired by IDNN. Fig 1 illustrates our VAE model. The model is designed to predict a log-mel vector from its neighbor frames. Given the 16x128 input data, a 2D Convolutional Neural Network (CNN) is performed as an encoder to obtain 256 feature maps each of which size is 1x8 (see also Fig. 1(b)). By applying a Global Max Pooling (GMP), a 256-dimensional vector is subsequently calculated. GMP has been widely used nowadays in many DNN architectures. We consider each channel implies a particular pattern appeared in normal data, hence, it is expected that applying GMP simply summarizes the result. In addition, GMP can reduce model parameters compared to conventional pooling models. We then compute mean and variance variables by means of Feed Forward Neural Network (FFNN), to obtain a 8-dimensional latent vector z.

To enhance the whole model, we employ two models in this work. The first one is a 1D-CNN-based decoder shown in Fig. 1(c), followed by an additional convolution layer to predict the log-mel vector. Another is a simple FFNN-based classifier, which estimates a machine section. In the classifier, an accepted latent vector is directly converted into classification outputs without any hidden



(a) An overview (b) 2DCNN (c) 1DCNN

Figure 1: Our VAE model.

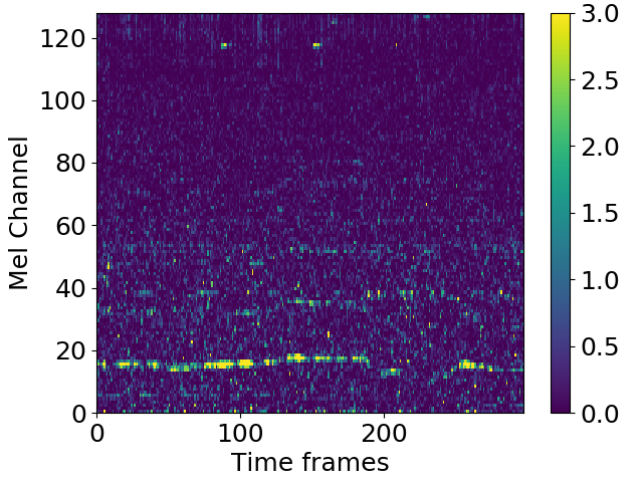


Figure 2: An example of reconstruction error visualization results.

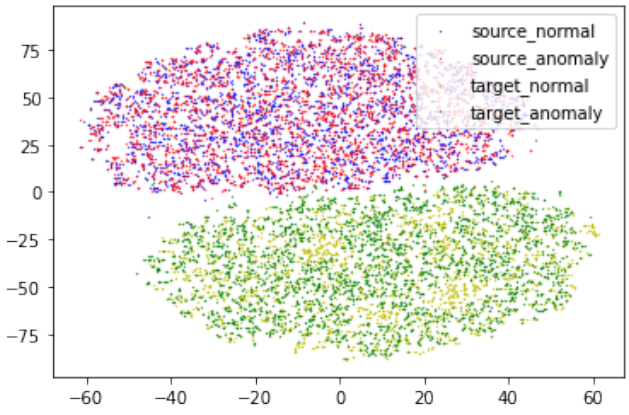


Figure 3: A t-SNE visualization result of VAE latent vectors.

layer. To optimize all parameters in these models, we adopt a mean-square-error loss function for the decoder and a categorical-cross-entropy loss function for the classifier.

We repeat the above process for each frame. Since we cannot prepare the input data fir the beginning frames and ending frames, we can obtain 297 predicted frames in one audio clip.

2.3. Anomaly score calculation

2.3.1. System 1

There are several frameworks we can choose to estimate anomaly scores. The most common approach is to measure a reconstruction error, by comparing an estimated vector with the original one. Let us denote original and reconstructed data by $x(i, j)$ and $y(i, j)$ in an i-th frame ($1 \leq i \leq 297$) at a j-th frequency bin ($1 \leq j \leq 128$), respectively. The reconstruction error can be calculated as:

$$E_1 = \sum_{i,j} \{x(i, j) - y(i, j)\}^2 \quad (1)$$

We can then carry out anomaly detection by comparing to a predefined threshold.

2.3.2. System 2

We further investigated the data in order to achieve better performance. First, we visualized the reconstruction error. Fig. 3 shows an example of the reconstruction result from a Pump normal data set. The horizontal axis indicates a frame index, and the vertical one means a frequency bin. We can see some large errors in the lower frequency part. Because we sometimes observed such the errors in normal data, therefore, we thus cut off any error below the particular frequency bin. This time we set the threshold bin as 20, roughly equivalent to 450Hz.

In some machines such as Valve, we often have periodic sounds in audio clips. To identify any anomaly, it may be good to firstly calculate the error within a frame:

$$E_i^C = \frac{1}{M - 20} \sum_{j=20}^{M} \{x(i, j) - y(i, j)\}^2 \quad (2)$$

Table 1: AUC [%] of baseline and our systems.

Domain	System	ToyCar	ToyTrain	fan	gearbox	pump	slide rail	valve
Source	Baseline	**67.98**	71.74	66.11	**62.58**	**71.26**	**78.19**	54.59
	System1	62.26	75.85	**68.00**	61.99	65.09	75.89	57.41
	System2	55.41	**82.07**	64.47	62.26	60.77	77.62	**75.52**
Target	Baseline	58.39	54.25	**61.96**	**70.94**	**56.04**	60.12	**52.89**
	System1	**64.76**	58.11	57.83	57.37	53.77	55.17	51.85
	System2	58.58	**64.82**	57.20	60.07	50.09	**60.44**	51.35

Table 2: pAUC [%] of baseline and our systems.

Domain	System	ToyCar	ToyTrain	fan	gearbox	pump	slide rail	valve
Source	Baseline	**53.08**	59.79	53.64	51.90	**58.52**	59.30	50.70
	System1	52.77	57.84	**55.03**	52.07	57.66	58.17	51.01
	System2	50.52	**63.66**	52.33	**52.76**	55.75	**64.05**	**59.00**
Target	Baseline	51.75	50.00	**53.51**	**53.70**	**50.95**	53.50	50.51
	System1	**52.77**	51.01	53.15	51.22	50.35	50.79	**50.70**
	System2	52.33	**53.10**	53.42	51.68	49.96	**54.25**	50.66

where N is the total number of frames, i.e. $N = 297$, and M is the number of frequency bins, i.e. $M = 128$. Similarly, if there is a stationary anomaly noise caused by any failure, we can always see the error in particular frequency bins. We thus also compute the average among all the frames at each frequency bin:

$$E_j^R = \frac{1}{N}\sum_{i=1}^{N}\left\{x(i,j) - y(i,j)\right\}^2 \tag{3}$$

We consider in such the above scenes choosing the whole average of reconstruction error, represented in Eq.(1), is not appropriate. We simply perform sorting to all E_i^C and E_j^R, and choose the top K values, followed by summing them up as an anomaly score E_2. Note that we set K as 15 in this paper.

2.4. Domain adaptation

In this challenge, for each machine and each section, a training data set is prepared including 1,000-clip source-domain data and 3-clip target-domain data only consisting of normal data [5] [6]. All the source data are used for building a VAE model in our scheme. It is thus expected to easily model the source data. On the other hand, there may be a difficulty to correctly predict a latent vector in the target domain. Therefore, before applying our method to the target domain, we conduct fine tuning to the decoder and FFNNs by using target data only.

To confirm this, we conducted visualization to source and target z. Fig. 3 depicts a visualization result of ToyTrain Section02 using t-SNE [4]. It is easily found that both distribution are quite different. That means a model built from source data is hardly suitable for target data to estimate anomaly score, and it indicates we need to adapt a model.

3. RESULT

Table 1 shows Area Under Curve (AUC) results of the autoencoder-based baseline [7] and our systems. Here we calculated a mean score among the three sections (Section00, 01 and 02). Regarding System 2, that is our improved method, improvements were observed in ToyTrain and valve. Checking audio clips and results, we found ToyTrain, slider rail and valve, in which our scheme was better than or almost the same as the baseline, have periodic sounds. It is considered that the approach to use top-K frame-by-frame error is thus effective. In contrast, our method unfortunately could not improve performance in fan, gearbox and pump. In audio clips of these machines we can find stationary sounds, so we tried to deal with frequency-dependent anomaly sounds. However, out system could not detect some of failures because our autoencoder did not sufficiently take care of the frequency domain. In terms of target data, our systems could not achieve significant improvement except ToyTrain. This might be caused because of the insufficient data amount of adaptation. It was also observed that a variance of AUCs became larger after fine tuning. These facts indicate that, in spite that we carefully tried to design models and parameters as few as possible, however, 3-clip target data might be too small for adaptation.

Table 2 shows pAUC results. Different from Table 1, our systems achieved better performance than the baseline for source data except ToyCar and pump. That indicates our systems can identify normal data with lower false positive acceptance.

4. REFERENCES

[1] http://dcase.community/challenge2021/task-unsupervised-detection-of-anomalous-sounds.

[2] D. Kingma and M. Welling, "Auto-encoding variational Bayes," *Proc. ICLR*, 2014.

[3] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," *Proc. ICASSP*, 2020.

[4] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol.9, pp.2579-2605, 2008.

[5] 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.

[6] 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.

[7] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, "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.