UNSUPERVISED ANOMALOUS SOUND DETECTION USING INTERMEDIATE REPRESENTATION OF TRAINED MODELS AND METRIC LEARNING BASED VARIATIONAL AUTOENCODER

Technical Report

Hiroki Narita

Aichi Institute of Technology
Graduate School of Business
Administration and Computer Science
Aichi, Japan

Akira Tamamori

Aichi Institute of Technology
Department of Information Science
Aichi, Japan

ABSTRACT

This paper is a technical report of DCASE Challenge2021 Task2. The objective of the DCASE Challenge2021 Task2 is unsupervised anomalous sound detection under domain shift. Our method consists of feature extraction using a pre-trained model and Center-Loss VAE (CL-VAE) based on Center-Loss and Variational Auto-Encoder (VAE). In feature extraction with pre-trained models, ResNet38 trained on acoustic data is used as a feature extractor to obtain intermediate representations. The CL-VAE is trained with the intermediate representations as input and is trained to minimize the Center-Loss of the section labels and the loss function of the VAE. As a result of validation on the development dataset, we confirmed that the performance of CL-VAE is superior to that of Conditional VAE (CVAE) using baseline models and section labels.

Index Terms— Transfer learning, deep metric leaning, center-loss, variational auto-encoder

1. INTRODUCTION

Anomalous sound detection is a technique to determine whether a machine is normal or anomalous based on its sound. It is difficult to identify a failure from the outside of a machine with a complex internal structure. However, if we can detect anomalous sounds emitted by a machine, we can quickly detect a failure. In the past, the DCASE Challenge2020 Task2 was held as a competition for anomalous sound detection [1]. The Challenge was a very difficult task that required the classification of normal or anomalous from only normal sound data with various sounds. Various methods for detecting anomalous sounds, including the Outlier Exposure (OE) approach, have been developed, and research on anomalous sound detection has made significant progress [2] [3]. The objective of this year's DCASE Challenge2021 Task2 is unsupervised anomalous sound detection under domain shift [4]. As in Challenge 2020, participants are only allowed to use normal data for training. In addition, each section has a Source / Target domain shift[1], and the model needs to be created using only the Source data and a small amount of Target data. Our proposed method consists of feature extraction based on [5] and CL-VAE to capture the section distribution. In [5], the representation in the pre-trained model of normal data is modeled by multivariate normal distribution (MVG), and the Mahalanobis distance from MVG is used for anomaly detection. However, in the Challenge2021 Task2 dataset, it was not easy to detect anomalies from a single MVG because the data distribution was assumed to be different between sections and domains. Therefore, we propose a VAE that introduces center-loss to capture the distribution of each section label. Our method improves the anomaly detection performance by learning to form clusters for each section label in the output layer of the encoder.

[1] *e.g.*, factory noise variations between domains

2. PROPOSED METHOD

2.1. Feature Extraction

We applied the same extraction method as [5]. While the authors in [5] utilized the architecture of EfficientNet [6] as feature extractor, we applied PANNs ResNet38 [7] architecture . The weight parameters of the network are provided as pre-trained model.

In the feature extraction, the output of each convolutional layer is treated as a feature vector. The shape of the output from each convolutional layer is $(N, C, H, W)^2$, and the values are averaged over H and W to take representative values for each channel.

$$(N, C, average(H \times W)) = (N, C) \tag{1}$$

These feature vectors are concatenated over convolutional layers and we can obtain the final feature vector with shape (N, d). The dimension d is calculated as follows:

$$d = \sum_{i=1}^{n} C_i, \tag{2}$$

where n is the number of convolutional layers, C_i is the number of channels in i-th convolutional layer. To reduce the dimensionality d, we used the output from BasicBlock[3] to create a $d = 3776$ feature vector.

2.2. Center-Loss Variational Auto-Encoder

As shown in Figure 1, the architecture of CL-VAE has a typical VAE structure. The difference between them is that center-loss [8]

[2] N : Samples, C : Channel, H : Height, W : Width

[3] (Conv2D, BatchNorm2D, ReLU)$\times 2$

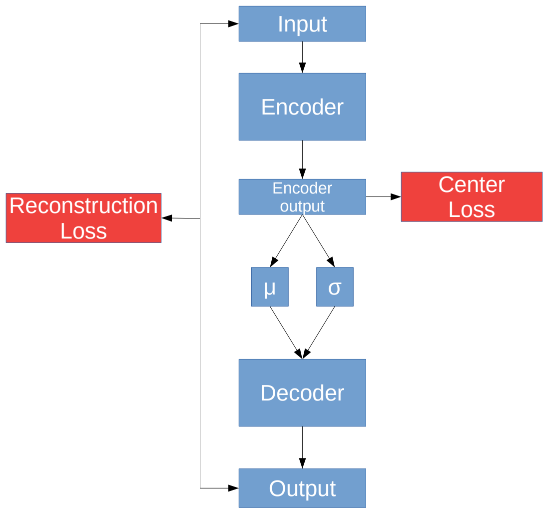

Figure 1: Architecture of CL-VAE

is calculated in encoder output. The center-loss L_C, a loss function, can be written as follows:

$$L_C = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2, \tag{3}$$

where i is the class label and m is the number of sections. The L_C can be utilized to minimize the distance between the center c_{y_i} of class y_i and feature vector x_i.

In addition, the center c_{y_i} is updated based on the following equation for each mini-batch.

$$\Delta c_j = \frac{\sum_{i=1}^{m} \delta(y_i = j) \cdot (c_j - x_i)}{1 + \sum_{i=1}^{m} \delta(y_i = j)} \tag{4}$$

$$c_j^{t+1} = c_j^t - \alpha \cdot \Delta c_j^t \tag{5}$$

where j is the class label in the mini-batch, c_j is the center of each class for each mini-batch, and α is a hyperparameter. δ is a function that is set to 1 when a label is matched.

The loss function of CL-VAE can be obtained with the reconstruction loss L_{rec}, Kullback-Leibler regularizer L_{kld}, and the center-loss L_C.

$$L_{rec} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \tag{6}$$

$$L_{kld} = -\frac{1}{2}\sum_{i=1}^{N}(1 + \log(\sigma_i) - \mu_i^2 - \sigma_i^2) \tag{7}$$

$$L_{clvae} = L_{rec} + L_{kld} + \lambda L_C \tag{8}$$

where x_i is the input of i-th mini-batch, $\hat{x}_i$ is the corresponding reconstruction, and N is the size of mini-batch. μ and σ are the parameters of the Gaussian distribution assumed in the VAE. λ is a hyperparameter for the center-loss weight.

The following two combinations were applied for anomaly scores, and the parameters of CL-VAE is shown in Table 1.

$$L_{rec} + L_{kld} \tag{9}$$

$$L_{rec} + L_{kld} + \lambda L_C \tag{10}$$

Table 1: Parameters of CL-VAE

	Architecture
Encoder	Input(3776) FCBlock(1024) $\times 3$ FCBlock(512) Reparameterization(512), CenterLossLayer(6)
Decoder	FCBlock(512) FCBlock(1024)$\times 3$ FCBlock(3776) (activation : ReLU)
FCBlock : Linear, Batchnorm, ReLU	
center-loss λ : 150	
center-loss α : 1	

2.3. Preprocessing

When inputting the data into the feature extractor described in Section 2.1, we set the parameters to match those of PANNs ResNet38 as shown in Table 2 and generated a log mel-spectrogram. In addition, the audio file in the dataset had a sampling rate of 16k, so we resampled it to 32k to fit the pre-trained model.

Table 2: Parameters of preprocessing

Parameter	**Value**
sample rate	32000
window size	1024
hop size	320
mel bins	64
fmin	50
fmax	14000

2.4. Postprocessing

For the anomaly detection threshold, the anomaly score of the normal data for each machine type was fitted with a gamma distribution as in the baseline system, and the 90th percentile was set as the anomaly.

3. DATASET

The DCASE Challenge2021 Task2 dataset consists of the MIMII DUE [9] and ToyADMOS2 [10] datasets, which have seven machine types. In addition, each machine type has five different labels, called ”sections”, and each section has a corresponding domain shift.

- ToyCar (ToyADMOS2)
- ToyTrain (ToyADMOS2)
- Fan (MIMII DUE)
- Gearbox (MIMII DUE)
- Pump (MIMII DUE)
- Slide rail (MIMII DUE)
- Valve (MIMII DUE)

4. RESULTS

We have submitted the following two systems as our submissions. The detailed scores are shown with reference to the organizer's overview paper [11]. The performance of the autoencoder-based anomaly detection system from [11] is referenced and shown in Table 3. The performance of CVAE, which was not used in the submission but used section labels for comparison, is shown in Table 4. This CVAE used the same input features as the CL-VAE.

- System 1 (Narita_AIT_task2_1)
  - Performance is shown in Table 5
  - Results predicted by Eq.(9).
- System 2 (Narita_AIT_task2_2)
  - Performance is shown in Table 6
  - Ensemble of Eq.(9) and Eq.(10). However, we multiplied Eq. (10) by 0.01 to adjust the scale.

5. REFERENCES

[1] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, *et al.*, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. 5th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE).*, 2020, pp. 81–85.

[2] R. Giri, S. V. Tenneti, K. Helwani, F. Cheng, U. Isik, and A. Krishnaswamy, "Unsupervised anomalous sound detection using self-supervised classification and group masked autoencoder for density estimation," DCASE2020 Challenge, Tech. Rep., July 2020.

[3] P. Primus, "Reframing unsupervised machine condition monitoring as a supervised classification task with outlier-exposed classifiers," DCASE2020 Challenge, Tech. Rep., July 2020.

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

[5] O. Rippel, P. Mertens, and D. Merhof, "Modeling the distribution of normal data in pre-trained deep features for anomaly detection," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 6726–6733.

[6] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[7] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[8] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[9] 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, 14*, 2021.

[10] 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.

[11] 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," 2021.

Table 3: Results of the AE-based baseline (Official Score)

Section			AUC [%]		pAUC [%]	
			Source	Target	Source	Target
ToyCar	Dev.	00	67.63	54.50	51.87	50.52
		01	61.97	64.12	51.82	52.14
		02	74.36	56.57	55.56	52.61
ToyTrain	Dev.	00	**72.67**	56.07	69.38	50.62
		01	72.65	51.13	62.52	48.60
		02	69.91	55.57	47.48	50.79
Fan	Dev.	00	66.69	69.70	57.08	55.13
		01	67.43	49.99	50.72	48.49
		02	64.21	**66.19**	53.12	56.93
Gearbox	Dev.	00	56.03	74.29	51.59	55.67
		01	72.77	72.12	52.30	51.78
		02	58.96	66.41	51.82	53.66
Pump	Dev.	00	67.48	58.01	61.83	51.53
		01	82.38	47.35	58.29	49.65
		02	63.93	**62.78**	55.44	51.67
Slide rail	Dev.	00	74.09	**67.22**	52.45	57.32
		01	82.16	66.94	60.29	53.08
		02	78.34	46.20	65.16	50.10
Valve	Dev.	00	50.34	47.12	50.82	48.68
		01	53.52	56.39	49.33	53.88
		02	59.91	**55.16**	51.96	48.97
MEAN			67.49	59.23	55.28	51.99

Table 4: Results of the CVAE

Section			AUC [%]		pAUC [%]	
			Source	Target	Source	Target
ToyCar	Dev.	00	72.18	69.87	54.58	55.68
		01	55.80	78.81	48.47	51.95
		02	79.31	67.90	60.58	59.21
ToyTrain	Dev.	00	70.66	53.78	57.00	52.42
		01	59.91	56.47	57.16	53.11
		02	56.61	59.37	48.26	58.58
Fan	Dev.	00	62.33	67.92	55.89	51.53
		01	54.35	46.09	49.74	48.89
		02	48.90	44.91	50.21	50.05
Gearbox	Dev.	00	61.52	84.65	60.66	76.62
		01	89.62	92.84	80.10	84.03
		02	71.54	72.63	60.36	60.13
Pump	Dev.	00	66.91	41.98	63.47	49.05
		01	45.93	43.22	49.05	48.58
		02	63.86	57.35	60.11	51.53
Slide rail	Dev.	00	54.00	57.33	50.37	55.21
		01	68.12	41.22	64.47	47.79
		02	62.39	61.59	55.34	50.89
Valve	Dev.	00	53.24	48.45	52.84	50.16
		01	47.43	67.56	49.84	61.63
		02	57.19	49.94	52.58	49.79
MEAN			61.78	58.69	56.74	55.75

Table 5: Results of CL-VAE (System 1)

Section			AUC [%]		pAUC [%]	
			Source	Target	Source	Target
ToyCar	Dev.	00	73.89	79.72	67.47	68.16
		01	76.84	93.54	60.47	82.16
		02	86.16	81.70	64.89	69.95
ToyTrain	Dev.	00	68.72	**67.79**	61.95	**58.32**
		01	72.42	**69.02**	70.00	60.58
		02	82.75	**82.63**	**51.47**	74.00
Fan	Dev.	00	**67.56**	**71.75**	**60.05**	**61.58**
		01	78.87	64.77	71.21	54.32
		02	63.06	58.28	54.16	54.32
Gearbox	Dev.	00	67.20	91.10	61.72	83.11
		01	96.26	94.67	87.57	84.57
		02	78.10	80.04	**70.36**	68.67
Pump	Dev.	00	71.41	62.17	59.32	57.63
		01	90.58	71.50	75.95	57.63
		02	70.04	55.97	**63.47**	53.74
Slide rail	Dev.	00	79.77	66.36	61.89	53.89
		01	93.46	68.92	**80.89**	61.57
		02	77.09	52.93	68.34	49.97
Valve	Dev.	00	77.99	66.15	**63.37**	63.63
		01	72.49	75.31	63.74	69.68
		02	80.45	46.14	72.00	49.79
MEAN			77.39	71.45	66.20	63.68

Table 6: Results of CL-VAE ensemble (System 2)

Section			AUC [%]		pAUC [%]	
			Source	Target	Source	Target
ToyCar	Dev.	00	**75.46**	**79.80**	**68.21**	**68.42**
		01	**82.18**	**94.31**	**66.32**	**85.21**
		02	**87.87**	**82.28**	**70.53**	**72.84**
ToyTrain	Dev.	00	71.48	65.25	**62.74**	53.84
		01	**74.44**	68.88	**70.89**	**61.32**
		02	83.23	82.24	48.05	**74.05**
Fan	Dev.	00	65.55	70.59	59.89	61.37
		01	**80.13**	**68.42**	**73.21**	**54.74**
		02	**74.46**	57.78	**66.11**	**54.74**
Gearbox	Dev.	00	**76.82**	**94.27**	**66.40**	**86.24**
		01	**96.48**	**95.86**	**89.93**	**87.15**
		02	**81.35**	**81.25**	67.55	**70.01**
Pump	Dev.	00	**72.46**	**71.17**	**62.05**	**61.95**
		01	**93.43**	**74.17**	**79.89**	**62.63**
		02	**73.66**	59.34	60.11	**55.79**
Slide rail	Dev.	00	**80.94**	61.02	**63.05**	**55.00**
		01	**93.91**	**69.51**	78.84	**63.42**
		02	**81.94**	**63.59**	**76.64**	**54.61**
Valve	Dev.	00	**78.85**	**66.73**	63.00	**64.47**
		01	**76.07**	**84.84**	**66.26**	**75.00**
		02	**82.22**	51.50	**72.11**	**50.63**
MEAN			**80.14**	**73.47**	**68.18**	**65.40**