OUTLIER-AUGMENTED CONTRASTIVE CLUSTERING FOR ANOMALY SOUND DETECTION WITH UNBALANCED DOMAIN

Technical Report

You-Siang Chen and *Mingsian R. Bai*

National Tsing Hua University, Hsinchu 30013, Taiwan,
s108033851@m108.nthu.edu.tw and msbai@pme.nthu.edu.tw

ABSTRACT

In this report, we developed a deep neural network (DNN) that can perform the deep clustering for the embedding vectors of machine sounds. The time-dilated convolutional neural network (TDCN) with attention mechanism was exploited to extract important features related to the time sequence. In addition, frequency masking is applied to the non-target sections of the machine sound to further increase the data size of the outliers. The results show that by applying the data augmentation to the outliers, the AUC performance can be improved. Furthermore, the deep clustering is able to contrastively attract and separate the machine sounds with unbalanced domain.

***Index Terms*—** Time-dilated convolutional network, attention mechanism, frequency masking, deep clustering

1. INTRODUCTION

DCASE task 2 [1] aims to perform the anomalous sound detection (ASD) under the unsupervised scenario. That is, the normal sound is the only available data. Like the DCASE2021 challenge task 2, the domain shift problem is also considered in this year. In DCASE2021, several methods of the self-supervised classifier [2-5] were used to deal with the domain-shifted ASD. However, we found that most of methods do not effectively improve the performance of the target domain where the data is extremely few in the training set. In our submission, we rearrange the classifier-based network to the clustering-based architecture where the embedding vectors of machine sounds in different sections is generated and clustered in an end-to-end scheme. For the generation of the embedding vectors, TDCN [6] with attention mechanism [7] is used to extract the important features according to the contextual relations of the machine sounds. In order to improve the domain generalization, the deep clustering (DC) loss [8] is utilized to attract the data of source and target domain within a specific section and separate those not in the same sections. Furthermore, frequency masking [9] is applied to the outliers in the training set to increase the data size.

2. NETWORK ARCHITECTURE

The DNN utilized for deep clustering is shown in figure 1. The architecture consists of the time-dilated convolutional network (TDCN) and the attention-based structure. The TDCN is constructed based on multiple 1-D convolution blocks where the detailed structure is illustrated on the left in figure 1. The repeated convolutional block can further extend the receptive field of the feature extraction. Then, the attention mechanism with transformer structure is applied to attain the important fractions of the sound features. After that, the statistical pooling is employed to obtain the mean and variance along the time sequence. Finally, the embedding vectors of different sections can be extracted through the linear layers. The important hyper-parameters are summarized in Table 1 and 2.

Table 1. Hyper-parameters applied in the TDCN

Symbol	Parameter	Description
M	128	Log-mel bins
B	64	Channels in bottleneck
H	128	Channels in convolutional blocks
P	3	Kernel size in convolutional blocks
D	6	Convolutional blocks in each repeat
R	3	Number of repeats

Table 2. Hyper-parameters applied in the attention mechanism with transformer structure

Symbol	Parameter	Description
q	32	Query in attention
k	32	Key in attention
v	32	Value in attention
h	1024	Hidden layers in Feed Forward

In the transformer structure, only 1 layer and 1 head are applied and the dimension of the embedding vector *N* is selected to be 64.

3. LOSS FUNCTION

The training objective is to minimize the DC loss [8] that is defined as

$$\mathrm{L}_{\mathrm{DC}} = \left\| \hat{\mathbf{Y}}\hat{\mathbf{Y}}^T - \mathbf{Y}\mathbf{Y}^T \right\|_F^2 = \left\| \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} \right\|_F^2 - 2\left\| \hat{\mathbf{Y}}^T\mathbf{Y} \right\|_F^2 + \left\| \mathbf{Y}^T\mathbf{Y} \right\|_F^2 \quad (1)$$

in which $\hat{\mathbf{Y}}$ is constituted with the embedding vectors $\hat{\mathbf{y}}$, $\mathbf{Y}$ is constructed by the one-hot vectors corresponding to the section IDs,

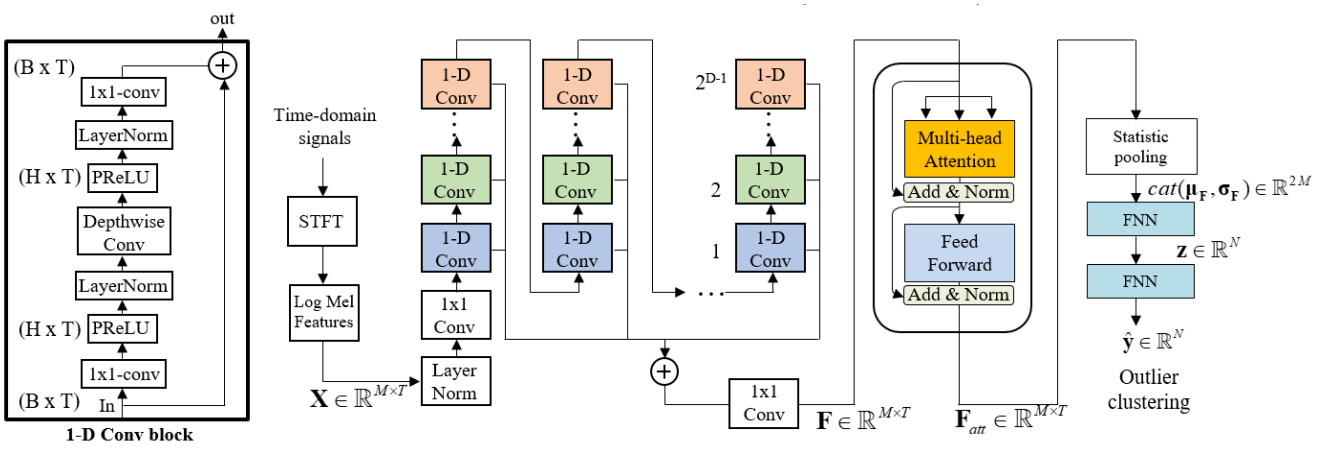

Figure 1. The workflow of the TDCN with attention mechanism

and $\| \bullet \|_F$ represents the Frobenius norm. The DC loss performs inner-product to every embedding vector making the in-class data attract together and out-class data separate apart. In doing so, the clustered vectors of the source and target domain can be generated with similar features. The following anomaly score is defined to calculate the state of the machine sound:

$$\text{Anomaly score} = 1 - \frac{\mathbf{y}_i^T \mathbf{y}_c}{\|\mathbf{y}_i\| \|\mathbf{y}_c\|} \quad (2)$$

where the $\mathbf{y}_c$ is the centroid vector of the target section ID that is calculated based on the training set, and $\mathbf{y}_i$ is the output embedding vector of the testing data. In order to make sure the result of 0 represents normal state and 1 stands for anomalous state, the activation function of the final layer is set to be ReLU.

4. DATASETS

We only use the development set of DCASE2022 task2. [10, 11] The sampling rate of the input audio is 16 kHz. 10-second audio file is firstly performed with 512 points short-time Fourier transform (STFT). Then, the log-Mel spectrogram with 128 bins is applied to reduce the dimension of the feature. To increase the variety of the outlier samples, the mel features of the non-target section are applied with frequency masking, while the data in the target section remains the same.

5. SUMMARY OF RESULTS

The evaluation of the ASD is based on area under curve (AUC). In Table 1, the result of the Toy Car is shown to demonstrate the efficacy of applying the frequency masking. The AUC performance is improved the section ID0 and ID1. As a result, the overall AUC score for Toy Car is improved by 10 percent.

Table 3. AUC score (%) of ToyCar in the development set

Data augmentation	ID0 AUC	ID1 AUC	ID2 AUC	Ave. AUC	Ave. pAUC
w/o masking	62.02	71.94	82.24	72.06	57.50
w masking	66.42	85.72	82.82	78.32	67.17

Table 4. Harmonic mean of the AUC score (%) for every machine type in the development set

	ToyCar	ToyTrain	Fan	gearbox	bearing	slider	valve
Source h-mean AUC	78.96	75.20	78.40	89.56	76.52	97.34	87.74
Target h-mean AUC	77.83	53.99	63..04	76.25	74.73	82.39	89.38
h-mean AUC	78.54	64.91	71.19	83.38	75.84	90.25	88.66
h-mean pAUC	61.23	55.08	61.33	65.22	65.16	76.54	79.05

The AUC scores for each machine type are summarized in Table 4. The result reveals that the data samples in target domain can achieve comparable performance with those in source domain for the cases of Toy Car, bearing and valve. For the other machine types, although the performance gap between source and target domain is around 15 to 20 percent, the overall results for the AUC and pAUC score is still good compared to the baseline models [12].

6. CONCLUSIONS

In this report, we have presented a clustering-based neural network that generate the embedding vectors for the ASD with unbalanced domain data. The DC loss was applied to attract the in-class data including the source and target domain, while the out-class data would be further separated apart from the required section. In addition, frequency masking was applied to the non-target sections of the machine sound to further increase the data size of the outliers. The results show that the frequency masking is effective and the proposed framework can improve the generalization of the unbalanced data domain.

7. REFERENCES

[1] "DCASE2022 Challenge challenge on detection and classification of acoustic scenes and events," https://dcase.community/contact-us, accessed: 2022-06-15.

[2] P. Primus, V. Haunschmid, P. Praher, and G. Widmer, "Anomalous sound detection as a simple binary classification problem with careful selection of proxy outlier examples," *arXiv preprint arXiv:2011.02949*, 2020.

[3] R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, "Self-supervised classification for detecting anomalous sounds." in *DCASE*, 2020, pp. p. 46-50.

[4] J. A. Lopez, G. Stemmer, P. Lopez-Meyer, P. Singh, J. A. del Hoyo Ontiveros, and H. A. Cordourier, "Ensemble of complementary anomaly detectors under domain shifted conditions." in *DCASE*, 2021, pp. 11–15.

[5] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in *neural information processing systems*, vol. 30, 2017.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[8] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv prerint arXiv:1904.08779*, 2019.

[9] 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 generalization task," *arXiv preprint arXiv:2205.13879*, 2022.

[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] N. Michael, S. Shen, K. Mohta, Y. Mulgaonkar, V. Kumar, K. Nagatani, Y. Okada, S. Kiribayashi, K. Otake, K. Yoshida, K. Ohno, E. Takeuchi, and S. Tadokoro, "Collaborative mapping of an earthquake-damaged building via ground and aerial robots," *Journal of Field Robotics*, vol. 29, no. 5, pp. 832–841, 2012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21436

[12] 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 ," *arXiv preprint arXiv*: *2206.05876*, 2022.