JLESS SUBMISSION TO DCASE2022 TASK2: BATCH MIXING STRATEGY BASED METHOD WITH ANOMALY DETECTOR FOR ANOMALOUS SOUND DETECTION

Technical Report

Jisheng Bai[1,2]*, Yafei Jia*[1]*, Siwei Huang*[1]*, Mou Wang*[1]*, Jianfeng Chen*[1,2]

[1] Joint Laboratory of Environmental Sound Sensing,
School of Marine Science and Technology,
Northwestern Polytechnical University, Xi'an, China
[2] LianFeng Acoustic Technologies Co., Ltd. Xi'an, China
{baijs, jyf2020260709, hsw838866721, wangmou21}@mail.nwpu.edu.cn, chenjf@nwpu.edu.cn

ABSTRACT

Anomaly detection has a wide range of applications such as finding fraud cases in industry or indicating network intrusion in network security. Anomalous sound detection (ASD) for machine condition monitoring can detect anomalies in advance and prevent causing damage. However, the operational conditions of machines often change, leading to the different acoustic characteristics between training and test data. Domain generalization techniques are required to adapt the model to different conditions. In this paper, we present a self-supervised method for ASD using batch mixing strategy with margin loss and anomaly detector. The proposed batch mixing strategy randomly mixes the data from source and target domains in a mini-batch to adapt the model between different domains. Moreover, we adopt a self-supervised method using machine IDs with additive angular margin loss to extract acoustic representations. Finally, we use the acoustic representations to train anomaly detectors to detect anomalous sound. Experimental results on the development dataset of DCASE2022 taks2 show that our method outperforms the baseline systems.

Index Terms— Anomalous sound detection, domain generalization, batch mixing, self-supervised learning

1. INTRODUCTION

Anomaly detection has been valued by researchers in recent years, and has been widely used in surveillance and mechanical equipment monitoring. A damaged bearing may not be found visually, but it can be detected acoustically through different acoustic manners. In addition, the acoustic monitoring system is cheap and easy to develop. Developing a reliable anomalous sound detection (ASD) system for the early detection of abnormal events can optimize and save a lot of resources in industrial production.

In DCASE2022 task2 ("Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques") [1], it is necessary to detect the abnormal sound of the machine using unsupervised methods, because we can only get the normal sound of the machine for training. In the task, the test data is mixed with samples not only affected by data in source domain, but also affected by data in target domain, and whether each sample belongs to the source domain or the target domain is not specific. Therefore, the model must detect anomalies with the same threshold without considering the domain (i.e., domain generalization)

This paper is organized as follows. In Section 2, we introduce the proposed method for ASD. In chapter 3, we describe the experimental settings. In chapter 4, we give the results we are submitting. In chapter 5, we summarize this report.

2. PROPOSED METHOD

2.1. Batch mixing

In this task, the source domain and the target domain of the training data are extremely unbalanced. When loading training data, the batch may not contain the data from target domain. This leads to the overfitting on the source domain of the final ASD model. Therefore, the proposed batch mixing strategy randomly mixes the data from source and target domains in a mini batch for better modeling different domains. The procedure of batch mixing strategy is shown in Figure 1.

Data augmentation is an effective way to improve generalization and prevent overfitting of the neural networks. In our system, we introduce mixup and FMix in batch mixing strategy to improve the performance on different domains.

Mixup [2]: Mixup is to mix samples by interpolating two graphs in proportion.The mixup operations on the training samples are as follows:

$$\widetilde{x} = \lambda x_i + (1-\lambda)x_j \tag{1}$$

$$\widetilde{y} = \lambda y_i + (1-\lambda)y_j \tag{2}$$

where x_i and x_j are the input features, y_i and y_j are the corresponding target labels and $\lambda \in [0, 1]$ is a random number drawn from the beta distribution.

FMix [3]: FMix is to cut out parts with arbitrary shapes from random images and paste them onto relevant images. The mask is obtained by sampling the low-frequency images in Fourier space.

$$\widetilde{x} = M_\lambda \times x_i + (1 - M_\lambda) \times x_i \tag{3}$$

where x_i is the input feature, M_λ is obtained by threshold low-frequency images sampled in Fourier space.

2.2. Model

MobileFaceNet(MFN) [4]: We use the log-Mel energies and spectrograms as the input of the network. The MFN is trained using

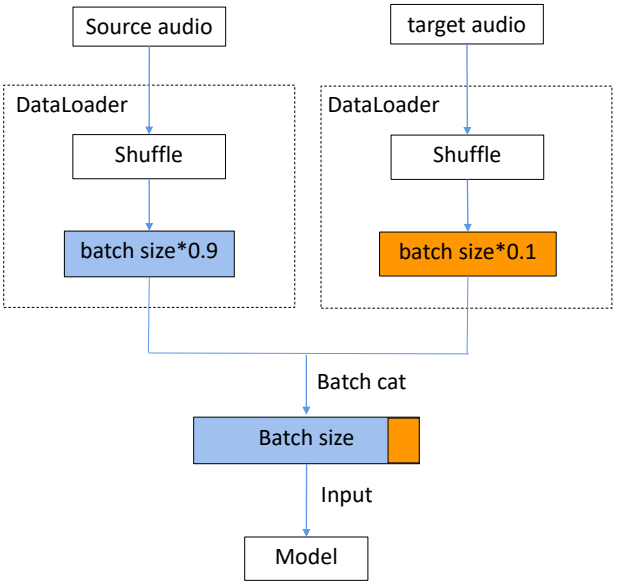

Figure 1: The procedure of batch mixing strategy

a self-supervised method with machine IDs. We get the output through the full connection layers. The probability output of softmax is used to calculate the loss. The network architecture of MFN using log-Mel energies is shown in table 1, the architecture of MFN using spectrograms is the same as [5].

Input	Operator	t	c	n	s
128x64x1	con2d 3x3	-	64	1	2
64x32x64	depthwise con2d 3x3	-	64	1	1
64x32x64	bottleneck	2	64	5	2
32x16x64	bottleneck	4	128	1	2
16x8x128	bottleneck	2	128	6	1
16x8x128	bottleneck	4	128	1	2
8x4x128	bottleneck	2	128	2	1
8x4x512	conv2d 1x1	-	512	1	1
1x1x512	linear GDConv 4x8	-	512	1	1
1x1x512	linear conv2d 1x1	-	128	1	1
1x1x128	linear	-	6	-	-

Table 1: MobileFaceNet Architecture

Dual-Path Transformer (DPT) [6, 7]: The DPT is a novel dual-path Transformer-based neural network for ASD. DPT can learn temporal and frequency dependencies and model interactive information by stacked Transformer encoders. Moreover, we introduce two self-supervised strategies to train DPT. The first strategy randomly masks the area of the input features and reconstructs them. The second strategy uses machine IDs to train the model in a supervised manner.

2.3. Loss function

In the experiment, we employed different loss functions for different models.

Additive Angular Margin Loss (ArcFace)[8]: ArcFace loss is a loss function that uses margin to expand the distance between different classes.

$$L_{arc} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n} e^{scos\theta_j}} \quad (4)$$

where s is the radius of the hyper sphere, representing the feature scale, N and n are the size of batch and the number of categories respectively, m is the angular distance between the two categories.

Focal loss[9]: Focal loss was originally used in the image field to solve model performance problems caused by data imbalance.

$$p_t = \begin{cases} \widetilde{p} & \text{if y=1} \\ 1-\widetilde{p} & \text{otherwise} \end{cases} \quad (5)$$

$$L_{ft} = -(1-p_t)^{\gamma} log(p_t) \quad (6)$$

where $\widetilde{p}$ is the predicted probability, p_t is the closeness to category y, the larger p_t is, the closer it is to category y, that is, the more accurate the classification is, γ is the adjustable factor.

2.4. Anomaly Detector (AD)

We use the Local Outlier Factor (LOF), cosine distance and log-softmax as anomaly detectors.

Local Outlier Factor (LOF)[10]: This method is a density-based local outlier detection algorithm, which is suitable for data with different cluster densities. In this report, we calculate the mean embedding vector for each audio clip, and use the embedding vectors for each section as the input of LOF. We use LOF to calculate the anomaly score for each test sample. The number of neighbors is set to 4 in our experiments.

Cosine distance: We also adopt cosine distance as the anomaly detector, the anomaly score is got by calculating the distance between the mean vector of all the training samples in a section and the test sample.

Log-softmax: We assume that the output of the network softmax follows the gamma distribution. The parameters of the gamma distribution are estimated according to the histogram of the softmax output of the network, and the anomaly detection threshold is determined as the 90th percentile of the gamma distribution.

3. EXPERIMENTAL SETTINGS

3.1. Dataset

The dataset of task2 consists of seven types of machines: toyCar, toyTrain, bearing, fan, gearbox, slider and valve [1, 11, 12].

The development dataset consists of three sections for each machine, and the sounds in each section contain around 990 normal recordings in source domain and 10 normal recordings in a target domain for training, and around 100 clips each of normal and anomalous recordings in the source and target domain for testing. Each recording is a 10-second audio that records the running sounds of a machine and its environmental noise.

The additional training dataset provides the other three sections for each machine type. Each section consists of around 990 normal

recordings in source domain 10 ten normal recordings in a target domain for training. Around 100 clips each of normal and anomalous recordings in the source and target domain from the three sections are used as evaluation dataset. The overview of the task2 dataset is shown in figure 2.

3.2. Features

The sample rate used in the experiments is 16KHz, and we applied STFT with a window size of 1024 and a hop length of 512. For STFT spectrogram, we use the n_frames of 32, and for log-Mel spectrograms, we use 128 Mel-bins and n_frames of 64. We use STFT and log-Mel spectrograms as input for different models.

3.3. Experimental architectures

To verify the performance, we compared the following models:

Baseline: The organizers provide a MobileNetV2(MNV2)-based baseline. This baseline takes log-Mel spectrogram with bands of 128 to identify from which section the observed signal is generated.

MFN: We trained the MFN network using Margin Loss and Focal Loss as loss functions. The STFT spectrogram of 512×32 and the log-Mel spectrogram of 128×64 are used as inputs to the model.

DPT: We trained the DPT network using mean square error (MSE) as reconstruction loss and cross-entropy as the classification loss. The log-Mel spectrograms of 128×64 are used as input of the model.

4. RESULTS

We conducted our experiments using the development and additional training dataset of DCASE2022 task2. The harmonic mean scores of experiments are shown in Table 2. The final submissions are conducted on the evaluation dataset of task2. In this task, because the data distribution of different machines is quite different, a single model can not show good detection performance on all machines. Therefore, the system we finally submitted combines the results of multiple systems. The ensemble strategy is shown in table 3. The results submitted are shown in table 4.

5. CONCLUSIONS

In this paper, we propose an ASD self enhancement method based on batch mixing strategy, which has margin loss and anomaly detector. The batch mixing strategy randomly mixes the data of the source domain and the target domain into a small batch to adapt to the models between different domains. In addition, we adopt a self-supervised method to extract the acoustic representation using the machine ID with additional angular margin loss. Finally, we use acoustic representation to train the anomaly detector to detect abnormal sound.The experimental results on the DCASE2022 taks2 development data set show that our method is superior to the baseline system.

6. 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," *arXiv preprint arXiv:2206.05876*, 2022.

[2] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[3] E. Harris, A. Marcu, M. Painter, M. Niranjan, A. Prügel-Bennett, and J. Hare, "Fmix: Enhancing mixed sample data augmentation," *arXiv preprint arXiv:2002.12047*, 2020.

[4] S. Chen, Y. Liu, X. Gao, and Z. Han, "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 428–438.

[5] K. Morita, T. Yano, and K. Tran, "Anomalous sound detection using cnn-based features by self supervised learning," DCASE2021 Challenge, Tech. Rep., July 2021.

[6] J. Bai, M. Wang, and J. Chen, "Dual-path transformer for machine condition monitoring," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 1144–1148.

[7] J. Bai, Z. Wang, and J. Chen, "Dptrans: Dual-path transformer for machine condition monitoring," DCASE2021 Challenge, Tech. Rep., July 2021.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[10] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

[11] 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.

[12] 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.

[13] https://dcase.community/challenge2022/task-unsupervised-anomalous-sound-detection-for-machine-condition-monitoring.

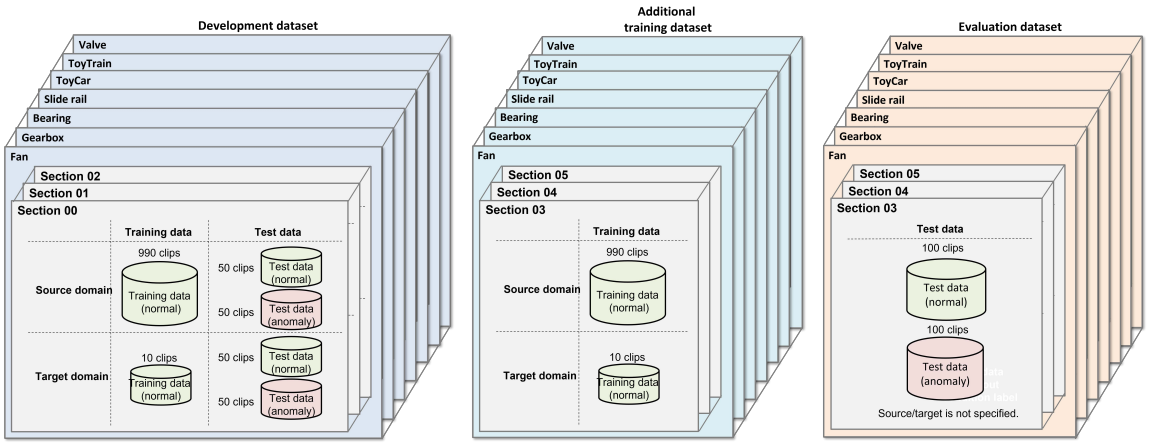

Figure 2: The overview of the task2 dataset [13].

Method	AD	h-AUC	**ToyCar**	**ToyTrain**	**Bearing**	**Fan**	**Gearbox**	**Slider**	**Value**
MNV2	log-softmax	source	0.5912	0.5726	0.6058	0.7075	0.6921	0.6515	0.6709
Log-Mel		target	0.5196	0.4590	0.5994	0.4822	0.5619	0.3823	0.5722
MFN	Cosine(Bea., Gea., Sli.)	source	**0.8947**	**0.6958**	0.6670	**0.7592**	0.6600	0.8960	**0.9381**
STFT	LOF(others)	target	**0.7552**	**0.5946**	0.8111	**0.5863**	0.6800	0.7600	**0.8946**
MFN	LOF(Car., Tra., Gea.)	source	0.7509	0.6878	0.6888	0.7262	**0.8697**	**0.9542**	0.8895
Log-Mel	log-softmax(others)	target	0.7171	0.5545	0.7652	0.4859	**0.8656**	**0.7804**	0.8571
DPT	log-softmax	source	0.6155	0.5716	**0.7618**	0.5220	0.7193	0.9332	0.8060
Log-Mel	&MSE	target	0.7154	0.5124	**0.8347**	0.4896	0.7164	0.7423	0.6331

Table 2: Harmonic mean AUC scores of experiments

	ToyCar	**ToyTrain**	**Bearing**	**Fan**	**Gearbox**	**Slider**	**Value**
Feature	STFT	STFT	log-Mel	STFT	log-Mel	log-Mel	log-Mel
Model	MFN	MFN	DPT	MFN	MFN	MFN	MFN
Loss	arcface	arcface	focal-loss	arcface	focal-loss	focal-loss	arcface
Augmentation	mixup	fmix mixup	mixup	fmix mixup	fmix mixup	fmix mixup	mixup
Train-sections	0,1,2,3,4,5	0,1,2,3,4,5	0,1,2,3,4,5	0,1,2,3,4,5	0,1,2,3,4,5	0,1,2,3,4,5	0,1,2,3,4,5
Test-sections	0,1,2	0,1,2	0,1,2	0,1,2	0,1,2	0,1,2	0,1,2
AD	LOF	LOF	log-softmax&MSE	LOF	LOF	log-softmax	LOF

Table 3: The overview of the ensemble strategy

		ToyCar	**ToyTrain**	**Bearing**	**Fan**	**Gearbox**	**Slider**	**Value**
Submission3	MFN	**0.7228**	**0.5956**	0.6384	**0.6671**	**0.8151**	0.6655	**0.8541**
Submission2	MFN	0.6925	0.5768	0.6619	0.6056	0.8021	**0.8174**	0.8447
	DPT	0.6191	0.5233	**0.7231**	0.5042	0.6816	0.7559	0.7053
Submission1	Ensemble	**0.7228**	**0.5956**	**0.7231**	**0.6671**	**0.8151**	**0.8174**	**0.8541**

Table 4: Omega scores of experiments