ANOMALOUS SOUND DETECTION USING CNN-BASED FEATURES BY SELF SUPERVISED LEARNING

Technical Report

Kazuki Morita[*], *Tomohiko Yano*[*], *Khai Q. Tran*[*]

Intelligent Systems Laboratory, SECOM CO.,LTD.
{morita-ka,tomo-yano,ku-chan}@secom.co.jp

ABSTRACT

We propose a detection method for the anomalous sound detection task of DCASE2021 task2 in this report. This is the task of anomalous sound detection for machine condition monitoring, and it is required to detect unknown anomalous sound only from normal sound data. We use the normal sound of the machine and its section index to train the Convolutional Neural Network (CNN) in a self-supervised learning manner. Then, we detect anomalous sound by using feature vectors extracted from CNN. As a result, for the development dataset we show the detection performance of 78.05% in Area Under Curve (AUC) and 68.09% in partial AUC (pAUC).

***Index Terms*—** Anomalous Sound Detection, Convolutional Neural Network

1. INTRODUCTION

In DCASE2021task2 "Unsupervised Anomalous Sound Detection for Machine Condition Monitoring under Domain Shifted Condition"[1], it is required to detect the anomalous sound of the machine. Since we can only obtain the normal sound of a machine, anomalous sound detection is an unsupervised problem. In DCASE2021 task, a new condition that the acoustic characteristics of the training data and the test data are different(i.e., domain shift) is newly added.

In DCASE2020 task2, we only used the conventional detection methods, and we found that it was important to extract more effective features. Therefore, in DCASE2021 task2, we extract features from the sound of the machine based on the Convolutional Neural Network(CNN). Furthermore, we use conventional anomaly detection methods same as the last year.

This paper is organized as follows. In chapter 2, we describe our anomalous sound detection method. In chapter 3, we show the evaluation experiments and the results. In chapter 4, we summarize this report. In chapter 5, we describe the model we are submitting.

2. ANOMALOUS SOUND DETECTION METHOD

2.1. Audio Processing

We transform all audio clip into spectrograms. The frame size for STFT is 128 ms, and hop size is 32 ms. We set these parameters experimentally. We use spectrograms as input for CNN.

2.2. Feature extractor using CNN

By using spectrograms and section indices, we train a CNN such as MobileNetV2(MNv2)[2] and MobileFaceNet(MFN)[3]. Aditionally, we use Additive Angular Margin Loss [4] as a loss function. A spectrogram of 1024 dimensions × 32 frames is used as a processing unit, and the unit is shifted by 16 frames in the audio clip. Each model structure is shown in Table 1 and 2. As a result, we obtain a 128 dimensions vector per an unit.

Table 1: MobileNetV2 Architecture

Input	Operator	t	c	n	s
1024×32×1	conv2d 3×3	-	32	1	2
512×16×32	bottleneck	1	16	1	1
512×16×16	bottleneck	6	24	2	2
256×8×24	bottleneck	6	32	3	2
128×4×32	bottleneck	6	64	4	2
64×2×64	bottleneck	6	96	3	1
64×2×96	bottleneck	6	160	3	2
32×1×160	bottleneck	6	320	1	1
32×1×320	conv2d 1×1	-	1280	1	1
32×1×1280	Ave Pool 16×1	-	-	1	-
1×1×1280	conv2d 1×1	-	128	-	

Table 2: MobileFaceNet Architecture

Input	Operator	t	c	n	s
1024×32×1	conv2d 3×3	-	64	1	2
512×16×64	depthwise conv2d 3×3	-	64	1	1
512×16×64	bottleneck	2	64	5	2
256×8×64	bottleneck	4	128	1	2
128×4×128	bottleneck	2	128	6	2
64×2×128	bottleneck	4	128	1	2
32×1×128	bottleneck	2	128	2	1
32×1×128	conv2d 1×1	-	512	1	1
32×1×512	linear GDConv16×1	-	512	1	1
1×1×512	linear conv2d 1×1	-	128	1	1

2.3. Anomaly Detector

We apply Local Outlier Factor(LOF) for source domain and k-Nearest Neighbors(k-NN) for target domain. We merge embedding vectors in an audio clip using mean or standard deviation.

[*] Equal contribution.

Local Outlier Factor(LOF)[5]

This method is based on local density, which is the density of k-neighboring feature values. When a feature is anomalous, the difference is large between the local density of the anomaly and the neighboring feature. In this report, we use the outputs of LOF as the anomaly score. We set the number of neighbors to 4.

k-Nearest Neighbors(k-NN)[6]

This method is based on the distance of k-neighboring features. In k-NN, the larger the distance to the selected neighborhood, the more deviated from normal. In this report, we use the mean of cosine distance as the anomaly score, and we set the number of neighbors to 1.

3. EVALUATION EXPERIMENTS

3.1. Experimental Condition

10-sec length audio (monaural, 16 kHz) was sampled from machinery sound sources. There are seven types of machines (Machine Type); ToyCar, ToyTrain[7], fan, gearbox, pump, slider and valve[8]. For each Machine Type, there are 3 sections in development dataset and 3 sections in additional dataset. We trained a CNN using 6 sections datasets in a Machine Type, and an anomaly detector using embedding vectors per section. We used librosa[9] and scikit-learn[10] for the implementation. When we evaluated sound clips in the source domain, we only used training data in the source domain, and when in the target domain, we used training data in the source domain and target domain.

In the experiment, we compared the following:

- Feature extractor model: MobileNetV2, MobileFaceNet
- Anomaly detector model: LOF, k-NN
- Feature merge method: mean, standard deviation(std)

3.2. Results

The results are shown in Table 3, Table 4, Table 5, and Table 6. Table 3 and Table 4 show the results for source domain and Table 5 and Table 6 for target domain. Each value is a harmonic mean of AUC or pAUC overall sections.

4. CONCLUSION

In this paper, we used the normal sound of the machine and its section index to train the CNN in a self-supervised learning manner. Then, we detect anomalous sound by using feature vectors extracted from CNN. The performance of 78.05% for AUC and 68.09% for pAUC was shown for the development dataset.

5. SUBMISSIONS

In this report, we submit three anomalous sound detection systems. Table 7 shows the conditions we used.

6. REFERENCES

[1] 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," *In arXiv e-prints: 2106.04492, 1 – 5*, 2021.

[2] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," pp. 4510–4520, 2018.

[3] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices," pp. 428–438, 2018.

[4] 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.

[5] 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*. ACM, 2000, pp. 93–104.

[6] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. ACM, 2000, pp. 427–438.

[7] 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.

[8] 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, 1 – 4*, 2021.

[9] B. McFee, V. Lostanlen, A. Metsai, M. McVicar, S. Balke, C. Thom é , C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, J. Mason, D. Ellis, E. Battenberg, S. Seyfarth, R. Yamamoto, K. Choi, viktorandreevichmorozov, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Here ñ ú , F.-R. St ö ter, P. Friesch, A. Weiss, M. Vollrath, and T. Kim, "librosa/librosa: 0.8.0," July 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3955228

[10] O. Grisel, A. Mueller, Lars, A. Gramfort, G. Louppe, P. Prettenhofer, M. Blondel, V. Niculae, J. Nothman, A. Joly, T. J. Fan, J. Vanderplas, manoj kumar, H. Qin, N. Hug, N. Varoquaux, L. Est è ve, R. Layton, J. H. Metzen, G. Lemaitre, A. Jalali, R. (Venkat) Raghav, J. Sch ö nberger, R. Yurchak, W. Li, C. Woolam, T. D. la Tour, K. Eren, J. du Boisberranger, and Eustache, "scikit-learn/scikit-learn: scikit-learn 0.24.1," Jan. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.4450597

Table 3: Harmonic Mean of AUC in the source domain of Development Dataset(%)

CNN	Detector	merge	ToyCar	ToyTrain	fan	gearbox	pump	slider	valve	total
baseline(MNv2)			55.80	67.89	61.02	70.21	65.48	72.67	55.95	63.53
MNv2	LOF	mean	72.68	63.48	82.18	78.31	75.46	82.02	74.31	74.99
		std	62.11	60.64	69.09	63.88	59.10	76.47	88.67	67.31
	k-NN	mean	75.59	67.95	83.61	79.39	77.99	89.65	75.43	78.01
		std	54.64	63.49	71.68	66.20	61.55	79.48	93.38	68.20
MFN	LOF	mean	91.06	86.13	90.36	77.76	82.52	90.83	75.37	84.42
		std	81.55	76.10	54.51	53.92	53.93	70.56	91.32	66.06
	k-NN	mean	89.37	81.50	85.59	79.52	83.37	91.97	71.31	82.73
		std	75.53	73.36	51.69	53.76	53.53	72.43	95.87	64.97
Our best			91.06	86.13	90.36	77.76	82.52	90.83	95.87	87.42

Table 4: Harmonic Mean of AUC in the target domain of Development Dataset(%)

CNN	Detector	merge	ToyCar	ToyTrain	fan	gearbox	pump	slider	valve	total
baseline(MNv2)			57.42	50.18	62.35	64.35	59.08	51.21	57.25	56.98
MNv2	LOF	mean	64.48	53.77	64.12	77.88	62.15	67.22	67.22	64.58
		std	59.65	53.50	62.41	59.07	59.53	60.07	73.46	60.63
	k-NN	mean	66.64	53.62	64.76	80.78	62.58	63.54	67.39	64.80
		std	54.82	56.67	66.08	64.24	60.84	58.95	78.30	62.08
MFN	LOF	mean	60.27	51.68	73.28	81.37	75.86	53.64	63.44	63.94
		std	61.94	46.37	53.67	47.09	49.09	49.33	65.92	52.48
	k-NN	mean	70.54	54.08	72.67	84.80	74.39	67.07	63.10	68.34
		std	61.62	42.55	53.85	47.96	47.55	49.49	78.67	52.59
Our best			70.54	54.08	72.67	84.80	74.39	67.07	78.67	70.50

Table 5: Harmonic Mean of partial AUC in the source domain of Development Dataset(%)

CNN	Detector	merge	ToyCar	ToyTrain	fan	gearbox	pump	slider	valve	total
baseline(MNv2)			58.64	51.87	65.79	61.45	59.24	59.50	52.17	58.01
MNv2	LOF	mean	60.12	57.23	75.39	64.94	64.26	74.41	62.82	65.00
		std	55.23	55.44	64.50	53.99	52.10	62.51	79.35	59.34
	k-NN	mean	63.48	57.64	75.80	64.67	65.34	78.68	64.07	66.43
		std	53.85	54.17	67.70	54.27	54.40	71.31	82.70	61.02
MFN	LOF	mean	78.25	65.36	79.66	67.67	66.50	83.05	59.41	70.48
		std	63.27	58.72	50.91	51.07	51.93	58.35	79.19	57.81
	k-NN	mean	74.76	58.65	70.11	67.43	68.74	82.59	58.03	67.69
		std	61.03	54.98	50.74	50.70	52.47	58.05	84.78	57.33
Our best			78.25	65.36	79.66	67.67	66.50	83.05	84.78	74.24

Table 6: Harmonic Mean of partial AUC in the target domain of Development Dataset(%)

CNN	Detector	merge	ToyCar	ToyTrain	fan	gearbox	pump	slider	valve	total
baseline(MNv2)			54.44	51.38	60.84	57.48	55.73	53.17	53.23	55.03
MNv2	LOF	mean	56.12	51.21	67.55	66.33	57.41	59.25	55.90	58.62
		std	53.94	50.39	58.04	52.05	53.43	54.99	60.50	54.58
	k-NN	mean	57.74	51.35	67.82	65.57	58.06	58.98	56.13	58.93
		std	52.39	51.22	64.82	52.67	54.19	56.80	64.41	56.18
MFN	LOF	mean	56.91	52.45	64.81	69.04	62.38	50.55	57.13	58.40
		std	52.64	48.64	49.32	49.22	51.41	51.97	56.80	51.30
	k-NN	mean	59.92	53.10	66.67	72.49	65.42	62.01	57.28	61.84
		std	52.50	48.66	49.35	49.40	51.48	52.13	64.12	52.12
Our best			59.92	53.10	66.67	72.49	65.42	62.01	64.12	62.88

Table 7: Submission of our System

		Anomaly detector		
Model name	Feature extractor	Source domain	Target domain	Merge Method
Morita_SECOM_task2_1	MFN	LOF	LOF	std(valve), mean(otherwise)
Morita_SECOM_task2_2		k-NN	k-NN	
Morita_SECOM_task2_3		k-NN(valve), LOF(otherwise)	k-NN	