AN ENSEMBLE APPROACH FOR ABNORMAL SOUND DETECTION WITH DATA AUGMENTATION

Technical Report

Bo Cheng Chan

National Taipei University
Taiwan
cbc94meng@gmail.com

Chung Li Lu

National Taiwan University
Taiwan
chungli59@ gmail.com

ABSTRACT

In this paper, we present the task description and discuss the results of DCASE 2021 Challenge Task 2: Unsupervised Anoma-lous Sound Detection for Machine Condition Monitoring under Domain Shifted Conditions. The task is identifying whether the sound emitted from a machine is normal or anomalous in test dataset. The training dataset does not contain any abnormal ma-chine sounds. Our approached is based on MobileNetV2 and ResNetV2-50 with data augmentation mix up to identify abnormal sounds in each machine.

Index Terms— Anomalous Sound Detection, MobileNetV2, ResNetV2-50, data augmentation, mix up

1. INTRODUCTION

The DCASE 2021 Challenge Task 2 is concerned to solve the problem of normal sounds being incorrectly judged as anomalous due to changes within the normal conditions (i.e., domain shift) [1]. We know there are many different conditions in real world, such like production demand changes and environmental condition changes. The fundamental question is to find an appropriate representation of machine that is enough to distinguish the difference of normal and abnormal machine.

The production demand changes is operation speed, machine load, viscosity, heating temperature operating speed, environmental noise, SNR, etc changes in different working season. E.q. A pump suffering from a small leakage, a slide rail that has no grease or a fan undergoing voltage changes might appear intact when inspected visually but when monitored acoustically, reveal its actual condition through distinct sound patterns. The machinery's surrounding noise under realistic industrial conditions may lead to low signal to noise ratio, thus impairing our ability to detect anom-alous operations.

Deep learning can learn hierarchical discriminative features from data, which is different from traditional machine learning and generally applied in various scenes [2]. Through deep learning, we don't need have sufficient engineering knowledge of system and process the dynamics of faults are unknown [3].

There are three main challenges as follows:

First, we search for a good deep neural network architecture for abnormal sound detection. Second, gather sufficient amount and variety of data for training. Third, we train the developed architec-ture in an end-to-end manner tailored for this task

2. DCASE 2020 CHALLENGE TASK 2 DATABASE AND SETUP

The DCASE 2021 Task 2 dataset is consists of ToyADMOS2 [4] and MIMII DUE [5] that includes both the 10 seconds audios of a machine and its associated equipment as well as environmental sounds. There seven types of machine categories. ToyCar and ToyTrain are from the ToyADMOS2[4]. Valve, Pump, Fan and Slider are from the MIMII DUE [5]. As show in Table 1, there are 2 domains in each section and each section is a complete set of training and test data. For each section, this training dataset provides around 1,000 clips of normal sounds in a source domain and 3 clips of normal sounds in a target domain. The testing data contains 100 clips of normal sounds and 100 clips of real anomalous sound in source and target domain.

Table 1: Overview of domain dataset in training and testing.

Dataset / Domain	Source		Target	
	Normal	Abnormal	Normal	Abnormal
Training	1000	0	3	0
Testing	100	100	100	100

The anomaly score calculator A with parameter θ. The input of $\mathbf{A}$ is the audio clip x and additional information, and one anomaly score $A_\theta(x)$ is output. Then, the machine is determined to be anomalous when the anomaly score $A_\theta(x)$ exceeds a pre-defined threshold value $\emptyset$ as

$$\text{Decision} = \begin{cases} Anomaly & (A_\theta(x) > \emptyset) \\ Nomal & (otherwise) \end{cases} \quad (1)$$

This task is evaluated with the AUC and the pAUC. The pAUC is an AUC calculated from a portion of the ROC curve over the pre-specified range of interest. In our metric, the pAUC is calculated as the AUC over a low false-positive-rate (FPR) range [0, p]. The reason for the additional use of the pAUC is based on practical requirements. If an ASD system frequently gives false alarms frequently, we cannot trust it. Therefore, it is important to increase the true-positive rate under low FPR conditions. In this task, we will use p = 0.1.

3. APPROACH

In this section, we first describe the members of our ensemble separately, and then describe how these are combined for the final scoring

3.1.1. Audio representation

For spectrogram, we used the tool kit librosa to extract our Mel spectrogram from input audio. Each audio sample is represented by mel-spectrogram of 128 frequency bins, a window size length 1024 ms and 50% overlap. The result of spectrogram is log-scaled. Acoustic feature is obtained by concatenating before/after several frames (5 frames) of log-mel-filterbank outputs. 512 and number of mel bands = 128, and after computing the log Mel spectrogram of each clip of signal, acoustic feature is obtained by concatenating before/after several frames (5 frames) of log-mel-filterbank outputs

3.1.2. Data Augmentation

For data augmentation, we tried as follow:

- The random noise from data:

$$y[n] = x[n] + 0.005 * z[n] \quad (2)$$

where z is an random uniform [0,1] noise from audio, y is the audio signal from machine ID's signal and z.

- The mixed output is combined by audio and Gaussian noise that RMS of the noise generated:

$$y[n] = x[n] + \mathrm{RMS}_{noise} \quad (3)$$

$$\mathrm{RMS}_{noise} = \sqrt{\frac{RMS_{signal}^2}{10^{SNR/10}}} \quad (4)$$

- The time stretching is stretch a spectrogram in time for a given rate r:

$$y_{speed}[n] = \frac{x_{speed}[n]}{r} \quad (5)$$

- For frame shift, we used the tool kit numpy to shift 16000 points to right:

$$y[n] = x[n + 16000] \quad (6)$$

- Mix up

$$y[n] = \alpha * x_1[n] + (1 - \alpha) * x_2[n] \quad (7)$$

where y is a mixture features of x_1 and x_2 are log-mel-filterbank outputs of machine section ID. We also multiply the one-hot encoded label with α.

3.1.3. Model Archecture

- AutoEncoder

The autoencoder (AE)-based model released by [6] that is same as the DCASE 2020 task 2. The corresponding hidden units in each layer follows this structure: [128, 128, 128, 128, 32, 128, 128, 128, 128], that compresses 5 frames of 128 mel-energies into an 8-dimensional space, with BatchNorm and ReLU non-linearities, and trained with Adam.

- Classifier

The MobileNetV2 [7] model released by the DCASE 2021 task 2. The ReNetV2-50[8] that are widely applied in computer vision tasks such as object detection, classification and semantic segmentation. These model compresses 5 frames of 128 mel-energies into an 8-dimensional space. The ADAM optimizer is used with the learning rate as 0.00001. We stop the training process after 20 epochs, and the batch size is 32. We train models independently for each machine type, using normal clips from all sections of that machine type.

4. EXPERIMENTAL RESULTS

There are 4 systems based on the development and addition training dataset. We tried different systems with data augmentation, additional training data and ensemble models. The arithmetic mean AUC scores for each machine type for the 4 different system are shown in Table 2 and Table 3.

In data augmentation, we found mix up has a better performance in some machine type, On the other hand, models training with dev and additional data are worse than the baseline. We also tried the settings of spectrogram varied with each machine type, but the results are not better than the default audio representation. Furthermore, we found time stretching had a good performance in ToyCar with training dev and others data augmentations (noise data, Gaussian noise and frame shift) did not improve the accuracy.

Final outputs we submitted 4 systems, based on the best performance for each machine type on development dataset and additional dataset.

5. REFERENCES

[1] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Descritpion and discussion on DCASE 2021 challenge task 2 : unsupersied anomalous sound detection for machine condition monitoring under domain shifted conditions." arXiv e-prints arXiv:2106.04492, 1-5, 2021.

[2] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019.

[3] G. S. Galloway, V. M. Catterson, T. Fay, A. Robb, and C. Love, "Diagnosis of tidal turbine vibration data through deep neural networks," in Proceedings of the 3rd European Conference of the Prognostics and Health Management Societry, 2016.

[4] 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.

[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," arXiv preprint arXiv:2105.02702, 2021

Table 2: Arithmetic mean AUC for baseline and system. The results of AE-Baseline1, MobileNetV2-Baseline, System1 (MobileNetV2 with mix up), System2 (ResNet50V2 with mix up) and System3 (MobileNetV2 and ResNet50V2 with mix up) are in the same dev set. The results of System4 (MobileNetV2 and ResNet50V2 with mix up) that training set is composed of dev set and additional set.

Machine	**Ae-Baseline**		**MobileNetv2 Baseline**		**System1**		**System2**		**System3**		**System4**	
Score	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
ToyCar	62.81	52.59	56.17	57.32	57.94	58.02	58.96	55.84	60.56	56.88	56.27	61.17
ToyTrain	62.93	55.01	60.55	51.74	54.02	50.23	49.83	50.86	51.67	51.09	47.84	51.93
fan	64.30	53.56	64.32	65.57	68.66	67.97	66.96	67.10	68.04	67.75	37.04	53.40
gearbox	66.63	52.26	61.73	54.26	63.39	53.95	65.41	58.24	69.81	62.02	40.63	49.25
pump	64.21	54.65	65.95	59.96	65.97	60.14	73.48	64.38	71.86	63.41	29.81	49.27
slider	69.59	56.31	64.65	57.39	63.33	57.40	58.99	60.70	65.62	57.67	35.66	51.02
valve	53.68	50.78	57.88	55.88	59.15	55.79	63.48	55.57	61.61	55.69	41.56	49.12

Table 3: Harmonic mean AUC for baseline and system. The results of AE-Baseline1, MobileNetV2-Baseline, System1 (MobileNetV2 with mix up), System2 (ResNet50V2 with mix up) and System3 (MobileNetV2 and ResNet50V2 with mix up) are in the same dev set. The results of System4 (MobileNetV2 and ResNet50V2 with mix up) that training set is composed of dev set and additional set.

Machine	**Ae-Baseline**		**MobileNetv2 Baseline**		**System1**		**System2**		**System3**		**System4**	
Score	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
ToyCar	62.17	52.53	53.91	56.26	54.65	56.63	56.34	54.90	58.54	55.81	51.58	60.00
ToyTrain	61.79	53.87	56.66	51.59	52.36	50.08	47.46	50.71	49.77	50.90	45.61	51.53
fan	63.50	53.38	61.09	63.62	64.27	65.75	60.51	64.67	62.93	65.55	30.26	52.78
gearbox	65.73	52.23	59.75	53.62	59.41	53.04	61.53	57.30	66.53	60.93	40.28	49.22
pump	62.85	54.38	64.68	59.09	64.94	59.16	72.20	63.26	70.75	62.32	20.28	49.22
slider	67.35	55.79	61.40	56.28	59.65	56.51	53.90	59.80	63.47	56.96	27.45	50.73
valve	53.35	50.69	57.36	55.58	58.51	55.26	62.83	55.21	60.82	55.21	36.27	49.06

[6] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, “Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection,” arXiv preprint arXiv:1909.09347, 2019.

[7] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlen ecks,” in Proc. 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520

[8] He, K., Zhang, X., Ren, S., Sun, J. “Identity mappings in deep residual networks,” in European conference on computer vision, Springer, 2016 pp. 630–645