ANOMALOUS SOUND DETECTION WITH AUTOENCODER AND IMAGE MOBILENET USING OUTLIER EXPOSURE APPROACH

Technical Report

Sophia Kazakova, Andrey Semenov, Andrey Surkov, Sergei Astapov

ITMO University
Department of Speech Information Systems
Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia
sophie.a.kaz@gmail.com

ABSTRACT

This technical report describes the autoencoder and MobileNetV2-based approach for DCASE 2022 Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques task [1]. Firstly, a basic Autoencoder-based architecture was developed. Then the outlier exposure approach was tested on DCASE 2022 Challenge Task 2 Development Dataset [2]. Proven its effectiveness, it then was used as a part of an image MobileNetv2 system. To tackle the challenge of domain shift and make the dataset more balanced in terms of source/target classes we used data augmentation with TimeStretch and PitchShift. Audio files then were transformed with STFT and saved as images. The MobileNetv2-based architecture was fine-tuned with those spectrograms and used for anomaly detection.

As a result, we are proposing three system configurations. The first one is solely Autoencoder with the Outlier Exposure approach. The second one is exclusively image MobileNetv2 spectrogram anomaly detection. The third is a combination of the previous two approaches, in which we use the best of two models for each machine class.

Index Terms— Autoencoder, Outlier Exposure, Image MobileNet, spectrogram, augmentation, anomaly detection

1. INTRODUCTION

The DCASE 2022 Challenge 2 is based on the task of determining abnormal operation sounds of different types of machines [1]. Special attention is paid to the classification of data into normal and abnormal under domain shift conditions. This task is crucial in terms of its practical application, as in real industry those machines and equipment around them are operated in various modes and conditions. The domain shift adaptation methods should level up the current state of the art in this field and significantly shift the accuracy of anomaly detection for the best. Fundamentally, the challenge of this track is to create a machine learning-based algorithm which would form an efficient representation sufficient to distinguish a normal event from an abnormal one.

Deep machine learning models are capable to define more complex dependencies in data, in contrast to basic machine learning models. Owning to deep architectures, there is no need to dive deep into the physics of the machine sound origin to fine-tune the model to extract and process the features.

Nevertheless, understanding the data itself is vital. The authors of the task propose a dataset consisting of 10-second recordings of the operation of 7 types of devices: fan, gearbox bearing, rail, toy car, toy train, and valve [3], [4]. The complexity of this data lies in the domain shift, i.e. changing conditions for 'recording' and operation itself. For instance, a few of the machines have different operating modes, various loads, and multiple versions of the environment. These changes should be taken into account at the training stage. To expand the dataset for consideration of other potential domain shifts, we use augmentation in the frequency and time domains. An example of this augmentation is described below.

2. PROPOSED APPROACH

For this submission we used two models, continuing the ideas proposed in the baselines [1]. The first model is a basic AutoEncoder on PyTorch with the same architecture as in the Baseline. At first, it was supposed to test the outlier exposure method [2], [5]. Nevertheless, during further experiments, it turned out to exceed the second model's results.

The second model is based on the pre-trained image MobileNetv2. We use spectrogram images as embeddings for audio and try to separate normal images from anomalies. Via visual analysis of random samples, we confirmed the hypothesis that it is possible to see the anomaly in the picture.

The three submissions are formed this way – the first one is fully autoencoder-based with the outlier exposure method, the second one is fully image MobileNetv2, and the third contains the best results of both models for each machine class as follows:

- Fan – AutoEncoder,
- Slider – MobileNetv2,
- Toy Car – AutoEncoder,
- Toy Train – AutoEncoder,
- Gearbox – MobileNetv2,
- Valve – MobileNetv2,
- Bearing – AutoEncoder.

2.1. Data Preprocessing & Training

For the MobileNetv2 model, we decided to focus on experiments to establish whether or not is it possible to use even fewer data than provided. In the first stage, 500 random events were selected from the training dataset for transfer learning. This data was augmented to expand the range of possible domain shifts in the training dataset and to create extra samples.

For the augmentation, we used the audiomentations library. Experiments were held to test the effectiveness of three different approaches – PitchShift, TimeStretch, and AddBackgroundNoise. The latter showed a rapid decrease in the result. We used the IDMT-ISA-ELECTRIC-ENGINE heavy load dataset as the added noise, and the model seemed to recognize it as the target sound [6]. The other two approaches seemed to work well. In the final system, we firstly apply the frequency domain augmentation via PitchShift Function. Then the same data goes through time-domain augmentation with TimeStretch. The events that exceeded the original duration are trimmed to 10 seconds. After the augmentation, the summary of 1000 events generated is added to the training dataset.

As the outlier exposure method implementation we then added to the training samples of another class. The outlier classes for each machine type are presented in the table in section 3.

Then the librosa library was used to extract the spectrogram. To form the features, the STFT method was implemented with a window size of 256 ms and an overlap of 50%. Then the spectrograms were reduced to a size of 320 by 320 pt. These pictures are the final data representation used to train and fine-tune the model. The pretrained neural network Mobile Net v2 initially is trained to work with 90 classes or images.

The AutoEncoder model uses the usual mel-spectrograms. The training data consists of 3000 samples of the target class (e. g. fan) and 3000 outliers (e. g. slider).

2.2. Anomaly scores

For the AutoEncoder architecture, an algorithm for calculating the sample anomaly score by the error of sound restoration is implemented. The value of the loss function is taken as the recovery error. Then the difference between the expected value of the loss function and the one calculated during the operation of the neural network can be taken as an abnormality. Let A be the level of abnormality, and X be a sample of the data processed by the algorithm. Then $loss_{true}$ and $loss_{pred}$ are the true and calculated values of the loss function, the derived ratio is reflected in 1.

$$A(X) = |loss_{true} - loss_{pred}| \tag{1}$$

We assume that for normal records the loss function tends to be equal to zero, while for abnormal records it tends to be close to one. Then, in the case of testing on unknown data, all $loss_{true}$ values are assumed to be 0, i. e. normal, and the greater the $loss_{pred}$, the higher the abnormality of this sample. The boundary by which a record is assigned a class (normal/anomaly) is calculated empirically. For our experiments, the threshold was taken as 90%.

For the MobileNetv2 architecture, the approach differs. At the output of the model, the probability of the record belonging to the target class is provided. For this probability, it is necessary to calculate the threshold value, starting from which events can be classified as anomalies. The calculation of the probability threshold from which the event will be assigned to the normal class is based on the training samples results. 500 random spectrograms from the training sample are put into the trained model and the threshold is calculated as 2, where L is the number of samples, and $MobileNetV2\,(sample_i)$ is the probability of belonging to the target class for the specific audio $sample_i$.

$$Threshold = (1 - \alpha) \cdot \frac{\sum_{i=0}^{L} MobileNetV2(sample_i)}{L} \tag{2}$$

Experiments have shown that the confidence of the model in the case of normal records is higher than for abnormal ones. Calculating the probability threshold, starting from which the event will be assigned to the normal class, is calculated based on the training sample and then decreases by the constant α, where $0 \leq \alpha \leq 0.1$. The optimal threshold for α is experimentally determined as $\alpha = 0.04$. The need for the introduction of α is justified by the specificity of the problem. The model is supposed to be trained exclusively on one class of normal recordings, and if the recording conditions change (for example, in the case of a serious breakdown of one of the devices), there is a high probability of a decrease in confidence, including for 'normal' samples.

3. EXPERIMENTS

3.1. Autoencoder

The AutoEncoder experiment was to test the outlier exposure method. The results of it for the fan class are described in table 1. AUC 0 means the arithmetic mean for AUC section 0, similarly to AUC 1 and AUC 2. AUC and pAUC are arithmetic means for total AUC and pAUC correspondingly.

	AUC 0	AUC 1	AUC 2	AUC	pAUC
clear AE	40.33	44.74	53.49	49.19	50.29
OE (Gearbox)	64.39	53.09	61.62	59.70	51.75
OE (Bearing)	62.11	**53.22**	**61.63**	58.98	51.49
OE (Slider)	**65.45**	51.49	60.29	**59.07**	**52.71**
OE (ToyCar)	58.37	51.52	61.18	57.02	51.17
OE (ToyTrain)	61.36	52.86	61.31	58.51	51.82
OE (Valve)	58.61	52.47	61.24	57.45	51.36

Table 1: Outlier exposure efficiency test for FAN class.

Clear AE is the same code without outlier exposure method implementation. OE (Gearbox) means that the gearbox class was used as the outlier for target one (fan for this table).

Based on the success of the outlier exposure method, we have tested it on the other classes. The best results are presented in table 2.

3.2. MobileNetv2

Visual analysis proved that it is possible to see some types of anomalies on a spectrogram. An example of normal audio of the

target class	outlier class	avg AUC	avg pAUC
ToyCar	gearbox	56.6	53.2
ToyTrain	slider	52.4	52.8
bearing	gearbox	48.8	53.8
fan	slider	58.9	54.6
gearbox	fan	50.2	53.1
slider	valve	52.3	56.4
valve	slider	14.3	49.3

Table 2: AutoEncoder + Outlier Exposure system results.

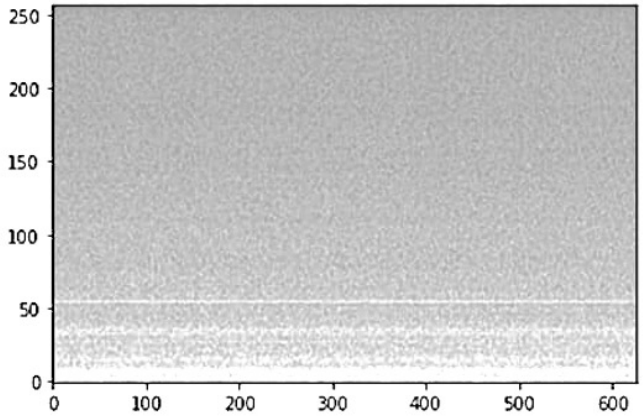

Figure 1: Normal audio spectrogram, class fan.

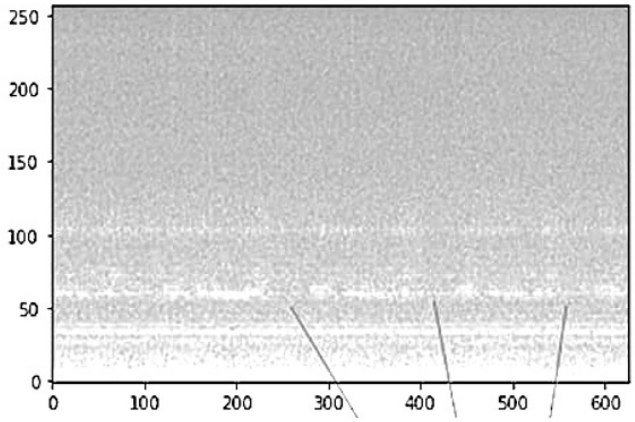

Figure 2: Abnormal audio spectrogram, class fan.

fan class is presented in Fig. 1. Fig. 2 shows an anomaly of the same machine class. Dark lines point to the discontinuities of the line of the spectrogram. There is nothing like that on the normal audio spectrograms.

Based on this hypothesis, experiments were held. Results are presented in table 3.

target class	outlier class	avg AUC	avg pAUC
ToyCar	gearbox	52.2	50.6
ToyTrain	slider	51.7	51.3
bearing	gearbox	50.7	50.6
fan	gearbox	55.2	50.6
gearbox	fan	64.8	51.6
slider	valve	57.0	53.2
valve	slider	54.3	52.9

Table 3: Image MobileNetV2 + Outlier Exposure system results.

4. REFERENCES

[1] Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Takashi Endo, Masaaki Yamamoto, and Yohei Kawaguchi. Description and discussion on DCASE 2022 challenge task 2: unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. In arXiv e-prints: 2206.05876, 2022.

[2] Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606, 2018.

[3] Kota Dohi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, Masaaki Yamamoto, Yuki Nikaido, and Yohei Kawaguchi. MIMII DG: sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task. In arXiv e-prints: 2205.13879, 2022.

[4] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Masahiro Yasuda, and Shoichiro Saito. ToyADMOS2: another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions. In Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021), 1–5. Barcelona, Spain, November 2021.

[5] Primus, Paul et al. CP-JKU Submission to DCASE'21: Improving Out-of-Distribution Detectors for Machine Condition Monitoring with Proxy Outliers & Domain Adaptation via Semantic Alignment. DCASE2021 Challenge, 2021.

[6] Sascha Grollmisch, Jakob Abeßer, Judith Liebetrau, Hanna Lukashevich. Sounding Industry: Challenges and Datasets for Industrial Sound Analysis, Proceedings of the 27th European Signal Processing Conference (EUSIPCO), A Coruña, Spain, 2019.