DCASE CHALLENGE 2021: UNSUPERVISED ANOMALOUS SOUND DETECTION OF MACHINERY WITH LENET ARCHITECTURE

Technical Report

Lam Pham, Anahid Jalali, Olivia Dinica, Alexander Schindler

Austrian Institute of Technology, Vienna, Austria,
{lam.pham, anahid.jalali, olivia.dinica, alexander.schindler}@ait.ac.at

ABSTRACT

In this study, we present an unsupervised anomalous sound detection framework trained on the DCASE2021 audio dataset. We use LeNet architecture to classify the machine IDs and use the classification loss as a threshold for detecting the anomalies in an unsupervised manner. We train our classifier on log-mel-bands and use the Mixup approach to augment our training set. Our framework outperforms both DCASE2021 benchmarks: the dense autoencoder and the MobileNet. The dense autoencoder has a harmonic mean of AUC of 61.92 and pAUC of 53.26 and the MobileNet has a harmonic mean of AUC of 59.72 and pAUC of 56.37. Our framework achieved the harmonic mean AUC of 66.72 and pAUC of 60.59, over all the machines, which shows an improved performance of 7.75% and 13.76%, AUC- and pAUC-harmonic-mean respectively from the dense autoencoder. The improved performance of our approach from the Mobilenet baseline is 11.72% and 7.48%, AUC- and pAUC-harmonic-mean respectively.

Index Terms— anomaly detection, anomalous sound detection, machine learning

1. INTRODUCTION

Automatic Anomalous Sound Detection (ASD) is a system that identifies abnormal sounds emitted from specific equipment and is considered an essential technology in industry 4.0 [1]. Such systems are often used for machine condition monitoring and aim to detect unknown anomalous sounds. In real-world cases, anomalies are infrequent and take many different forms. An extensive and time consuming data collection process would be needed to capture all the variations of anomalies from a machine. If, on the other hand, only data from the machinery in normal condition are collected, the system can be trained to only learn the natural routine of the targeted equipment. Deviations from this routine are then identified as abnormal behaviour.

Additionally, real-world cases often involve different machine operating conditions between the training and testing phases. For instance, changes in the seasonal demand of many products will lead to variations in the sound of the machines producing these products. Consequently, using training data and test data that are different in operating speed, machine load, environmental noise, etc. (i.e., contain a domain shift) will more properly capture these complications.

The DCASE2021 challenge of unsupervised anomalous sound detection [2] focuses on these 2 issues (unsupervised training and domain shift), where participants are asked to use the provided audio dataset and submit their results.

The audio dataset provided by organizers of this task contains recordings of 7 different types of machines that are parts of the ToyADMOS[3] and MIMII [4] datasets: Pump, Fan, Slider, ToyCar, ToyTrain, Gearbox and Valve. Each machine type consists of three sections. The dataset is available under 3 different releases:

- Development set: contains a training set and a testing set for each machine
- Extra training set: contains more training data for each machine
- Evaluation set: contains evaluation data for each machine

Furthermore, the DCASE community provides two baseline systems [1]: a dense autoencoder with 8 layers (4 encoding and 4 decoding layers) each with 128 units. The bottleneck of this architecture has 8 units with a rectified linear unit (ReLU) activation function. Each layer of the autoencoder is followed by a batch normalization layer, then a dense layer of size 640 (number of features), defined as its output layer. the second baseline system is the MobileNet, which classifies the machine conditions, also called as machine IDs (or sections). The classification loss between the input and the predictions are used to calculate a gamma point distributed anomaly threshold, which detects the machine anomaly in an unsupervised manner. Both models are trained on 5-consecutive (2*P+1, where P is the context window size) frames of log Mel band energies of size 128 x 64 ms analysis window (50% hope size) resulting in an input with the dimension of 640. Evaluation metrics used for this task are the Area Under Receiver Operating Characteristic (ROC) curve (AUC) and the partial AUC (pAUC) as illustrated in equations 1 and 2. The official score Ω is calculated using the harmonic mean of the AUC and pAUC as in 4.

$$AUC = \frac{1}{N_{-}N_{+}} \sum_{i=1}^{N_{-}} \sum_{j=1}^{N_{+}} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)) \tag{1}$$

$$pAUC = \frac{1}{\lfloor \frac{pN_{-}}{2} \rfloor N_{+}} \sum_{i=1}^{\lfloor pN_{-} \rfloor} \sum_{j=1}^{N_{+}} \mathcal{H}(\mathcal{A}_\theta(x_j^+) - \mathcal{A}_\theta(x_i^-)) \tag{2}$$

where

$$\mathcal{H}(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

and

$$\Omega = h\{\text{AUC}_{m,n,d}, \text{pAUC}_{m,n,d} | \; m \in \mathcal{M}, n \in \mathcal{S}(m), d \in \{\text{source,target}\}\} \tag{4}$$

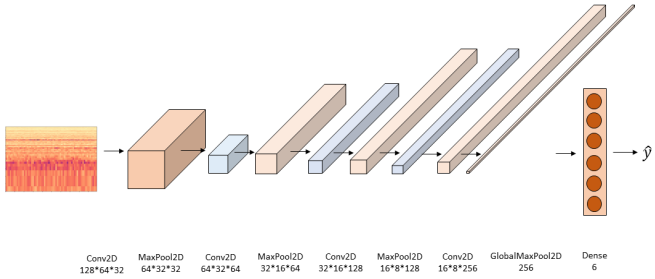

Figure 1: Overview of Our LeNet Architecture

where $h\{\cdot\}$ represents the harmonic mean (over all machine types sections, and domains), $\mathcal{M}$ represents the set of the machine types, and $\mathcal{S}(m)$ represents the set of the sections for the machine type m.

In this work, we consider the a classifier using the LeNet architecture with convolutional layers. This idea is motivated by the Mobile-Net benchmark [2] provided by the DCASE2021 organizers. We use mel-spectrograms for our model's input as they have proven to be robust in capturing audio features and appropriate input for training neural networks [5]. We further use the Mixup technique as our data augmentation approach. It increases the size of the train set to four times the given development set. Our proposed framework outperforms both baselines by 11.72% and 7.48%, AUC- and pAUC-harmonic-mean. We provide more details of our results compared to both baselines for each machine type and machine id in section 3.

The rest of this report is organized as follows. We present our model architecture in section 2) and our experimental results in section 3.

2. METHODOLOGY

Our methodology is motivated by the DCASE2021 MobileNet baseline, where a classification task is used to classify the machine IDs in different domains.

The idea of using such classification model is to calculate the classification loss between the train and the predictions to achieve a gamma-point-distribution for the anomaly threshold. We use this idea and evaluate the performance of LeNet architecture on DCASE2021 machine anomaly dataset.

Our LeNet architecture has four 2-dimensional convolutional layers with 32, 64, 128 and 256 filters, kernel size of 3, stride size of 1 and relu activation function. each convolutional layer is followed by a batch normalization, a 2-dimensional MaxPooling layer of the size 2 and 1% dropout. the output of the last convolutional layer is passed via a global max pooling layer to the dense layer with softmax activation function and 6 units, each unit for one machine condition. We depict our architeture in Figure 1.

We further apply Mixup [6] method to augment our train set and create new data from blended spectrograms. This data augmentation technique has proven to be robust for image and acoustic data, when aiming at regularizing the machine learning models.

.

3. RESULTS

As features for our model, we use 128 log mel-bands that are extracted from a 0.025 second analysis time window with a 0.012 second overlap over 64 time steps. Our LeNet architecture has four convolutional layers and one fully-connected layer resulting in 391302 total parameters with 960 non-trainable parameters. The activation function in each layer is a Relu function. Additionally, a dropout of size 0.1 is set at each encoding layer. We use the Adam optimization algorithm with 0.01 learning rate to compile the model. The model is trained on 80% of the train set and evaluated on the remaining 20%, over 100 epochs. Furthermore, we monitor the evaluation loss at each training epoch and use early stopping, setting the patience to 20. We stop at the x^{th} training epoch (where $x <= patience\_value$), if we observe no improvement in the evaluation loss[7]. We augment the training data using the Mixup technique with the alpha value set as 0.4 over 100 batches of the input spectrograms.
The results of our experiments compared to the DCASE2021 baseline system are presented in the following tables.

Table 1: Results of LeNet Architecture compared to the baselines on ToyCar

ToyCar						
MachineID	Dense-AE		MobileNet		LeNet	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	67.63%	51.87%	66.56%	66.47%	63.95%	61.36%
Source-01	61.97%	51.82%	71.58%	66.44%	53.37%	51.10%
Source-02	74.36%	55.56%	40.37%	47.48%	52.45%	48.89%
Target-00	54.50%	50.52%	61.32%	52.61%	64.96%	51.10%
Target-01	64.12%	52.14%	72.48%	63.99%	53.65%	53.73%
Target-02	56.57%	52.61%	45.17%	48.85%	69.58%	59.78%
Arithmetic mean	63.19%	52.42%	59.58%	57.64%	59.65%	54.32%
Harmonic mean	62.49%	52.36%	56.04%	56.37%	58.91%	53.94%

Table 2: Results of LeNet Architecture compared to the baselines on ToyTrain

ToyTrain						
MachineID	Dense-AE		MobileNet		LeNet	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	72.67%	69.38%	69.84%	54.43%	89.79%	82.52%
Source-01	72.65%	62.52%	64.79%	54.09%	88.65%	81.15%
Source-02	69.91%	47.48%	69.28%	47.66%	78.80%	47.36%
Target-00	56.07%	50.62%	46.28%	51.27%	54.90%	51.73%
Target-01	51.13%	48.60%	53.38%	49.60%	57.42%	53.63%
Target-02	55.57%	50.79%	51.42%	53.40%	62.74%	59.57%
Arithmetic mean	63.00%	54.90%	59.16%	51.74%	72.05%	62.66%
Harmonic mean	61.71%	53.81%	57.46%	51.61%	69.22%	59.80%

4. CONCLUSION

In this work, we proposed a framework for an unsupervised anomaly detection, which uses the Mixup data augmentation approach on log-mel bands as input and the LeNet Architecture to classify the machine sections. We used the classification loss between the inputs and model predictions to estimate an anomaly

Table 3: Results of LeNet Architecture compared to the baselines on Fan

Fan						
MachineID	Dense-AE		MobileNet		LeNet	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	66.69%	57.08%	43.62%	50.45%	65.46%	51.94%
Source-01	67.43%	50.72%	78.33%	78.37%	85.45%	82.73%
Source-02	64.21%	53.12%	74.21%	76.80%	68.52%	72.05%
Target-00	69.70%	55.13%	53.34%	56.01%	34.10%	48.15%
Target-01	49.99%	48.49%	78.12%	66.41%	83.33%	75.89%
Target-02	66.19%	56.93%	60.35%	60.97%	60.56%	64.42%
Arithmetic mean	64.03%	53.58%	64.66%	64.84%	66.23%	65.77%
Harmonic mean	63.24%	53.38%	61.56%	63.02%	60.35%	63.30%

Table 4: Results of LeNet Architecture compared to the baselines on Gearbox

Gearbox						
MachineID	Dense-AE		MobileNet		LeNet	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	56.03%	51.59%	81.35%	70.46%	84.43%	67.75%
Source-01	72.77%	52.30%	60.74%	53.88%	73.66%	63.16%
Source-02	58.96%	51.82%	71.58%	62.23%	49.75%	49.36%
Target-00	74.29%	55.67%	75.02%	53.96%	64.77%	68.72%
Target-01	72.12%	51.78%	56.27%	53.30%	75.10%	57.80%
Target-02	66.41%	53.66%	64.45%	55.58%	48.50%	50.31%
Arithmetic mean	66.76%	52.80%	68.24%	60.03%	68.91%	59.51%
Harmonic mean	65.97%	52.76%	66.70%	59.16%	65.46%	58.48%

Table 5: Results of LeNet Architecture compared to the baselines on Pump

Pump						
MachineID	Dense-AE		MobileNet		LeNet	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	67.48%	61.83%	64.09%	62.40%	63.41%	58.84%
Source-01	82.38%	58.29%	86.27%	66.66%	91.40%	76.21%
Source-02	63.93%	55.44%	53.70%	50.98%	67.18%	58.36%
Target-00	58.01%	51.53%	59.09%	53.96%	54.67%	54.15%
Target-01	47.35%	49.65%	71.86%	62.69%	82.01%	64.68%
Target-02	62.78%	51.67%	50.16%	51.69%	64.79%	58.68%
Arithmetic mean	63.66%	54.74%	64.20%	58.06%	70.57%	61.82%
Harmonic mean	61.92%	54.41%	61.89%	57.37%	68.55%	61.08%

threshold. Our framework outperformed both baseline systems provided by the challenge organizers with 11.72% and 7.48% AUc- and pAUc-harmonic mean over all machine types. We further would like to focus on the transparency of all three systems, baselines and LeNet classifier, to justify the outcome of the models and why they achieve different results on the same inputs.

5. ACKNOWLEDGMENTS

We would like to thank the Austrian Research Promotion Agency (FFG) for funding this work. It is part of the industrial project under the name DeepRUL, project ID 871357.

Table 6: Results of LeNet Architecture compared to the baselines on Slider

Slider						
MachineID	Dense-AE		MobileNet		LeNet	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	74.09%	52.45%	61.51%	53.97%	96.38%	88.78%
Source-01	82.16%	60.29%	79.97%	55.62%	74.96%	56.73%
Source-02	78.34%	65.16%	79.86%	71.88%	83.30%	81.33%
Target-00	67.22%	57.32%	51.96%	51.96%	80.10%	57.89%
Target-01	66.94%	53.08%	46.83%	52.02%	52.53%	51.70%
Target-02	46.20%	50.10%	55.61%	55.71%	59.21%	53.79%
Arithmetic mean	69.16%	56.40%	62.62%	56.86%	74.41%	65.03%
Harmonic mean	66.74%	55.94%	59.26%	56.00%	71.31%	62.24%

Table 7: Results of LeNet Architecture compared to the baselines on Valve

Valve						
MachineID	Dense-AE		MobileNet		LeNet	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	50.34%	50.82%	58.34%	54.97%	73.14%	65.26%
Source-01	53.52%	49.33%	53.57%	50.09%	96.02%	83.84%
Source-02	59.91%	51.96%	56.13%	51.69%	89.95%	77.47%
Target-00	47.12%	48.68%	52.19%	51.54%	71.28%	66.89%
Target-01	56.39%	53.88%	68.59%	57.83%	61.54%	57.52%
Target-02	55.16%	48.97%	53.58%	50.86%	78.90%	58.52%
Arithmetic mean	53.74%	50.61%	57.07%	52.83%	78.47%	68.25%
Harmonic mean	53.41%	50.54%	56.51%	52.64%	76.76%	66.97%

6. REFERENCES

[1] http://dcase.community/challenge2021/.

[2] 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.

[3] 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.

[4] 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.

[5] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 1996–2000.

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

[7] https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping.