DCASE CHALLENGE 2022, TASK 2: VARIATIONAL DENSE AUTOENCODER FOR UNSUPERVISED ANOMALOUS SOUND DETECTION OF MACHINERY

Technical Report

Anahid Jalali, Lam Pham, Clemens Heistracher, Denis Katic, Alexander Schindler

Austrian Institute of Technology, Vienna, Austria,
{anahid.jalali, lam.pham, clemens.heistracher, denis.katic, alexander.schindler}@ait.ac.at

ABSTRACT

In this study, we present an unsupervised anomalous sound detection framework trained on the DCASE2022 audio dataset. We use variational dense autoencoder to reconstruct the machine's healthy (normal) state and use the reconstruction loss as a threshold for detecting the anomalies in an unsupervised manner. Our framework outperforms DCASE2021 benchmarks in target domains. The dense autoencoder has a harmonic mean of AUC of 72.10% (source), and 45.49% (target) and pAUC of 54.09%. Our framework achieved the harmonic mean AUC of 68.78 and pAUC of 53.96, over all the machines. Our target domain arithmetic average, however, achieved 47.77% (baseline: 45.49%) which shows an slight improved performance from the dense autoencoder.

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

1. INTRODUCTION

Automatic Anomalous Sound Detection (ASD) identifies abnormal sounds of specific equipment and is considered an essential technology in industry 4.0 [1]. This systems are often used to monitor the machine conditions and aim to detect unknown anomalous sounds, which might also lead to machine outage and health degradation. DCASE2022, same as the previous two years, emulates the real life scenario, where 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 DCASE2022 challenge of unsupervised anomalous sound detection [2] focuses on three issues: unsupervised training, domain shift, and domain generalization, where participants are asked to use the provided audio datasets and submit their results.
The audio datasets 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: Bearing, Fan, Slider, ToyCar, ToyTrain, Gearbox and Valve. Each machine type consists of six sections. The dataset is available under 3 different releases:

- Development set: contains a training set and a testing set for each machine (sections 00, 01 and 02)
- Extra training set: contains more training data for each machine (sections 03, 04 and 05)
- Evaluation set: contains evaluation data for each machine (sections 03, 04 and 05)

Furthermore, the DCASE community provides two baseline systems [1, 5]: 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

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

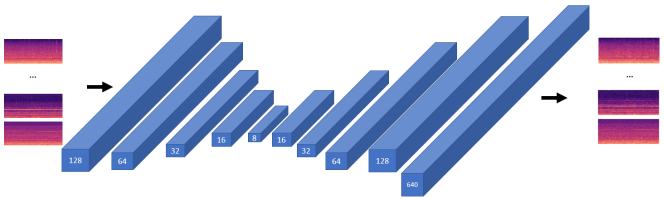

Figure 1: Overview of Our Variational Dense Autoencoder

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.

This year, we consider a variational dense autoencoder. This idea is motivated by the Vannila dense autoencoder benchmark [5] provided by the DCASE2022 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 [6].

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 vanilla dense autoencoder baseline, which the healthy recordings are used to reconstruct the wave data, and the reconstruction error is used to detect the anomalous sounds.

Our variational dense autoencoder has four dense layers with 128, 64, 32 and 18 neurons, each have 20% dropout-rate, and is followed by a batch normalization. The bottleneck has 8 neurons, and the output layer has the size of 640.

The idea behinde using the variational autoencoder is to have abstract information of the audio spectrogram in each layer and create the input data from the extracted and abstract embedding.

.

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 variational autoencoder has four dense layers and one fully-connected layer resulting in 188680 total parameters with 187704 trainable parameters. The activation function in each layer is a Relu function. Additionally, a dropout of size 0.2 is set at each encoding layer. We use the Adam optimization algorithm with 0.001 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 10. We stop at the x^{th} training epoch (where $x <= patience\_value$), if we observe no improvement in the evaluation loss[7].

The results of our experiments compared to the DCASE202 baseline system are presented in the following tables.

Table 1: Results of Variational Dense Autoencoder Architecture compared to the baselines on ToyCar

ToyCar						
MachineID	Dense-AE		MobileNet		Var-AE	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	86.42%	51.31%	47.40%	49.96%	86.74%	54.00%
Target-00	41.48%		56.40%		34.28%	
Source-01	89.85%	54.08%	62.02%	50.92%	91.59%	51.05%
Target-01	41.93%		56.38%		32.83%	
Source-02	98.84%	52.96%	74.19%	56.51%	98.58%	51.31%
Target-02	26.50%		45.64%		38.00%	
Source Arithmetic mean	91.70%	52.79%	61.21%	52.46%	92.30%	52.78%
Target Arithmetic mean	36.64%		52.81%		35.04%	
Source Harmonic mean	90.41%	52.74%	59.12%	52.27%	92.05%	52.75%
Target Harmonic mean	34.81%		51.96%		34.90%	

Table 2: Results of Variational Dense Autoencoder Architecture compared to the baselines on ToyTrain

ToyTrain						
MachineID	Dense-AE		MobileNet		Var-AE	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	67.54%	52.72%	46.02%	50.25%	67.76%	52.47%
Target-00	33.68%		49.41%		30.84%	
Source-01	79.32%	50.64%	71.96%	52.97%	77.90%	50.31%
Target-01	29.87%		45.14%		26.12%	
Source-02	84.08%	48.33%	63.23%	51.54%	80.98%	49.36%
Target-02	15.52%		44.34%		32.72%	
Source Arithmetic mean	76.98%	50.56%	60.40%	51.59%	75.54%	50.71%
Target Arithmetic mean	26.36%		46.30%		29.89%	
Source Harmonic mean	76.32%	50.48%	57.26%	51.52%	75.10%	50.68%
Target Harmonic mean	23.35%		45.90%		29.62%	

4. CONCLUSION

In this work, we proposed a framework for an unsupervised anomaly detection, which uses the log-mel bands as input and the Variational Dense Autoencoder Architecture to reconstrict the healthy (normal) machine state. We used the reconstruction loss between the inputs and model predictions to estimate an anomaly

Table 3: Results of Variational Dense Autoencoder Architecture compared to the baselines on Bearing

Bearing						
MachineID	Dense-AE		MobileNet		Var-AE	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	57.48%	51.49%	67.85%	54.41%	56.98%	50.63%
Target-00	63.07%		60.17%		60.48%	
Source-01	71.03%	55.85%	59.67%	55.09%	60.73%	51.89%
Target-01	61.04%		64.65%		53.83%	
Source-02	42.34%	49.18%	61.71%	64.18%	43.76%	50.00%
Target-02	52.91%		60.55%		52.68%	
Source Arithmetic mean	56.95%	52.18%	63.7%	57.89%	53.82%	51.73%
Target Arithmetic mean	59.01%		61.79%		55.68%	
Source Harmonic mean	54.42%	51.98%	60.58%	57.14%	52.75%	51.72%
Target Harmonic mean	58.38%		59.94%		55.48%	

Table 4: Results of Variational Dense Autoencoder Architecture compared to the baselines on Fan

Fan						
MachineID	Dense-AE		MobileNet		Var-AE	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	84.69%	59.95%	71.07%	55.40%	60.20%	59.10.00%
Target-00	39.35%		62.13%		44.82%	
Source-01	71.69%	51.12%	76.26%	52.14%	74.24%	51.31%
Target-01	44.74%		35.12%		41.16%	
Source-02	80.54%	62.88%	67.29%	65.14%	78.34%	67.73%
Target-02	63.49%		58.02%		67.00%	
Source Arithmetic mean	78.97%	57.98%	71.54%	57.56%	70.92%	58.38%
Target Arithmetic mean	49.19%		51.76%		50.99%	
Source Harmonic mean	78.59%	57.52%	70.75%	56.90%	70.01%	57.85%
Target Harmonic mean	47.18%		48.22%		48.75%	

Table 5: Results of Variational Dense Autoencoder Architecture compared to the baselines on Gearbox

Gearbox						
MachineID	Dense-AE		MobileNet		Var-AE	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	64.63%	60.93%	63.54%	62.12%	61.55%	62.05%
Target-00	64.79%		67.02%		64.74%	
Source-01	67.66%	53.74%	66.68%	56.85%	66.12%	53.00%
Target-01	58.12%		66.96%		55.68%	
Source-02	75.38%	61.51%	80.87%	50.62%	73.80%	59.42%
Target-02	65.57%		43.15%		63.57%	
Source Arithmetic mean	69.22%	58.72%	70.37%	56.53%	67.16%	58.15%
Target Arithmetic mean	62.83%		59.04%		61.33%	
Source Harmonic mean	68.93%	58.49%	69.21%	56.03%	66.78%	57.90%
Target Harmonic mean	62.64%		56.19%		61.05%	

Table 6: Results of Variational Dense Autoencoder Architecture compared to the baselines on Slider

Slider						
MachineID	Dense-AE		MobileNet		Var-AE	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	81.92%	61.65%	87.15%	71.57%	72.25%	60.26%
Target-00	58.04%		80.77%		51.68%	
Source-01	67.85%	53.06%	49.66%	48.21%	59.42%	52.63%
Target-01	50.3%		32.07%		45.88%	
Source-02	86.66%	53.44%	72.70%	49.69%	84.46%	53.42%
Target-02	38.78%		32.94%		41.55%	
Source Arithmetic mean	78.81%	56.05%	69.84%	56.49%	72.04%	55.23%
Target Arithmetic mean	49.04%		48.59%		46.37%	
Source Harmonic mean	77.95%	55.78%	65.15%	54.67%	70.57%	55.23%
Target Harmonic mean	47.67%		38.23%		46.00%	

threshold. Both baseline systems slightly outperformed our approach, where our approach had the arithmetic mean over all sections of source domain 68.78% AUC (baselines= 72.10), and pAUC of 53.96% (baseline 54.09). However, the average of our target source of 47.77, where baselines achieved an average arithmetic mean of 45.49%. We further would like to focus on the transparency of all three systems, baselines and our auto encoder, 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.

6. REFERENCES

[1] http://dcase.community/challenge2022/.

[2] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, “Description

Table 7: Results of Variational Dense Autoencoder Architecture compared to the baselines on Valve

Valve						
MachineID	Dense-AE		MobileNet		Var-AE	
	AUC	pAUC	AUC	pAUC	AUC	pAUC
Source-00	54.24%	52.15%	75.26%	55.37%	50.82%	52.52%
Target-00	52.73%		43.60%		49.43%	
Source-01	50.45%	49.78%	54.78%	54.69%	49.20%	49.47%
Target-01	53.01%		60.43%		52.58%	
Source-02	51.56%	49.24%	76.26%	85.74%	49.01%	49.57%
Target-02	43.84%		78.74%		42.38%	
Source Arithmetic mean	52.09%	50.39%	68.77%	65.27%	49.67%	50.52%
Target Arithmetic mean	49.86%		60.92%		48.13%	
Source Harmonic mean	52.01%	50.36%	67.09%	62.42%	49.66%	50.48%
Target Harmonic mean	49.46%		57.22%		47.73%	

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," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 1–5.

[4] 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," *In arXiv e-prints: 2205.13879*, 2022.

[5] 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," *In arXiv e-prints: 2206.05876*, 2022.

[6] 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.

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