TASK 2 DCASE 2020: ANOMALOUS SOUND DETECTION USING UNSUPERVISED AND SEMI-SUPERVISED AUTOENCODERS AND GAMMTONE AUDIO REPRESENTATION

Technical Report

Javier Naranjo-Alcazar[1,2]*, Sergi Perez-Castanos*[1]*, Pedro Zuccarello*[1]*, Maximo Cobos*[2]*,*

[1] Visualfy, Benisanó, Spain {javier.naranjo, sergi.perez, pedro.zuccarello}@visualfy.com
[2] Universitat de València, Burjassot, Spain, {maximo.cobos}@uv.es

ABSTRACT

Anomalous sound detection (ASD) is one of the fields of machine listening that is attracting most attention among the scientific community. Unsupervised detection is attracting a lot of interest due to its immediate applicability in many fields. For example, related to industrial processes, the early detection of malfunctions or damage in machines can mean great savings and an improvement in the efficiency of industrial processes. This problem can be solved with an unsupervised ASD solution since industrial machines will not be damaged simply by having this audio data in the training stage. This paper proposes a novel framework based on convolutional autoencoders (both unsupervised and semi-supervised) and a Gammatone-based representation of the audio. The results obtained by these architectures substantially exceed the results presented as a baseline.

Index Terms— Deep Learning, CNN, ASD, autoencoder, unsupervised learning

1. INTRODUCTION

Anomaly sound detection (ASD) has received much interest from the scientific community in recent years. The early detection of these events can mean a substantial improvement in systems that face this problem such as surveillance [1, 2] via audio or predictive maintenance [3, 4]. This last case is related to the industrial process and the early detection of a possible failure in the process machinery can mean a great advance and savings in the production of industrial products.

The ASD problem can be separated into two categories. Those problems in which the anomalous event to be detected is available in the training phase (supervised-ASD) [5] and those problems in which no such sound event is available (unsupervised-ASD) [6, 7]. Supervised-ASD can be defined as a kind of sound event detection (SED) but with some peculiarities like the duration or the nature of the sound event, like for example, a gunshot. On the other hand, in the unsupervised-ASD problem the objective is the detection of unknown or anomaly sound events without the system being aware of their existence, i.e. no anomalous events are available in the training data set. This is a case of real application in the industry today because it is unthinkable to purposely damage machines of great economic cost to obtain audio samples. A good unsupervised-ASD system should be able to recognize the anomaly by training only with samples from non-anomalous, or normal, sound events.

As it can be seen, this problem cannot be dealt as a classic classification problem like Acoustic Event Classification [8] or Audio tagging [9]. In this problem, there is a class, called *unknown* or *anomaly*, that must be recognized without the existence of positive samples of that class in the training set. In the case of engines or industrial machinery, the samples belonging to the *anomaly* class, or anomalous samples, are audio clips recorded when the machine is not working in the expected normal regime. The assumption is that this anomalous sounds show a different pattern than the ones produced with the machine working in normal regime. Therefore, if only one kind of training is available, a typical way of dealing with this kind of problem would be an outlier-detection scheme, that is, calculating the deviation, or difference, between the normal samples and the anomalies, this value is known as anomaly score. If this value exceeds a certain threshold, the sample is considered anomalous.

The first approaches to the unsupervised-ASD problem were made using classic machine learning techniques such as Gaussian mixture model [4] or Support vector machine [2]. In the last few years, due to the availability of larger amounts of data, Deep Learning techniques have become the state of the art in this field. As the main objective is to obtain a value, anomaly score, which provides us with information about the anomaly, the proposal of autoencoders seems to be a reasonable solution. Different architectures such as unsupervised autoencoders [10, 11, 12, 13] have been proposed in the state of the art. These solutions often implement recurrent or Dense layers in their solutions rather than convolutional. A different strategy may be the use of generative adversarial networks (GAN) [14]. This type of network is composed of two modules: the generator and the discriminator. The first one is in charge of generating false samples and the second one of discerning if the sample is false or real.

This work aims to propose a novel sound detection of anomalies based on a trained convolutional autoencoder with a 2D audio representation. As the information about the type of machine is available, one of the autoencoders proposed has a semi-supervised architecture. The other architecture is unsupervised, i.e. such information is not taken into account. The simplest approach would be to calculate an anomaly score per machine, that is, to train as many autoencoders as there are machines available. In this way, the autoencoder would be specialized to this type of machine. However, our approach is more complex since we obtain a single anomaly detector (autoencoder in this case) for all the machines.

2. PROPOSED METHOD

The proposed method is constituted by two steps: a 2D audio representation and a convolutional autoencoder with a bottleneck layer that serves as a divider between the encoder and the decoder. It is

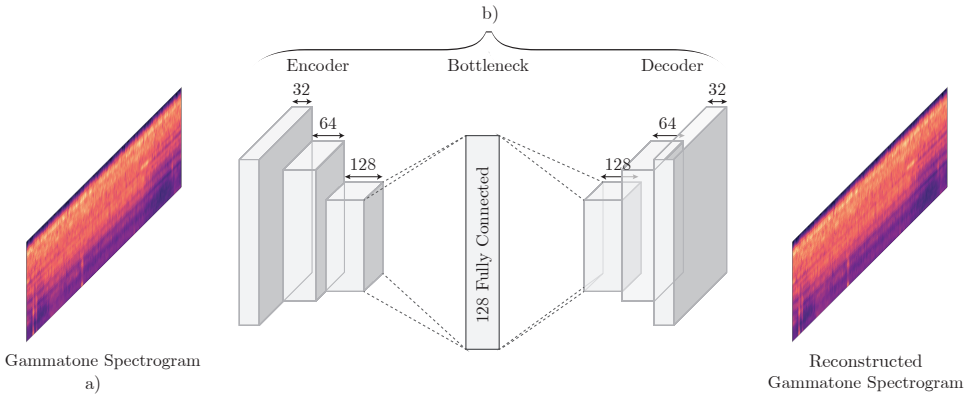

Figure 1: Full framework for ASD based on a Convolutional Autoencoder. Step a) shows the chosen audio representation and step b) the designed autoencoder architecture. The numbers indicate the number of filters in each convolutional block.

important to emphasize that a single autoencoder is trained for all available machines. As mentioned in the task description, this solution is much more challenging than proposing one autoencoder per machine type.

2.1. Audio representation

The 2D audio representation used in this framework is based on Gammatone filters [15]. This filter bank has shown promising results in the task of audio classification, surpassing the representation based on Mel filters [16], proposed, for example, in the MIMII dataset baseline [17]. Temporal bins are calculated with a window size of 40 ms and an overlap of 50%. The number of filters or frequency bins is set to 64. Once the representation is obtained, the logarithm is calculated and a normalization of mean 0 and standard deviation 1 is performed for each frequency bin with all available data. Therefore, the representation has a size of $64 \times T \times 1$, where T corresponds to the temporal bins according to the duration of the audio.

2.2. Autoencoder architecture

The autoencoder is made up of convolutional layers and a Dense layer that acts as a bottleneck. As can be seen in Figure 1, the encoder and decoder have a symmetric architecture. As can be recognized, each one is composed of 3 convolutional blocks. Each convolutional block is actually composed of 7 layers. The convolutional layer, the batch normalization (BN) layer and the activation layer, in this case ReLU. This set of 3 layers is repeated twice and followed by a pooling layer. In the case of the decoder, the pooling layer is replaced by an upsampling layer. The bottleneck layer corresponds to a Dense layer of 128 neurons with linear activation. This layer is the least dimensional representation that the encoder makes of the input signal and from which the decoder must be able to reconstruct to obtain the same input signal. Unlike the encoder, the decoder has an extra convolutional layer with 1 filter and linear activation that is responsible for reconstructing the representation of the input.

The architecture explained previously corresponds to an unsupervised autoencoder, that is, its only purpose is to reconstruct the input without taking into account extra information such as the class to which the input belongs. Therefore, the cost function to be optimized in this architecture is the mean squared error (MSE).

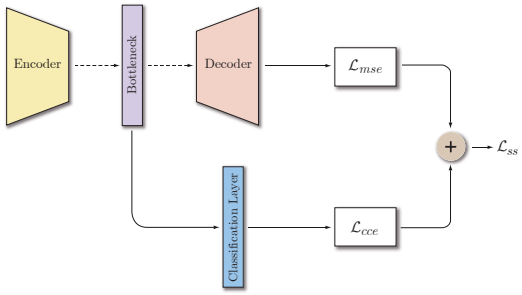

Figure 2: Semi-supervised autoencoder architecture

By having the information of what type of machine is associated with each audio, the autoencoder can be modified to take this into account. In this case, semi-supervised is understood because the supervised information corresponds to the type of machine, not if the sample is anomalous or normal as defined in [18]. The change consists of adding a classification layer by means of a Dense layer with the number of units equal to the number of machine classes available in the dataset. This layer is connected to the bottleneck layer, which, as previously mentioned, is where the information is most compressed. Thus, a semi-supervised architecture of the autoencoder is made. This modification can be seen in Figure 2. With this change, the cost function is affected and the classification error is now taken into account by means of the categorical crossentropy loss (CCE):

$$\mathcal{L}_{ss} = \alpha \mathcal{L}_{mse} + \beta \mathcal{L}_{cce} \tag{1}$$

where $\mathcal{L}_{mse}$ corresponds to the mean squared error and $\mathcal{L}_{cce}$ represents the categorical crossentropy loss. α and β are weighting factors such that $\alpha + \beta = 1$.

Framework	AUC					
	ToyCar	ToyConveyor	fan	pump	slider	valve
B	78.77±1.03	72.53±0.67	65.83±0.53	72.89±0.70	84.76±0.29	66.28±0.49
U	**95.67**	**96.63**	79.87	81.51	80.86	82.85
U FD	91.12	93.36	**80.40**	**82.61**	**81.16**	**83.19**
SS-0.7-0.3	87.27	90.35	78.63	80.33	78.94	80.94

Table 1: Results AUC (%) obtained by the proposed frameworks compared to the baseline proposed with the dataset. Baseline is denoted by B. The unsupervised autoencoder is represented with U and the semi-supervised one with SS followed by α and β values, i.e. SS-0.7-0.3 corresponds to the semi-supervised architecture with $\alpha = 0.7$ and $\beta = 0.3$. FD denotes that full dataset was used in training stage (1st and 2nd release).

Framework	pAUC					
	ToyCar	ToyConveyor	fan	pump	slider	valve
B	67.58±1.04	60.43±0.74	52.45±0.21	59.99±0.77	66.53±0.62	50.98±0.15
U	**87.14**	**90.45**	70.78	70.99	**70.69**	71.62
U FD	73.41	80.32	**72.56**	**72.23**	69.94	**72.34**
SS-0.7-0.3	74.21	81.50	71.26	70.94	70.08	70.83

Table 2: Results pAUC (%) obtained by the proposed frameworks compared to the baseline proposed with the dataset. Notation is explained in Table 1

3. RESULTS

The dataset used to train and evaluate models is the one used in Detection and Classification of Acoustic Scenes and Events (DCASE) 2020 Task2, focused on ASD. It consists in subsets of ToyADMOS [19] and MIMII [17] datasets. From the first one, car and conveyor classes are combined with valve, pump, fan and slide rail classes from the second one. In this context, a class corresponds to a machine type.

Results obtained in this task are shown in Tables 1 and 2. As it can be appreciated, all the proposed frameworks exceed the results presented as a baseline [20] except the class *slider*. The architecture that shows a better result is the unsupervised one. However, some machines show a better result when the training set corresponds only to the portion released in the first release.

As it can be observed, the improvement is substantial in all machines obtaining the lowest improvement in the *pump* class of about 10 percentage points. On the other hand, *ToyConveyor* is the machine that has been most improved with about 24 more percentage points compared to the baseline. As far as the slider machine is concerned, a decrease of about 4 perceptual points is obtained.

As for the semi-supervised architecture, it seems that this extra information during training adds noise to the internal representations generated by the convolutional layers. However, a much more comprehensive grid-search is needed for both α and β.

Table 3 shows the name relationship between the submission name and the results shown in this work.

Autoencoder used	**Data used**	**Submission name**
Unsupervised	1st release	Naranjo-Alcazar_Vfy_task2_1
Unsupervised	Full data	Naranjo-Alcazar_Vfy_task2_2
Semi-supervised	1st release	Naranjo-Alcazar_Vfy_task3_3

Table 3: Relationship between the name of the submission and the implementation explained in this paper.

4. CONCLUSION

The state of the art in the field of Anomalous sound detection has shown the great potential that solutions based on autoencoders have for mitigating the problems related to this task. Different architectures have been proposed, such as variational autoencoders. However, it is not so common the appearance of autoencoders with convolutional layers. Therefore, this paper shows the potential of such layers to extract relevant information when reconstructing the audio in order to obtain the necessary anomaly score to discern whether the sample is anomalous or not. In addition, it is also studied how a semi-supervised architecture behaves in this kind of problems. Regarding the audio representation, the choice was made to use the Gammatone representation instead of the one based on Mel filters or instead of converting the audio into a one dimensional vector as it is proposed in several state of the art solutions.

5. ACKNOWLEDGMENT

The participation of Javier Naranjo-Alcazar and Dr. Pedro Zuccarello in this work is partially supported by Torres Quevedo fellowships DIN2018-009982 and PTQ-17-09106 respectively from the Spanish Ministry of Science, Innovation and Universities.

6. REFERENCES

[1] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Probabilistic novelty detection for acoustic surveillance under real-world conditions," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 713–719, 2011.

[2] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE transactions on intelligent transportation systems*, vol. 17, no. 1, pp. 279–288, 2015.

[3] A. Yamashita, T. Hara, and T. Kaneko, "Inspection of visible and invisible features of objects with image and sound signal processing," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2006, pp. 3837–3842.

[4] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Optimizing acoustic feature extractor for anomalous sound detection based on neyman-pearson lemma," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 698–702.

[5] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2007, pp. 21–26.

[6] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[8] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[9] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. D. Plumbley, "Convolutional gated recurrent neural network incorporating spatial features for audio tagging," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 3461–3466.

[10] 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.

[11] E. Marchi, F. Vesperini, F. Weninger, F. Eyben, S. Squartini, and B. Schuller, "Non-linear prediction with lstm recurrent neural networks for acoustic novelty detection," in *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–7.

[12] Y. Kawaguchi and T. Endo, "How can we detect anomalies from subsampled audio signals?" in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2017, pp. 1–6.

[13] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212–224, 2018.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[15] S. Tabibi, A. Kegel, W. K. Lai, and N. Dillier, "Investigating the use of a gammatone filterbank for a cochlear implant coding strategy," *Journal of neuroscience methods*, vol. 277, pp. 63–74, 2017.

[16] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 356–367.

[17] 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.

[18] Y. Kawachi, Y. Koizumi, and N. Harada, "Complementary set variational autoencoder for supervised anomaly detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2366–2370.

[19] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "Toyadmos: A dataset of miniature-machine operating sounds for anomalous sound detection," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 313–317.

[20] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, *et al.*, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2006.05822*, 2020.