VISION TRANSFORMER BASED EMBEDDINGS EXTRACTOR FOR UNSUPERVISED ANOMALOUS SOUND DETECTION UNDER DOMAIN GENERALIZATION

Technical Report

Antonio Almudévar, Alfonso Ortega, Luis Vicente, Antonio Miguel, Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
{almudevar, ortega, lvicente, amiguel, lleida}@unizar.es

ABSTRACT

Anomalous sound detection (ASD) is the task of identifying if a sound is normal or anomalous with respect to a given reference. In most scenarios, we have a large amount of normal data to design our model, but little or no anomalous data. When this situation occurs, the problem can be approached in an unsupervised manner, i.e., only normal data is used for design. In this report we present a solution for the DCASE2022 task 2 (Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques), which aims to address the ASD problem under domain generalization. This means that the data to develop the system belongs to the source domain, while the test data can belong to this domain or to a different one (target domain). The presented solution proposes an embeddings extractor based on a Vision Transformer (ViT) and makes use of the k-Nearest-Neighbor (k-NN) algorithm to obtain the anomaly score.

***Index Terms*—** unsupervised anomaly detection, vision transformer, arcface, domain generalization

1. INTRODUCTION

Unsupervised anomaly detection (UAD) is the task of identifying whether an event is normal or anomalous with respect to a given reference using only normal data for its design. In particular, in the field of Deep Learning, the concept of unsupervised refers to the fact that no labels are used to train the models. For the case of anomaly detection, this means that only normal data are used to train the models. The solutions proposed in the literature for the anomaly detection (AD) problem focus on being able to obtain an anomaly score, which is a number that should be higher for anomalous data than for normal data [1]. Subsequently, a threshold is set and a sample is considered to be anomalous if and only if its anomaly score is higher than that threshold.

In this paper we present a proposal to solve the unsupervised anomalous sound detection problem, that is, to try to identify whether a given audio is normal or anomalous using only normal audios for its design. This problem has been particularly relevant in recent years, being the second of the six tasks of the DCASE Challenge in 2020 and 2021 [2, 3]. To evaluate the performance of our system, we use the dataset of task 2 of DCASE2022 and show its performance in terms of AUC and pAUC, as proposed in the challenge.

To solve this problem there has been multiple proposals with different approaches. Some of the most relevant are the following. In DCASE2020, an autoencoder with dense layers and whose input was 5 consecutive frames of the mel-spectrogram was proposed as baseline. Mean squared error between input and output was used as anomaly score. In DCASE2021, in addition to the previous one, another baseline was proposed using MobileNetV2 [4] that took the mel-spectrograms as inputs. In this case, the negative logit was used as the anomaly score. The 2020 winner [5] proposed to use a Group Masked Autoencoder, which is an adaptation of the Masked for Distribution Estimation (MADE) [6] in which its inputs were the mel-spectrogram frames, instead of scalars. The 2021 winner [7] combined three different systems to achieve the best possible performance. The first system obtained x-vectors [8] and calculated the cosine and Mahalanobis distances between the test embedding and the average training embedding as an anomaly score. The second system used a WaveNet [9], but instead of using a few convolutions, they used an x-vector component. The last one used a Normalizing Flow [10] to estimate the distribution of an n-bin segment of a spectrogram conditioned to the remaining bins.

Our method combines a classifier with generative models to detect anomalous sounds. In particular, the proposed system is a combination of three networks. The first one is a embedding extractor. The second one is a classifier that discriminates to which section the embeddings coming from the previous network belong. Finally, the third one is a normalizing flow that tries to estimate the probability distribution that the embeddings follow. The characteristics of each of the three networks, the training process and how the proposed anomaly scores are defined are detailed below. As far as we believe, this is the first time such a system has been used for anomalous sounds detection.

This paper is organized as follows. Section 2 presents the dataset. Section 3 describes the proposed approach to solve the problem. Section 4 explains the three systems presented to the Challenge. In section 5 the results for the three systems are presented. Finally, in section 6, conclusions obtained are summarized.

2. DATASET

The DCASE2022 task 2 dataset is composed of data from ToyADMOS2 [11] and MIMII DUE [12] and contains the sounds emitted by seven machines operating normally and when broken (abnormally). In addition, for each machine there are sounds belonging to six sections. A section is defined as a subset of the dataset for calculating performance metrics. Each section is dedicated to a specific type of domain shift. The machines are: ToyCar, ToyTrain, fan, gearbox, bearing, slide rail, and valve and sections 0 to 5. The recordings are 10 seconds long, single channel and sampled at 16 kHz. The way we split the database is as follows:

Training Dataset: Consists of three sections for each machine type (Sections 0, 1, and 2), and each section is a complete set of the training and test data. For each section, this dataset is composed of 990 clips of normal sounds in a source domain for training, 10 clips of normal sounds in a target domain for training and 100 clips each of normal and anomalous sounds including data from both domains for the test. Source/target information is provided in the test data in the development dataset. Attributes that represents operational or environmental conditions are also provided.

Additional Development Dataset: Provides three sections identical to the evaluation dataset (Sections 3, 4, and 5). Each section consists of (i) around 990 clips of normal sounds in a source domain for training and (ii) only 10 clips of normal sounds in a target domain for training.

Evaluation Dataset: Provides test clips for three sections (Sections 3, 4, and 5). Each section consists of 100 test clips, none of which have a condition label (i.e., normal or anomaly) or the domain information (i.e., source or target). Attributes are not provided in this dataset.

3. PROPOSED APPROACH

This section presents the proposed method for solving the UAD problem under domain generalization. In this method, given an audio, an embedding is obtained from an embedding extractor based on the Vision Transformer (ViT) and the cosine distance between this embedding and the nearest neighbor of the training audios is used as the anomaly score. The architecture used as embeddings extractor and the way to define the anomaly scores are detailed below.

3.1. Vision Transformer based Embeddings Extractor

ViT was presented in [13] and was the first approach to image processing that outperformed Convolutional Neural Networks (CNN) in the image recognition task. The idea is to split an image into 16×16 patches, embed them linearly, add position embeddings and pass the resulting arrays as input to a standard Transformer encoder. To perform the classification task, they add an extra learnable ”classification token” to the sequence. Subsequently it has been successfully used for other tasks such as object detection [14] or semantic segmentation [15].

As with computer vision systems based on CNNs, the Vision Transformer has been used to process audio by taking the spectrogram, mel-spectrogram or similar as input. In particular, the Audio Spectrogram Transformer proposed in [16] stands out. This system divides the spectrogram into 16×16 patches, just like ViT, and performs the identical process to the one described in the previous paragraph. However, in the solution presented here, instead of taking patches of size 16×16 from the spectrogram, we take each of the time-frames, embed them linearly, add position embeddings and pass the resulting arrays as input to a standard Transformer encoder. As in the ViT, to perform the classification task, we add an extra learnable ”classification token” to the sequence. Next we formally define the process to extract the embeddings.

To obtain the inputs to this network, first, we calculate the log-mel-spectrogram of the signal $X = \{X_t\}_{t=1}^{T}$, where $X_t \in R^F$ and F and T are the number of mel-filters and time-frames, respectively. As input to the embedding extractor, $\psi_t = (X_t, .., X_{t+P-1}) \in R^{P\times F}$ is used. The context window is shifted L frames. So there are $N = \lceil \frac{T-P}{L} \rceil$ inputs for each signal X. For the given results, a frame size of the short-term-Fourier transform (STFT) of 1024 samples (64 ms) and a hop size of 50% is used. In addition, $F = 128$, $P = 64$ and $L = 4$ have been taken. This means that for 10 second inputs and a sampling rate of 16kHz, $T = 313$ and $N = 63$. The output of the network is $e(\psi_t) = e_t \in R^S$ where e is the ViT based embeddings extractor, whose details are shown in table 1.

Layers	Hidden size D	MLP size	Heads	Params
12	384	1536	6	22M

Table 1: Details of Vision Transformer used in this work

Name	Sections
all	0, 1, 2, 3, 4, 5
dev	0, 1, 2
eval	3, 4, 5
0_3	0, 3
1_4	1, 4
2_5	2, 5

Table 2: Different subsets of data used to train the Embeddings Extractor

Finally, for the classifier we have used the Additive Angular Margin Loss (ArcFace) [17] to enhance discriminative power for cosine distance.

3.2. Training of Embeddings Extractor

To train the embeddings extractor the objective is that the system described in the previous section to be able to classify the machine from which the sound comes from. We have used the crossentropy as a cost function, we have used the Adam optimizer with a learning rate of 1e-5 and a batch size of 128. Mixup with α equal to 0.5 has been used as data augmentation. In addition, we have trained the same network with different training subsets to analyze how training with different sections influences each machine. Table 2 shows all the training subsets used. In all of them the data are normal and in the source domain. The reason for this is the way in which the attributes are modified according to the machines. figure 1 shows a t-SNE representation of the embeddings obtained for network trained with `all` subset.

3.3. Anomaly Scores

Once the network has been trained, to obtain the anomaly scores, all the embeddings of the Training Dataset (both in the source and target) are calculated. Once all these embeddings are obtained, to calculate the anomaly score corresponding to an audio, the cosine distances with the nearest neighbors of the training embeddings are taken for the N inputs of each audio. Subsequently, the average of these N cosine distances is used as the anomaly score. The fact of using training embeddings in both source and target domains solves the domain generalization problem.

4. SYSTEMS DESCRIPTION

This section presents the three proposed systems, which differ in the use of the different training subsets described in the table 2. The training data subsets used for each system, as well as the number of epochs used, are described below.

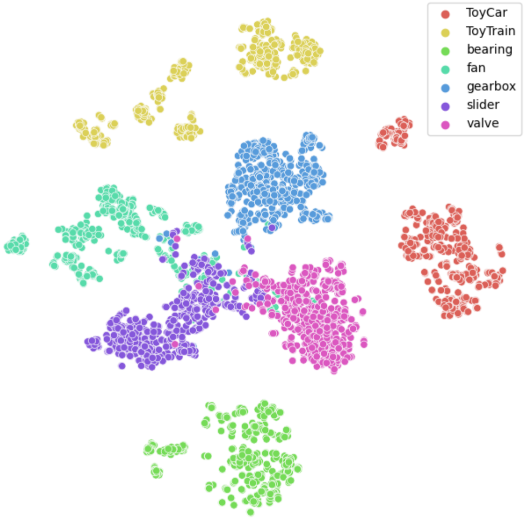

Figure 1: t-SNE representation of the embeddings for network trained with `all` subset

	Section					
	0	1	2	3	4	5
ToyCar	dev	dev	dev	eval	eval	eval
ToyTrain	dev	dev	dev	eval	eval	eval
bearing	all	all	all	all	all	all
fan	0_3	1_4	2_5	0_3	1_4	2_5
gearbox	all	all	all	all	all	all
slider	dev	dev	dev	eval	eval	eval
valve	0_3	1_4	2_5	0_3	1_4	2_5

Table 3: Training subsets used in system 3

System 1: In this system the subset `all` is used and the network has been trained for 15 epochs for all machines and sections.

System 2: In this system the subset `dev` is used for the entries of sections 0, 1 and 2 of all machines and the subset `eval` for the entries of sections 3, 4 and 5 of all machines. In this case, the networks have been trained for 30 epochs.

System 3: In this case, all the networks trained with the different subsets of the table 2 have been combined to obtain a higher performance than in the previous systems. The way these subsets have been used is shown in table 3. Likewise, the number of epochs is different depending on the machine and section, as reflected in table 4.

	Section					
	0	1	2	3	4	5
ToyCar	30	30	30	30	30	30
ToyTrain	25	25	25	25	25	25
bearing	15	15	15	15	15	15
fan	15	15	25	15	15	25
gearbox	10	10	10	10	10	10
slider	25	25	25	25	25	25
valve	20	25	25	20	25	25

Table 4: Number of epochs in system 3

5. RESULTS

The harmonic mean of AUC Source, AUC Target and pAUC for each machine in both the two baseline systems [18] and the three systems proposed in the previous section are shown in table 5. Several conclusions can be drawn from this table.

The first is that in the target domain the systems proposed in this work outperform the baseline systems by far in all machines. On the other hand, in the source domain in the ToyCar, ToyTrain and fan, the baseline AE system outperforms the systems proposed here, although the difference is minimal for ToyCar and fan. In the rest of the machines, the proposed systems outperform the baseline systems by a considerable margin, except in the case of valve, where the result of the MobileNetV2 Baseline system is similar to System 3.

6. CONCLUSION

In this paper we have presented a method to detect anomalous events using a ViT based Embeddings Extractor and tested its performance with the DCASE 2022 task2 dataset. To the best of our knowledge, this is the first time it has been used that each of the time-frames has been passed as input to a ViT. Notably, as demonstrated, this way of dealing with mel-spectrograms works and outperforms other systems for the anomaly detection task.

When working with transformers, it is common to use pretrained models with rich databases such as Imagenet [19] or AudioSet [20] and subsequently perform fine-tuning with the task's own data. This allows implementing larger and more expressive models, obtaining better results in a multitude of tasks [16, 21]. Therefore, a future line of work consists of taking pretrained ViT with databases such as those mentioned above and performing fine-tuning with the data of this task and analyzing if a better performance is achieved. On the other hand, another future line of work consists of using this embedding extractor for tasks other than unsupervised anomaly detection.

7. ACKNOWLEDGMENT

Supported by EU H2020 Marie Skłodowska-Curie Grant 101007666; MCIN/AEI/10.13039/501100011033 and "NextGenerationEU"/PRTR PDC2021-120846-C41. Government of Aragón (Reference Group T36_20R). Antonio Almudévar is supported by Government of Aragón predoctoral grant (2021–2025).

8. REFERENCES

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

[2] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, November 2020, pp. 81–85.

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

		AE Baseline	MobileNetV2 Baseline	System 1	System 2	System 3
ToyCar	AUC Source	**90.41**	59.12	87.09	87.46	87.46
	AUC Target	34.81	51.96	71.18	**77.79**	**77.79**
	pAUC	52.74	52.27	54.10	**56.20**	**56.20**
ToyTrain	AUC Source	**76.32**	57.26	62.95	67.89	66.64
	AUC Target	23.35	45.90	49.87	52.43	**54.26**
	pAUC	50.48	51.52	51.29	51.18	**52.08**
bearing	AUC Source	54.42	60.58	**67.74**	58.18	**67.74**
	AUC Target	58.38	59.94	**74.79**	69.53	**74.79**
	pAUC	51.98	57.14	**59.34**	59.15	**59.34**
fan	AUC Source	**78.59**	70.75	73.92	75.65	77.81
	AUC Target	47.18	48.22	63.80	65.25	**66.05**
	pAUC	57.52	56.90	56.94	60.20	**60.55**
gearbox	AUC Source	68.93	69.21	83.02	79.07	**86.10**
	AUC Target	62.64	56.19	75.87	68.68	**75.88**
	pAUC	58.49	56.03	**61.48**	55.34	60.56
slider	AUC Source	77.95	65.15	91.15	**94.06**	93.36
	AUC Target	47.67	38.23	79.70	85.60	**86.72**
	pAUC	55.78	54.67	68.20	**77.74**	77.19
valve	AUC Source	52.01	67.09	63.74	64.55	**67.95**
	AUC Target	49.46	57.22	56.16	**59.00**	57.33
	pAUC	50.36	**62.42**	53.33	53.07	57.05

Table 5: Harmonic mean for AUC Source, AUC Target and pAUC for all the machines in baseline and proposed systems

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.

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[5] R. Giri, S. V. Tenneti, K. Helwani, F. Cheng, U. Isik, and A. Krishnaswamy, "Unsupervised anomalous sound detection using self-supervised classification and group masked autoencoder for density estimation," DCASE2020 Challenge, Tech. Rep., July 2020.

[6] M. Germain, K. Gregor, I. Murray, and H. Larochelle, "MADE: Masked autoencoder for distribution estimation," in *International Conference on Machine Learning*. PMLR, 2015, pp. 881–889.

[7] J. Lopez, G. Stemmer, and P. Lopez-Meyer, "Ensemble of complementary anomaly detectors under domain shifted conditions," DCASE2021 Challenge, Tech. Rep., July 2021.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[9] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[10] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3964–3979, 2020.

[11] 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.

[12] 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.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," *arXiv preprint arXiv:2012.09958*, 2020.

[15] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262–7272.

[16] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[17] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[18] 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.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[20] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[21] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.