DCASE CHALLENGE 2022 : SELF-SUPERVISED LEARNING PRE-TRAINING, TRAINING FOR UNSUPERVISED ANOMALOUS SOUND DETECTION

Technical Report

Ismail Nejjar[1,2]*, Jean Meunier-Pion*[1,3]*, Gaetan Frusque*[1]*, Olga Fink*[1]*,*

[1] EPFL, IMOS, Lausanne, Switerzland, {ismail.nejjar, gaetan.frusque, olga.fink}@epfl.ch
[2] ETH Zürich, Chair of Intelligent Maintenance Systems, Zürich, Switerzland, {inejjar}@ethz.ch
[3] CentraleSupélec, Gif-sur-Yvette, France, {jean.meunier-pion}@student-cs.fr

ABSTRACT

This technical report presents our proposed approaches for Task 2 of the DCASE 2022 Challenge, Unsupervised anomalous sound detection (ASD) for machine condition monitoring by applying domain generalization techniques. The main objective of this challenge is to detect anomalous machine sounds regardless of the domain shifts. Our approach introduces a two-step learning process, where normal sounds of each specific machine type are used to pre-train a Convolutional Neural Network (CNN) in a self-supervised way. Three objectives are thereby pursued: (1) reveal the impact of attributes on the data by enforcing embeddings in the same batch to be different (2) obtain uncorrelated embedding features containing specific information, (3) respecting defined geometrical constraints between the different domains. The model trained in an unsupervised way is then fine-tuned on the labels of the section indices. Ultimately, anomalous sounds are detected by using the feature vectors extracted from the CNN and applying k-NN to them. As a result, for the development set, it is shown that the presented framework significantly outperforms both baselines.

Index Terms— Self-supervised Learning, Domain Generalization, Anomalous Sound Detection, Convolutional Neural Network

1. INTRODUCTION

The DCASE 2022 Challenge Task 2 [1] aims to tackle the problem of unsupervised anomaly detection for machine condition monitoring using domain generalization techniques.

This task presents two main challenges:

- Source and target domains are strongly imbalanced
- Machines have a large variability in operating and recording conditions.

The performance of the models trained on the source domain data degrades severely when applied to the target domain is. This is due to domain shifts between the source and the target domain caused by other factors than anomalies.

Two baseline methods are provided for Task 2. The first method is using an autoencoder that performs well for unsupervised anomaly detection but struggles with the domain generalization problem. The second method is based on MobileNetV2 and is less sensitive to the gap between the source and target domains.

1.1. Dataset

The dataset used for this task was generated from MIMII DG [2] and ToyADMOS2 [3] datasets consisting of normal and anomalous operating sounds of seven types of toy/real machines. ToyCar and ToyTrain machines types are extracted from ToyADMOS2 dataset while fan, gearbox, bearing, slide rail, and valve are extracted from MIMII DG dataset.

Each recording is single-channel, 10-second audio sampled at 16 kHz. The signals result from a mixture between machine sounds and environmental noise samples at several real-world factories. Each machine type contains three sections in the development dataset and three sections in the additional dataset. In this report, all the training data in the development dataset as well as the additional training dataset are used for training the models. The model's performance is evaluated on the test data in the development dataset.

2. METHOD

We propose a novel approach inspired by Variance-Invariance-Covariance Regularization (VICReg), a self-supervised learning method proposed in [4]. An overview of the proposed approach is provided in Fig 1. The framework comprises a self-supervised learning algorithm, a subsequent fine-tuning step and detection based on kNN.

The objective of the self-supervised task is to learn an encoder providing meaningful representations of audio samples. Then, the pre-trained encoder is extended and fine-tuned to perform supervised classification of the section ID. Finally, the embeddings produced by this encoder are used to calculate the anomaly score with k-NN.

2.1. Audio Prepocessing

Each audio sample given as a time series signal is transformed into a log-mel spectrogram. Before applying this transformation, the mean of each individual sample is subtracted from the raw audio signal. Similar to the MobileNetV2-based baseline, the input to our model is a two dimensional image-like feature $\psi_t \in \mathcal{R}^{P \times F}$. The frame size of short-time Fourier transform (STFT) is 64 ms, and the hop size is 32 ms. We also set the number of Mel bins to 128. The number of frames of the context window P was fixed to 64. The context window is shifted by L frames resulting in B extracted images, with $B = \lceil \frac{T-P}{L} \rceil$ with $L = 8$. Given the previous parameters the total spectrogram size T is equal to 313.

2.2. CNN Architectures

In this work, we used the MobileNetV2 [5] backbone trained from scratch. The off-the-shelf Pytorch [6] implementation of MobileNetV2 is used. The width multiplier parameter was set to 0.5 and the last layers were adapted to obtain a 320 dimensions vector per input image. A detailed summary of the applied CNN architecture is provided in Table 1.

Input	Operator	t	c	n	s
$128 \times 64 \times 3$	conv2d 3×3	-	16	1	2
$64 \times 32 \times 16$	bottleneck	1	8	1	1
$64 \times 32 \times 8$	bottleneck	6	16	2	2
$32 \times 16 \times 16$	bottleneck	6	16	3	2
$16 \times 8 \times 16$	bottleneck	6	32	4	2
$8 \times 4 \times 32$	bottleneck	6	48	3	1
$8 \times 4 \times 48$	bottleneck	6	80	3	2
$4 \times 2 \times 80$	bottleneck	6	160	1	1
$4 \times 2 \times 160$	conv2d 1×1	-	320	1	1
$4 \times 2 \times 320$	conv2d 4×2	-	320	1	1
$1 \times 1 \times 320$	conv2d 1×1	-	320	1	

Table 1: Modified MobileNetV2 architecture used for all experiments. Each row represents the sequence of layers, repeated n times, with c channels, and stride s

2.3. Self-Supervised Pre-training

The first step of our approach is based on a self-supervised learning algorithm, inspired by VICReg. An overview of this step is provided in Fig 1. For each machine type, we trained a Siamese architecture where the three branches are similar and share the same weights. Each branch is composed of an encoder f_θ which corresponds to the modified MobileNetV2 presented in Table 1, followed by an expander h_ϕ. The expander is composed of three fully-connected layers of size 1280. Each of the layers is followed by a batch normalization layer [7] and a ReLU [8] activation function.

In order to mitigate the gap between the source and target domains, we propose to improve upon the VICReg framework [4] by using Mixup [9] to augment the target domain. A novel loss is proposed to take into account the added mixup branch, acting as a regularization term and improving domain generalization.

While in the original VICReg approach, a data augmentation approach is applied, we propose to impose similarity between the samples representations. Given all the S log Mel-spectrograms from both the source and target domains of all sections for each machine type, two different samples X and X' are selected. For each such pair of samples, a linear combination with respect to λ is obtained. This combination gives rise to a new sample denoted as X_λ. Formally, λ is a realisation of a beta distribution $Beta(\alpha, \beta)$ and represents the mixup rate. In our case we set $\alpha = \beta = 0.5$.

First X, X' and X_λ are encoded by f_θ resulting in Y, Y' and Y_λ, and then mapped by the expander onto the embeddings, Z, Z' and Z_λ. The loss is composed of three terms and computed at the embedding level on Z, Z' and Z_λ.

Given a batch of size N, we denote $Z = [z_1, ..., z_N]$, with $z \in \mathbb{R}^D$.

The loss comprises three parts. The first two parts follow the VicReg implementation, while we propose to substitute the third part (representing the invariance criterion in VicReg) by a term that enforces the embedding feature of the source and target domains to be distinguishable. Thereby, we are able to learn a specific representation for both domains. This is achieved by computing the mean squared error between the produced embedding vector of the mixed input X_λ to the linear combination with respect to λ of the produced embeddings for X and X':

$$s(Z_\lambda, Z, Z', \lambda) = \frac{1}{N}\sum_{i=1}^{N} \|z_{\lambda,i} - (\lambda z_i + (1-\lambda) z'_i)\|_2^2 \quad (1)$$

The second term forces the variance inside each batch to be equal to 1, preventing a mode collapse.

$$v(Z) = \frac{1}{D}\sum_{i=1}^{D} max(0, 1 - S(z^j)) \quad (2)$$

where z^j is the jth row of the matrix Z in the batch, and $S(z) = \sqrt{Var(z)}$ is the standard deviation.

Finally, the last term aims to learn uncorrelated features for each embedding, by forcing the off-diagonal elements to be zero, resulting in a rich embedding. The covariance matrix is defined as :

$$C(Z) = \frac{1}{N-1}\sum_{i=1}^{N} (z_i - \bar{z})(z_i - \bar{z})^T \quad (3)$$

with $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$ representing the mean embedding over a mini-batch.

$$c(z) = \frac{1}{d}\sum_{i \neq j} [C(Z)]_{i,j}^2 \quad (4)$$

The final loss used to train the model is:

$$\gamma s(Z_\lambda, Z, Z', \lambda) + \mu(v(Z) + v(Z')) + \nu(c(Z) + c(Z')) \quad (5)$$

γ, μ, ν are hyper-parameters that we set in this report to $25, 25, 1$ respectively. The networks are trained using the LARS [10] optimizer, with a learning rate of 0.8, weight decay of 10^{-4} and a batch size of 1024 for 100 epochs. In addition, 10 warmup epochs were used and the learning rate followed a cosine decay schedule starting from 0 and finishing at 0.002.

After the pre-training, only the encoder model used for the downstream classification task presented in the next section.

2.4. Fine-tuning on Section ID Classification

Similar to the baseline method, the model is fine-tuned to identify the section ID of an audio sample. The pre-trained encoder is used and a classifier composed of 2 fully-connected (320-128-6) is added. To improve the robustness of the model a mixup strategy on source and target data for each section is used to generate augmented data of intra-domain and inter-domain samples. The KL divergence loss between the classifier output and the mixed section ID is used for this task along with the geometrical constraint presented in equation 1 as a regularization term. However, this time it is directly applied to the encoder . The networks are finetuned using AdamW [11] optimizer, with a learning rate of 10^{-4}, weight decay of 10^{-4} and a batch size of 64.

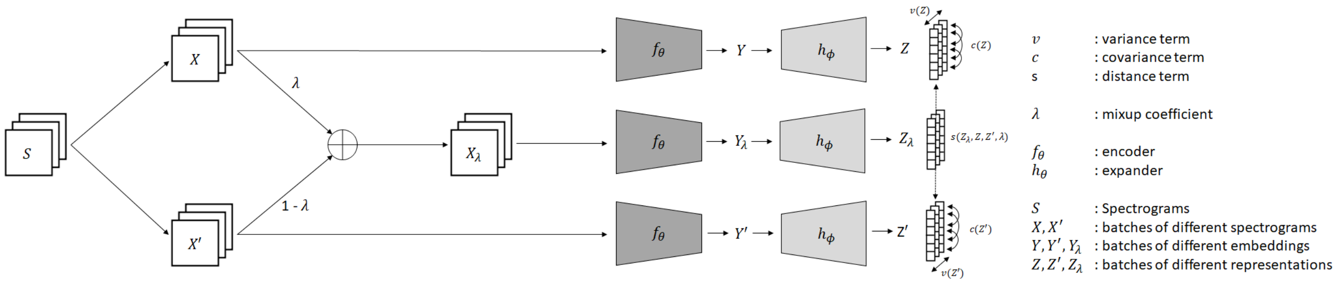

Figure 1: Self-Supervised Pre-training

2.5. Anomaly Detection

After the fine-tuning step, we apply k-Nearest Neighbors (k-NN) [12] to compute the anomaly score. We use the mean embedding vector from the 10-s audio recording as input feature to the k-NN algorithm. In this research, we use the standard Euclidean metric as the anomaly score, and the number of neighbors is set to 1. In other words, the larger the distance from the training embeddings, the more abnormal the sample is.

3. RESULTS

As instructed by the challenge organizers, in this section we only report results using the development set. In Table 2 and 3, we report the harmonic mean of AUC calculated for all seven machines and for the source and target domain respectively. In Table 4, the harmonic mean of pAUC (p = 0.1) is presented for all seven machines across all domains. The reported baseline results are taken from [1].

4. CONCLUSION

In this paper, we proposed a sound anomaly detection framework composed of a self-supervised pre-training algorithm, a supervised training phase (section ID classification), and an unsupervised phase (k-NN) for anomaly detection.

In this work, our goal was to develop a unified framework that is robust and performs well across all machine types. To achieve this goal, we used the same hyperparameters for each machine type. Experimental evaluation shows that the proposed approach significantly outperforms baseline systems.

5. SUBMISSIONS

The challenge allowed us to create up to four different submissions. We uploaded two different submissions in accordance with our goal to develop only one general method to detect anomalies.

The results on the development dataset showed that the approach presented above outperformed the baseline, except for a specific section which is section 00 for fan. For that peculiar section we thus decided to take the opposite sign for the anomaly scores. We, thus, have two submission : (1) one with the normal anomaly scores for each section and (2) one with the opposite of the anomaly scores for fan 00.

6. REFERENCES

[1] 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.

[2] 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.

[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] A. Bardes, J. Ponce, and Y. LeCun, “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning,” 2021. [Online]. Available: https://arxiv.org/abs/2105.04906

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” 2018. [Online]. Available: https://arxiv.org/abs/1801.04381

[6] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” in *NIPS-W*, 2017.

[7] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[8] A. F. Agarap, “Deep Learning using Rectified Linear Units (ReLU),” *arXiv preprint arXiv:1803.08375*, 2018.

[9] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” *CoRR*, vol. abs/1710.09412, 2017. [Online]. Available: http://arxiv.org/abs/1710.09412

[10] Y. You, I. Gitman, and B. Ginsburg, “Scaling SGD Batch Size to 32k for ImageNet Training,” *CoRR*, vol. abs/1708.03888, 2017. [Online]. Available: http://arxiv.org/abs/1708.03888

Model	ToyCar	ToyTrain	Bearing	Fan	Gearbox	Slider	Valve	total
Auto-Encoder Baseline	90.41	76.32	54.42	78.59	68.93	77.95	52.01	68.84
MobileNetV2 baseline	59.12	57.26	60.58	70.75	69.21	65.15	67.09	64.73
System 1	93.26	84.93	66.96	83.38	88.95	94.5	85.1	84.35

Table 2: Harmonic Mean of AUC in the source domain of Development Dataset (in %)

Model	ToyCar	ToyTrain	Bearing	Fan	Gearbox	Slider	Valve	total
Auto-Encoder Baseline	34.81	23.35	58.38	47.18	62.64	47.67	49.46	42.27
MobileNetV2 baseline	51.96	45.90	59.94	48.22	56.19	38.23	57.22	51.06
System 1	81.73	43.4	82.87	76.73	82.74	67.35	81.96	70.34

Table 3: Harmonic Mean of AUC in the target domain of Development Dataset (in %)

Model	ToyCar	ToyTrain	Bearing	Fan	Gearbox	Slider	Valve	total
Auto-Encoder Baseline	52.74	50.48	51.98	57.52	58.49	55.78	50.36	53.38
MobileNetV2 baseline	52.27	51.52	57.14	56.9	56.03	54.67	62.42	55.8
System 1	68.34	53.88	64.75	70.76	71.23	68.92	71.66	66.48

Table 4: Harmonic Mean of pAUC for the Development Dataset (in %)

[11] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2017. [Online]. Available: https://arxiv.org/abs/1711.05101

[12] E. Fix and J. L. Hodges, “Discriminatory analysis. nonparametric discrimination: Consistency properties,” *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.