Anomaly Sound Detector Based on Variational Autoencoder with Hyperparameter Optimization Strategy

Technical Report

Lars C. Gleichmann, Yeremia G. Adhisantoso, Alexander Lange, Quy Le Xuan

Gottfried Wilhelm Leibniz Universität Hannover
Institute for Information Processing (TNT),
Appelstraße 9a, 30167 Hanover, Germany
gleichmann@tnt.uni-hannover.de

ABSTRACT

The second task of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2024 challenge addresses the difficulties of domain adaptation and generalization in Anomalous Sound Detection (ASD). We present two types of Variational Autoencoders (VAEs) to overcome these challenges. Linear prediction coefficients provide a sparse and meaningful representation of the original raw audio clips for our models. This report also introduces two optimization strategies for setting reasonable hyperparameters for anomalous sound detectors.

Index Terms— domain shift, anomaly sound detection, (vector quantized) variational autoencoders, linear prediction coefficients, convolution

1. INTRODUCTION

Anomaly Sound Detection (ASD) is a promising area of predictive maintenance because low-cost, non-intrusive sensing enables health monitoring for a wide range of applications. The Detection and Classification of Acoustic Scenes and Events (DCASE) 2024 challenge encourages the development of anomalous sound detectors easily adaptable from one application to another. The second task of the challenge [1], entitled "First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring", introduces 16 different machine types that necessitate a well generalizable solution. The ToyADMOS2 [2] and MIMII DG [3] datasets provide the seven machine types in the development dataset. Outperforming the provided baseline model [4] is also part of the challenge in Task 2.

2. PROPOSED METHODS

2.1. Dataset

We consider only the development and evaluation dataset for the second task of the DCASE 2024 challenge [2, 3]. To speed up the training of our models, we consistently evaluate 160,000 samples from each clip. Consequently, we either crop this number of samples from the center or apply symmetric zero-padding, depending on the original number of samples.

2.2. Data Preprocessing

In each epoch of our training, we pass all training clips from the development dataset at least once. However, the target domain makes up just one percent of the original data pool. To balance this, we randomly sample clips from the target domain that are repeated several times in training. This ensures that one third of the training samples originate from the target domain. Following this compromise, we compensate for the imbalance, but also avoid overfitting on the ten clips from the target domain.

To facilitate the anomaly detection, we provide our network with meaningful Linear Prediction Coefficients (LPC). They are computed according to the algorithm proposed by Larry Marple [5] and describe the envelope of the spectrum calculated for a specified window from the audio clip.

Each clip is decomposed into windows of 2,000 audio samples described by 100 LPCs in the operation mode with fixed hyperparameters. In another operation mode, the number of coefficients and window size are tuned.

2.3. Models

We propose Variational Autoencoders (VAEs) [6] as anomaly detectors. In contrast to vanilla classification networks, VAEs are trained exclusively with normal samples of a healthy machine. Therefore, they are predestined for our task.

VAEs encode normal inputs into a compact representation. This latent embedding is seven-dimensional in our case. From this representation, the original inputs are reconstructed as best as possible by a decoder. Without anomalies, the evaluated samples are similar to the normal training samples. Therefore, the reconstruction loss is expected to be in the same range as that observed in the training data. Otherwise, an increase in the reconstruction loss indicates the presence of anomalies.

We apply two types of VAEs. The standard VAE is encouraged to map all normal inputs to a standard multivariate Gaussian distribution. Deviations from the typical acoustic fingerprint of a healthy machine, on which the encoder is trained, lead directly to deviations from the learned distribution.

The second proposed model is a Vector Quantified Variational Autoencoder (VQ-VAE) [7]. It stores multiple vectors with size of the latent dimension. For each encoding, it selects the vector closest to the embedding and uses that vector for decoding. While normal VAEs tend to suppress details to map all inputs into one distribution, the VQ-VAE clusters information into multiple centers. As a result, different details of an input are preserved in different modes or at a particular step in a sequence. The number of vectors in the codebook is set to 40 in the fixed hyperparameter mode, and a tunable number of clustering centers otherwise.

Our decoder is a mirrored architecture of our encoder. The first layer of the encoder consists of a skip connection and two independent convolutions of the current window and its neighbors with kernel sizes of 7 and 13, respectively. The filters use a dilation of three and seven. The information obtained serves as input to multiple fully connected layers, all sharing the same number of parallel neurons. In the case of fixed hyperparameters, the network consists of three hidden layers with 160 neurons in each layer. With tuned hyperparameters, the number can vary from one to five layers with up to 256 neurons each. The dimension of the embedding is seven. PReLUs act as activation functions.

2.4. Loss Functions

Our loss function for the standard VAE is a linear combination of the reconstruction error measured by the Mean Squared Error (MSE) and Kullback-Leibler (KL) divergence [8] evaluating the deviation from the target distribution:

$$L = \alpha \cdot L_{\mathrm{MSE}} + \beta \cdot L_{\mathrm{KL}}, \tag{1}$$

Where L_{MSE} represents the mean of the squared differences of the original inputs $\hat{Y}_i$ and their corresponding reconstructions Y_i:

$$L_{\mathrm{MSE}} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2. \tag{2}$$

The VQ-VAE also applies this loss to measure the differences between the vectors from its codebook and the embeddings of its encoder. For this architecture, the MSE between the nearest codebook vector and the embedding replaces L_{KL} in (1). For the VAE, the KL divergence (3) describes the deviation of the distribution of the embeddings $P(x)$ from the targeted multivariate gaussian normal standard distribution $S(x)$:

$$L_{\mathrm{KL}} = \int_{-\infty}^{+\infty} P(x) \cdot log\left(\frac{P(x)}{S(x)}\right) dx. \tag{3}$$

We refer to this as hidden loss. It should be noted that applying the Wasserstein distance [9] instead of the KL divergence had no significant impact on the results. We train all types of machines separately and use ADAM as optimizer with a learning rate of 0.001.

2.5. Anomaly Score and Detection

The k-highest hidden and reconstruction losses are selected from all windows derived from an audio clip. The mean is calculated over both subsets separately. In this way, our system can detect peaks that represent anomalies (k=1) or continuous deviations of an anormal signal when k equals the number of windows from the audio clip. In the case of fixed hyperparameters, we set k to 45. Otherwise, we optimize it for its target machine. The final anomaly score is a linear combination of both means.

Based on the score calculated on the normal samples from training, we heuristically determine that at least 95% of them shall be detected as normal. According to this rule, we choose the threshold for finally labelling the detections.

3. HYPERPARAMETER OPTIMIZATION AND RESULTS

The results displayed in Table 1 to Table 6 demonstrate the significant impact of the hyperparameter settings of our input representations and models on the resulting AUROCs. We use Optuna [10] for hyperparameter tuning.

Table 1 displays the optimal results for the VAE architecture, while Table 4 shows the optimal results for the VQ-VAE architecture. It is possible to reach these results by tuning the hyperparameters for each machine individually directly on the AUROC, which is calculable for the development dataset. However, the AUROC is not calculable on the evaluation dataset due to the lack of anomaly labels. The optimal hyperparameters vary considerably depending on the machine, making it challenging to identify common ones.

To meet the requirement of generalization, we determined feasible ranges for each machine and hyperparameter combination by considering not only the best configuration but also those leading to AUROCs that are up to 10% smaller. For the final evaluation, we determined the general hyperparameters so that they fall into the ranges of all machines from the development dataset. As anticipated, the results for the machines from the development dataset show a decline, as illustrated in Table 2 and Table 5 for the VAE and VQ-VAE, respectively.

For this reason, we propose an alternative approach to optimizing the hyperparameters for each machine in the evaluation dataset individually. While the AUROC is not directly computable for this evaluation data, the use of three alternative objectives allows for the maximization of the unknown AUROC.

Firstly, a suitable detector must exhibit a minimal average reconstruction loss for the normal training samples. The second objective is to ensure that the distribution of normal samples from training and the distribution of the normal test samples from the evaluation set, measured by the difference of their means, are as similar as possible. Once the optimal solution has been identified for the first two objectives, the third objective is to maximize the average distance between the normal and the anormal distribution of the evaluation dataset.

However, the evaluation dataset does not allow for the division of normal and abnormal samples due to missing labels. Therefore, a heuristic is required. Based on the distribution of reconstruction losses after each epoch, the 30% of samples with the lowest reconstruction loss are expected to be normal, while the 30% with the highest reconstruction loss are expected to be anomalous. This heuristic allows us to calculate the second and third objectives. Finally, we determine the optimal hyperparameters for each machine by optimizing the three defined objectives. Table 3 and Table 6 display the results for the development dataset.

4. SUBMISSION STRUCTURE

For solving the second task of the DCASE 2024 challenge, we apply the two models and two hyperparameter optimization strategies presented in this report. For comparison, we have submitted all four possible combinations of the aforementioned approaches:

1) VAE with KL divergence and optimized hyperparameters
2) VAE with KL divergence and fixed hyperparameters
3) VQ-VAE with optimized hyperparameters
4) VQ-VAE with fixed hyperparameters

5. CONCLUSION

VAEs and VQ-VAEs can learn the acoustic fingerprint of a machine and detect deviations from it as anomalies. However, they also tend to suppress low power signal components, even though these portions of the spectrum indicate the anomaly of interest. Therefore, it is required to choose hyperparameters carefully. Estimating a common configuration for multiple machines can only be a compromise. However, if assumptions about the evaluation data set are possible, the hyperparameters are tunable for a specific machine.

The objectives for this optimization proposed in this paper are reasonable and potentially lead to superior results. However, they are not sufficient to reliably discriminate between models that provide an excellent AUROC and those that do not. Therefore, further research is needed to identify a reliable function for estimating the AUROC of a trained model when it is not computable due to missing target labels.

6. REFERENCES

[1] Tomoya Nishida, Noboru Harada, Daisuke Niizumi, Davide Albertini, Roberto Sannino, Simone Pradolini, Filippo Augusti, Keisuke Imoto, Kota Dohi, Harsh Purohit, Takashi Endo and Yohei Kawaguchi. “Description and Discussion on DCASE 2024 Challenge Task 2: First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring” in *arXiv e-prints: 2406.07250*, 2024.

[2] Noboru Harada, Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Masahiro Yasuda and Shoichiro Saito. “ToyADMOS2: Another Dataset of Miniature-Machine Operating Sounds for Anomalous Sound Detection under Domain Shift Conditions” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1-5.

[3] Kota Dohi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, Masaaki Yamamoto, Yuki Nikaido and Yohei Kawaguchi. “MIMII DG: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection for Domain Generalization Task” in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, 2022.

[4] Noboru Harada, Daisuke Niizumi, Yasunori Ohishi, Daiki Takeuchi and Masahiro Yasuda. “First-Shot Anomaly Sound Detection for Machine Condition Monitoring: A Domain Generalization Baseline” in *2023 31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 191-195.

[5] Larry Marple. “A New Autoregressive Spectrum Analysis Algorithm” in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol 28, no. 4, 1980.

[6] Diederik P Kingma and Max Welling. “Auto-Encoding Variational Bayes” in *arXiv e-prints: 1312.6114*, 2013.

[7] Oord, A.v.d.; Vinyals, O.; Kavukcuoglu, K. “Neural Discrete Representation Learning” in *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS)*, Long Beach Convention Center, Long Beach, CA, USA, 4–9 December 2018.

[8] I. Csiszar. “I-Divergence Geometry of Probability Distributions and Minimization Problems” in *Ann. Probab. 3*, February 1975, pp. 146-158.

[9] R.L. Dobrushin. “Prescribing a system of random variables by conditional distributions” in *Theor. Prob. Appl.*, 1970, pp. 458-486.

[10] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. “Optuna: A Next-generation Hyperparameter Optimization Framework” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Results for KL-VAE with Best Hyperparameters

Machine	Source AUROC	Target AUROC	pAUROC
Bearing	72%	66.44%	55.32%
Fan	69.32%	80.36%	56.32%
Gearbox	70.76%	72.28%	53.95%
Slider	98.8%	92.12%	91.26%
Toy Car	59.36%	59.48%	53.58%
Toy Train	86.88%	57.44%	55.47%
Valve	69.96%	79.92%	59.37%

Table 1 - Results for the VAE trained with KL loss with best hyperparameters optimized directly for best AUROC.

Results for KL-VAE with Fixed Hyperparameters

Machine	Source AUROC	Target AUROC	pAUROC
Bearing	59.4%	45.525	52.05%
Fan	45.84%	55.44%	49.37%
Gearbox	63.08%	56.84%	52.79%
Slider	73.88%	59.2%	55.74%
Toy Car	62.24%	48.48%	52.32%
Toy Train	69.48%	41.04%	53%
Valve	68.04%	53.76%	50.26%

Table 2 - Results for the VAE trained with KL loss with fixed and generalized hyperparameters.

Results for KL-VAE with Tuned Hyperparameters

Machine	Source AUROC	Target AUROC	pAUROC
Bearing	67.48%	56.84%	52.32%
Fan	46.32%	62.8%	51.74%
Gearbox	58.2%	60.56%	53.16%
Slider	80.12%	66.72%	57.74%
Toy Car	51.8%	62.88%	50.05%
Toy Train	77.8%	41.32%	52.68%
Valve	64.56%	52.4%	49.95%

Table 3 - Results for the VAE trained with KL loss with tuned hyperparameters according to the reconstruction losses in training and the differences of the anormal and normal distributions.

Results for VQ-VAE with Best Hyperparameters

Machine	Source AUROC	Target AUROC	pAUROC
Bearing	75.04%	64.44%	55.74%
Fan	62.84%	68.84%	50.79%
Gearbox	81.44%	69%	51.26%
Slider	98.12%	96.12%	90.58%
Toy Car	59.16%	54.8%	55.53%
Toy Train	85.96%	58.2%	53.95%
Valve	62.72%	64.4%	51.95%

Table 4 - Results for the VQ-VAE with best hyperparameters optimized directly for best AUROC.

Results for VQ-VAE with Fixed Hyperparameters

Machine	Source AUROC	Target AUROC	pAUROC
Bearing	67.24%	55.04%	54.79%
Fan	46.56%	60.4%	48.89%
Gearbox	58.52%	61.08%	52.94%
Slider	66.04%	51.84%	50.79%
Toy Car	58.51%	48.88%	53.16%
Toy Train	67.76%	43.12%	52.89%
Valve	75.6%	56.64%	51.16%

Table 5 - Results for the VQ-VAE with fixed and generalized hyperparameters.

Results for VQ-VAE with Tuned Hyperparameters

Machine	Source AUROC	Target AUROC	pAUROC
Bearing	73.2%	56.92%	53.26%
Fan	55.84%	53.08%	51%
Gearbox	68.56%	62.96%	51.63%
Slider	88.88%	79.12%	68.05%
Toy Car	51.6%	63.96%	58.16%
Toy Train	72.64%	44.4%	53.42%
Valve	67.16%	60.36%	52.89%

Table 6 - Results for the VQ-VAE with tuned hyperparameters according to the reconstruction losses in training and the differences of the anormal and normal distributions.