AN ENSEMBLE METHOD FOR UNSUPERVISED ANOMALOUS SOUND DETECTION

Technical Report

Qinwen Hu, Kai Chen, Jing Lu

Key Laboratory of Modern Acoustics, Nanjing University,
Nanjing, 210008, China
qinwen.hu@smail.nju.edu.cn, chenkai@nju.edu.cn,
lujing@nju.edu.cn

ABSTRACT

This report describes our submitted system for DCASE2022 challenge task2 (Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques) [1] in detail. The system is composed of two modules, a hierarchical recurrent variational autoencoder and a self-supervised classifier, and the final score is a weighted average over the normalized results of the two systems. The anomaly scores are all calculated in the latent/embedding space.

Index Terms— DCASE, unsupervised anomalous sound detection, representation learning

1. INTRODUCTION

The DCASE2022-task2 [1][2][3] is an unsupervised anomalous sound detection competition. Participants are expected to design a system to discriminate between sounds emitted by normal machines and unhealth machines. The main challenges lie in three aspects: (a) Only normal sounds are provided in the training stage, putting the task in an unsupervised manner; (b) Half of the test data are recorded in different acoustic conditions from training data, so the system faces a domain shift challenge; (c) In the test stage, the domain information of the test data is not given, which is different from DCASE2021-task2 [4], and significantly more difficult.

In DCASE2021-task2[4], models used by the participating teams [5][6][7][8][9][10] include self-supervised classifiers, autoencoders or variational autoencoders (VAE), and normalizing flows, etc. In the classifier-based methods, the output of the Softmax layer or the embedding distances are usually calculated as anomaly scores. Reconstruction errors are used in autoencoders. The normalizing flows are directly used as density estimators. Many participating teams used ensemble systems to balance the models' preferences to further improve the performance.

Our proposed system is an ensemble of two modules, including a hierarchical recurrent variational autoencoder, abbreviated as HRVAE, and a classifier. HRVAE is trained in two stages to extract more meaningful embeddings of the machine sounds. In both modules, the anomaly detection is conducted in the embedding space. The final score is a weighted sum of the normalized scores from the two modules.

2. METHODS

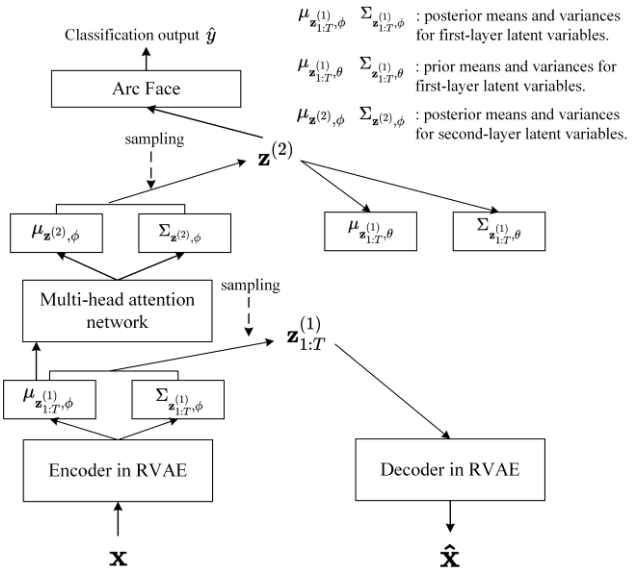

Figure 1: The architecture of HRVAE.

2.1. HRVAE-based method

The VAE [11] is a generative model assuming that the high-dimensional data can be generated from the low-dimensional latent variables. The aim of VAE is to learn the latent representation and generate data based on the learned representation, using a pair of encoder and decoder. The encoder inferences the latent variables $\boldsymbol{z}$ of the given data $\boldsymbol{x}$, and the decoder reconstructs the data $\boldsymbol{x}$ based on $\boldsymbol{z}$. VAE assumes that the latent variables follow a simple prior distribution $p(\boldsymbol{z})$ in the latent space, usually $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Given the conditional likelihood $p_\theta(\boldsymbol{x}|\boldsymbol{z})$, which is usually complex proper Gaussian distribution when the data is short time Fourier transform (STFT) coefficients $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}_c\big(\boldsymbol{x}; \mu(\boldsymbol{z}), \boldsymbol{\Sigma}(\boldsymbol{z})\big)$, VAE uses a multivariate Gaussian distribution as the variational distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{z}; \mu(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x})\big)$ to approximate the posterior distribution, and the encoder and decoder are jointly trained to maximize the evidence lower bound (ELBO) as

$$\mathcal{L}_{ELBO} = \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[logp(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})\middle\|p(\boldsymbol{z})\right), \quad (1)$$

where D_{KL} means the Kullback-Leibler (KL) divergence.

Recurrent variational autoencoder (RVAE) [12] induces a posterior dynamic over the latent variables, and can better model the temporal relationships of both input features and latent

variables. The proposed HRVAE has a two-layer hierarchy, which adds a second stochastic layer on the RVAE to extract latent variables on different scales, depicted as $\boldsymbol{z}^{(1)}$ and $\boldsymbol{z}^{(2)}$ in Fig. 1. The input feature is a *T*-frame STFT spectrogram, and then the first layer of HRVAE extracts the latent variables $\boldsymbol{z}_{1:T}^{(1)}$ for each single frame, and the second layer extracts the latent variable $\boldsymbol{z}^{(2)}$ for the whole *T*-frame segment. HRVAE introduces a conditional prior distribution over the latent variables $p_\theta(\boldsymbol{z}_{1:T}^{(1)}|\boldsymbol{z}^{(2)})$, and assumes that the second layer latent variables $\boldsymbol{z}^{(2)}$ still follow a standard normal distribution.

The loss of HRVAE can be described as

$$\begin{aligned}\mathcal{L}_{ELBO} = \mathbb{E}_{\boldsymbol{z}_{1:T}^{(1)},\boldsymbol{z}^{(2)}\sim q_\phi^{(1)},q_\phi^{(2)}}\left[logp(\boldsymbol{x}|\boldsymbol{z}_{1:T}^{(1)},\boldsymbol{z}^{(2)})\right] \\ -D_{KL}\left(q_\phi^{(1)}(\boldsymbol{z}_{1:T}^{(1)}|\boldsymbol{x})\middle\| p(\boldsymbol{z}_{1:T}^{(1)}|\boldsymbol{z}^{(2)})\right) \\ -D_{KL}\left(q_\phi^{(2)}(\boldsymbol{z}^{(2)}|\boldsymbol{x},\boldsymbol{z}_{1:T}^{(1)})\middle\| p(\boldsymbol{z}^{(2)})\right),\end{aligned} \quad (2)$$

where $q_\phi^{(1)}$ and $q_\phi^{(2)}$ denotes the posterior distribution of $\boldsymbol{z}_{1:T}^{(1)}$ and $\boldsymbol{z}^{(2)}$, respectively.

To force the HRVAE to learn the features of normal machines instead of other varying factors including environmental noises, the HRVAE is then re-trained on an auxiliary classification task to discriminate the auxiliary information, for example the machine IDs and the operation velocities. The loss is modified as

$$\mathcal{L} = \mathcal{L}_{ELBO} + \alpha\mathcal{L}_{auxi}, \quad (3)$$

where α is a hyperparameter.

The RVAE encoder and decoder structures are the same as those in [12], with the output dimension of 128 for all the layers except the output layer of the encoder and the decoder. The second layer encoder is a network with two multi-head attention layers [13]. The number of attention heads is 2. The classification task is trained with an ArcFace loss [14].

In the test stage, the anomaly score is calculated based on the mean of cosine distances of the *K*-nearest-neighbours (KNN) in the latent space. The number of neighbours *K* is set to 1.

2.2. Classifier-based method

STgram-MFN[15] is used as the classifier backbone in our system, which combines the information in the raw wav signals with that in the log-Mel spectrograms. The detailed structure of the STgram-MFN is the same as that in [15]. Two ArcFace losses are applied in training the classifier. One is to classify each section, and the other is to distinguish different auxiliary information. The output before the ArcFace layers are seen as embeddings.

During the testing stage, the cosine distances between the test data embedding and the centres of different sub-classes in a single section are calculated, and the negative of the minimum distance is used as the anomaly score.

3. EXPERIMENTS AND RESULTS

3.1. Training Configurations

All audios are used at a sampling rate of 16 kHz. When training HRVAE, 50-frame STFT spectrograms are used as the input. The FFT length is 1024, and the hop length is 512, so the input feature has a size of 50×513. The latent vector dimension is 16 for $\boldsymbol{z}^{(1)}$, and 64 for $\boldsymbol{z}^{(2)}$.

For STgram-MFN, the STFT configuration is the same as that for HRVAE, and the number of Mel-frequency bins is 128. The whole spectrogram is fed into the network, and each input feature of a 10 s long audio clip has a size of 313×128.

HRVAE is trained for 500 epochs with an early stopping patience of 10, and a batch size of 256. The hyperparameter α is set to 10.

The STgram-MFN is trained for 200 epochs with an early stopping patience of 10, and a batch size of 64. To avoid overfitting and improve the model's performance, Mixup [16] is used in training with the two ArcFace losses.

We train a separate model on each machine type, using all 6 sections in a machine type.

The four submitted systems use different weights on the scores of two modules.

3.2. Results

Table 1 shows the scores of the best submitted system compared with two baseline models on the development dataset, including area under the receiver operating characteristic curve (AUC) and partial-AUC (pAUC) scores. The scores of each machine type are harmonic mean over test data from section_00, section_01, and section_02. On machine type ToyCar, ToyTrain, Slider, Gearbox, and Fan, our proposed system achieves significantly better performance over both challenge baselines. The performance on Bearing and Valve is worse than Baseline2. Note that the performance gap between the source domain and the target domain is not very large, and is even reversed on Bearing and Valve.

4. CONCLUSIONS

A system combining two anomaly detection modules, HRVAE and a STgram-MFN, is proposed. The anomaly scores are calculated in the embedding space based on distances. The proposed system achieves significantly better average performance over the baseline models.

5. REFERENCES

[1] Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Takashi Endo, Masaaki Yamamoto, and Yohei 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] 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 arXiv e-prints: 2205.13879*, 2022.E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography*, London, UK: Academic Press, 1999.

[3] 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 6th Detection and Classification of

Acoustic Scenes and Events 2021 Workshop (DCASE2021), 1–5. Barcelona, Spain, November 2021.

[4] Kawaguchi, Yohei, et al. "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions." *arXiv preprint arXiv:2106.04492* (2021).

[5] Ontiveros, Juan Del Hoyo, and Hector Courdourier. "ENSEMBLE OF COMPLEMENTARY ANOMALY DETECTORS UNDER DOMAIN SHIFTED CONDITIONS Technical Report Jose A. Lopez, Georg Stemmer, Paulo Lopez-Meyer, Pradyumna S. Singh."

[6] Morita, Kazuki, Tomohiko Yano, and Khai Tran. "Anomalous sound detection using CNN-based features by self supervised learning." *Tech. Rep., DCASE2021 Challenge* (2021).

[7] Wilkinghoff, Kevin. "Utilizing sub-cluster AdaCos for anomalous sound detection under domain shifted conditions." *Tech. Rep., DCASE2021 Challenge* (2021).

[8] Kuroyanagi, Ibuki, et al. *Anomalous sound detection with ensemble of autoencoder and binary classification approaches*. DCASE2021 Challenge, Tech. Rep, 2021.

[9] Sakamoto, Yuya, and Naoya Miyamoto. "Combine Mahalanobis distance, interpolation auto encoder and classification approach for anomaly detection." *In other words* 10 (2021): 4.

[10] Zhou, Qiping. *Ensemble of ArcFace based systems for unsupervised anomalous sound detection under domain shift conditions*. DCASE2021 Challenge, Tech. Rep, 2021.

[11] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114* (2013).

[12] Leglaive, Simon, et al. "A recurrent variational autoencoder for speech enhancement." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[13] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[14] Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[15] Liu, Youde, et al. "Anomalous Sound Detection Using Spectral-Temporal Information Fusion." *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[16] Zhang, Hongyi, et al. "mixup: Beyond empirical risk minimization." *arXiv preprint arXiv:1710.09412* (2017).

Table 1: Performances on the development dataset of baseline systems and the proposed system. Total score for each machine type means the harmonic mean over AUC (source), AUC (target), and pAUC. 'All' means the harmonic mean over these three scores from all the machine types.

System	Score	ToyCar	ToyTrain	Slider	Gearbox	Bearing	Valve	Fan	All
Baseline1 (AE)	AUC(source)(%)	90.41	76.32	77.95	68.93	54.42	52.01	78.59	
	AUC(target) (%)	34.81	23.35	47.67	62.64	58.38	49.46	47.18	
	pAUC(%)	52.74	50.48	55.78	58.49	51.98	50.36	57.52	
	Total(%)	51.06	39.61	57.99	63.07	54.80	50.58	58.47	52.61
Baseline2 (MobileNetV2)	AUC(source)(%)	59.12	57.26	65.15	69.21	60.58	67.09	70.75	
	AUC(target) (%)	51.96	45.90	38.23	56.19	59.94	57.22	48.22	
	pAUC(%)	52.27	51.52	54.67	56.03	57.14	62.42	56.90	
	Total(%)	54.26	51.14	50.17	59.89	**59.18**	**61.98**	57.20	55.94
Best submitted system (system_1)	AUC(source)(%)	81.72	70.07	84.77	72.32	53.37	58.63	60.01	
	AUC(target) (%)	70.83	61.04	77.45	68.66	61.60	66.90	59.95	
	pAUC(%)	55.98	53.96	64.49	60.50	55.01	57.83	56.22	
	Total(%)	**67.85**	**61.12**	**74.60**	**66.78**	56.44	58.67	**60.85**	**63.25**