SELF-SUPERVISED LEARNING METHODS USING ST-GRAM FOR ANOMALY MACHINE SOUND DETECTION

Technical Report

Wonki Cho

Department of Computer Engineering, Sogang University, Republic of Korea

ABSTRACT

It is difficult to apply supervised learning to anomaly detection due to absence of abnormal data. Therefore, in anomaly detection, Therefore, we use unsupervised anomaly detection, which assumes that most of the data is a normal sample and learns without obtaining a label. In this paper, A self-supervised learning method is proposed for unsupervised anomaly detection. This network performs self-supervised classification using metadata associated with the audio files and compare to labeled normal and abnormal. To better extract the characteristics of the machine sounds, We adopt ST-gram for Spectral-temporal feature fusion and compared performance with some CNN Networks for DCASE 2022 Challenge task2: Unsupervised Anomaly Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques. As a result, Our method wasn't performed well except 'Slider' machine sounds.

Index Terms— Anomaly Sound Detection, ST-gram, Self-Supervised Learning

1. INTRODUCTION

Anomaly Detection is important task in diverse applications, because there aren't exits enough abnormal samples and hard to labeling data for classifying normal or abnormal. Basically, Unsupervised Anomaly Detection System based Auto-Encoder are widely used. Auto-Encoder can learn important features by compressing input data and reconstruct data from decoder network. but reconstruction based method such as Auto-Encoder, GANS[1] are unstable and very sensitive to model hyperparameters. Other method to use in unsupervised anomaly detection is the representation learning method[2]. The performance of machine learning models depends on the representation of the data being learned. By using representation learning method, We can learn various features on normal data intensively and learn with a model optimized for distribution on normal data.

Anomaly detection system is widely used in the manufacturing task, and DCASE 2022 task 2 challenge[3] focuses on detecting abnormalities in machine sound. And just for the DCASE challenge definition, Anomalous sound detection is the task of identifying whether the sound emitted from a target machine is normal or anomalous. In real-world factories, anomalies rarely occur and are highly diverse. Therefore, exhaustive patterns of anomalous sounds are impossible to create or collect and unknown anomalous sounds that were not observed in the given training data must be detected. Additionally, Domain generalization[4] is an important part of this year's challenge. Each data can have different attributes and domains, ideal detection performance should be shown even at the same threshold when designing an anomaly detection system.

In order to solve this challenge, We would like to use a self-supervised learning technique using Spatio-Temporal Mel spectrogram. Log-mel spectrograms are generally used as input data in deep learning-related sound tasks. Through the Mel Filter Bank, we can see how much energy is around the low frequency in the spectrogram, and the disadvantage is that the width of the filter increases as the frequency goes up, and the high frequency is rarely considered. Therefore, it is intended to sufficiently reflect the high frequency components of machine sound through ST-gram[5]. Since the unsupervised learning method does not have a label, the performance of learning data features is not significantly better than that of having a label, so we would like to approach it as a self-supervised learning method. Additionally, Mixup[6] for data augmentations is used to lead domain generalization between source and target data.

2. PROPOSED METHODS

2.1. Spectral-temporal feature fusion networks

Usually, Mel spectrograms are used as input feature data in Audio and Sound tasks. However, The mel spectrogram reflects more of the lower frequency values by filtering out high-frequency components. It is difficult to sufficiently reflect the high frequency components of machine sounds with mel spectrogram, So we adopt ST-gram, Spectral-temporal feature gram as input for anomaly sound detection. ST-gram is created by combining a log mel spectrogram containing spectral elements and another spectrogram having temporal elements. To compensate for the missing anomaly information from the log-Mel spectrogram, we apply a CNN-based network to extract the temporal feature from the raw input signal.

2.2. MixUp Augmentations

Mixup is a technique of data augmentations that can create new data by mixing two data and labels at a certain rate. This method shows that not only improves the performance of existing single-label classification, but also significantly improves the performance in multi-label classification.

$$\widetilde{x} = \lambda x_i + (1 - \lambda) x_j$$
$$\widetilde{y} = \lambda y_i + (1 - \lambda) y_j$$

X means raw input datas and Y is label values for x. By mixing two input datas and labels, it can get a normalization effect by preventing overfitting. In Anomaly detection, We can only use normal data, so we have an label, normal. For domain generalization in

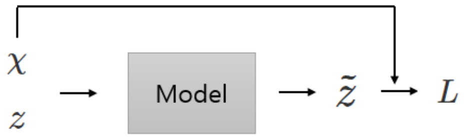

Figure 1: Overall process of Self Supervised Learning

Anomaly detection, Our model have to show similar performance at the same decision threshold value for source data and target data.

2.3. Self Supervised Learning

Basically, Self-Supervised learning use psuedo labels from pretext task about unlabeled input data. Through Self-Supervised learing, Our model can learn feature representations without annotated datasets. In Figure 1, psuedo label Z is from pretext task ouputs. But, We create different kinds of labels from metadatas[7]. In DCASE2022 task2, Datasets are consist of six sections(00, 01, 02, 03, 04, 05, 06), each with different characteristics. We use the section value as a label for self supervised method. And with the one-class classification[8] concept, we would like to compare the performance by using only normal and abnormal labels.

2.4. Arcface

Arcface[9] is a loss function that emerged to increase performance in facial recognition tasks, but it is widely used in many challenges to increase performance. Using additive angular margin penalty, ArcFace lead our model in the direction of learning well feature representation that has high intra-class similarity and low inter-class similarity, leading to find an accurate decision boundary between normal and abnormal sounds.

$$L = -\frac{1}{N}\sum_{i=1}^{N} log(\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))} + \sum_{j=1,j\neq y+i}^{n} e^{s\cos(\theta_j)}})$$

3. EXPERIMENTS

About DCASE 2022 Task 2 (Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques) datasets, it has consist of development, additional training, evalauation sets. It has seven classes(i.e.,Fan, Bearing, Gearbox, Slider, Toycar, ToyTrain, Valve) and consists of task2 datasets, combining MIMII DG[10] and ToyADMOS2[11]. and each data is 10 seconds long. Since the training set consists of only normal data, how to learn the feature of normal sound is important. And the main focus of this challenge is domain generalization techniques that mainly use the source domain data to learn common features across different domains so that the model can generalize to both the source and target domain in the test data.

3.1. Experiments Setup

For this task, we train models separately for different machines, and Adam is selected as the model optimizer. To calculate the Log-Mel spectrogram, the frame size is set to 1024, the hop size is set to 512, and the number of Mel filter banks is set to 128. As train parameters, we train the network for 200 epochs of every mechaine type, and the learning rate is initialized as 0.0001. For Arcface, margin is set to 0.7 and scale value is set to 30.

3.2. Evaluation Metrics

DCASE 2022 task2 is evaluated with the AUC and the pAUC. The pAUC is an AUC calculated from a portion of the ROC curve over the pre-specified range of interest. The anomaly detector is expected to work with the same threshold for domain generalization.

$$\mathrm{AUC}_{m,n,d} = \frac{1}{N_d^- N_n^+}\sum_{i=1}^{N_d^-}\sum_{j=1}^{N_n^+}\mathcal{H}\left(\mathcal{A}_\theta\left(x_j^+\right) - \mathcal{A}_\theta\left(x_i^-\right)\right),$$

$$\mathrm{pAUC}_{m,n} = \frac{1}{\lfloor pN_n^- \rfloor N_n^+}\sum_{i=1}^{\lfloor pN_n^- \rfloor}\sum_{j=1}^{N_n^+}\mathcal{H}\left(\mathcal{A}_\theta\left(x_j^+\right) - \mathcal{A}_\theta\left(x_i^-\right)\right)$$

AUC stands for ”Area under the ROC Curve.” That is, AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). And, pAUC is calculated as the AUC over a low false-positive-rate (FPR) range [0,p]. The reason for the additional use of the pAUC is based on practical requirements. In task 2, we will use p value as 0.1 for pAUC.

3.3. Experiment Results

Table 1 shows a comparison between the baseline and the results proposed in this report. We used MobileNetV3[12], MobileFaceNet[13] for compare with baseline results. The performance of learning through labels made using section values resulted in similar results to the baseline in addition to the performance on Slider. In the case of MobileNetV3, learning was conducted through a pre-trained model in Pytorch Image Models[14], and on average, it was confirmed that the results were the worst.

4. CONCLUSION

In this paper, we tried a self-supervised anomaly sound detection method to handle this challenge. For Sound Acoustic features, We attempted to reflect more of the characteristics of machine sound through ST-gram, which is made by combining the temperature information extracted using CNN and the spectral information in the spectrum. When comparing the performance of each class, it was not an effective method to detect abnormalities only with the proposed method, and using machine condition as a label like the baseline system did not significantly differ in performance. However, the performance of the slider class has improved a lot. The proposed method exploits complementary spectral-temporal information from the normal sound, and for domain generalization, self-supervised learning was performed using section information as a label.

5. REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.

Table 1: Performance comparison(Harmonic Mean) on Development datasets

	ToyCar			ToyTrain			Bearing			Fan			Gearbox			Slider			Valve		
Method	**AUC(s)**	**AUC(t)**	**pAUC**	**AUC(s)**	**AUC(t)**	**pAUC**	**AUC(s)**	**AUC(t)**	**pAUC**	**AUC(s)**	**AUC(t)**	**pAUC**	**AUC(s)**	**AUC(t)**	**pAUC**	**AUC(s)**	**AUC(t)**	**pAUC**	**AUC(s)**	**AUC(t)**	**pAUC**
MobileNetV2(Baseline)	59.12	51.96	52.27	57.26	45.90	51.52	60.58	59.94	57.14	70.75	48.22	56.9	69.21	56.19	56.03	65.15	38.23	54.67	67.09	57.22	62.42
ST-gram + MobileNetV3	49.45	52.90	53.07	62.64	44.29	51.17	56.76	47.81	54.85	53.88	45.31	54.53	64.20	58.79	56.15	59.39	53.07	52.44	58.34	58.02	53.11
ST-gram + MobileFaceNet	40.35	60.58	50.31	51.33	54.42	50.95	64.55	49.35	55.79	69.89	44.18	53.14	65.50	55.94	53.28	89.23	62.03	63.06	66.92	58.47	58.50

[2] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[3] 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.

[4] K. Muandet, D. Balduzzi, and B. Schölkopf, “Domain generalization via invariant feature representation,” in *International Conference on Machine Learning*. PMLR, 2013, pp. 10–18.

[5] Y. Liu, J. Guan, Q. Zhu, and W. Wang, “Anomalous sound detection using spectral-temporal information fusion,” *arXiv preprint arXiv:2201.05510*, 2022.

[6] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.

[7] 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,” *Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2020 Challenge), Tech. Rep*, 2020.

[8] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, “Deep one-class classification,” in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.

[9] 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.

[10] 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.

[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] B. Koonce, “Mobilenetv3,” in *Convolutional Neural Networks with Swift for Tensorflow*. Springer, 2021, pp. 125–144.

[13] S. Chen, Y. Liu, X. Gao, and Z. Han, “Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices,” in *Chinese Conference on Biometric Recognition*. Springer, 2018, pp. 428–438.

[14] R. Wightman, “Pytorch image models,” https://github.com/rwightman/pytorch-image-models, 2019.