A CONTRASTIVE SEMI-SUPERVISED LEARNING FRAMEWORK FOR ANOMALY SOUND DETECTION

Xinyu Cai[1,2], Heinrich Dinkel[2], Zhiyong Yan[2], Yongqing Wang[2], Junbo Zhang[2], Zhiyong Wu[1], Yujun Wang[2]

[1]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Xiaomi Corporation, Beijing, China
caixy19@mails.tsinghua.edu.cn
{dinkelheinrich,yanzhiyong, zhangjunbo1,wangyongqing3,wangyujun}@xiaomi.com
zywu@sz.tsinghua.edu.cn

ABSTRACT

Anomaly Sound Detection (ASD) is a popular topic in deep learning and has attracted the attention of numerous researchers due to its practical applications within the industry. In the case of unsupervised conditions, how to better discover the inherent consistency of normal sound clips has become a key issue in ASD. In this paper, we propose a novel training framework that jointly trains two different feature extractors using contrastive loss to obtain a better representation of normal sounds in the latent space. We evaluate our framework on the development dataset of DCASE 2021 challenge task 2. Our framework is a combination of two baseline systems from the challenge: 1) An AutoEncoder-based model and 2) a MobileNetV2-based model. Our approach trains two models, whereas during inference only model 2) is used. Experimental results indicate that the MobileNetV2-based model trained under our proposed training framework exceeds the baseline model in terms of the official score metric. Since we participated in the challenge and submitted the system trained on the proposed framework with some data augmentation methods, we also analyze the results of DCASE 2021 challenge task 2 and discuss the effect of the median filter as a data augmentation technique. Notably, our proposed approach achieves the first place for anomaly detection for the machine type "Fan" with an AUC of 90.68 and a pAUC of 79.99.

Index Terms— Unsupervised anomaly sound detection, autoencoder, convolutional nerual network, contrastive learning

1. INTRODUCTION

Anomaly sound detection (ASD) is the task of identifying whether the sound emitted from an object is normal or anomalous. It has a wide range of applications, such as machine condition monitoring and home monitoring.

In this paper, we focus on ASD in an unsupervised setting, which means that only normal (positive) sound samples can be accessed during the training phase, while during evaluation abnormal (negative) samples need to be ascertained. These settings commonly occur in real-world scenarios, where diverse anomalous sounds rarely occur. Therefore, collecting a dataset that contains exhaustive anomalous patterns is hard.

The main idea of unsupervised ASD is to learn the inherent consistency of the normal sounds, and then classify samples as anomalous or normal by the deviation of a sample from normal sound properties. Early researchers adopted statistic-based methods such as Hidden Markov Model [1] (HMM) and Gaussian Mixture Model [2](GMM) to model the probability distribution of normal sound. Anomalous sounds are usually outside of the normal sound distribution, thus we can determine whether the sound is abnormal by its posterior probability. Other researchers used generative models such as Non-negative Matrix Factorization [3] (NMF) and Autoencoder approaches [4]. These models are trained to compress and reconstruct normal sounds to learn a normal sound's properties in latent space. If an abnormal sample is fed into a generative model, the model will likely produce large reconstruction errors, meaning that the sample has not been seen during training and thus is abnormal.

Recently in the DCASE challenges, the classifier-based method showed promising performance [5, 6, 7]. Supervised training is made possible since the challenge training data is composed of normal sounds from different operating conditions with different section IDs. Classifier based ASD method uses the section ID as a label and then performs classification on latent features. Since we have access to the section ID during inference, a classifier could perform anomaly sound detection by identifying misclassified samples (wrong section ID) as anomaly sounds.

As we can see from previous works, for deep learning based anomaly sound detection methods, a key issue to improve the performance is to obtain better latent space features of normal sounds, both for the widely used Autoencoder method and classifier-based method. Inspired by the recent success of contrastive learning approaches for self-supervised audio pretraining [8, 9, 10], we aim to enhance the model's capability to detect unseen events by linking multiple views together. Our proposed learning framework is a novel combination of two mainstream anomaly detection models trained with an additional contrastive loss function.

The paper is structured as follows: In Section 2 we introduce our proposed learning framework and its components. Further, in Section 3 details regarding the dataset and experimental setup are provided. Results are provided in Section 4 and the conclusion is given in Section 5.

2. PROPOSED APPROACH

During the training phase, our approach jointly trains two individual models: an unsupervised AE-based model combined with a supervised convolutional neural network (CNN). Once the loss converges, inference can be performed using either model independently. The architecture can be seen in Figure 1.

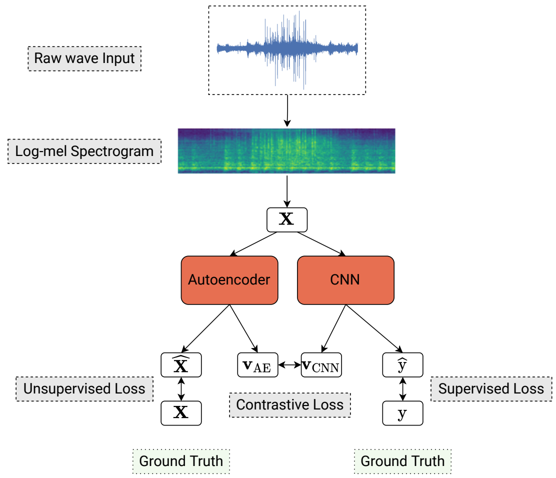

Figure 1: The proposed architecture used in this work. A spectrogram feature is first extracted from the input waveform. Then the feature is fed into two separate models: an Autoencoder (AE) and a Convolutional neural network (mainly MobileNetV2). The models are jointly optimized to reconstruct the input spectrogram, predict the section label and minimize the contrastive loss between the two models' hidden representations.

2.1. Autoencoder-based unsupervised classification

Our AE baseline model is a fully connected neural network with a bottleneck structure and trained to reconstruct a given input sound (normal sound). Ideally, a well-trained AE will produce a low error if a new data sample has been seen during the training phase (normal sample) and a large error when it encounters unseen anomalous sounds.

Formally, let x be an input sample and AE be the autoencoder, our training objective follows:

$$\begin{aligned} \mathrm{AE}(x) &\mapsto \hat{x}, \\ \mathcal{L}_{\text{unsup}}(\cdot) = \mathcal{L}_{\text{AE}}(x) &= \mathcal{L}_{\text{MSE}}(\hat{x} - x), \end{aligned} \tag{1}$$

where the training loss is chosen to be the mean square error (MSE).

2.2. MobileNet-based supervised classification

Our supervised approach uses the provided section ID as classification targets and predicts each section's probability. Formally, for a sample x and corresponding one-hot target y, we compute the standard cross-entropy (CE) loss, as seen in Equation (2).

$$\begin{aligned} \mathrm{CNN}(x) &\mapsto \hat{y}, \\ \mathcal{L}_{\text{sup}}(\cdot) = \mathcal{L}_{\text{CE}}(\hat{y}, y) &= -\frac{1}{N}\sum_{i}^{N} y_i \log \hat{y}_i, \end{aligned} \tag{2}$$

where CNN represents the CNN-based classifier and N the number of samples. Then the anomaly score $A(x)$ is calculated as:

$$A(x) = \log\left(\frac{1-\hat{y}_i}{\hat{y}_i}\right), \tag{3}$$

where $\hat{y}_i$ is the softmax output for the correct section. Note that if the sample x is divided into consecutive segments $(x_1, x_2, ..., x_P)$, the anomaly score will be $\frac{1}{P}\sum_i^P A(x_i)$.

2.3. Proposed contrastive semi-supervised learning

We train these models with an additional contrastive loss [11]. The contrastive loss $\mathcal{L}_{\text{contrastive}}$ is added between the hidden representations of both models ($\mathbf{v}_{\text{AE}}, \mathbf{v}_{\text{CNN}}$) as:

$$\begin{aligned} \mathbf{p} &= \mathbf{v}_{\text{AE}}, \\ \mathbf{u} &= \mathbf{v}_{\text{CNN}}, \\ \mathcal{L}_{\text{contrastive}}(\cdot) &= -\sum_i \log \frac{\exp(\langle \mathbf{u}_i, \mathbf{p}_i \rangle/\rho)}{\sum_{j\neq i}\exp(\langle \mathbf{u}_i, \mathbf{p}_j\rangle/\rho)}, \end{aligned} \tag{4}$$

where $\langle , \rangle$ represents inner product, $\rho \in \mathbb{R}$ is a scalar hyperparameter and $\mathbf{p}, \mathbf{u} \in \mathbb{R}^{256}$ are hidden vector representations obtained by both models via projection. Concretely speaking, we transform the output vector of Autoencoder's bottleneck layer and CNN's feature layer into same dimension by linear transformation, then map representations to the space where the contrastive loss is applied via a shared MLP projection layer with one hidden layer. In most cases, the dimension of the bottleneck layer in the Autoencoder is much smaller than the dimension of the feature layer in the CNN model (8 *vs.* 1280 in this paper). We assume that the bottleneck layer output in the AE tends to represent the general structure of normal sound clips, while CNN extracted feature represents their microscopic structure. Our approach aims to obtain two different representations of a single sample, which is reminiscent of SimCLR [8], unsupervised data augmentation (UDA) [12] and other semi and self-supervised approaches.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{unsup}} + \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{contrastive}} \tag{5}$$

The final loss for optimization can be seen in Equation (5).

2.4. Data Augmentation

One of our contributions is the exploration of data augmentation techniques. Regarding conventional techniques, we explore the use of Mixup [13] along with time masking [14] and frame-shifting for model training during the DCASE challenge. Further, our intuition is that the input audio data contains large amounts of short-time noise, thus an input feature might contain a surplus of unreliable information, which can affect the performance of our supervised training method. We propose a median filtering approach applied on the input spectrogram feature along the frequency axis aiming to reduce the influence of distracting noise.

3. EXPERIMENTAL SETUP

Log Mel-spectrogram (LMS) features are chosen as the default front-end feature for the task. Overall, seven models are trained in our approach, one for every machine type.

For the supervised CNN training, each 128-filter LMS is extracted from a 64 ms window with a stride of 32 ms. We follow the baseline approach by concatenating 64 consecutive frames with a shift of 8 frames, resulting in an 128×64 dimensional input tensor. If segments are shorter than 10 seconds (or 311 samples), we zero-pad the input to the longest sample within a batch.

Regarding the AE training, we flatten the input tensor to a single input vector of size 8192 ($128 * 64$). All experiments are run for 100 epochs, with the learning rate halving every 30 epochs. The batchsize is set to 32 for training and we set the hyperparameter $\rho = 0.07$ for the contrastive loss. Our median filtering approach

Table 1: Performance of our models in comparison to other participants in the challenge on the official evaluation dataset. Best results are highlighted in bold.

Model	Official Score	Fan	Gearbox	Slider	Toy Train	Toy Car	Pump	Valve
AE Baseline	56.375	60.68	65.49	57.22	68.51	65.93	58.30	51.87
MBv2 Baseline	54.770	64.96	51.14	72.92	42.91	42.73	67.97	53.13
1st	66.798	61.01	63.07	83.18	**69.15**	**75.27**	**86.76**	65.36
2nd	64.956	86.48	67.45	83.05	45.60	60.88	85.04	**71.49**
3rd	64.201	88.98	57.75	**86.84**	57.50	69.83	74.82	62.74
4th	63.745	66.60	62.53	86.27	61.79	61.70	74.60	62.36
5th	62.593	68.98	**67.74**	79.88	61.71	73.32	71.87	63.73
6th	62.239	82.65	57.20	83.76	53.43	58.67	85.54	60.54
7th	61.480	87.68	56.56	76.66	48.24	70.60	72.54	60.70
8th	61.186	73.17	64.70	69.89	51.71	68.23	78.65	53.93
Ours best	60.966	**90.68**	58.00	77.34	47.49	53.81	77.82	53.53

uses a window size of 31 frames (i.e., 1 second) for each filter bank respectively.

PyTorch [15] was used as the default neural network toolkit[1].

3.1. Evaluation metrics

The evaluation metrics used in the challenge is the area under curve (AUC) and partial-AUC (pAUC) scores respectively [16]. The final official score Ω is computed as the harmonic mean of the AUC and pAUC scores.

3.2. Dataset

The data used for this task consists of running sounds of seven machine types being “ToyCar”, “Fan”, “ToyTrain”,“Valve”, “Gearbox”, “Silder” and “Pump”, including two recent machine audio datasets, ToyADMOS [17] and MIMII [18].

Notably, all provided data samples by the challenge authors have a length of 10 seconds, and each section, as well as machine type, has a near uniformly distributed duration. The overall data length is 70 hours of which the large majority belongs to the source domain.

Model	Fan	Gearbox	Slider	Toy Train	Toy Car	Pump	Valve	Score
MBv2	60.30	57.43	59.43	51.10	53.60	56.17	55.19	56.01
+ CL	60.61	58.87	60.70	50.92	52.51	56.90	54.38	56.18
+ MF	64.08	65.38	59.83	49.69	55.38	59.50	53.74	57.75
+ CL, MF	64.45	67.16	58.66	51.89	56.15	57.27	53.46	57.99

Table 2: Main results proposed in our work for the DCASE 2021 Task2 challenge on the held-out development dataset in regards to the main evaluation metric Ω (see [16]). “C” represents adding contrastive learning and “M” the addition of median filtering. Note that a single model is trained for each machine type.

The two models used in this work are described. First, our AE is the same as the one provided by the challenge baseline. Each hidden block has 128 units except for the bottleneck block, which has 8 units. Second, the MobileNetV2 (MBv2) architecture is directly taken from [19], where our approach differs from the standard architecture by using global average and max pooling (GAMP) as our aggregation method compared to the standard global average pooling (GAP). During training, both the AE and MBv2 models are jointly optimized given the total loss Equation (5), whereas during evaluation only the MBv2 model is used.

[1]The source code is available at `https://github.com/bibiaaaa/SmallRice_DCASE2021Challenge`

4. RESULTS

Our model’s performance on the held-out development set is displayed in Table 2. As it can be seen, our MBv2 model trained in the proposed training framework shows improvement over the baseline model in some machine types such as “Fan” and “Gearbox”.

For the DCASE challenge, we trained an EfficientNet-B0 based model under our proposed training framework along with median filter and other data augmentation techniques such as Mixup [13] and time masking. For the challenge, our method ranked 9th out of 27 participated methods. As shown in Table 1, our method lacks behind an absolute of 6 % against the winning system.

It is worth mentioning that Table 3 shows that our method performed best on the Fan dataset, especially from the perspective of pAUC metric, leading by a large margin of around 9% compared to the 2nd result. We believe that it contributes to the median filter applied on the log-mel spectrogram along the time axis since it can erase short-time noise and improve the generalization ability of the model.

Model	Fan (AUC)	Fan (pAUC)
AE Baseline	60.68	50.50
MBv2 Baseline	64.96	58.14
2nd	90.22	71.19
3rd	88.98	70.20
4th	88.09	70.84
Ours	**90.68**	**79.99**

Table 3: Top 5 best results in the Fan dataset in the challenge. Our result ranks 1st both in AUC and pAUC.

5. CONCLUSION

This paper proposes a novel contrastive loss training framework for anomaly sound detection. Experimental results indicate that the MobileNetV2-based model trained under our proposed training framework exceeds the baseline model for some machine types in the DCASE 2021 challenge task 2, while no additional parameters are introduced during inference. Notably, our model achieves the best performance for the "Fan" machine type. We conclude that anomaly sounds greatly vary between different machine types, thus finding a universal anomaly sound detection method suitable for machine condition monitoring is still a problem worthy of research.

6. REFERENCES

[1] E. Dorj and E. Altangerel, "Anomaly detection approach using hidden markov model," in *Ifost*, vol. 2. IEEE, 2013, pp. 141–144.

[2] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Probabilistic novelty detection for acoustic surveillance under real-world conditions," *IEEE Transactions on Multimedia*, vol. 13, no. 4, pp. 713–719, 2011.

[3] A. Sasou and N. Odontsengel, "Acoustic novelty detection based on ahlac and nmf," in *2012 International Symposium on Intelligent Signal Processing and Communications Systems*, 2012, pp. 872–875.

[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[5] P. Primus, "Reframing unsupervised machine condition monitoring as a supervised classification task with outlier-exposed classifiers," DCASE2020 Challenge, Tech. Rep., July 2020.

[6] 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.

[7] P. Daniluk, M. Gozdziewski, S. Kapka, and M. Kosmider, "Ensemble of auto-encoder based systems for anomaly detection," DCASE2020 Challenge, Tech. Rep., July 2020.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[9] L. Wang, K. Kawakami, and A. van den Oord, "Contrastive Predictive Coding of Audio with an Adversary," in *Proc. Interspeech 2020*, 2020, pp. 826–830. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1891

[10] L. Wang, P. Luc, A. Recasens, J. Alayrac, and A. van den Oord, "Multimodal self-supervised learning of general audio representations," *CoRR*, vol. abs/2104.12807, 2021. [Online]. Available: https://arxiv.org/abs/2104.12807

[11] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised Contrastive Learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 18 661–18 673. [Online]. Available: http://arxiv.org/abs/2004.11362

[12] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised Data Augmentation for Consistency Training," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2019, pp. 6256–6268. [Online]. Available: http://arxiv.org/abs/1904.12848

[13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[14] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September. International Speech Communication Association, 2019, pp. 2613–2617.

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8026–8037.

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

[17] 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," *arXiv preprint arXiv:2106.02369*, 2021.

[18] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, "Mimii due: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," *arXiv preprint arXiv:2105.02702*, 2021.

[19] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: http://proceedings.mlr.press/v97/tan19a.html