UNSUPERVISED ANOMALOUS DETECTION BASED ON TRANSFORMER AUTOENCODER MODEL AND PRETRAINED AUTOENCODER MODEL

Technical Report

Zongmu Lin, Yihao meng, Yuanhang Qian, Yuankai Zhang, Yujie Zhu, Gongping Huang

School of Electronic Information, Wuhan University, Wuhan, China

ABSTRACT

Automatic detection of machine anomaly remains challenging for machine learning. Unsupervised models have been widely applied in lots of scenarios successfully. This technical report outlines our solutions to Task 2 of the DCASE2025 challenge.The objective is to detect audio recording containing anomalous machine sounds in a test set, when the training dataset itself does not contain any examples of anomalies. Our approaches are based on transformer auto-encoder model and use pretrained model to extract the multi-scale features.

Index Terms—Anomaly detection, Transformer, auto-encoder

1. INTRODUCTION

Anomalous sound detection (ASD)for machine condition monitoring has been a task at the DCASE challenge for several years[1][2]. Anomaly detection aims to identify anomalous samples from normal samples when only normal samples are provided. Audio-based anomaly detection remains challenging for both traditional and deep learning-based models.In practical applications, the types of machines may be entirely new, and the available test data are often very limited. Therefore, it is impractical to rely on the test data from the development option set to fine-tune hyperparameters for each type of machine. Moreover, there is a significant imbalance between the source-domain and target-domain data. When a model trained on source-domain data is applied to target-domain data, the model's performance can degrade significantly due to domain shifts caused by factors other than anomalies, such as environmental noise and equipment differences.

For Task 2, [3] provides two baseline methods.The dataset used in this task is derived from the MIMII DG[4] and ToyADMOS2 dataset[5], which includes normal and abnormal operating sounds of various types of machinery. However, in real-world scenarios, collecting data and training models for new machines is highly challenging, especially when the number of machines is very small. The development dataset and the evaluation dataset do not share the same types of machines, and only a single part of each machine type is available for use. To address these challenges, we aim to develop an audio encoder with strong generalization capabilities to avoid overfitting on limited training data, thereby enhancing the robustness of the model across different machine types and data distributions.Encoder-decoder structures have been widely used in the field of anomaly detection because they can easily learn normal patterns in an unsupervised learning environment and calculate a score to identify abnormalities through a reconstruction error indicating the difference between input and reconstructed .

Transformer[6] has a higher capacity to represent global information, which gives it the potential to surpass AE and become a new reconstruction network foundation for anomaly detection.We designed a transformer-based autoencoder structure, which judges anomalies by reconstructing audio and calculating the reconstruction error. For different types of machines, we employed the dropout strategy to enhance the model's generalization capability.

2. PROPOSED METHOD

2.1. Transformer Model

Due to the different operating modes of various instruments, we transform their audio into frames of length 5 or 20, with 128 mel bins.Our model consists of an encoder, a bottleneck layer, and a reconstruction decoder. We employ 12 Transformer layers as the encoder to extract feature maps with different semantic scales. The bottleneck layer is implemented using a simple MLP (Multi-Layer Perceptron), and the decoder has a structure similar to the encoder, consisting of 8 Transformer layers. During training, the decoder learns to reconstruct the features extracted by the encoder by maximizing the similarity of the features. In the inference stage, the decoder can accurately reconstruct the normal regions of the feature maps, but it fails to reconstruct the abnormal regions.

Previous studies have attributed the performance degradation of UAD (Unsupervised Anomaly Detection) methods trained on multi-class samples to the identity mapping phenomenon. However, we argue that this is due to the model's overgeneralization. As the multi-class UAD setting increases the diversity of images and their patterns, the decoder may generalize its reconstruction capability to abnormal samples, leading to the failure of anomaly detection based on reconstruction error. In this study, we instead employ Dropout technology to enhance the generalization ability across different instrument types.We classified the data from the same equipment and calculated the anomaly scores separately.

2.2. HTS-AT AE

Pre-trained models are trained on large-scale unlabeled audio data, enabling them to learn complex features of audio signals, including information in the temporal domain, frequency domain,

and semantic level. As a result, they exhibit stronger generalization capabilities when applied to new datasets or scenarios.

HTS-AT[7] is an efficient and light-weight audio transformer with a hierarchical structure and has only 30 million parameters. It achieves new state-of-the-art (SOTA) results on AudioSet[8] and ESC-50[9], and equals the SOTA on Speech Command V2. It also achieves better performance in event localization than the previous CNN-based models.In this system, we used a pre-trained model based on HTS-AT as the feature extractor.Then, we constructed an autoencoder to reconstruct the embedding.In the encoder part, we used three independent large-kernel convolutional layers to extract features. Subsequently, we transformed the features through three multi-scale channel attention modules, where the multi-scale features were fused via the channel attention modules.In the decoder part, we used Linear Attention[10] instead of the Softmax operation. Linear Attention does not have the nonlinear reweighting operation of Softmax, so it cannot focus on specific regions of the input like Softmax Attention. This makes it more difficult for the reconstruction decoder to handle abnormal regions, thereby improving the performance of anomaly detection.

3. RESULTS

In this task, to achieve unsupervised anomaly detection, we utilized a Transformer autoencoder and an unsupervised pre-trained model for anomaly detection. Compared with the baseline system based on autoencoders, it can achieve superior results on the validation set. Additionally, we introduced perturbations in this task to augment the data simulating different operational states. Finally, we integrated our subsystems using score-level fusion, and we selected the fusion results with four different weights as our final submission.

4. REFERENCES

[1] Y. Koizumi, et al., “Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring,” in 5th Workshop on Detection and Classification of Acoustic Scenes and Events, 2020, pp. 81–85.

[2] Y. Kawaguchi, et al., “Description and discussion on dcase 2021 challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted condi- tions,” in 6th Workshop on Detection and Classification of Acoustic Scenes and Events, 2021, pp. 186–190.

[3] 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 2025 challenge task 2: first-shot unsupervised anomalous sound detection for machine condition monitoring. In arXiv e-prints: 2506.10097, 2025.

[4] 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, November 2022.

[5] 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), 1–5. Barcelona, Spain, November 2021.

[6] Vaswani, Ashish et al. “Attention is All you Need.” Neural Information Processing Systems (2017).

[7] Chen, Ke et al. “HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection.” ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2022): 646-650.

[8] Jort F. Gemmeke, Daniel P. W. Ellis, and Dylan Freed- man et al., “Audio set: An ontology and human-labeled dataset for audio events,” in ICASSP 2017.

[9] Karol J. Piczak, “ESC: dataset for environmental sound classification,” in ACM MM 2015. ACM.

[10] Guo, Jia et al. “Dinomaly: The Less Is More Philosophy in Multi-Class Unsupervised Anomaly Detection.” ArXiv abs/2405.14325 (2024): n. pag.