A DUAL-STREAM CNN WITH SUB-CLUSTER ADAPTIVE COSINE LOSS FOR ANOMALOUS SOUND DETECTION

Technical Report

*Hà Mạnh Dũng**

Posts and Telecommunications
Institute of Technology
Faculty of Information Technology 1
Hanoi, Vietnam
DungHM.B22CN125@stu.ptit.edu.vn

ABSTRACT

This report describes our system for the DCASE 2025 Challenge Task 2: "Unsupervised Anomalous Sound Detection for Machine Condition Monitoring" [1]. Our approach is based on a dual-stream Convolutional Neural Network (CNN) architecture designed to extract robust features from raw audio signals. One stream processes frequency characteristics via a Fast Fourier Transform (FFT), while the second stream analyzes time-frequency features from a magnitude spectrogram. To enhance model generalization, we employ two data augmentation techniques: Mixup [2] and SpecAugment [3]. The core of our system is a metric learning approach using the Sub-Cluster AdaCos (SCAdaCos) loss function, inspired by AdaCos [4], to learn highly discriminative embeddings. Anomaly scores are calculated based on the cosine similarity between test sample embeddings and pre-computed class centroids from the training data. Our results on the development set show that the system has a foundational capability for anomaly detection, with performance metrics surpassing the random guess baseline.

Index Terms— Anomalous Sound Detection, DCASE 2025, Convolutional Neural Networks, Metric Learning, Data Augmentation

1. INTRODUCTION

The objective of DCASE 2025 Challenge Task 2 is to identify anomalous machine sounds under conditions where the acoustic characteristics of the environment may change between training and testing (domain shift) [1]. This requires a system that is not only capable of modeling the distribution of normal sounds but also robust to variations in operational conditions.

This paper presents a system built upon a dual-stream deep learning model. We leverage both raw waveform and spectrogram representations to create rich feature embeddings. The training is guided by an adaptive cosine-based loss function, and data augmentation is used extensively to improve generalization. The final anomaly score is determined using a distance-based metric in the learned embedding space.

2. PROPOSED SYSTEM

Our system follows a deep metric learning paradigm. It first learns to map audio samples into a high-dimensional embedding space where normal sounds from the same machine type form tight clusters. Anomaly detection is then performed by measuring the distance of a test sample to these normal-condition clusters.

2.1. System Overview

The core of our system is a dual-stream CNN model, model_emb_cnn. It processes audio data through two parallel branches to capture a comprehensive set of features, which are then concatenated to form a final embedding vector.

2.2. Feature Extraction

FFT Branch: This stream takes the raw audio waveform, applies a Fast Fourier Transform to convert the signal to the frequency domain, and then feeds the absolute values into a series of 1D Convolutional layers to learn frequency-based patterns.

Spectrogram Branch: This stream first converts the raw waveform into a magnitude spectrogram. The spectrogram is then processed by a deep 2D CNN with residual connections, inspired by the ResNet architecture [5], to learn complex time-frequency patterns.

2.3. Feature Extraction

To improve model robustness and prevent overfitting, we apply two forms of data augmentation during training:

Mixup [2]: This technique creates new training samples by taking a linear combination of two random samples and their corresponding labels. This encourages the model to learn smoother decision boundaries.

SpecAugment [3]: Applied to the spectrogram branch, this method randomly masks out vertical (frequency) and horizontal (time) bands of the spectrogram. This forces the model to learn more robust and less localized features.

2.4. Loss Function and Training

We employ the Sub-Cluster AdaCos (SCAdaCos) [4] loss function. This is an adaptive metric learning loss that aims to maximize inter-class distance while minimizing intra-class variance. It dynamically adjusts its scaling factor during training to create more discriminative embeddings, which is crucial for our distance-based anomaly scoring.

The model is trained using the Adam optimizer with a Cosine Annealing learning rate schedule. This schedule helps the model converge effectively over a fixed number of epochs by starting with a higher learning rate and gradually decreasing it.

2.5. Anomaly Score Calculation

After training, the model is used as a feature extractor to generate a 256-dimensional embedding for each audio clip. For each machine type, we compute centroids (mean embeddings) for all normal samples in the training set, distinguishing between source and target domains.

The anomaly score for a given test sample is calculated as the minimum cosine distance (1 - cosine similarity) to the set of pre-computed normal centroids corresponding to its machine type. A higher score indicates a higher probability of being an anomaly.

3. EXPERIMENTAL RESULTS

3.1. Dataset and Setup

We used the DCASE 2023 Challenge Task 2 Development Dataset for training and validation. The system was trained with a batch size of 64 for 10-20 epochs, depending on the experiment. Due to the implementation challenges of the AdaCos loss function, the model was trained with the run_eagerly=True flag in TensorFlow.

3.2. Performance

The following table shows the performance of our system on the "bearing" machine type from the development set, which is representative of the system's current capabilities.

Machine Type	AUC (Source)	AUC (Target)	pAUC (Mean)
bearing	51.00%	54.56%	53.95%

As shown, all metrics are above the 50% baseline of a random classifier, confirming that the model has learned to distinguish anomalous sounds. The performance on the target domain is slightly higher than on the source domain, which suggests a degree of successful generalization. However, the overall scores indicate that there is significant room for improvement.

4. CONCLUSION

We have presented a dual-stream CNN system for anomalous sound detection. The combination of multi-view feature extraction, advanced metric learning loss, and strong data augmentation techniques creates a robust foundation. The current results are promising and validate the overall approach. Future work will focus on extensive hyperparameter tuning, exploring different model architectures to reduce complexity, and refining the anomaly scoring logic to further enhance performance.

5. REFERENCES

[1] DCASE 2025 Challenge, http://dcase.community/challenge2025/.

[2] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” in Proc. ICLR, 2018.

[3] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Quoc V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition,” in Proc. Interspeech, 2019, pp. 2613-2617.

[4] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, “AdaCos: Adaptively Scaling Cosine Logits for Effectively Learning Deep Face Representations,” in Proc. IEEE CVPR, 2019, pp. 10823-10832.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE CVPR, 2016, pp. 770-778.

[6] T. Nishida et al., "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.

[7] N. Harada et al., "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), 2021, pp. 1-5.

[8] K. Dohi et al., "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), 2022.

[9] N. Harada et al., "First-Shot Anomaly Detection for Machine Condition Monitoring: A Domain Generalization Baseline," in Proceedings of 31st European Signal Processing Conference (EUSIPCO), 2023, pp. 191-195.