SEMI-SUPERVISED ACOUSTIC SCENE CLASSIFICATION BASED ON PRETRAINED MODEL WITH ATTENTION MECHANISM

Haiyue Zhang, Xichang Cai✉, Jingxuan Chen, Liangxiao Zuo, Menglong Wu, Ziyi Liu, Yige Zhang

North China University of Technology, Beijing, China
caixichang@ncut.edu.cn

ABSTRACT

In this technical report, we present our submission system for IEEE ICME 2024 Grand Challenge: Semi-supervised Acoustic Scene Classification based on Pretrained Model with attention mechanism. Facing the challenge of domain bias in Acoustic Scene Classification (ASC) in different geographical contexts, we propose an improved model based on the semi-supervised learning framework. The model introduces the Coordinate Attention mechanism and the dual-branch attention mechanism based on the baseline model SE-Trans to enhance the model's ability to localize and recognize audio events. Furthermore, considering the limitations of the training data, we adopted noise addition techniques. By merging features with noise with original features, we successfully expanded the dataset. Our systems finally achieve the accuracy of 0.983 on the validation set.

Index Terms— Acoustic Scene Classification, semi-supervised learning, pretrained model，attention mechanism.

1. INTRODUCTION

With the widespread application of intelligent devices in daily life, Acoustic Scene Classification (ASC) has emerged as a significant research field within computational auditory scene analysis. ASC aims to recognize unique sound features in the environment and provide technical support for applications such as intelligent sound recognition, environmental monitoring, and urban soundscape analysis. However, a major challenge for ASCs is the problem of domain shift, which the distributional differences between training and testing data, especially when these data originate from diverse geographical backgrounds. To address this challenge, the IEEE ICME 2024 Grand Challenge has introduced the problem of "domain shift in semi-supervised Acoustic Scene Classification," aiming to develop more robust ASC models under domain shift conditions through innovative semi-supervised learning techniques. Semi-supervised learning methods utilize a large amount of unlabeled data while being guided by a small amount of labeled data, which is an effective way to solve the problem of scarce and costly labeled data.

In this context, this study proposes an improved model based on the baseline SE-Trans, which enhances the feature extraction and scene classification ability of the model by introducing the Coordinate Attention mechanism and the dual-branch attention mechanism. Furthermore, considering the limitations of the training data, we have explored a novel data augmentation strategy that expands the dataset by incorporating noise addition techniques, thereby enhancing the model's adaptability to data from diverse domains. The following sections will detail the methodology, experimental design, and the results obtained.

2. BLIND REVIEW

2.1. Data Augmentation

To enhance the model's robustness to various environmental noises and increase data diversity, this study employs a data augmentation technique involving the addition of noise. By overlaying randomly selected noise audio onto original acoustic scene recordings, new noisy audio samples are generated, simulating a variety of auditory environments in the real world. Furthermore, to further enrich the model's input features, the generated noisy features are concatenated with the original audio features, employing the feature-level fusion strategy to increase the information content of the input data.

2.2. Pretrained Settings

This study employed the TAU [1] and Cochl [2] datasets as required by the competition for pre-training. The Cochl dataset contains 13 categories, totaling over 76,000 audio recordings. The TAU dataset includes 10 categories, with more than 23,000 audio recordings. Due to the different

numbers of categories and volumes of data, to fully leverage the potential of these two datasets, We adopt a "big-to-small" phased training method. Specifically, we first trained the model in depth for 200 training cycles using the Cochl dataset. After completing the training with the Cochl dataset, we continued training the model using the TAU dataset. It is worth noting that, before transitioning to training with the TAU dataset, we removed the model's fully connected layer (fc parameter) to avoid parameter mismatch issues due to different numbers of categories.

2.3. Network Architecture

The baseline system is based on a semi-supervised framework with a Squeeze-and-Excitation and Transformer (SE-Trans) model pre-trained on the TAU Urban Acoustic Scenes (UAS) 2020 Mobile development dataset. The training model in this reported is improved based on the official baseline. The model improvement is mainly realized by introducing Coordinate Attention mechanism [3] and Double Branch Attention mechanism [4],[5].

Specifically, the model consists of two key components: the convolutional operation block incorporating Coordinate Attention, and the two-branch attention structure. Each convolution-Coordinate Attention block employs a layout alternating between two layers of convolution and two layers of Coordinate Attention (CA) mechanisms. Batch Normalization (BN) and ReLU activation functions are applied after each layer of convolution. It not only promotes deeper feature extraction, but also enhances the nonlinear feature representation of the model. To effectively capture both global and local features in the audio signal, we further introduced the dual-branch attention mechanism. This mechanism enables fine-grained analysis of acoustic features by combining convolutional and attention layers, allowing the model to more accurately extract information that is critical for classification decisions.

3. EXPERIMENTS

3.1. Dataset

The dataset utilized for training in this paper is the CAS 2023[6], which is divided into a development set and an evaluation set. This dataset encompasses 10 common acoustic scenes, with a total duration exceeding 130 hours. Each audio clip is 10 seconds long and includes metadata about the recording location and timestamp. The development dataset, approximately 24 hours in length, features recordings from 8 cities, and 20% of the data is labeled with scene tags.

3.2. Experimental Setup

The log-mel spectrum serves as the input feature for the Acoustic Scene Classification (ASC) system. To optimize the training process, we have configured the batch size to 32 and selected an Adam optimizer with a learning rate of 0.001 for parameter adjustment. Furthermore, the model's pre-training and training phases are set to 200 and 20 cycles, respectively.

3.3. Experimental Results

In this study, accuracy is employed as the metric to evaluate the performance of the proposed model, the experimental results presented in Table 1. The official baseline accuracy for Acoustic Scene Classification (ASC) model is 0.956. The accuracy improved to 0.963 following enhancements to the pre-trained model. Upon this foundation, we further introduced noise, resulting in an additional increase in accuracy to 0.973.

For the proposed model, accuracy escalated to 0.979 following the enhancement of both the training model and the pre-trained model. After the introduction of noise, the accuracy increased to 0.983.

Table 1. Performance of proposed model

Method	Accuracy
Baseline	0.956
Baseline+ Pretrained	0.963
Baseline+ Pretrained+ noises	0.973
Proposed+ Pretrained	0.979
Proposed+ Pretrained+ noises	0.983

4. CONCLUSIONS

This technical report presents our submission to the IEEE ICME 2024 Grand Challenge, a semi-supervised Acoustic Scene Classification system. This system is based on the pre-trained model and incorporates both the Coordinate Attention mechanism and the dual-branch attention mechanism. Through strategies such as data augmentation with noise addition and a "large to small" pre-training approach, the system effectively enhances the accuracy of acoustic scene classification. Particularly in addressing the issue of domain shift across different geographical backgrounds, our improved model demonstrates enhanced precision in capturing both global and local features. The experimental results on the validation set, achieving an accuracy of 0.983, validate the effectiveness of our proposed method.

5. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, “A multidevice dataset for urban acoustic scene classification,” in Proceedings of the Detection and

Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), November 2018, pp. 9–13.

[2] Jeong, Il-Young, and Jeongsoo Park. "CochlScene: Acquisition of acoustic scene data using crowdsourcing." 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2022.

[3] Hou, Qibin, Daquan Zhou, and Jiashi Feng. "Coordinate attention for efficient mobile network design." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

[4] Bai, Jisheng, et al. "Description on IEEE ICME 2024 Grand Challenge: Semi-supervised Acoustic Scene Classification under Domain Shift." ar**v preprint ar**v:2402.02694 (2024).

[5] Peng, Yifan, et al. "Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding." International Conference on Machine Learning. PMLR, 2022.

[6] Kim, Kwangyoun, et al. "E-branchformer: Branchformer with enhanced merging for speech recognition." 2022 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2023.