THE NERCSLIP-USTC SYSTEM FOR ACOUSTIC SCENE CLASSIFICATION OF ICME 2024 GRAND CHALLENGE

Qing Wang[1], *Guirui Zhong*[1], *Hengyi Hong*[1], *Lei Wang*[2], *Mingqi Cai*[3], *Xin Fang*[3], *Jun Du*[1],

[1] University of Science and Technology of China, Hefei, China
[2] National Intelligent Voice Innovation Center, Hefei, China
[3] iFLYTEK, Hefei, China

ABSTRACT

In this technical report, we present our submission system for the acoustic scene classification (ASC) task in the ICME 2024 Grand Challenge. To address the domain shift problem, we propose a two-stage training strategy based on fully convolutional neural network (FCNN). We first pre-train FCNN models using two publicly released datasets, which are then combined with the labeled data from Chinese Acoustic Scene (CAS) development dataset to fine-tune the models in the previous stage. For semi-supervised learning, we generate convincing pseudo labels for unlabeled data within the development dataset according to the confidence of different models. Furthermore, we train a three-class classifier besides the ten-class ASC system, which recognizes an input audio scene as one of three main classes, including in-door, out-door, and transportation. In addition, we adopt manifold mixup augmentation during the model training processing. Evaluated on a larger dataset containing both CAS 2023 development and two publicly released datasets, our proposed approach achieve an accuracy of 82.2%.

Index Terms— Acoustic scene classification, semi-supervised learning, convolution neural network, data augmentation

1. INTRODUCTION

Acoustic scene classification (ASC) is a task that involves analyzing and identifying audio recordings based on the environmental sounds they contain. The goal is to automatically classify audio recordings into pre-defined classes, such as shopping malls and urban parks. Potential applications of ASC technique include environmental monitoring and smart devices.

Different from the ASC task in the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenges, the ICME 2024 Grand Challenge focuses two critical factors influencing the performance of ASC task : domain shift and scarcity of labeled data. The problem of domain shift between different recording regions is explored in this challenge. Another key issue is utilizing abundant unlabelled data to train robust ASC systems.

In this technical report, we present our approach for the ASC task in the ICME 2024 Grand Challenge. Fully convolutional neural network (FCNN) models trained using TAU Urban Acoustic Scenes (UAS) 2020 Mobile development dataset [1, 2] and CochlScene dataset [3], are utilized as the pre-trained models. These two datasets are combined with the labeled data from Chinese Acoustic Scene (CAS) development dataset to fine-tune on the pre-trained models. Then pseudo labels are generated for unlabeled data from CAS development dataset. Finally, we utilized the pseudo-labeled data and the combined dataset for fine-tuning on the pre-trained FCNNs, resulting in the ten-class and the three-class ASC models.

2. DATASETS

The TAU UAS 2020 Mobile development dataset [2] and CochlScene dataset [3] are used to pre-train FCNN models. TAU UAS 2020 Mobile contains 23,040 samples, with each sample provided in binaural and 48 kHz sampling rate. CochlScene contains 76,115 single-channel audio files with 44.1 kHz sampling rate. Since these two datasets include different types of acoustic scenes, we removed certain types of scenes and merged others to generate a new pre-training dataset. The number of audio recordings for each scene is shown in Table 1. The data in Table 1 are used to pre-train FCNNs.

The development dataset of CAS 2023 contains 8,700 audio clips, and 20% of the data have ground truth. The labeled data set is combined with the new pre-training dataset, which is named initial labeled dataset. Meanwhile, 50 labeled audio recordings from each scene category in the CAS 2023 development dataset are merged with the validation set in the new pre-training dataset. To utilize the unlabeled data, we propose a two-step pseudo label generation method. In the first step, the pre-trained FCNN is fine-tuned on the initial labeled dataset and then used to assign pseudo labels to unlabeled data in the development dataset. The predicted posterior probabilities of unlabeled data are sorted from high to low, and we

Table 1. The number of audio recordings for each scene in the new generated pre-training dataset.

Scene	Number of audio recordings	
	Training	Validation
Airport	1393	296
Bus	6053	1170
Car	4676	584
Metro	6096	885
Metro station	6094	884
Public square	1427	297
Restaurant	4725	590
Shopping mall	1373	297
Traffic street	6007	869
Urban park	6022	867
Total	43866	6739

select the first half of the pseudo-labeled data. In the second step, the initial labeled dataset and the pseudo-labeled data selected in the previous step are utilized to fine-tune the pre-trained FCNN. This FCNN and the SE-Trans released by [4] are employed to generate pseudo labels for the left half of the unlabeled data in the development dataset. Audio samples predicted to belong to the same scene category by both of these two models are selected as reliable pseudo-labeled data. The pseudo-labeled data from the development dataset are combined with the initial labeled dataset to form final labeled dataset, which are used to train our submission system.

3. PROPOSED APPROACH

3.1. Network Architecture

In the proposed system, FCNN [5, 6] as shown in Table 2 is adopted, which achieved the second place in the ASC task of DCASE 2021 Challenge. FCNN is a VGG-like model [7] consisting of 9 stacked convolutional layers. Each convolutional layer is followed by a batch normalization and a ReLU activation function. Dropout is also used to alleviate overfitting problem. A 2 × 2 max-pooling layer is applied after the second, fourth, and eighth ReLU layers. Channel attention is applied before the final global average pooling layer. Finally, a 10-way softmax activation is used at the output layer.

3.2. Data Augmentation

We applied manifold mixup [8] augmentation to increase the diversity of training data. Layers at which we perform mixup include input and three hidden layers behind max-pooling operation. The parameter alpha is set to 1.0. Two batches of feature are randomly mixed in each step along with the corresponding labels when training models.

Table 2. Configuration of our FCNN architecture.

Block name	Configuration
Input	LMFB
BatchNorm	Learn γ and β
Block1	Conv 5 × 5 @ 144, BN, ReLU Conv 3 × 3 @ 144, BN, ReLU Max Pooling 2 × 2
Block2	Conv 3 × 3 @ 288, BN, ReLU Conv 3 × 3 @ 288, BN, ReLU Max Pooling 2 × 2
Block3	Conv 3 × 3 @ 576, BN, ReLU, Dropout Conv 3 × 3 @ 576, BN, ReLU, Dropout Conv 3 × 3 @ 576, BN, ReLU, Dropout Conv 3 × 3 @ 576, BN, ReLU, Dropout Max Pooling 2 × 2
Conv	Conv 3 × 3 @ 10 / 3, BN, ReLU
BatchNorm	Learn γ and β
ChanAttn	Channel Attention
Output	Adaptive Average Pooling

3.3. Two-stage Classification

We adopt a two-stage classification strategy to train the ASC system. A ten-class classifier and a three-class classifier are build based on the FCNN architecture. The three-class FCNN model aims to classify an input audio clip into one of three main classes, including in-door, out-door and transportation. For the three-class classifier, the number of output channels for the last convolutional layer is set to 3 as shown in Table 2. The final prediction of scene class is obtained by score fusion of these two classifiers, which is expressed as follows,

$$c = \arg\max_{c, i \supset c} y_c^1 * y_i^2 \tag{1}$$

where y_c^1 is the probability for the c-th class predicted by the ten-class classifier, y_i^2 is the probability for the i-th class predicted by the three-class classifier, $c \in \{1, 2, ..., 10\}$, $i \in \{1, 2, 3\}$. And $i \supset c$ means that class i is a super set of class c. For example, the class of out-door scene is the super set for public square, urban park, traffic street and construction site.

3.4. Training Setup

We first resampled the audio recordings in the TAU UAS 2020 and CAS 2023 datasets to 44.1 kHz. All audio clips have a fixed-length of 10 seconds. Log-Mel filter bank (LMFB) were extracted as audio features by using Librosa [9] library with 2048 short-time Fourier transform (STFT) points, a window size of 2048 samples, and a frame shift of 1024 samples. We applied 128 Mel-filter bands on the spectrograms. Then log-mel delta and delta-delta operations without padding, which generates a feature tensor shape of 423 × 128 × 3. Dropout

rate is set to 0.3. We trained our model using Adam [10] optimizer. Batch size is set to 64 and learning rate is set to 0.001. All our models are trained using PyTorch toolkit [11].

The size of the data from the challenge dataset is much smaller than that from the pre-training dataset. To ensure that the model sees the challenge data more frequently during fine-tuning, we employ a strategy where half of the data in each batch comes from the challenge dataset. By using this data mixing strategy during training, the model is supposed to achieve good performance on the challenge data while also maintaining robustness under domain shift.

4. RESULTS

This section reports the class-wise accuracy of our submission system for the validation set in our generated labeled dataset. The experimental results are shown in Table 3.

Table 3. The performance of the proposed ASC system for each scene class on the validation set in our generated labeled dataset.

Scene	Accuracy
Airport	85.8%
Bar	100.0%
Bus	86.5%
Construction site	100.0%
Metro	91.1%
Public square	14.4%
Restaurant	95.5%
Shopping mall	62.8%
Traffic street	92.8%
Urban park	93.4%
Average	82.2%

5. CONCLUSION

In this technical report, we detailed our approach to tackle the ASC task of the ICME 2024 Grand Challenge. We utilize publicly released datasets during pre-training and fine-tuning to address the domain shift problem. We employed two ASC models, namely FCNN and SE-Trans, to generate reliable pseudo-labeled data. Additionally, we use manifold mixup augmentation and two-stage classification method to obtain robust prediction results.

6. REFERENCES

[1] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen, "Acoustic scene classification in DCASE 2020 Challenge: generalization across devices and low complexity solutions," in *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2020, pp. 56–60.

[2] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen, "TAU urban acoustic scenes 2020 mobile, development dataset," 2020, [Online]. Available: https://doi.org/10.5281/zenodo.3685828.

[3] Il-Young Jeong and Jeongsoo Park, "Cochlscene: Acquisition of acoustic scene data using crowdsourcing," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 17–21.

[4] Jisheng Bai, Mou Wang, Haohe Liu, Han Yin, Yafei Jia, Siwei Huang, Yutong Du, Dongzhe Zhang, Mark D Plumbley, Dongyuan Shi, et al., "Description on ieee icme 2024 grand challenge: Semi-supervised acoustic scene classification under domain shift," *arXiv preprint arXiv:2402.02694*, 2024.

[5] Qing Wang, Jun Du, Siyuan Zheng, Yunqing Li, Yajian Wang, Yuzhong Wu, Hu Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Yannan Wang, et al., "A study on joint modeling and data augmentation of multi-modalities for audio-visual scene classification," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2022, pp. 453–457.

[6] Hu Hu, Chao-Han Huck Yang, Xianjun Xia, Xue Bai, Xin Tang, Yajian Wang, Shutong Niu, Li Chai, Juanjuan Li, Hongning Zhu, et al., "A two-stage approach to device-robust acoustic scene classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 845–849.

[7] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[8] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International conference on machine learning*. PMLR, 2019, pp. 6438–6447.

[9] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python.," in *SciPy*, 2015, pp. 18–24.

[10] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.