A SEMI-SUPERVISED MODEL FOR ACOUSTIC SCENE CLASSIFICATION

Zhuo Liu, Gong Chen, Min Zhao, Zhao Zhou, Fan Zhang, Yun Wang, Danlei Chen

Northwestern Polytechnical University, Xi' an, China,
{zhuoliu, chengong2022, minzhao, zhao.zhou}@mail.nwpu.edu.cn
{felixzhangchn, wangyunjeff, aalalei1023}@gmail.com

1. INTRODUCTION

The task of acoustic scene classification is to tag a given audio recording as belonging to a specific category of acoustic scenes from the real world (Bus, Restaurant, Urban park, etc). Distinguishing these specific acoustic scenarios is crucial for context-aware computing [1], intelligent wearable devices [2], and security monitoring [3]. Due to the complexity of the acoustic environments and the diversity of noise, supervised learning methods [4, 5, 6] usually require a large number of labeled recordings to obtain high-performance classifiers. However, large-scale labeled datasets are difficult to obtain considering both economic cost and time consumption of manually annotating. Meanwhile, the domain shift problem caused by differences across recording equipment and regions also poses great challenges to acoustic scene classification task. In the *IEEE ICME 2024 Grand Challenge of Semi-supervised Acoustic Scene Classification under Domain Shift*, participants were encouraged to use semi-supervised learning (SSL) as well as domin generalization (DG) methods to develop acoustic scene classifiers using a small number of labeled recordings while utilizing a large number of unlabeled recordings. This technical report describes the system of our submission to the challenge.

2. METHODOLOGY

Figure 1 shows the framework of our proposed acoustic scene classification architecture. First, the feature extraction stage is used to transfer the audio recordings into acoustic features. Then a data augmentation strategy is applied to enhance the number of data for training. We utilize an efficient convolutional neural network, MoblieNet, as our network architecture to recognize the acoustic features under different situations. During the training process, first a supervised stage is used to train the model with the labelled data, then we train our network with a semi-supervised strategy on the development dataset.

2.1. Dataset

The dataset contains two parts. First, the audio recordings in *TAU Urban 2020 Mobile* and *cochlscene* with the acoustic scene labeled in 13 classes are used as the dataset to pretrain the model. We name this dataset as $\mathcal{D}_1$. Then the data selected from the development dataset with 1740 labeled and 6920 unlabeled recordings, named as $\mathcal{D}_2 = \{\mathcal{T}, \mathcal{S}\}$, are applied to train the semi-supervised method. We denote the dataset in $\mathcal{D}_2$ with scene labeled as $\mathcal{T}$ and the unlabeled recording dataset as $\mathcal{S}$.

2.2. Feature Extraction

We employ log Mel spectrograms as our input features. These spectrograms are generated using the short-time Fourier transform (STFT) with a Hanning window of 1024 and a hop size of 512. Following the extraction of the spectrogram, we apply 128 Mel-filter bands and subsequently perform a logarithmic operation to compute the log Mel spectrograms. Each log Mel spectrogram is with the size of 512×128, with 512 representing the number of time frames and 128 indicating the number of frequency bins.

2.3. Data Augmentation

There are 3 ways used in our method for data augmentation.

- **Randomly trimming:** In this strategy, we randomly trim the 10s recording in to 6s recording to augment the dataset.
- **FMix:** The technique proposed in [4] is applied to enhance the generalization of our proposed method.
- **Mix up:** This augmentation operation is defined as

$$\tilde{\mathbf{x}} = \alpha\mathbf{x} + (1-\alpha)\mathbf{x}' \tag{1}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are two audio recordings, and $\alpha \in (0, 1)$ is a random parameter.

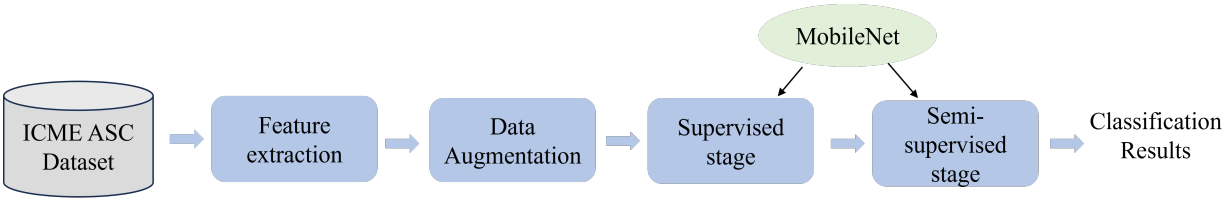

Fig. 1. The framework of our proposed method.

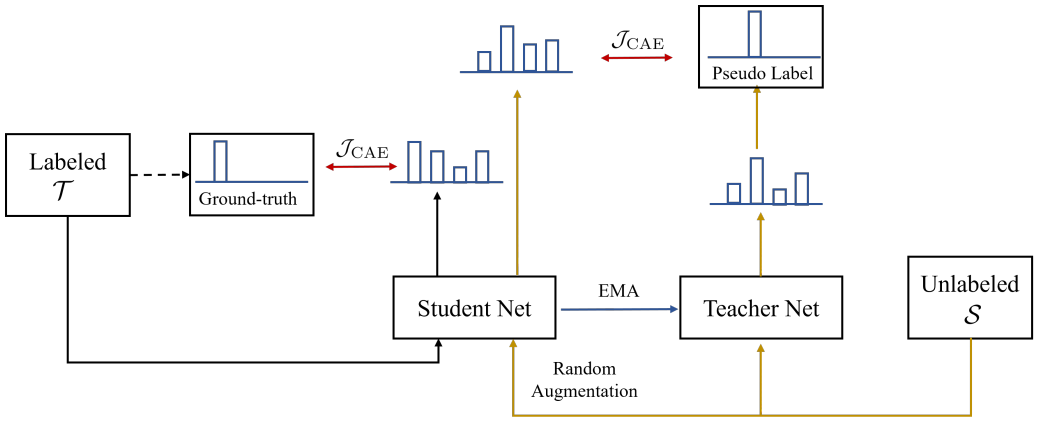

Fig. 2. The scheme of the semi-supervised method.

2.4. Network Architecture

We use MobileNet as our network architecture for acoustic scene classification. This architecture is a lightweight deep neural network and founded upon a streamlined approach utilizing depthwise separable convolutions. Several groups of convolution layer and depthwise separable convolution layer are used to formulate the main part of our network. Batch normalization and ReLU are used after each layer. A fully connected layer is used as the classifier head of our network to map the embedding features to the estimated class.

2.5. Supervised Stage

In our method, the transfer learning strategy is used to improve the performance of our models. We firstly pretrain our network architecture with the data from $\mathcal{D}_1$ in a supervised manner. Then we transfer it to the training of $\mathbf{D}_2$. The data are firstly augmented using the strategies mentioned above. Then the model is trained with the learning rate set to 1×10^{-3}, and we set the batch size to 32. The categorical cross-entropy loss is used to train our model, which is defined as:

$$\mathcal{J}_{\text{CAE}} = -\sum_{i=1}^{N} \mathbf{x}_n \log \hat{\mathbf{x}}_n \tag{2}$$

where N is the number of data samples. To enhance the distance of the output value, we apply a `Softmax` operator on $\hat{\mathbf{x}}_n$.

2.6. Semi-supervised Stage

Our semi-supervised method is based on the mean teacher strategy with its scheme shown in Figure 2. It consists of two MobileNet (teacher and student models) with the same architecture. The student model is trained with the strong labels (get from $\mathcal{T}$) and pseudo labels (get from $\mathcal{S}$). The teacher model averages the parameters of the student models on the training steps to obtain the parameters of the teacher models, which is defined as:

$$\mathbf{\Theta}^{'} = \lambda \mathbf{\Theta}^{'} + (1-\lambda)\mathbf{\Theta} \tag{3}$$

where $\mathbf{\Theta}^{'}$ and $\mathbf{\Theta}$ are the parameters of teacher model and student model, respectively. The teacher model is used to generate the pseudo label. A random augmentation strategy is used to the unlabeled data to enhance the generality of our model. The total training loss function contains two parts, namely strong label loss function and pseudo label loss function, which is defined as

$$\mathcal{J}_{\text{total}} = \mathcal{J}_{\text{strong}} + \beta \mathcal{J}_{\text{pseudo}} \tag{4}$$

where $\mathcal{J}_{\text{strong}}$ and $\mathcal{J}_{\text{pseudo}}$ are categorical cross-entropy loss between the estimation and corresponding labels, and β is the parameter to control the importance of these two parts.

3. CONCLUSION

In this technical report, we propose a novel framework for semi-supervised acoustic scene classification. Our framework

contains four steps. The first step extract useful features for classification. The second step augments the datasets to improve the generalization of our model. The third step trains the model with labeled data in $\mathcal{D}_1$ for transfer learning. Then the model is fine-tuned on $\mathcal{D}_2$ with a specially designed semi-supervised manner.

4. REFERENCES

[1] Bill Schilit, Norman Adams, and Roy Want, "Context-aware computing applications," in *1994 first workshop on mobile computing systems and applications*. IEEE, 1994, pp. 85–90.

[2] Darius Nahavandi, Roohallah Alizadehsani, Abbas Khosravi, and U Rajendra Acharya, "Application of artificial intelligence in wearable devices: Opportunities and challenges," *Computer Methods and Programs in Biomedicine*, vol. 213, pp. 106541, 2022.

[3] Sutrisno Warsono Ibrahim, "A comprehensive review on intelligent surveillance systems," *Communications in science and technology*, vol. 1, no. 1, 2016.

[4] Jisheng Bai, Jianfeng Chen, Mou Wang, Muhammad Saad Ayub, and Qingli Yan, "A squeeze-and-excitation and transformer based cross-task model for environmental sound recognition," *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[5] Yerin Lee, Soyoung Lim, and Il-Youp Kwak, "Cnn-based acoustic scene classification system," *Electronics*, vol. 10, no. 4, pp. 371, 2021.

[6] Rohit Patiyal and Padmanabhan Rajan, "Acoustic scene classification using deep learning," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2016.