SEMI-SUPERVISED ACOUSTIC SCENE CLASSIFICATION WITH CRNN AND PROBABILISTIC PSEUDO-LABELING

Linfeng Feng[1,2]*, Zijun Huang*[2]*, Boyu Zhu*[1,2]*, Xiao-Lei Zhang*[1,2]*, Xuelong Li*[1]

[1]Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd, China
[2]School of Marine Science and Technology, Northwestern Polytechnical University, China

ABSTRACT

This technical report outlines our approach to the Semi-supervised Acoustic Scene Classification task in the IEEE ICME 2024 Grand Challenge. We developed a backbone model using a convolutional recurrent neural network (CRNN). Our training strategy consists of two stages: pre-training and fine-tuning. Initially, we pre-trained the model using the labeled dataset. In the fine-tuning stage, we utilized the pre-trained model to assign pseudo-labels to the unlabeled data. Samples with predicted peak probabilities higher than a predefined threshold were deemed reliable and combined with the labeled data. This enlarged dataset was then used for fine-tuning the model. Our experimental results demonstrate a significant improvement in performance compared to the baseline method.

Index Terms— Acoustic scene classification, semi-supervised, convolutional recurrent neural network, pseudo label

1. INTRODUCTION

Acoustic Scene Classification (ASC) refers to the task of automatically classifying an acoustic environment or soundscape based on the audio signals it produces [1]. Acoustic scene sounds contain a wealth of information and rich content. However, for ASC, the relevant information may be sparse and scattered throughout an audio clip, making accurate scene prediction challenging. As a result, ASC has been a longstanding and appealing research field for decades. ASC is an important area of research in audio signal processing, machine learning, and artificial intelligence, with applications in various fields including surveillance, environmental monitoring, multimedia content analysis, and smart devices.

The goal is to identify and label the category of a given acoustic scene, such as bus, restaurant, or park, etc. In the current task, each piece of data belongs to one of ten categories, with no instances having multiple labels [2]. Only 20% of the data is labeled, presenting a challenge for participants to utilize semi-supervised and domain adaptation methods to tackle this issue.

We described our submitted system for this task. According to [2], the baseline model, pretrained on the TAU Urban Acoustic Scenes 2020 Mobile development dataset, achieved only 14% accuracy on the Chinese acoustic scene dataset. This indicates a significant domain-shift between the two datasets. Therefore, we decided not to consider using external datasets for pre-training or domain-adaptation in this task. Instead, we focused on leveraging semi-supervised learning techniques to make full use of the unlabeled subset within this development set. Specifically, we designed a model based on convolutional recurrent neural network (CRNN). By adding pseudo-labels to the unlabeled subset, we expanded the available dataset for model training.

The subsequent sections detail our model architecture, training methods, and validation experiments. We partitioned the development set into training and validation sets to confirm the effectiveness of the proposed method. The experimental results demonstrate that the proposed systems can achieve state-of-the-art performance for ASC.

2. METHOD

In this study, inspired by [3], we propose a CRNN model, as illustrated in Fig. 1. The convolutional layers capture local receptive fields, while the Bidirectional Long Short-Term Memory (BiLSTM) layers provide global temporal context. By integrating their respective strengths, we aim to extract more comprehensive features. Segment-level features are obtained by averaging the frame-level embeddings generated by the BiLSTM outputs across the temporal dimension, followed by two fully connected layers for final output generation.

We employ log mel spectrograms as the input to our model. The Short-Time Fourier Transform (STFT) is computed using a Hanning window, with 2048 points for the Fast Fourier Transform (FFT), a frame length of 2048, and a frame shift of 1024. We apply 40 mel-filter bands followed by a base-10 logarithmic operation to compute the log mel spectrograms. Given that each audio sample has a sampling rate of 48 kHz and a duration of 10 seconds, the shape of each log mel spectrogram is $1 \times 40 \times 469$, representing 1 channel, 40 frequency bins, and 469 frames. We denote a log mel spectrogram as $\mathbf{x}$.

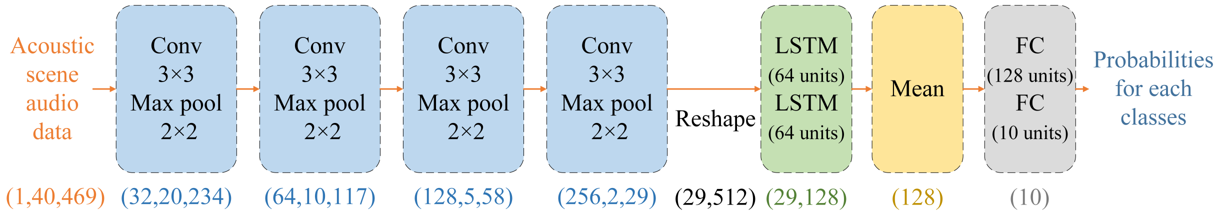

Fig. 1. Architecture of the CRNN model.

Given a labeled training set $\mathcal{X}$ and an unlabeled training set $\mathcal{U}$, first, a pre-training model is trained on $\mathcal{X}$, denoted as $f^1(\cdot)$. In the second step, samples from $\mathcal{U}$ are filtered using $f^1(\cdot)$. The criterion for filtering is based on the peak probability output by $f^1(\cdot)$. Samples with peak probabilities higher than a predefined threshold δ are considered reliable and assigned pseudo-labels, while those below this threshold are discarded. The aforementioned process can be formalized as follows:

$$\begin{aligned}\overline{\mathcal{U}} = \{(\mathbf{x}_n, \hat{y}_n) \mid \\ &\mathbf{x}_n \in \mathcal{U}, \\ &f^1(\mathbf{x}_n) = \hat{\mathbf{y}}_n, \\ &\max(\hat{\mathbf{y}}_n) > \delta, \\ &\hat{y}_n = \arg\max_i \{\hat{\mathbf{y}}_{n,i}\}_{i=1}^{I}\}\end{aligned} \tag{1}$$

The purpose of taking the argmax operation in Eq. (1) is to subsequently perform one-hot encoding on $\hat{y}_n$ during fine-tuning. Next, merging the two datasets and fine-tuning the model can be formalized as follows:

$$\mathcal{W} = \mathcal{X} \cup \overline{\mathcal{U}} \tag{2}$$

The loss function will use the cross-entropy, combined with one-hot encoding. The cross entropy loss of a single sample can be formalized as follows:

$$\mathcal{L}^{\mathrm{CE}} = -\sum_{i=0}^{I} y_i \log \hat{y}_i \tag{3}$$

3. EXPERIMENTS

3.1. Experimental setup

We shuffled the labeled development set $\mathcal{X}^{\mathrm{dev}}$, setting the random seed to 0. It was then divided into training set $\mathcal{X}^{\mathrm{tr}}$ and validation set $\mathcal{X}^{\mathrm{val}}$ in a 9:1 ratio. The training set $\mathcal{X}^{\mathrm{tr}}$ contains 1566 samples, while the validation set $\mathcal{X}^{\mathrm{val}}$ contains 174 samples. Our objective was to find the best-performing model on $\mathcal{X}^{\mathrm{val}}$. During the fine-tuning stage, the training set $\mathcal{X}^{\mathrm{tr}}$ will be merged with the filtered unlabeled set $\overline{\mathcal{U}}^{\mathrm{dev}}$.

For all experiments, we employed the AdamW optimizer [4] with a maximum of 100 training epochs. The batch size was set to 32. The learning rate was initialized at 0.001. If the validation loss did not decrease for 5 epochs, the learning rate was multiplied by 0.5, with a lower bound set to 0.0001. Training was stopped early if the validation loss on the validation set did not decrease for 20 epochs. The model with the minimum cross-entropy loss on the validation set was chosen for evaluation.

Table 1. Results on the validation set.

method	accuracy (%)	cross-entropy
baseline	98.28	0.050
pre-train	98.85	0.035
$\delta = 0.90$	98.85	0.029
$\delta = 0.95$	98.85	0.017
$\delta = 0.99$	98.85	0.036

3.2. Results

Table 1 presents the effectiveness of our proposed method. The baseline, trained under the training strategy outlined in Section 3.1, achieves a accuracy of 98.28% on the validation set, i.e. only 3 misclassified samples. Thus, in the current work, accuracy as a criterion for model selection is not sufficiently smooth, which led us to opt for using cross-entropy loss to select model. It can be observed that incorporating pseudo-labels consistently improves model performance compared to the pre-trained model. When $\delta = 0.95$, the cross-entropy loss is minimized.

4. REFERENCES

[1] Biyun Ding, Tao Zhang, Chao Wang, Ganjun Liu, Jinhua Liang, Ruimin Hu, Yulin Wu, and Difei Guo, "Acoustic scene classification: a comprehensive survey," *Expert Systems with Applications*, p. 121902, 2023.

[2] 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.

[3] Lauréline Perotin, Romain Serizel, Emmanuel Vincent, and Alexandre Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[4] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.