SCUT SYSTEM FOR IEEE ICME 2024 GRAND CHALLENGE: SEMI-SUPERVISED ACOUSTIC SCENE CLASSIFICATION UNDER DOMAIN SHIFT

Guoqing Chen, Jiaxin Tan, Yanxiong Li

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
15205923882@163.com, tanjiaxin02@126.com, eeyxli@scut.edu.cn

ABSTRACT

This report describes our submission for the IEEE ICME 2024 grand challenge: semi-supervised acoustic scene classification under domain shift. Our system is based on a semi-supervised model with a pre-trained convolutional neural network. We limit the effective receptive field of the convolutional layers of the model for improving the performance of the model. The TAU UAS 2020 Mobile development dataset is adopted to train the model with data augmentations. Experimental results show that our system obtains a macro-average accuracy of 0.973 on the official CAS 2023 development dataset.

Index terms— Acoustic scene classification, domain shift, semi-supervised learning, mix-style

1. INTRODUCTION

Acoustic scene classification (ASC) is a task to classify each input audio recording into one class of pre-given acoustic scenes. ASC systems have the potential to benefit numerous applications, including wearable devices, robotics, and smart home devices. In recent years, the challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) has received significant attention. ASC, as a core task of the DCASE challenge, has attracted considerable interest and undergone extensive research [1]-[4]. Notably, deep learning algorithms have emerged as the predominant approach for enhancing the ASC performance. However, ASC methods based on deep learning typically require substantial training data to achieve satisfactory performance, which raises two crucial considerations in the development of deep learning approaches for ASC: domain shift and scarcity of labeled data.

The issue of domain shift is common in ASC systems based on deep learning. Previous work has studied the mismatch between devices and cities in ASC [5], [6]. However, the concentration of recording cities in Europe leads to a dataset where people tend to share more similar living environments compared to other continents, such as Asia and Africa. When directly evaluating the performance of an ASC model trained on the recordings from TAU UAS 2020 Mobile Development Dataset on the recordings collected in Chinese cities, a significant drop in accuracy is observed. This underscores the challenge of applying existing datasets to environments with broader differences in culture, language, and infrastructure.

Deep learning methods require abundant data for training, and collecting a large-scale dataset is time consuming and labor-intensive. Semi-supervised learning, as an effective machine learning approach, can leverage both labeled and unlabeled data, reducing the dependence on labeled data [7].

2. THE METHOD

To address the above two issues of domain shift and scarcity of labeled data, we first pre-train an ASC model using the TAU UAS 2020 Mobile Development Dataset after performing a series of operations of data augmentation. Then, we fine-tune the pre-trained ASC model using labeled data from the development dataset of the IEEE ICME 2024 grand challenge. Next, we use the finetuned ASC model to assign pseudo-labels to unlabeled data in the development dataset. Finally, we fine-tune the CP-Res Damp [8] using the pseudo-labeled data, and obtain the final ASC model for evaluation. The following sections of this report will introduce the details of the method, results and discussions.

Our experiments are built upon the ResNet architecture explained in [9]. Initial experiments indicate that the optimal receptive field (RF) size for the task is achieved with $\rho = 4$. ρ is a coefficient that controls the maximum RF in the model. The ResNet architecture is referred to as CP-Res Damp [8]. Each convolutional neuron possesses a limited receptive field within its layer input. To mitigate the influence of neuron inputs that are farther away from the center of the receptive field, we introduce a damped CNN. In practical implementation, we achieve this damping effect by element-wise multiplication of the weights of the convolutional layer with a non-trainable constant matrix $\boldsymbol{C} \in \boldsymbol{R}^{T\times F}$, termed the "damping matrix". The damping matrix aligns with the spatial shape of the filters and functions by attenuating the impact of the outermost filter elements on the output across the frequency dimension. Consequently, in the resulting network, every convolution operation of the form $O_n = W_n * Z_{n-1} + B_n$ is replaced by $O_n = (W_n \oplus C_n) * Z_{n-1} + B_n$, where $*$ denotes the convolutional operation

and $\oplus$ denotes element-wise multiplication. Z_{n-1} is the output from the previous layer, W_n denotes the trainable weight of the filter, and B_n represents the bias. The damping matrix has a value of 1 at its center and decays linearly to reach a value of λ. In our submissions, λ equals 0.1.

3. EXPERIMENTS

In this section, we will describe experimental data, setup, and data augmentation in detail.

3.1. Experimental data

In pre-training ASC model, we conduct our experiments on the TAU Urban Acoustic Scene 2020 Mobile development dataset (TAU20) which consists of audio clips acquired by mobile devices in urban environments. The dataset includes 230,35 audio clips. Each clip is with a duration of 10 seconds and a hard label of an acoustic scene. The dataset contains audio clips from 10 cities and 9 devices: 3 real devices (A, B, C) and 6 simulated devices (S1-S6). Audio clips recorded by devices B, C, and S1-S6 are composed of audio segments that are randomly selected from simultaneous recordings. Hence, all of these audio clips overlap with the audio clips from device A, but not necessarily with each other. The total amount of audio clips in the development dataset is 64 hours. There are 10 classes of acoustic scenes, including Airport, Metro station, Indoor shopping mall, Pedestrian street, Public square, Street with medium level of traffic, Travelling by a tram, Travelling by a bus, Travelling by an underground metro, and Urban park. Afterwards, the pre-trained ASC model is further fine-tuned using the official CAS 2023 dataset.

3.2. Experimental Setups

In pre-training ASC model, audio clips of TAU20 are split into frames via a Hamming window whose length is 2048 with 1/2 overlapping. Short-time Fourier transform is then performed on each frame for obtaining linear power spectrum which is smoothed with a bank of triangular filters for extracting log Mel-spectrum. In addition, the delta coefficients of log Mel-spectrum are calculated and concatenated with the log Mel-spectrum to form the input audio feature. We train the models for 100 epochs using the Adam optimizer [10] with batch size to 16. The learning rate is scheduled to linearly increase from 0 to 0.001 in ten epochs as a warmup, then decay to 0 with cosine annealing for the rest of epochs.

During the fine-tuning of the system, we utilize log Mel spectrogram as input feature, initially resampling all recordings to 44,100 Hz. We extract the spectrogram using the short-time Fourier transform with a Hanning window of 40 ms and a hop size of 20 ms. Subsequently, we apply 64 Mel-filter bands to the spectrograms, followed by a logarithmic operation to obtain the log Mel spectrogram. Each log Mel spectrogram has a shape of 500 × 64, where 500 represents the number of time frames and 64 denotes the number of frequency bins. For fine-tuning the baseline, we employ the Adam optimizer with a learning rate of 0.001 and a batch size of 4.

3.3. Data Augmentation

In pre-training ASC model, to prevent overfitting and enhance robustness, we employ various data augmentation methods during training in the time-frequency domain. These methods included:

• Mix-style: Introduced by Zhou et al. [11], Mix-style is an approach for manipulating instance-level feature statistics. It relies on probabilistic mixing of cross-source domain training samples. The application of Mix-style can be adjusted using the parameter p. The parameter p controls the likelihood of its application to a batch of recordings. Through careful parameter tuning, we achieve better performance for the model.

• SpecAugment [12]: SpecAugment is a widely used data augmentation technique in ASC, which encompasses functional warping, frequency channel masking blocks, and timestep masking blocks. We implement two masking lines for each dimension, with a maximum thickness of 2 for each line.

• Spectrum Modulation: In line with the DCASE 2022 challenge submission, where spectrum modulation is proved to be effective [13], we adopt the same method. As most of the provided datasets are recorded using device A, resulting in an imbalance of data, we address this issue by introducing a frequency energy difference to the data recorded by non-device A.

4. RESULTS

The evaluation metric for this competition is the macro-average accuracy, a commonly utilized metric in previous ASC challenges. This metric is computed as the average of class-wise accuracies. We evaluate and compare the performance of both the improved system and the baseline system on the development dataset. The results are given in Table 1.CP-Res system refers to the original ResNet architecture that did not introduce Damp. 1-acc refers to the macro-average accuracy of the pre-trained ASC model fine-tuned using labeled data from the development dataset of the IEEE ICME 2024 Grand Challenge. 2-acc refers to the macro-average accuracy of the model further fine-tuned by us using pseudo-labeled data.

Table 1 Results obtained by baseline and our systems

system	1-acc	2-acc
Basline	0.936	0.956
CP-Res system	0.976	0.965
CP-Res Damp system	0.972	0.973

Figure 1 and Figure 2 respectively show the highest

accuracy for each class in the corresponding validation data of the submitted system before and after fine-tuning the model with pseudo-labeled data. Among these ten acoustic scenes, the traffic street is the most likely to be confused with other scenes due to the diversity of sounds.

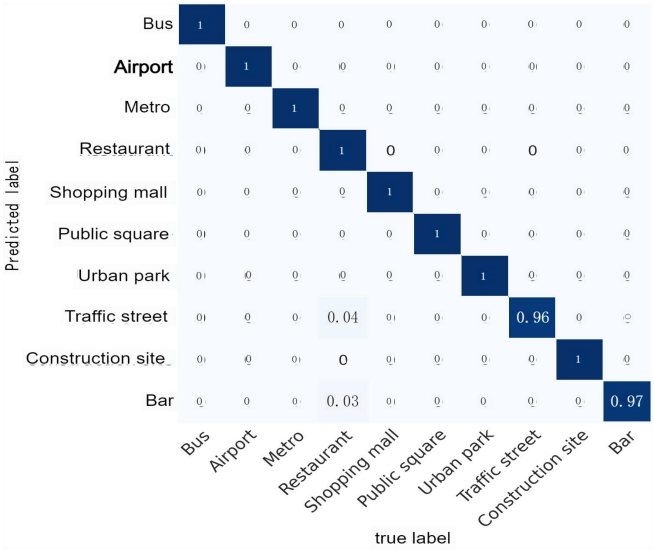

Figure 1: Confusion matrix of the pre-trained ASC model fine-tuned using labeled data from the development dataset on validation data.

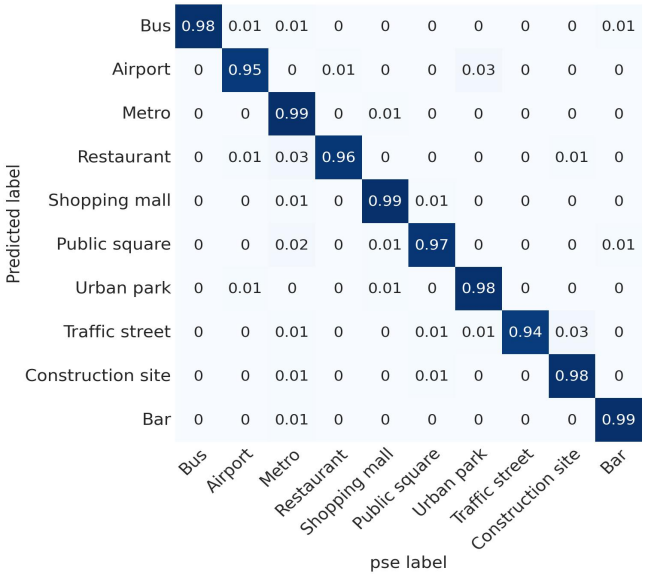

Figure 2: Confusion matrix of the ASC model further fine-tuned using pseudo-labeled data on validation data.

5. CONCLUSIONS

In this report, we described a system designed for the IEEE ICME 2024 grand challenge: semi-supervised acoustic scene classification under domain shift. The system we constructed was based on a semi-supervised model with a pre-trained convolutional neural network. We attempted to improve the performance of this system by applying data augmentation. The performance of the submitted system on the development dataset was improved compared to the baseline system.

6. REFERENCES

[1] J. Tan and Y. Li, , “Low-complexity acoustic scene classification using blueprint separable convolution and knowledge distillation,” in Tech. Rep of DCASE2023 Challenge, pp. 1-4.

[2] Y. Li, X. Li, Y. Zhang, W. Wang, M. Liu and X. Feng, “Acoustic scene classification using deep audio feature and BLSTM network,” in Proc. of ICALIP, 2018, pp. 371-374.

[3] W. Xie, Q. He, Z. Yu and Y. Li, “Deep mutual attention network for acoustic scene classification,” Digital Signal Processing, vol.123, 103450, 2022.

[4] H.K. Chon, Y. Li, W. Cao, Q. Huang, W. Xie, W. Pang and J. Wang, “Acoustic scene classification using aggregation of two-scale deep embeddings,” in Proc. of IEEE ICCT, 2021, vol. 4, pp. 1341-1345.

[5] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification." arXiv preprint arXiv:1807.09840, 2018.

[6] Y. Tan, H. Ai, S. Li, and M. D. Plumbley, "Acoustic scene classification across cities and devices via feature disentanglement." IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 1286-1297, 2024.

[7] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments." arXiv preprint arXiv:1807.10501, 2018.

[8] J. Bai, M. Wang, H. Liu, et al, "Description on IEEE ICME 2024 Grand Challenge: Semi-supervised Acoustic Scene Classification under Domain Shift." arXiv preprint arXiv:2402.02694, 2024.

[9] K. Koutini, F. Henkel, H. Eghbal-zadeh, and G. Widmer. "CP-JKU submissions to dcase'20: low-complexity cross-device acoustic scene classification with rf-regularized CNNs." in DCASE2020 Challenge, Technical Report, June 2020.

[10] D. Kingma, and J. Ba, “Adam: A method for stochastic optimization.” in *Proc. of ICLR*, 2015, pp. 1-15.

[11] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, “Domain generalization with mixstyle.” in Proc. of International Conferenceon Learning Representations, 2021.

[12] Y. Li, W. Cao, W. Xin Xie, Q. Huang, W. Pang, and Q. He, "Low-complexity acoustic scene classification using data augmentation and lightweight resnet." in Proc. of 2022 16th IEEE International Conference on Signal Processing (ICSP), vol. 1, pp. 41-45, 2022.

[13] R. Sugahara, R. Sato, M. Osawa, Y. Yuno, and C. Haruta, "Self-ensemble with multi-task learning for low-complexity acoustic scene classification." in Technical Report of DCASE2022 Challenge, , June 2022.