ACOUSTIC SCENE CLASSIFICATION BASED ON MULTI-TEACHER KNOWLEDGE DISTILLATION AND SERFR-CNN

Technical Report

Hongbo Fei, Xing Li, Jie Jia

vivo Mobile Commun co Ltd, China
feihongbo@vivo.com

ABSTRACT

In this technical report, we describe our low-complexity acoustic scene classification algorithm submitted in DCASE 2023 Task 1a. We focus on knowledge distillation strategy and network innovation, multi-teacher knowledge distillation method and SERFR-CNN is proposed, which aims at the problems of insufficient classification accuracy and adaptability of current models. Based on traditional knowledge distillation method, combined with the model ensemble strategy, and then t multi-teacher knowledge distillation method is proposed. In terms of audio feature extraction, we use Log-Mel spectrograms and Time-frequency masking algorithm. In order to further improve system performance, virtual data generation technology is adopted. Finally, use the trained model for transfer learning. By using proposed systems, we achieved a classification accuracy of 59.3% on the officially provided evaluation dataset, which is 16.4% over than the baseline system.

Index Terms— Low-complexity acoustic scene classification, multi-teacher knowledge distillation, SERFR-CNN, Time-frequency masking algorithm

1. INTRODUCTION

Audio carry a large amount of life scenes and physical events in the city [1], which plays an important role in our life. From the perspective of human cognition, auditory cognition is an important part of artificial intelligence. In the study of cognitive science, auditory cognition is often regarded as the second perception system second only to vision. Obviously, auditory cognition, as an important way to perceive the environment, its research value and development potential are self-evident. Acoustic scene classification(ASC) aims to classify sounds into one of predefined classes [2]. The audio scene classification competition and related conferences are also in full swing with the development of ASC. The DCASE Challenge was organized and launched by the University of London Queen Mary College Digital Music Center and Tampere University of Technology in 2013. It is currently the most authoritative competition in the field of acoustic events. Since 2016, the DCASE Challenge has been accompanied by a seminar which is held once a year, and many experts and scholars participate in it every year.

A high-quality dataset is an important prerequisite for testing whether the sound scene classification system is excellent. The DCASE challenge releases new dataset every year. From the DCASE 2016 challenge to the DCASE 2017 challenge, the length of each audio sample has been reduced from 30s to 10s [3]. The dataset released by the DCASE 2018 Challenge records high-quality binaural audio from 6 European cities as samples [4]. The dataset released by the DCASE 2019 Challenge ensures that each audio sample is recorded by the same device. With the release of the DCASE 2020 challenge, audio samples recorded by multiple devices and scenes are added to the data set for the first time, thereby further improving the quality of the data set. From the DCASE 2021 challenge to the DCASE 2022 challenge, the length of each audio sample has been reduced from 10s to 1s. The competition also sets limits on the complexity and computation required to submit models.

In this report, we introduce an low-complexity acoustic scene classification model based on SERFR-CNN, which is a more powerful convolutional neural network model improved by RFR-CNN[5, 6]. As for knowledge distillation strategy, multi-teacher knowledge distillation method is used to improve the generalization ability of low-complexity models.

2. DATA PREPROCESSION

This section describes our method of converting audio samples into acoustic features, and the method of data enhancement by generating virtual samples after feature stitching is completed.

2.1. Acoustic Feature

The sampling rate of the audio samples is the original 16000Hz, the length of the Fourier change window is set to 2048-samples and the frame-shift is set to 512-samples, and then each audio sample is divided into 28 frames.

Using the short-time Fourier transform to obtain the spectrogram after the audio samples that have been framed, and then pass through the 256-bit Mel filter bank, and finally take the logarithmic processing to obtain (N, 28, 256) shape Log-Mel spectrogram features [7].

$$mel(f) = \frac{1000}{log\,2} log(1 + \frac{f}{1000}) \quad (1)$$

2.2. Data Augmentation

In recent years, the number of samples in the dataset released by the DACSE challenge has increased year by year, but in order to improve the generalization ability of the model, it is not enough to just use the given samples. Therefore, the use of data augmentation to generate additional virtual data has gradually become the mainstream.

The mixup method is a form of neighborhood risk minimization [8]. This is an unconventional data enhancement method. Its principle is to extract additional virtual samples from the neighborhood distribution of training samples to expand the support of the sample distribution. The training distribution is expanded by fusing linear interpolation of feature vectors.

Mixup augmented data is obtained as follows:

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j \quad (4)$$

$$\tilde{y} = \lambda y_i + (1-\lambda)y_j \quad (5)$$

where (x_i, y_i) and (x_j, y_j) are two acoustic scenes randomly chosen from the training data and $\lambda \in (0,1)$. λ is acquired from the beta distribution and $\beta \in (0.1, 0.9)$

3. NETWORK FRAMEWORK

In this part, firstly we introduce a SERFR-CNN that we propose as student model, and then shows multi-teacher knowledge distillation method [9,10] which uses CNN mobilev2 and BCResNet model. The ensemble strategy is applied to help student model learn more detail information in 1s audio.

3.1. Teacher model

We build a model based on MobileNetv2, which is known for audio classification task. Meanwhile the teacher model is pre-trained on AudioSet with great generalization performance. Then fine-turning the model to DCASE 2020 and DCASE 2023 dataset will be efficient.

Except for MobileNetv2, we also applied CNN and BCResnet model as teacher model. In the course of the experiment we found that different teacher models show different effect on 10 classes and 11 devices. Based on the above conclusions, We tend to combine different models through ensemble, which helps student model learn more information from different teacher models. Teacher models are trained on 10s dataset reconstituted by 1s dataset(TAU Urban Acoustic Scenes 2023 Mobile develop dataset), which helps teacher models learn more acoustic feature information and get higher accuracy to teach students model.

3.2. Student model

We build a model based on Receptive Field Regularized Convolutional Neural Network (RFR-CNN) which performed well in DCASE dataset. In addition, we introduced the channel attention mechanism by adding the se module. In the course of the experiment we found that channel attention strategy can effectively improve the classification accuracy of high noise audio.

Table 1: The architecture of SERFR-CNN

Layer	Description
Convolution(1)	5×5 conv-32-BN-RELU
Pooling	2×3 average pool
RFR Block1 (1)	$\begin{bmatrix}1\times1\text{ conv}\\3\times3\text{ conv}\end{bmatrix}\times3$
SE Block	-
Transition Layer (1)	1×1 conv + RES 2×2 max pool
RFR Block2 (2)	$\begin{bmatrix}1\times1\text{ conv}\\1\times1\text{ conv}\end{bmatrix}\times3$
SE Block	-
Transition Layer (2)	1×1 conv + RES 2×2 max pool
RFR Block3 (3)	$\begin{bmatrix}1\times1\text{ conv}\\1\times1\text{ conv}\end{bmatrix}\times3$
SE Block	-
Transition Layer (3)	1×1 conv + RES 2×2 max pool
Convolution(2)	1×1 conv-256
Pooling	Global Average pool
Dense	Dense (10, activation='softmax')

4. EXPERIMENTS AND RESULTS

4.1. Datasets

The dataset for this task is TAU Urban Acoustic Scenes 2023 Mobile. The dataset contains recordings from 12 European cities in 10 different acoustic scenes using 4 different devices. Additionally, synthetic data for 11 mobile devices was created based on the original recordings. Of the 12 cities, two are present only in the evaluation set.

4.2. Training strategy

We use the officially provided fold 1 procedure to evaluate our systems' performance. Then the systems are retrained on the whole development data for submission. The train set is split into the train and evaluation set. The classifiers were trained on the trainset in maximum 800 epochs. We use Adam [24] with a weight decay of 0.001 and a learning rate schedule: for the first 300 epochs the learning rate is exponentially increased to 0.001, followed by a linear decrease to 1e-5 until epoch 800.

4.3. Result

Results of experiments of various teacher models on the 10s fold 1 evaluation set is described in Table 2.

Analyzing the experimental results in Table 2, it can be found that the accuracy of CNN is higher than the Mobilev2 and BCResNet. Meanwhile the accuracy after ensemble can increase 3.28% compared to the best single model.

Student models with different mounts of Parameters are applied to test the accuracy of SERFR-CNN. Aiming to build a low-complexity network with less Parameters and computation

and great classification precision. Based on the above, we provide four system to the constitutor.

Submission 1: student model is SERFR-CNN-32 and 16kHz dataset is trained, mixup and KD strategy is applied

Submission 2: student model is SERFR-CNN-24 and 16kHz dataset is trained, mixup and KD strategy is applied

Submission 3: student model is SERFR-CNN-32 and 16kHz dataset is trained, mixup, time stretching and KD strategy is applied

Submission 4: student model is SERFR-CNN-24 and 16kHz dataset is trained, mixup, time stretching and KD strategy is applied

Table 2: The results of teacher models experiments on 10s dataset

Model	Classification accuracy (%)
Modilev2	72.16
CNN	77.34
BCResNet	70.74
Ensemble	80.62

Table 3: The results of submit system experiments on 1s dataset

system	accuracy (%)	Log loss
Baseline	42.9	1.575
Submission 1	59.3	1.145
Submission 2	56.7	1.204
Submission 3	58.0	1.231
Submission 4	55.4	1.284

5. CONCLUSIONS

In this technical report, we proposed a novel low-complexity network acoustic scene classification model based on SERFR-CNN. Combined with multi-teacher knowledge distillation method, the best reliable result achieves 59.3% at the time when this technical report is submitted.

6. REFERENCES

[1] Z. Huang, C. Liu, and H. Fei, et al. Urban sound classification based on 2-order dense convolutional network using dual features [J]. Applied Acoustics 2020, 164.

[2] A. Mesaros, T. Heittola, and T. Virtanen, “Tut database for acoustic scene classification and sound event detection,” pp.1128–1132, 2016.

[3] A. Mesaros, T. Heittola, and T. Virtanen, “A multi-device dataset for urban acoustic scene classification,” IEEE AASP Challenge on DCASE 2018 Technical Report, 2018.

[4] H. Yang, C. Shi, H. Li, “Acoustic scene classification using CNN ensembles and primary ambient extraction,” IEEE AASP Challenge on DCASE 2019 Technical Report, 2019.

[5] K. Koutini, H. Eghbal-zadeh, and G. Widmer, “Receptive field regularization techniques for audio classification and tagging with deep convolutional neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1987–2000, 2021.

[6] K. Koutini, H. Eghbal-zadeh, M. Dorfer, and G. Widmer, “The Receptive Field as a Regularizer in Deep Convolutional Neural Networks for Acoustic Scene Classification,” in Proceedings of the European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2019.

[7] T. Virtanen, M. D. Plumbley, D. Ellis, Computational Analysis of Sound Scenes and Events. [M] p. 76–78.

[8] H. Zhang, M. Cisse, Y. N. Dauphin,and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” in arXiv: 1710.09412, 2017.

[9] G. E. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” CoRR, vol. abs/1503.02531, 2015.

[10] J. Ba and R. Caruana, “Do deep nets really need to be deep?” in Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, 2014, pp. 2654–2662.