Data-Efficient Low-Complexity Acoustic Scene Classification Using Parallel Attention Broad-cast-Residual Network

Technical Report

Guoqing Chen

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
15205923882@163.com

Yanxiong Li[†]

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
eeyxli@scut.edu.cn

ABSTRACT

This technical report describes our proposed system for Task 1 of Detection and Classification of Acoustic Scenes and Events (DCASE) 2024. We propose a data-efficient low-complexity acoustic scene classification method, which utilizes a parallel attention broad-residual network that consists of four parts (i.e., the modules of pre-processing, fusion, global and local contextual information extraction). We integrate the broadcast residual learning into the network to enhance its ability for extracting local contextual information. To further improve accuracy and reduce complexity, we integrate other techniques into our method, such as knowledge distillation, data augmentation, adaptive residual normalization, and quantization-aware training. There are five training subsets that contain approximately 5%, 10%, 25%, 50%, and 100% of the audio snippets in the training dataset. Using a subset of the five training subsets above as training data to construct a system, we obtain five systems. The accuracy scores obtained by these five systems on the evaluation samples of the development dataset are 47.14%, 52.38%, 58.04%, 60.88%, and 63.7% respectively.

***Index Terms*—** Broadcast residual learning, attention mechanism, knowledge distillation, data augmentation, adaptive residual normalization

1. INTRODUCTION

The task of acoustic scene classification (ASC) is to classify each audio sample into one pre-defined class of acoustic scenes. ASC technique is beneficial for numerous applications, such as wearable devices, robotics, and smart home devices. In recent years, the DCASE challenge has received significant attention. ASC, as a main task of the DCASE challenge, has attracted considerable interest and undergone extensive research [1]-[5]. It is a critical pre-processing step for many audio-video tasks, such as sound event detection [6]-[9], video content analysis [10], [11], speaker diarization and recognition [12]-[15].

The methods using convolutional neural network (CNN) and its variants are dominant solutions for low-complexity ASC. Tan et al [16] designed a CNN with blueprint separable convolution [17]. The model with convolutional architecture is good at capturing local contextual information (LCI) from input audio samples, but lacks the ability to effectively extract global contextual information (GCI). These two kinds of information above are complementary to each other and have been proved to be beneficial for improving the performance in other audio processing tasks [18]. Therefore, based on [16], we design a parallel attention broad-residual network (PABRN) to extract both GCI and LCI and fuse them to improve the classification performance of the ASC method. The structure of PABRN is similar to that of parallel attention-convolution network (PACN) in [19], but they differ in the branches used to extract LCI.

This technical report describes our work for Task 1 of DCASE 2024. The rest of this report is organized as follows. Section 2 introduces our method for data-efficient low-complexity ASC. Experiments on the development data are presented in Section 3 and conclusions are drawn in Section 4.

2. THE METHOD

The basic steps of the proposed method are as follows. A large-size teacher model is first trained using audio clips in the training dataset. Then, a small-size student model is generated under the guidance of the pretrained teacher model. Namely, the knowledge distillation (KD) is used to train the student model which is learned from the teacher model. In addition, the data augmentation (DA) is applied to audio clips to increase the data diversity for training both teacher and student models. The student model is designed as a PABRN, while the teacher model is an ensemble model of Patchout faSt Spectrogram Transformer (PaSST) [20] and PACN.

2.1. Parallel Attention Broad-Residual Network

The framework of the proposed PABRN is illustrated in Fig. 1, which comprises four key components: pre-processing module, LCI extraction module, GCI extraction module and fusion module. By simultaneously utilizing GCI and LCI, PABRN aims to enhance the ASC performance. Specifically, it has two computationally efficient modules to capture these two types of information. The preprocessing module transforms the log-Mel spectrogram of each audio clip into features suitable for GCI and LCI

[†] Corresponding author (eeyxli@scut.edu.cn)

This work was partly supported by the national natural science foundation of China (62371195, 62111530145, 61771200), international scientific research collaboration project of Guangdong (2023A0505050116), Guangdong basic and applied basic research foundation (2022A1515011687), and Guangdong provincial key laboratory of human digital twin (2022B1212010004).

extraction. Finally, the fusion module integrates the extracted GCI and LCI, leveraging both for obtaining accurate classification results. The difference between the proposed PABRN and the PACN in [19] is the LCI extraction module. Hence, we describe the LCI extraction module of the PABRN here, and detailed descriptions of other three modules can be found in [19].

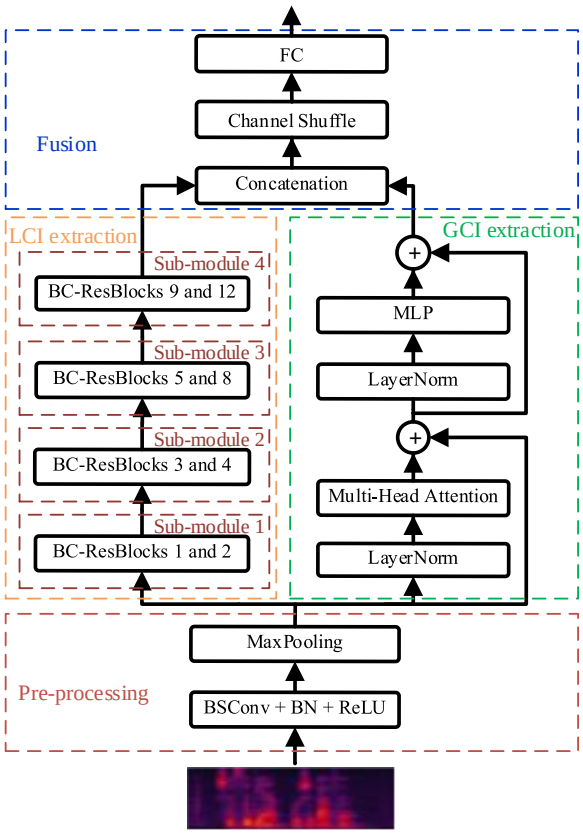

Fig. 1 The framework of the proposed parallel attention broadcast-residual network. BSConv: Blueprint separable convolution; ReLU: Rectified Linear Unit; ARN: Adaptive residual normalization; BC-ResBlocks: Broadcast residual blocks; MLP: Multilayer perceptron; FC: Fully-connected; LCI: Local contextual information; GCI: Global contextual information.

Broadcast residual learning [21] extracts two feature maps specific to frequency and temporal dimension through frequency-wise 2D and temporal-wise 1D convolution. Inspired by the success in [22], we design a new LCI extraction module based on BC-ResBlocks to replace the LCI extraction module in the PACN [18]. The LCI extraction module in the PABRN consists of twelve BC-ResBlocks. As illustrated in Fig. 1, these twelve BC-ResBlocks are contained in four sub-modules. Sub-modules 1, 2, 3 and 4 contains BC-ResBlocks 1 to 2, BC-ResBlocks 3 to 4, BC-ResBlocks 5 to 8, and BC-ResBlocks 9 to12, respectively.

The parameters settings of these twelve BC-ResBlocks are presented in Table 1. Features will undergo repeated averaging and expansion operations in each sub-module for effectively extracting LCI from log Mel-spectrum.

Table 1: The parameters settings of BC-ResBlocks in the LCI extraction modules. Each row is a sequence of one or more identical BC-ResBlocks repeated n times with the input shape of channel × frequency × time (i.e., $C\times F\times T$), total time steps W, and the number of output channels c. Changes in number of channels and down-sampling by stride s belong to the first block of each sequence of BC-ResBlocks. The temporal convolutions in all BC-ResBlocks use dilation of d.

Input	Sub-module	n	c	s	d
$C\times F\times T$	1	2	$C/2$	1	1×1
$C/2\times F\times T$	2	2	$3C/2$	2,1	(1, 2)
$3C/2\times F/2\times T$	3	4	C	2,1	(1, 4)
$C\times F/4\times T$	4	4	C	1	1×8

2.2. Adaptive Residual Normalization

The adaptive residual normalization (ARN) technique dynamically adjusts the weights of residual connections during the training process. As a result, the model can adapt to training data and optimization objectives, and thus the generalization performance and stability of the model can be improved. Batch-instance normalization (BIN) [23] is applied along the frequency dimension for generating device-specific features. Trainable parameters are incorporated for controlling the trade-off between normalization for different devices and frequencies. The introduction of ARN [24] allows the network to learn how to normalize input data based on specific tasks and input distributions.

2.3. Data Augmentation

In order to prevent overfitting and enhance robustness, we employ various DA methods during training in the time-frequency domain. These DA methods are presented as follows.

• Mix-style: It is an approach for manipulating instance-level feature statistics [25]. 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 can achieve better performance.

• SpecAugment: It encompasses functional warping, frequency channel masking blocks, and timestep masking blocks [26]. We implement two masking lines for each dimension, with a maximum thickness of 2 for each line.

• Spectrum Modulation: In one system of the DCASE 2022 challenge [27], spectrum modulation is proved to be effective. we adopt the same method of spectrum modulation. 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.

2.4. Knowledge Distillation

KD is proved to be effective for the ASC task for maintaining the performance even with low complexity. Using the same DA method on the five officially provided training subsets of different sizes, we train five high-complexity PACN models and five PaSST models respectively. The PACN and PaSST models trained on the same subset are grouped together as the teacher models corresponding to that subset. Student models use PACN and PABRN models with different parameter configurations. Finally, we obtain 20 models on five training subsets and each training subset is used to generate four models.

2.5. Quantization Aware Training

To reduce the number of model parameters, we employ the quantization-aware training (QAT) approach [28] which performs floating-point calculations during training. However, the QAT approach simulates the effect of INT8 through a fake quantization module that includes clamping and rounding. We utilize the QAT plugin provided by the PyTorch-lightning library for quantization-aware training. We use default values for all QAT settings and applied layer fusion to all convolution-batch normalization-ReLU sequences in the model. After completing QAT, we apply quantization to fix the model parameter variable type to INT8 and perform inference.

3. EXPERIMENTS

This section describes experimental setups and results in detail.

3.1. Experimental Setups

The development set contains data from 10 cities and 9 devices: 3 real devices (A, B, and C) and 6 simulated devices (S1-S6). Data from devices B, C, and S1-S6 is composed of randomly selected segments from the simultaneous recordings. Therefore, all overlap with the data from device A, but not necessarily with each other. The total amount of audio in the development set is 64 hours. Unlike previous competitions, participants are required to develop systems for five increasingly challenging scenarios that progressively limit the available training data. The organizers provide five predefined subsets/splits of the development-training dataset, which are 100%, 50%, 25%, 10%, and 5% of the original development-training set size. The 100% subset includes all segments of the development-training split.

Audio clips are split into frames by a Hamming window whose length is 4096 with 1/6 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. The final size of input audio feature is: 256×65×2, where 256, 65 and 2 denote numbers of frequency-band, frame and channel, respectively.

As shown in Table 2, we design three systems on five training subsets with various sizes. We train the models for 100 epochs by the Adam optimizer [29] with batch size of 16. Learning rate is set to linearly increase from 0 to a value in the first ten epochs, and then decays to 0 with cosine annealing for the rest epochs.

Table 2: The differences and complexities of the three different systems. *Dim* represents the feature dimension used to train the model. *Mlp* represents the expansion ratio of the layer of multi-layer perceptron. *Sta* represents the number of stages of the LCI extraction module based on broadcast residual learning. PN and MACs denote parameter number and multiply-accumulate operations and multiply-add operations, respectively.

Model	*Dim*	*Mlp*	*Sta*	PN (kilo)	MACs (M)
BC-PACN-48	48	2	4	69.78	10.06
BC-5-PACN-48	48	2	5	122.29	11.21
BC-PACN-64	64	2	4	117.87	16.59

As shown in Table 3, the configuration of the training parameters of the system is adjusted on different training subsets.

Table 3: The system has adjusted the basic hyperparameters for five training subsets with different sizes. *LR* stands for learning rate, while λ and T represent the weight coefficient and temperature coefficient of KD, respectively.

Subset	5%	10%	25%	50%	100%
LR	0.001	0.001	0.002	0.003	0.005
λ	0.226	0.22	0.2	0.18	0.165
T	1.5	1.5	1.5	2	2

3.2. Experimental Results

As shown in Table 4, we train three different models using each of five training subsets, including 5%, 10%, 25%, 50%, and 100%, respectively, and obtain a total of 15 models. The teacher models that are used for generating the student models are also trained on the corresponding training subsets only. We then evaluated these models on the validation set of the development dataset, which contain 29,680 audio clips. We calculate the overall accuracy of each model and compare it to the baseline system. The performance of the models trained on each subset is improved. Among them, the five models that are trained based on the BC-PACN-64 system have the best performance.

Table 4: on set. The accuracy obtained by the baseline and three proposed systems evaluated on the development validati

Subset	5%	10%	25%	50%	100%
Baseline	42.40%	45.29%	50.29%	53.19%	56.99%
BC-PACN-48	46.44%	52.61%	57.41%	59.75%	61.95%
BC-5-PACN-48	46.53%	53.13%	57.05%	60.18%	61.55%
BC-PACN-64	47.14%	52.38%	58.04%	60.88%	63.70%

Fig. 2 shows the accuracy of the best results for each class obtained by the BC-PACN-64 system on the 100% training subset. Although the accuracy scores for *bus*, *park*, and *street traffic* are relatively high, the *street pedestrians* are similar to other scenes due to the diversity and complexity of sounds, making them the most likely to be confused with other scenes.

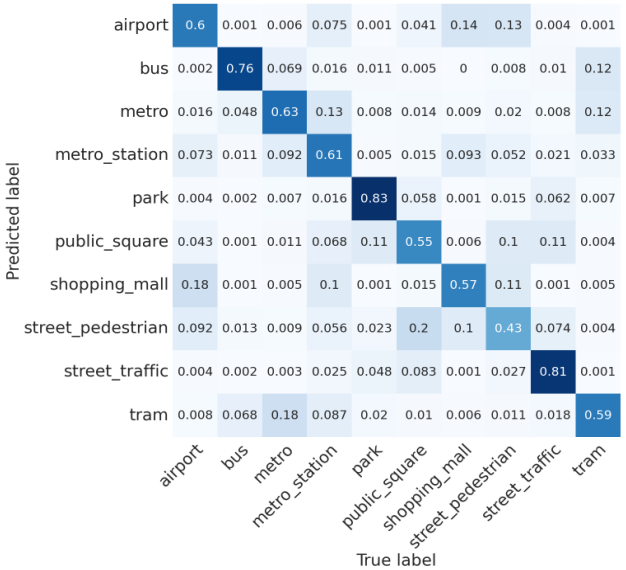

Fig. 2: Confusion matrix of the BC-PACN-64 system on the 100% training subset.

4. CONCLUSIONS

In this technical report, we describe three systems for Task 1 of the DCASE challenge 2024. We propose a data-efficient low-complexity ASC method primarily based on a PABRN that simultaneously extracts both LCI and GCI. We integrate many techniques, such as KD, DA, ARN, and QAT, into our method to enhance the system performance. Our method achieved classification accuracy higher than the baseline method while meeting complexity requirements.

5. REFERENCES

[1] Y. Li, M. Liu, W. Wang, Y. Zhang and Q. He, "Acoustic scene clustering using joint optimization of deep embedding learning and clustering iteration," *IEEE TMM*, vol. 22, no. 6, pp. 1385-1394, 2020.

[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] W. Xie, Q. He, H. Yan, and Y. Li, "Acoustic scene classification using deep CNNs with time-frequency representations," in *Proc. of IEEE ICCT*, 2021, vol. 4, pp. 1325-1329.

[6] Y. Li, Q. Wang, X. Li, X. Zhang, Y. Zhang, A. Chen, Q. He, and Q. Huang, "Unsupervised detection of acoustic events using information bottleneck principle," *Digital Signal Processing*, vol. 63, pp. 123-134, 2017.

[7] Z. Lin, Y. Li, Z. Huang, W. Zhang, Y. Tan, Y. Chen, and Q. He, "Domestic activities clustering from audio recordings using convolutional capsule autoencoder network," in *Proc. of IEEE ICASSP*, 2021, pp. 835-839.

[8] Y. Li, X. Li, Y. Zhang, M. Liu, and W. Wang, "Anomalous sound detection using deep audio representation and a BLSTM network for audio surveillance of roads," *IEEE Access*, vol. 6, pp. 58043-58055, 2018.

[9] Y. Li, M. Liu, K. Drossos, and T. Virtanen, "Sound event detection via dilated convolutional recurrent neural networks," in *Proc. of IEEE ICASSP*, 2020, pp. 286-290.

[10] W. Pang, Q. He, Y. Li, and N. Ahmed, "Detecting video anomalies by jointly utilizing appearance and skeleton information," *Expert Systems with Applications*, vol. 246, pp. 1-12, 2024.

[11] W. Pang, W. Xie, Q. He, Y. Li, and J. Yang, "Audiovisual dependency attention for violence detection in videos," *IEEE TMM*, vol. 25, pp. 4922-4932, 2023.

[12] Y. Li, W. Wang, M. Liu, Z. Jiang, and Q. He, "Speaker clustering by co-optimizing deep representation learning and cluster estimation," *IEEE TMM*, vol. 23, pp. 3377-3387, 2021.

[13] Y. Chen, G. Cheng, R. Yang, P. Zhang, and Y. Yan, "Interrelate training and clustering for online speaker diarization," *IEEE/ACM TASLP*, vol. 32, pp. 1352-1364, 2024.

[14] Y. Li, Z. Jiang, Q. Huang, W. Cao, and J. Li, "Lightweight speaker verification using transformation module with feature grouping and fusion," *IEEE/ACM TASLP*, vol. 32, pp. 794-806, 2024.

[15] Y. Li, Z. Jiang, W. Cao and Q. Huang, "Speaker verification using attentive multi-scale convolutional recurrent network," *Applied Soft Computing*, 2022, vol. 126, 109291, pp. 1-11.

[16] J. Tan and Y. Li, "Low-complexity acoustic scene classification using blueprint separable convolution and knowledge distillation," in *Tech. Rep of DCASE2023 Challenge*, 2023, pp. 1-4.

[17] Z. Li, Y. Liu, X. Chen, H. Cai, J. Gu, Y. Qiao, and C. Dong, "Blueprint separable residual network for efficient image super-resolution," in *Proc. of IEEE/CVF* CVPRW, 2022, pp. 832-842.

[18] Y. Li, H. Chen, W. Cao, Q. Huang and Q. He, "Few-shot speaker identification using lightweight prototypical network with feature grouping and interaction," *IEEE TMM*, vol. 25, pp. 9241-9253, 2023.

[19] Y. Li, J. Tan, G. Chen, J. Li, Y. Si and Q. He, "Low-complexity acoustic scene classification using parallel attention-convolution network," in *Proc. of Interspeech*, 2024, pp. 1-5.

[20] K. Koutini, J. Schluter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. of Interspeech*, 2022, pp. 2753-2757.

[21] B. Kim, S. Yang, J. Lee, and D. Sung, "Broadcasted residual learning for efficient keyword spotting," in *Proc. of INTERSPEECH*, 2021, p. 4538-4542.

[22] J. Lee, J. Choi, P. Byun, and J. Chang, "HYU submission for the DCASE 2022: fine-tuning method using device-aware data-random-drop for device-imbalanced acoustic scene classification," in *Tech. Rep of DCASE2022 Challenge*, 2022, pp. 1-4.

[23] H. Nam, and H.E. Kim, "Batch-instance normalization for adaptively style-invariant neural networks", in *Neural Information Processing Systems*, 2018, pp 2558-2567.

[24] B. Kim, S. Yang, J. Kim, and S. Chang, "QTI Submission to DCASE 2021: residual normalization for device-imbalanced acoustic scene classification with efficient design", in *Tech. Rep of DCASE2021 Challenge*, 2021, pp.1-5.

[25] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *Proc. of International Conference on Learning Representations*, 2021.

[26] 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 IEEE International Conference on Signal Processing (ICSP)*, vol. 1, pp. 41-45, 2022.

[27] 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.

[28] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, p. 2704-2713.

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