LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION USING BLUEPRINT SEPARABLE CONVOLUTION AND KNOWLEDGE DISTILLATION

Technical Report

Jiaxin Tan

School of Electronic and Information
Engineering, South China University of
Technology, Guangzhou, China
tanjiaxin02@126.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 in Detection and Classification of Acoustic Scenes and Events (DCASE) 2023. We design a teacher model based on blueprint separable Convolution (BSConv) with reference to the middle layer of the blueprint separable residual network. To meet the requirements of system complexity, we adopt knowledge distillation to teach student models from teacher model. Data augmentations (e.g., Mixstyle, SpecAugment, and spectrum modulation) are applied to prevent overfitting. When evaluated on the development data, one of the proposed systems obtains the accuracy score of 54.9% and has 73,386 parameters with 13.18 million multiply-and-accumulate operations.

Index Terms— Knowledge distillation, blueprint separable residual network, blueprint separable convolution, Mixstyle, acoustic scene classification

1. INTRODUCTION

Acoustic scene classification (ASC) is a task to classify each input audio recording into one class of pre-given acoustic scenes. As an important task in Detection and Classification of Acoustic Scenes and Events (DCASE), ASC has attracted a lot of attention from researchers in the community of audio and acoustic signal processing in recent years [1–4]. This task is a continuation of the ASC task from DCASE2022 challenge editions, with modified memory limit and added measurement of energy consumption.

Some lightweight networks (such as MobileNet) has deep separable convolution (DSC), which consists of depthwise (DW) and pointwise (PW) parts to extract special feature maps. The DSC has a lower number of parameters and computational costs than conventional convolution. The architectures based on DSC, implicitly rely on cross-kernel affinity. The researchers have found that standard convolution can be separated more efficiently based on correlations within the kernel [5]. Blueprint Separable Residual Network (BSRN) [6] also uses blueprint separable convolution (BSConv) to get more efficient performance.

This technical report describes our work to marry the concepts of BSRN-based teacher models and BSConv network in a knowledge distillation (KD) framework. The rest of this report is organized as follows. In section 2, we introduce the proposed method for low-complexity ASC. Experiments on the development data are presented in section 3 and conclusions are drawn in section 4.

2. THE METHOD

Convolutional neural networks with attention mechanisms are attracting more and more attention because of their efficiency and effectiveness. However, there are still redundant parts in convolution operations. Existing methods for ASC generally have the characteristics of large model size and heavy computational load. For real-world scenarios where practical inference speed is more important, lightweight and effective methods are preferred. The lightweight BSNR mainly aims at optimizing convolutional operations and introducing effective attention modules, and has achieved good results in the field of image super resolution, especially the feature extraction modules.

KD can be used to generate a small model using the supervision information of a large model with better performance. A pretrained model is used as the teacher model, and the trained teacher model is used to guide the training of the student model. In this work, We use KD in its original form. The framework of our method is shown in Figure 1.

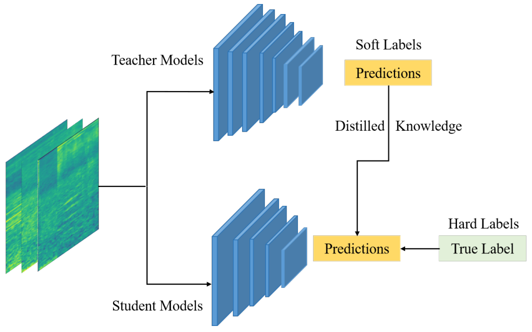


Figure 1: The framework of our method. We bring together BSRN teacher models and a low-complexity BSConv-CNN in the KD framework.

2.1. Teacher Model

The teacher model is Based on BSRN, whose architecture is shown in Figure 2. We designed a new blueprint separable residual structure layer (BSRNLayer). After inputing the Log-Mel spectrogram, simple feature extraction is performed by standard convolution, and then input into the BSRNLayer for deep feature extraction.

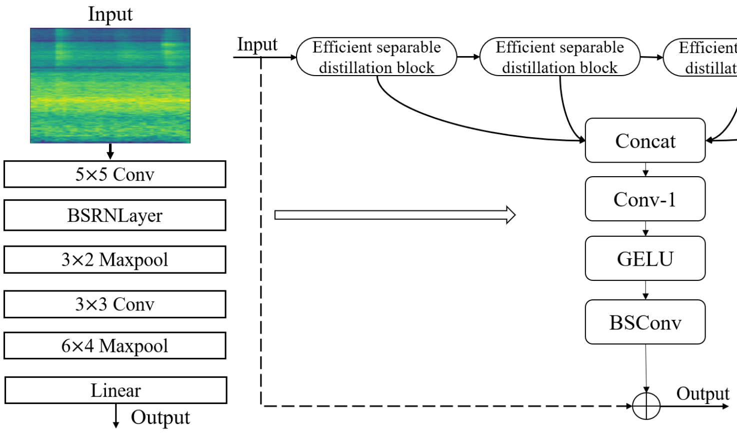

Input
5×5 Conv
BSRNLayer
3×2 Maxpool
3×3 Conv
6×4 Maxpool
Linear
Output
Input
Efficient separable distillation block
Efficient separable distillation block
Efficient separable distillation block
Efficient separable distillation block
Concat
Conv-1
GELU
BSConv
Output

Figure 2: Structure of teacher model.

In the BSRNLayer , the features are gradually refined through four feature distillation blocks, and then the features are spliced into 64 dimensions in the channel dimension. In order to make full use of all deep features, 1 × 1 convolution and Gaussian Error Linear Unit (GELU) activation are used after the tandem operation, and the features produced at different layers are fused and mapped. BSConv is used to refine the features, the input and distillation features are fused to the next layer. Finally the results are output by the fully connected layer.

An efficient separable distillation block generally consists of 3 stages: feature distillation, feature condensation and feature enhancement. In the first stage, for an input feature, the feature distillation can be formulated by BSConv, and this stage refines the coarse feature step by step. In the feature condensation stage, the distilled features are concatenated together and then condensed by a 1 × 1 convolution. For the last stage, to enhance the representational ability of the model while keeping efficiency, we introduce a contrast-aware channel attention block.

2.2. Student Model

The student model is a blueprint separable convolutions network, whose architecture is shown in Figure 3. The audio feature is fed into a 3 × 3 blueprint convolution with a 2 × 2 stride, and further transformed at several BSConvLayers and max pooling layers. Before the second max pooling, a 3 × 3 BSConv is adopted to improve the feature maps. Finally, the prediction vector is output by a fully-connected layer. The last fully-connected layer outputs a 10-dimensional vector.

DSC is a channel-by-channel convolution and a weighted combination in the deep direction. BSConv is a weighted combination of deep directions and a channel-by-channel convolution. Briefly, the order of DW and PW is swapped.

As shown in Figure 4, the BSConvLayer consists BSConv, batchnorm and Rectified Linear Unit (ReLU). After ReLU activation, the original input with the activated output will be together fed into the next layer.

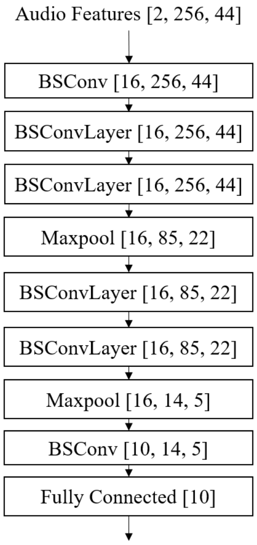


Figure 3: The architecture of student model.

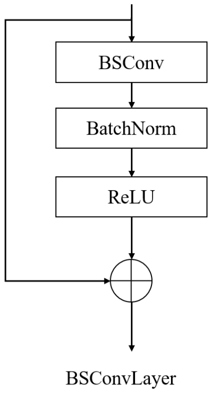


Figure 4: Structure of BSConvLayer.

3. EXPERIMENTS

3.1. Experimental Setup

The development dataset consists of training subset and validation subset. The development dataset contains audio recordings from 10 cities and 9 devices: 3 real devices (A, B, C) and 6 simulated devices (S1-S6). Audio recordings 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 recordings overlap with the audio recordings from device A, but not necessarily with each other. The total amount of audio recordings in the development dataset is 64 hours. Some devices appear only in the validation subset.

The file information of all audio recordings are as follows: 44.1 kHz sampling rate, and mono channel. Audio recordings are first divided into frames via a Hamming window whose length is 4096 with 25% overlapping. Short-time Fourier transform is then performed on each frame for obtaining linear power spectrum which is finally smoothed with a bank of triangular filters for producing log-Mel spectrogram. The center frequencies of these triangular filters are uniformly spaced on the Mel-scale. In addition, to enhance the discriminative ability of audio feature, the delta coefficients of the Log-Mel spectrogram are calculated. Then, they are stacked along channel axis and used as the input feature of the model. The final size of input features is: $256 \times 44 \times 2$, where 256, 44 and 2 represent numbers of frequency-band, frame and channel, respectively.

The teacher model is trained 100 epochs with a batch size of 32 whose size is set to twice the size of the student model. We train both the teacher model and student model using the entire development set on the original cross-validation setup.We use RAdam [7] with a weight decay of 0.001. We use the checkpoint with the highest validation accuracy as the best model.

To prevent overfitting and improve robustness, we have adopted several data augmentation (DA) methods. These methods are performed in the time-frequency domain during training.

- Mixstyle: Zhou et al. [8] proposed a new method for instance-level feature statistics based on probabilistic mixed cross-source domain training samples, called MixStyle. The effect of MixStyle can be regularized by a second parameter p, which controls the probability of whether MixStyle is applied to a batch of recordings. For both teacher and student models, we apply MixStyle to the input. We also set several parameters p to get the teacher model with better performance.
- SpecAugment [9]: SpecAugment is a commonly used DA technique in ASC, which includes functional warping, frequency channel masking blocks, and timestep masking blocks. We apply two masking lines for each dimension, and the maximum thickness of one line is 2.
- Spectrum Modulation: As the spectrum modulation was confirmed to be very effective in the submission of the DCASE 2022 challenge [10], we used the same method. Most of the provided datasets were recorded using device A. Therefore, the data are imbalanced. We dealt with this problem by applying a frequency energy difference to the data of non-device A.

3.2. Experimental Results

The validation set for the development dataset contains 29680 audio clips, and there are new devices. We calculate the overall accuracy and evaluation indexes, such as log-loss on development dataset. Table 1 shows results of teacher models trained downstream on TAU Urban Acoustic Scenes 2023 Mobile development dataset on the provided development set split. α and p denote the MixStyle configurations. MixStyle configuration (α = 0.4, p = 0.5) achieves the best results in terms of validation loss. Averaging the logits of multiple tearcher models of different configurations (denoted as teacher ENSEMBLE) further improves the results. We use model ensembles as teacher models.

We also try to change the input size for the student model as shown in the Figure 5, and better result is obtained when the input feature is with 256 dimensions. We observe that higher number of Mel bins (higher dimensions of frequency-band) leads to better results. Finally, we adopt 256 dimensions of frequency-band and deltas as inputs at 44.1 kHz.

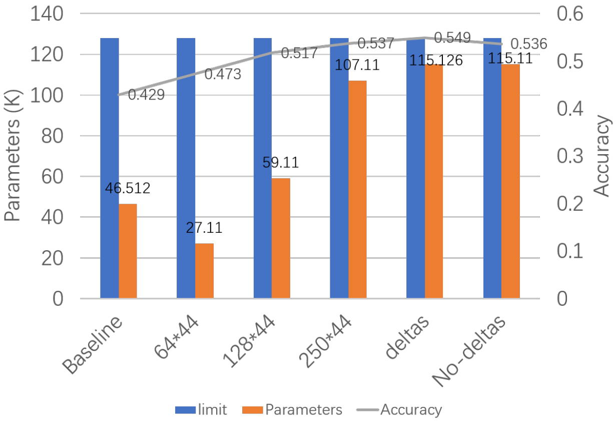


Figure 5: Mel spectrogram tuning. These results were obtained with 4096 STFT.

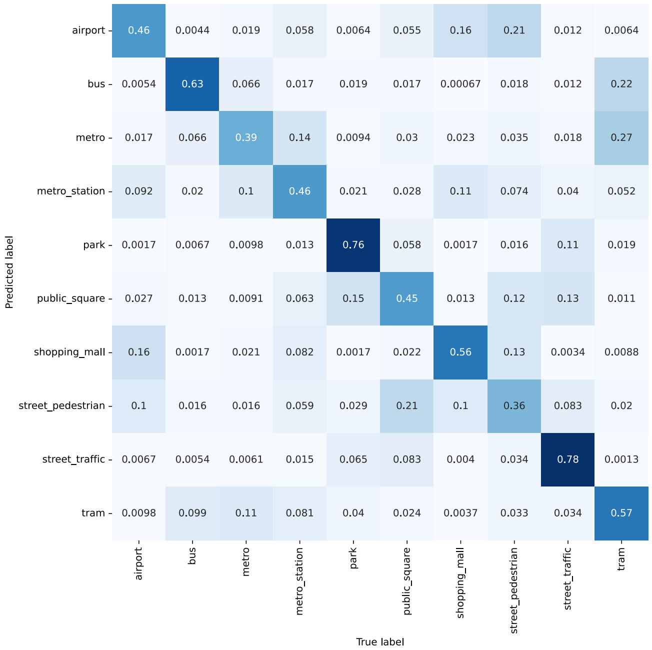
Table 1: Results of teacher models trained downstream on TAU Urban Acoustic Scenes 2023 Mobile development dataset based on the provided development set split.

Teacher models settings	Accuracy	Log Loss
α=0.3, p=0.2	0.605	1.182
α=0.3, p=0.5	0.601	1.177
α=0.4, p=0.2	0.608	1.190
α=0.5, p=0.2	0.606	1.166
Ensemble	0.627	1.071

The configurations and the final results on the development set split are reported in Table 2. Figure 6 shows the accuracy of the best results in Submission 1 for each class. Although the accuracy scores for *Bus*, *Park*, and *Street Traffic* are relatively higher, *Street pedestrian* is not accurately classified.

Table 2: Methods and results with regard to the development dataset for each system. MixStyle describes the respective configurations when training the student model. T and λ denote temperature and the weight of the distillation loss. We perform quantization aware training (QAT) to convert weights and all computations of the final low-complexity model to INT8 type.

Number	T, λ, KD	Mixstyle	QAT	Parameters	MACs	Accuracy	Log Loss
1	3, 250, ✓	$\alpha = 0.3, p = 0.2$	✓	73,386	13.180M	55.6%	1.182
2	✗	$\alpha = 0.3, p = 0.2$	✓	73,386	13.180M	54.3%	1.243
3	3, 250, ✓	$\alpha = 0.4, p = 0.2$	✓	73,386	13.180M	55.5%	1.320
4	✗	$\alpha = 0.4, p = 0.2$	✓	73,386	13.180M	54.5%	1.305



Predicted label \ True label	airport	bus	metro	metro_station	park	public_square	shopping_mall	street_pedestrian	street_traffic	tram
airport	0.46	0.0044	0.019	0.058	0.0064	0.055	0.16	0.21	0.012	0.0064
bus	0.0054	0.63	0.066	0.017	0.019	0.017	0.00067	0.018	0.012	0.22
metro	0.017	0.066	0.39	0.14	0.0094	0.03	0.023	0.035	0.018	0.27
metro_station	0.092	0.02	0.1	0.46	0.021	0.028	0.11	0.074	0.04	0.052
park	0.0017	0.0067	0.0098	0.013	0.76	0.058	0.0017	0.016	0.11	0.019
public_square	0.027	0.013	0.0091	0.063	0.15	0.45	0.013	0.12	0.13	0.011
shopping_mall	0.16	0.0017	0.021	0.082	0.0017	0.022	0.56	0.13	0.0034	0.0088
street_pedestrian	0.1	0.016	0.016	0.059	0.029	0.21	0.1	0.36	0.083	0.02
street_traffic	0.0067	0.0054	0.0061	0.015	0.065	0.083	0.004	0.034	0.78	0.0013
tram	0.0098	0.099	0.11	0.081	0.04	0.024	0.0037	0.033	0.034	0.57

Figure 6: Accuracy confusion matrix for the validation data of Submission 1.

4. CONCLUSION

In this technical report, we described a system for the low-complexity ASC Task 1 of DCASE challenge 2023. The network architecture is based on the BSRN-based models with knowledge distillation. We tried to improve the performance of the proposed systems by applying DA. The accuracy of the submitted system is 12% higher than the baseline on the development dataset.

5. REFERENCES

[1] Y. Li, X. Li, Y. Zhang, W. Wang, M. Liu, and X. Feng, "Acoustic scene classification using deep audio feature and blstm network," in *2018 International Conference on Audio, Language and Image Processing (ICALIP)*, 2018, pp. 371–374.

[2] 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 Transactions on Multimedia*, vol. 22, no. 6, pp. 1385–1394, 2020.

[3] 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 *2021 IEEE 21st International Conference on Communication Technology (ICCT)*, 2021, pp. 1341–1345.

[4] W. Xie, Q. He, Z. Yu, and Y. Li, "Deep mutual attention network for acoustic scene classification," *Digital Signal Processing*, vol. 123, p. 103450, 01 2022.

[5] D. Haase and M. Amthor, "Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets," 2020.

[6] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 832–842.

[7] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," 2021.

[8] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=6xHJ37MVxxp

[9] 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," *2022 16th IEEE International Conference on Signal Processing (ICSP)*, vol. 1, pp. 41–45, 2022.

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