DCASE 2021 TASK 1A: LOW-COMPLEXITY ACOUSTIC SCENE CLASSIFICATION

Technical Report

Yingzi Liu, Jiangnan, LiangLuojun Zhao, Jia Liu, Weiyu Liu, Kexin Zhao, Long Zhang, Tanyue Xu Chuang Shi

University of Electronic Science and Technology of China, Chengdu, China
Corresponding Email:shichuang@uestc.edu.cn

ABSTRACT

This technical report describes the systems for the task 1/subtask A of the DCASE 2021 challenge. In order to reduce the number of model parameters, we add the feature reuse units to the deep residual network. Also the 1-bit-per-weight convolution layer are used in this paper. The log-mel spectrograms, delta features and delta-delta features are extracted to train the acoustic scene classification model. The HRTF and spectrum correction are used to augment the acoustic features. Our system achieves higher classification accuracies and lower log loss in the development dataset than baseline system.

Index Terms— DCASE 2021, acoustic scene classification, deep residual network, data augmentation

1. INTRODUCTION

In daily life, there are all kinds of sounds, such as the sound of a car engine in the street, people talking in the mall, the radio in the airport, the song of birds in the park [1]. These sounds contain a lot of environmental information. For computers, it takes training and learning to perceive sound scenes through sound analysis. The research on this subject is called sound scene classification (ASC). ASC can be used in the design of context-aware services, intelligent wearable devices, robotics navigation systems, and audio archive management [2]. The DCASE Challenge (Detection and Classification of Acoustic Scenes and Events) is a technical competition sponsored by the audio and acoustic signal processing (AASP) technical committee, IEEE signal processing society (SPS). It is one of the most authoritative international evaluation and competition in the field of audio signal processing and focuses on Acoustic scene Classification, Acoustic event Detection and identification. Acoustic scene classification is a regular task in the DCASE challenge series, being present in each of its editions up until now.

In DCASE2021, there are two different subtasks of task 1. The subtask B is concerned with classification using audio and video modalities. The subtask A focus on classifying acoustic scenes with mismatched recording devices. It means that some devices appear only in the evaluation dataset. What's more, the model size of the ASC system is limited. The goal is to build a three-class classifier occupying no more than 128KB.

Each consecutive edition of the challenge has brought a new and larger dataset than previous edition, making it possible to use deep neural networks that rely on large amounts of data for training [3]. Past entries into DCASE challenges have used the spectrogram and its variants for CNN processing, such as the short-time Fourier transform (STFT), log mel spectrogram, mel frequency cepstral coefficients (MFCC), constant-Q transform(CQT) [4]. In DCASE 2020 challenge, most of participants chose the log mel spectrogram as the features in their system [5]. So we prefer to extract log mel spectrogram in our system. Moreover, we extract delta features and delta-delta features to capture dynamic features. We use 1-bit-per-weight networks to satisfy task requirement. The HRTF, spectrum correction, mix-up and temporal crop are used for data augmentation.

2. ARCHITECTURE

2.1. Network Architecture

2.1.1. 1-bit-per-weight CNNs with feature reuse unit

The CNNs used for acoustic scene classification in the past used 1-10 million convolution weights, with a very large number of parameters. In the DCASE2021 task requirements, the number of model parameters is no more than 128KB. 1-bit-per-weight method introduced by McDonnell is used in this technical report[6]. Figure 1 shows the difference between full precision convolution and 1-bit-per-weight convolution.

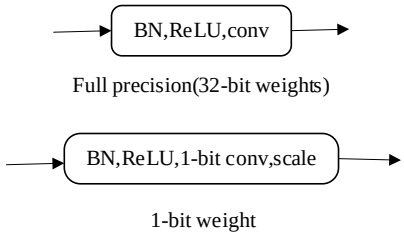

Figure1: Difference between the full-precision and 1-bit-per-weight networks[6].

We used the 1-bit-per-weight method to train the wide residual network, therefore, each convolutional weight was set to one of two values following training, and hence could be stored using a single bit[7]. The wide residual network architecture is shown in Fig 2.

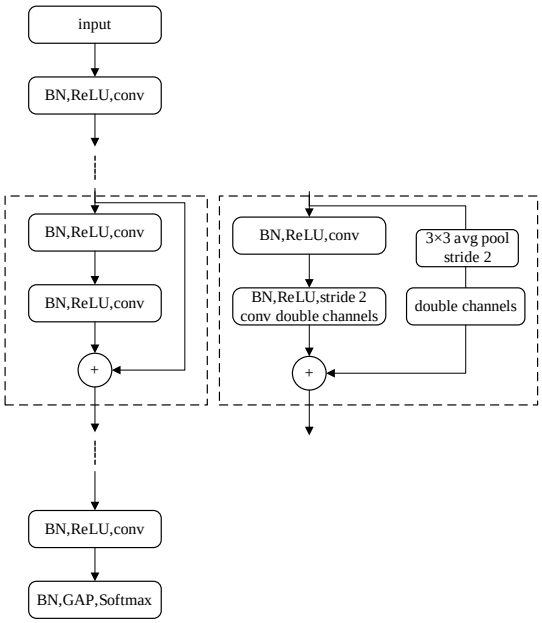

Figure2: Wide Residual Network architecture [7].

As shown in Fig3, Feature reuse units(FR-unit) are introduced to reduce the number of model parameters. After convoluting the input feature map, the input feature was combined with output feature. Finally, the combined feature was transferred to the next layer. Through the reuse of low-level features, the total number of extracted features would not change, so as to ensure that the accuracy of optimized network would not change.

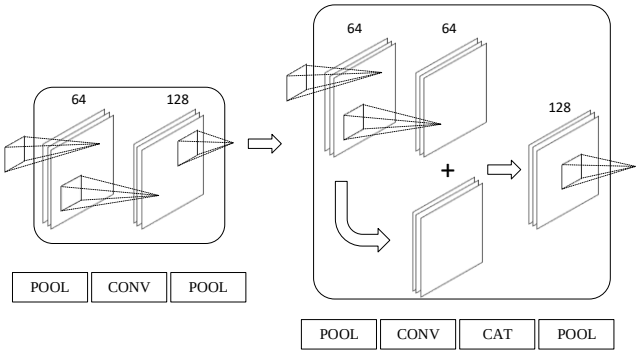

Figure3: Using FR-unit to optimize single convolution.

2.1.2. CNN-based Network with weight quantization

In the other network, we use the network structure of CNN, which is a typical network model for training. And we use the weight quantization to handle the complexity restrictions. Our model consists of 1 input layer, 3 convolution layers, 1 fully connected layer, and 1 output layer, as shown in table1. And batch normalization layers, two max pooling layer, and an activation layer with the Softmax function are also applied in this model. Here Batch normalization (BN) is used to accelerate the learning process and improve the baseline level by regularization terms.

Table 1: Network structure of the model

Input 128×431×1
3×3 Conv2D (pad=1, stride=1)-16-BN-ReLU
3×3 Conv2D (pad=1, stride=1)-16-BN-ReLU
3×3 Conv2D (pad=1, stride=1)-32-BN-ReLU
3×3 Conv2D (pad=1, stride=1)-32-BN-ReLU
5×5 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-32-BN-ReLU
3×3 Conv2D (pad=1, stride=1)-32-BN-ReLU
3×3 Conv2D (pad=1, stride=1)-32-BN-ReLU
12×86 MaxPooling2D
Dense ((64,100), activation='relu')
Dense ((100,10), activation='softmax')

2.2. Acoustic Features Extraction

The samples in the DCASE2021 task 1 / subtask A dataset are monaural and have a common sampling rate of 44.1 kHz. The Librosa library is used to extract the acoustic features. A Short Time Fourier Transform (STFT) with a hamming window size of 2048 and 50% overlap is used to extract the spectrogram. Then apply the log mel filter bank on the spectrogram to get the log mel spectrogram. There are 256 log mel filters in the filter bank that cover a frequency range from 0 to 22.05 kHz, yielding 431-frame spectrograms with 256 frequency bins. Also, the delta features and delta-delta features are added to form three-channel features. The size of the acoustic feature is (432,256,3). Finally, the features are normalized by subtracting the mean and dividing the standard deviation. In order to train the other CNN-based network with weight quantization, the single log mel spectrogram with the shape of (128, 431,1) is extracted.

2.3. Data Augmentation

Four data augmentation methods are used in this technical report ,including the HRTF[8,9], spectrum correction[10], mix-up and temporal crop[11]. In order to weaken the proportion of device A and reduce the error caused by device properties, only HRTF and spectrum correction methods are carried out on data of device A.

2.4. Model size

The size calculations of the Onebit_agm model, Onebit_noagm model and FR_agm model are given in Table 2 and Table 3.

The Onebit_agm model and the Onebit_ noagm model have the same structure, except that one uses data enhancement and the other does not. Trainable parameters of batch normalized layer are used only in BN_1 and BN_3. In Group0, Group1, and Group2, all convolutional layer parameters are stored as single bits, and the BN layer is stored as float32. The first convolution layer and the last convolution layer are stored using float32. The total size of the final model is 42.5KB.

Table 2: Size of onebit_agm model and onebit_noagm model

Layer		Non-zero parameters	Data type	Size
BN_1		13	float32	52 bytes
Conv0		432	float32	1728 bytes
Group0	conv	13824	1-bit	1728 bytes
	BN	198	float32	792 bytes
Group1	conv	50688	1-bit	6336 bytes
	BN	358	float32	1432 bytes
Group2	conv	202752	1-bit	25344 bytes
	BN	710	float32	2840 bytes
BN_2		129	float32	516 bytes
Conv_last		640	float32	2560 bytes
BN_3		41	float32	164 bytes
Total		269785		**42.5KB**

In the FR_agm model, Feature reuse units is added to reduce the number of parameters while doubling the number of channels in the convolution kernel. When the number of channels is different, the 1×1 convolution layer is used to change the number of channels.

Table 3: Size of FR_agm model

Layer		Non-zero parameters	Data type	Size
BN_1		13	float32	52 bytes
Conv0		864	float32	3456 bytes
Group0	conv	40704	1-bit	5088 bytes
	BN	873	float32	3492 bytes
Group1	conv	120064	1-bit	15008 bytes
	BN	1320	float32	5280 bytes
Group2	conv	480256	1-bit	60032 bytes
	BN	2632	float32	10528 bytes
BN_2		257	float32	1028 bytes
Conv_last		1280	float32	5120 bytes
BN_3		41	float32	164 bytes
Total		648304		**106.7KB**

3. EXPERIMENTS

3.1. Dataset

The dataset of the task 1/subtask A of the DCASE2021 challenge contains recordings from 12 European cities in 10 different acoustic scenes using 4 different devices. 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. Additionally, in order to simulate realistic recordings, 11 mobile devices S1-S11 are simulated using the audio recorded with device A, impulse responses recorded with real devices, and additional dynamic range compression. The development set contains data from 10 cities and 9 devices: 3 real devices (A, B, C) and 6 simulated devices (S1-S6). The total amount of audio in the development set is 64 hours. The evaluation dataset contains data from 12 cities, 10 acoustic scenes, 11 devices. There are five new devices: a real device D and simulated devices S7-S11. Evaluation data contains 22 hours of audio.

3.2. Results and Submissions

Models are trained using an Adam optimizer with a batch size of 32, and the cross-entropy function. Each model is trained for 256 epochs. The initial learning rate is set to 0.001 and decreased by a factor 0.5 every 35 epochs. Then, the model with the highest testing accuracy is saved.

Table 4 lists the models that we submit. The main metric for this task is the multiclass cross-entropy (Log loss). The macro-average accuracy (average of the class-wise accuracies) is used as a secondary metric. All submissions achieve lower Log loss in the development dataset than baseline system.

Table 4: Results of development dataset

Model	Log loss	Accuracy	Model size
Baseline	1.473	0.477	90.3KB
FR_agm	0.909	0.682	106.7KB
Onebit_agm	0.923	0.680	42.5KB
Onebit_noagm	0.990	0.650	42.5KB
weight_qz	1.434	0.454	119KB

- **FR_agm** is the one-bit-per-weight model with feature reuse unit that used data augmentation.
- **onebit_agm** is the one-bit-per-weight model that used data augmentation.
- **onebit_noagm** is the one-bit-per-weight model that not used data augmentation.
- **weight_qz** is the CNN-based model with weight quantization.

4. CONCLUSIONS

In this technical report, we have described the systems for the task 1/subtask A of the DCASE 2021 challenge. We use the HRTF, spectrum correction, mix-up and temporal crop to augment the acoustic features. Our system used a 1-bit-per-weight Resnet with the feature reuse unit. The experiment results over DCASE2021 development dataset targeting task 1A review that our method are effective to obtain the lower log loss.

5. REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Introduction to Sound Scene and Event Analysis in Computational Analysis of Sound Scenes and Events*, Applications. Springer, Cham, 2018.

[2] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May. 2015.

[3] K. Koutini1, H. Eghbal-zadeh, and G. Widmer, "CP-JKU submission to DCASE 2019: Acoustic scene classification

and audio tagging with receptive-field-regularized CNNs," Tech. Rep., 2019, DCASE 2019 technical reports.

[4] K. Michał, "Calibrating neural network for secondary recording devices," Tech. Rep., 2019, DCASE 2019 technical reports.

[5] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques", *Journal of Computing*, vol. 2, no. 3, pp. 138-143, Apr. 2010.

[6] M. D. McDonnell, "Training wide residual networks for deployment using a single bit for each weight," *Proceedings of the International Conference on Learning Representations, 2018*; arxiv: 1802.08530.

[7] M. D. McDonnell, H. Mostafa, R. Wang and A. Schaik, "Single-Bit-per-Weight Deep Convolutional Neural Networks without Batch-Normalization Layers for Embedded Systems," *Proceedings of the 4th Asia-Pacific Conference on Intelligent Robot Systems*, Nagoya, Japan, 2019, pp. 197-204.

[8] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, New Paltz, USA, 2001, pp.99-102.

[9] Y. Iwaya, M. Otani, T. Tsuchiya, and J. Li, "Virtual auditory display on a smartphone for high-resolution acoustic space by remote rendering," *Proceedings of the 2015 International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Adelaide, Australia, 2015, pp. 368-371.

[10] T. Nguyen, F. Pernkopf, and M. Kosmider, "Acoustic scene classification for mismatched recording devices using heated-up softmax and spectrum correction," *Proceedings of the ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, 2020, pp. 126-130.

[11] M. D. McDonnell and W. Gao, "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths," Tech. Rep., 2019, DCASE 2019 technical report.