LOW-COMPLEXITY FOR DCASE 2022 TASK 1A CHALLENGE

Technical Report

Yuanbo Hou

Xidian University, Xian, China
Corresponding Email:654358834@qq.com

ABSTRACT

This technical report describes the systems for task1/subtask A of the DCASE 2022 challenge. In order to reduce the number of model parameters and improve accuracy. In this work, I use a simple neural networks with causal convolution and bottleneck structure. The log-mel spectrograms are extracted to train the acoustic scene classification model. The mix-up and Spec-augmentation are used to augment the acoustic features. My system achieves higher classification accuracies and lower log loss in the development dataset than baseline system.

Index Terms— DCASE 2022, acoustic scene classification, data augmentation, neural network

1. INTRODUCTION

Acoustic scene classification(ASC) is a classification task of assigning predefined semantic labels to audio streams recorded in a certain environment by analyzing audio signals [1]. The DCASE 2020 challenge included two subtasks addressing different properties for ASC: 1) Subtask A requires the generalization to unknown devices, and 2) Subtask B demands a low-complexity solution in terms of model size (i.e., the number of parameters). 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.

Compared to previous challenges, this year's dataset becomes one second. This big change had a huge impact on the mission, which meant building models that were better suited to short audio categorization. Not only that, the model size of the ASC system is limited. The number of arguments (including zero arguments) is limited to 128K in size. These presents quite a challenge for sorting tasks.

This report describes our submissions for Task 1A Acoustic Scene Classification (ASC) in the DCASE-2021Challenge. The following sections include details of our model structure and training methods. Due to the model size limitation in subtask A, it necessary to simple the model based on neural network. The basic approach to building our final classifier is based on CNN using Logmel-band energies as features. The following sections describe the details of the proposed system and the experimental results and conclusions.

2. ARCHITECTURE

In the DCASE2022 challenge, the shorter audio time required the model to capture more detail information. We need to use more clever model design to cope with the increasingly demanding parameter size requirements.

2.1. Network Architecture

2.1.1. Causal convolution

Recently, recognition and classification tasks are more attentive to convolutional positions. The researchers prefer to retain the original position information during the convolution operation. In the field of audio classification, it is desirable to preserve the temporal order of convolution. Therefore, We expect to use causal convolution to replace the preservation of audio location information.

2.1.2. bottleneck structure

We use the two-end large and middle-small bottleneck structure proposed in Res-Net. This can help increase network depth while reducing the amount of parameters. We use two bottle-block in model.

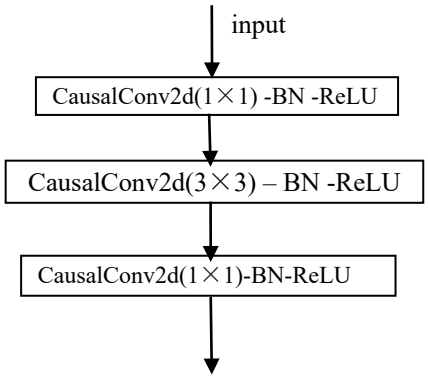

Figure 1. structure of bottle -block

2.1.3. Structure diagram

The model HYB_1 is composed of four causal convolutions and two Bottle-blocks. I use average pooling between two bottle-blocks. Before the final linear classification layer, use Adaptive-Avg-Pool. I use dropout before the first linear layer .Table 1 is the network structure of model HYB_1.

Table 1: Network structure of model HYB_1

Input
CausalConv2d (7*7,1,16)
CausalConv2d (3*3,16,16)
Bottle-block (16, 32)
AvgPool2d (2)
Bottle-block (32, 64)
CausalConv2d (1*1,64, 64)
CausalConv2d (1*1,64, 128)
dropout
Linear (128, 128)
AdaptiveAvgPool2d (2)
Linear (128, 10)
Output

2.2. Acoustic Features Extraction

DCASE2022 task1 dataset have a sample rate of 44.1kHz. The librosa library is used to extract the acoustic features. A SFTF with a hann 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 44-frame spectrograms with 256 frequency bins.

2.3. Data Augmentation

We use two data augmentation methods in this technical report, including Spec-Augmentation and mix-up. We use 13 of time drop width and 48 of frequency drop width. And 2 stripes numbers of both time and frequency axis. As for mix-up, we use 0.3 of alpha during training and get improve the log loss at test time.

Table 2: Size of model HYB_1

Model-block	**Size**
BN1	512bytes
CausalConv2d (7*7)	816 bytes
CausalConv2d (3*3)	2336 bytes
Bottle-block (16, 32)	3200 bytes
AvgPool2d (2)	0
Bottle-block (32, 64)	12544 bytes
CausalConv2d (64, 64)	4224 bytes
CausalConv2d (64, 128)	8448 bytes
Linear (128, 128)	16512 bytes
AvgPool2d (2)	0
Linear (128, 10)	1290 bytes
Total	49.882KB

2.4. Model size

Because Batch Normalization is used after each convolution block, the parameter size of each subsequent block includes convolution and BN. The size of model is given in Table 2. The Maximum number of MACS per inference is given in Table 3.

Table 3:Macs of model HYB_1

Model-block	**Size**
BN1	22.53KMac
CausalConv2d (7*7)	0.903MMac
CausalConv2d (3*3)	5.12MMac
Bottle-block (16, 32)	12.51MMac
AvgPool2d (2)	0
Bottle-block (32, 64)	3.34MMac
CausalConv2d (64, 64)	2.32MMac
CausalConv2d (64, 128)	4.15MMac
Linear (128, 128)	16.51KMac
AvgPool2d (2)	90.11KMac
Linear (128, 10)	1.29KMac
Total	28.51MMac

3. EXPERIMENTS

3.1. Dataset

The dataset of task1 of DCASE challenge 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. The dataset has exactly the same content as TAU Urban Acoustic Scenes 2022 Mobile, development dataset, but the audio files have a length of 1 second. Compared with the data set in 2020, the biggest change is that the audio duration is reduced from 10 seconds to 1 second, which is also the biggest difficulty in increasing the challenge.

The development dataset comprises 40 hours of data from device A, and smaller amounts from the other devices. Audio is provided in a single-channel 44.1kHz 24-bit format.

3.2. Training procedure

The model is trained for 120epoches, with a batch size of 128, a weight decay of 1e-5, using Adamw and a starting learning rate of 1e-3 and a cosine annealing LR scheduler with a 1e-5 eta_min. We use average pool layers in the first model: the first between two bottle-block and the second is in front of the classification layer. Also, we use dropout layer after final convolution. These layers drop each neuron with probability 0.3.

The system are trained by applying two linear layer on the finallogits and using the cross-entropy loss.

3.3. Results

The results show that my proposed model outperforms the baseline by 7%.

Table 3: Results of development dataset

Model	Log loss	Accuracy	Size
Baseline	1.575	42.9%	46.5KB
Model 1	1.449	49.7%	59.8KB

4. CONCLUSIONS

In this technical report, wo have described the systems for the task1/subtask A of DCASE 2022 challenge. We use SpecAugment and mix-up to augment the features. We use causal conv and bottle block in the model. my method improves the accuracy by 7% over the baseline and reduces the loss.

5. REFERENCES

[1] http://dcase.community/workshop2022/.

[2] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020), 56–60. 2020. URL: https://arxiv.org/abs/2005.14623.

[3] Yingzi Liu, Jiangnan, and LiangLuojun Zhao, “DCASE 2021 Task 1 Subtask A: Low-Complexity Acoustic Scene Classification,” Tech. Rep., 2021, DCASE 2021 technical reports.

[4] Gilles Puy, Himalaya Jain, and Andrei Bursuc, “Separable Convolutions and Test-Time Augmentations for Low-Complexity and Calibrated Acoustic Scene Classification,” Tech. Rep., 2021, DCASE 2021 technical reports.

[5] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[6] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “MixUp: Beyond empirical risk minimization,” in *International Conference on Learning Representations(ICLR)*, 2018.

[7] H. -j. Shim, J. -w. Jung, J. -h. Kim and H. -J. Yu, "Attentive Max Feature Map and Joint Training for Acoustic Scene Classification," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 1036-1040, doi: 10.1109/ICASSP43922.2022.9746091.

[8] X. Y. Kek, C. Siong Chin and Y. Li, "An Investigation on Multiscale Normalised Deep Scattering Spectrum with Deep Residual Network for Acoustic Scene Classification," 2021 IEEE/ACIS 22nd International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2021, pp. 29-36, doi: 10.1109/SNPD51163.2021.9704888.

[9] Hee-Soo, Heo and Jee-weon, “Clova Submission for the DCASE 2021 Challenge: Acoustic Scene Classification Using Light Architectures and Device Augmentation,” Tech. Rep., 2021, DCASE 2021technical reports.