DCASE 2021 CHALLENGE TASK1A TECHNICAL REPORT

Technical Report

Jiawang Zhang, Shengchen Li, Bilei Zhu

Beijing University of Posts and Telecommunications, Beijing, P. R. China
Xi'an Jiaotong-liverpool University, Suzhou, P. R. China
ByteDance, Shanghai, P. R. China
zhangjiawang@bupt.edu.cn,
shengchen.li@xjtlu.edu.cn,
zhubilei@bytedance.com

ABSTRACT

This report describes our method for Task 1a (Low-Complexity Acoustic Scene Classification with Multiple Devices) of the DCASE 2021 challenge. The task targets low complexity solutions for the classification problem. This report uses Residual Network (ResNet) model and uses Log Mel Spectrogram to process features. To compress system complexity, this report uses Post Training Static Quantization. Post Training Static Quantization are used to do the 8-bits quantization, this method can reduce the model size by four times. The accuracy of the method proposed in this report on the development data set is 73%, which is 25% higher than the baseline.

Index Terms— Low-Complexity Acoustic Scene Classification, Multiple devices, ResNet, Post Training Static Quantization.

1. INTRODUCTION

In our lives, sound carries a lot of information. We can judge the surrounding scene (acoustic scene) and what is happening (acoustic event) by the information carried by the sound we hear. In recent years, Deep neural networks have achieved the advanced results in many machines learning tasks [1]. This includes acoustic scene classification and acoustic event recognition. The accuracy of the judgment by the computer is gradually improving, even more than the human ear recognized. The development of algorithms that use computers to automatically extract this information has great potential in a variety of applications. For example, searching multimedia based on audio content, making context-aware mobile devices, intelligent monitoring systems that can sense hearing, robot hearing, unmanned cars.

However, a lot of research is still needed to reliably identify sound scenes and sound events in real sound scenes. In a real acoustic environment, there are usually multiple sounds at the same time, and they will be disturbed by environmental noise.

Queen Mary University of London (QMUL) organized the first DCASE (Challenge on Detection and Classification of Acoustic Scenes and Events) challenge in 2013 [2], paving the way for sound detection classifier performance evaluation. The DCASE Challenge aims to expand the most advanced technical sound scene and event analysis methods,

As a routine task of the DCASE Challenge, Acoustic Scene Classification (ASC) is a task to classifying the recorded data to the place that the data was recorded. In this year's DCASE2021 challenge, Task1a targets low complexity solutions for the classification problem in term of model size, and uses audio recorded with multiple devices. Each data corresponds to one class out of ten, there is no data with multiple labels. The data is ten seconds, and the useful information is rarely.

The main problem of subtask A was to design a classifier that would work stably on a variety of microphone. However, the development data set mostly includes data collected from specific microphones, while the evaluation data will include data recorded using microphones that did not appear in the development set. And the model size should under 128 kilobytes. This translates into 32768 parameters when using float32 (32-bit float) which is often the default data type (32768 parameter values * 32 bits per parameter / 8 bits per byte= 131072 bytes = 128 KB (kibibyte)). It is smaller than last year's limited.

This report builds a ResNet model for acoustic scene classification system. This report uses 11 layers ResNet model because the limited model size of the task. Because of the model size is very small, this report tests the appropriate parameters of frames and Mel bins by running comparative experiments. This report set five experiments, they are 28 frames 128 Mel bins, 28 frames 256 Mel bins, 43 frames 128 Mel bins, 43 frames 256 Mel bins and 28 frames 64 Mel bins. According to the results, the best parameters is 28 frames and 128 Mel bins. This report builds a small-size ResNet model at first, and then use post training static quantization method to compress the model. A model can be compressed to 1/4 of the original size by this method.

2. DATA PREPROCESSING

This section describes the methods about signal processing that this report to transform the audio sample into acoustic features.

2.1. Acoustic Feature

The audios in the dataset are mono audio files with 44.1 kHz sample rate. This report uses Log Mel spectrogram as audio feature. When extracting features from audio, Log Mel spectrogram is generally used, compared to MFCC, wavelet, Log Mel has better performance[3]. This report uses 2048 length hanning sample windows and the hop-size of 1536 sample to divide an audio sample into 280 frames. This report uses 1 channel and 128 Mel bins, so the Log Mel spectrogram in the shape of (280, 256, 1).

3. METHOD

This section introduces the methods used to classify the acoustic scene. This report uses ResNet models [4] and Log-Mel spectrogram as audio feature [5]. Post Training Static Quantization method is used to compress system complexity [6]. Finally, a 10-way SoftMax layer is used to classifying the acoustic scene.

3.1. Network Structure

ResNet11: This model we used is ResNet network proposed by Khaled, the input of the network is 5*5, and then has five residual block and an average pooling, the output to do the SoftMax [9]. A Residual block in this report has two convolution blocks, the layer order of each convolution block is Convolution-BatchNormalization-ReLU. Khaled showed that the proper size of the receptive field is important for the ASC task. They confirmed that the CNN with a large receptive field is overfitting for ASC data, and proposed a method to improve classification performance by limiting the receptive field.
This report does the Post Training Static Quantization on this model, to do the 8-bits quantization aim at reducing the model size by four times.

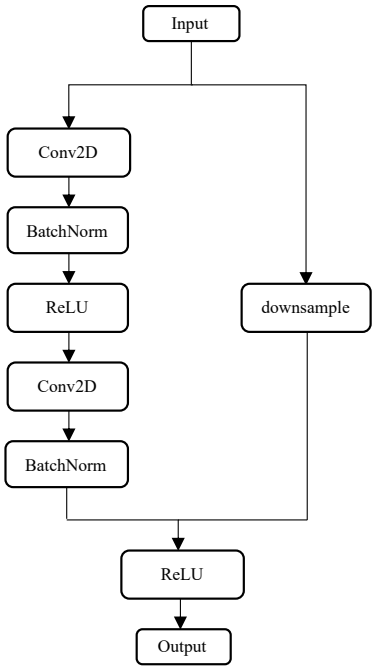

Figure 1: Residual block

Table 1: ResNet11

Input	Features
Conv2D	5 * 5, stride = 2
Residual Block	3 * 3, 1 * 1, p
Residual Block	3 * 3, 3 * 3, p
Residual Block	3 * 3, 1 * 1
Residual Block	1 * 1, 1 * 1
Residual Block	1 * 1, 1 * 1
Average Pooling	2 * 2
Output	SoftMax
P: 2 * 2 Max Pooling Residual Block 1 – 3: 32 channels Residual Block 4: 64 channels Residual Block 5: 128 channels	

3.2. Post Training Static Quantization

Quantization usually uses this method, this method quantifies the weight in advance, Post Training quantization is usually used for CNN [10]. The process of post training static quantification is: First, prepare the model and specify the activation values to be quantized and dequantized. The model cannot be reused. Convert operations that need to be quantified again into modules. Then fuse the operations and choose the configuration of the quantization methods. Finally use Pytorch function to quantify the model.
Post Training Static Quantization are used to do the 8-bits quantization, this method can reduce the model size by four times.

3.3. Attention

3.3.1. SENet

SE module first carries out squeeze operation on the feature map obtained by convolution, get the channel level global features, then excitation was carried out on the characteristics of global operation, study the relation between each channel, and get the weight of different channel, the last times characteristic graph are the final characteristics of the original [11]. In essence, SE module does attention or gating operation on channel dimension. This attention mechanism enables the model to pay more attention to channel features with the largest amount of information, while suppressing those features that are not important. Another point is that SE modules are generic, which means they can be embedded into existing network architectures.

Table 2: SENet

SENet	
AdaptiveAvgPool2d	output = 1
Linear	in_features = 32, out_features = 2
Linear	in_features = 2, out_features = 32
Sigmoid	

3.3.2. CBAM

CBAM (Convolutional Block Attention Module) calculates the attention map of feature map from two dimensions of channel and space, and then multiply the attention map with the input feature map to carry out adaptive feature learning [12]. CPAM is a lightweight universal module, which can be incorporated into various convolutional neural networks for end-to-end training.

Table 3: CBAM

CBAM	
channel attention	
AdaptiveAvgPool2d	output = 1
AdaptiveMaxPool2d	output = 1
Conv2d	1 * 1
Conv2d	1 * 1
Sigmoid	
spatial attention	
Conv2d	7 * 7
Sigmoid	

4. EXPERIMENTS AND RESULTS

4.1. Audio Dataset

The dataset that this report used is TAU Urban Acoustic Scenes 2020 Mobile [13]. This dataset contains 23,040 samples, these audios are used to train and validate the model. This dataset consists of two types of audios: three real devices and six simulated devices. Most of the data were collected form device A, and other devices are Samsung Galaxy S7 and iPhone SE. At the evaluation dataset, there is a new device GoPro Hero5 Session. The simulated devices are synthesized by processing the data of device A. By using various impulse responses and dynamic range compression. The development dataset has 13,965 samples in the training set and 2,970 samples in the test set.

4.2. Task1a Result

This section compares the different parameters and the different models.

4.2.1. Different Models

This report trains different models, they are CNN5, CNN9, MobileNetV2, ResNet18 and ResNet11.

Table 4: Different models

Model	Model Parameters	Accuracy
CNN5	4308480	0.655
CNN5_small	1554944	0.667
CNN9	4689344	0.712
MobileNetV2	1946994	0.555
MobileNetV2_relu6	1895386	0.618
ResNet18	11168005	0.658
ResNet11	49738	0.745

From the results, when reduce the size of the convolution kernel of the CNN5, this report reduces the 5 * 5 size to the 3 * 3, the model size was reduced by two thirds, and the accuracy is increased. As for MobileNetV2, the relu6 can improve the accuracy of the network. ResNet11 has the smallest network size and the highest accuracy. So, this report uses the ResNet11 as the train model.

4.2.2. Different methods

This report tests the different parameters about batch size and Mel bins, and also use the attention model of the ResNet11.

Table 5: Different methods and parameters

Model	Focal loss	Model Parameters	Accuracy
resnet_28frame_128mel	-	49738	0.782
resnet_28frame_256mel	-	49738	0.748
resnet_43frame_128mel	-	49738	0.779
resnet_43frame_256mel	-	49738	0.745
resnet_28frame_128mel	Y	49738	0.736
resnet_28frame_256mel	Y	49738	0.733
resnet_43frame_128mel	Y	49738	0.752
resnet_43frame_256mel	Y	49738	0.748
resnetSEnet_28frame_128mel	-	85706	0.712
resnetCBAM_28frame_128mel	-	53177	0.682

From the results, 28 frame and 128 Mel bins has the highest accuracy. Focal loss is mainly to solve the problem that the proportion of positive and negative samples is seriously unbalanced in one-stage target detection [14]. This loss function reduces the weight of a large number of simple negative samples in training and can also be understood as a kind of difficult sample mining. But focal loss has poor performance on ResNet11.
SENet and CBAM also has poor performance on this challenge. By using Post Training Static Quantization method can reduce the model size by four times.

4.3. Submission Systems

This part is the detailed information about the submitted systems.

4.3.1. Zhang_resnet_1

This system uses ResNet11, 28 frame per seconds, 128 mel bins and CE loss. 8-bit quantization is used to compressing the model. The trainable model parameters are 83572, the trainable model parameters non-zero is 49738. So, the model size is 49738 parameter * 8 bits per parameter / 8 bits per byte= 49738 bytes = 48.57 KB (kibibyte).

4.3.2. Zhang_resnet_2

This system uses ResNet11, 43 frame per seconds, 128 mel bins and CE loss. 8-bit quantization is used to compressing the model. The trainable model parameters are 83572, the trainable model

parameters non-zero is 49738. So, the model size is 49738 parameter * 8 bits per parameter / 8 bits per byte= 49738 bytes = 48.57 KB (kibibyte).

4.3.3. Zhang_resnet_cbam

This system uses ResNet11 and CBAM attention module. 28 frame per seconds, 128 mel bins and CE loss is used. The trainable model parameters are 87011, the trainable model parameters non-zero is 53177. So, the model size is 53177 parameter * 8 bits per parameter / 8 bits per byte= 53177 bytes = 51.93 KB (kibibyte).

4.3.4. Zhang_resnet_senet

This system uses ResNet11 and SE attention module. 28 frame per seconds, 128 mel bins and CE loss is used. The trainable model parameters are 86516, the trainable model parameters non-zero is 85706. So, the model size is 85706 parameter * 8 bits per parameter / 8 bits per byte= 85706 bytes = 83.70 KB (kibibyte).

5. CONCLUSION

This report concludes the methods for DCASE 2021 task1a. This report use ResNet model, uses Log-Mel spectrogram as audio feature. By use Post Training Static Quantization methods can reduce the model size, and almost no loss of accuracy. This Low-Complexity Acoustic Scene Classification is suit for mobile systems and embedded systems which have limited hardware resources.

6. REFERENCES

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[2] http://dcase.community/challenge2020/

[3] Muda L, Begam M, Elamvazuthi I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques[J]. arXiv preprint arXiv:1003.4083, 2010.

[4] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Thirty-first AAAI conference on artificial intelligence. 2017.

[5] Deng L, Droppo J, Acero A. Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise[J]. IEEE Transactions on Speech and Audio Processing, 2004, 12(2): 133-143.

[6] Jain S R, Gural A, Wu M, et al. Trained quantization thresholds for accurate and efficient fixed-point inference of deep neural networks[J]. arXiv preprint arXiv:1903.08066, 2019.

[7] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[8] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//European conference on computer vision. Springer, Cham, 2016: 630-645.

[9] Koutini K, Eghbal-Zadeh H, Dorfer M, et al. The Receptive Field as a Regularizer in Deep Convolutional Neural Networks for Acoustic Scene Classification[C]//2019 27th European Signal Processing Conference (EUSIPCO). IEEE, 2019: 1-5.

[10] Xu Y, Kong Q, Wang W, et al. Large-scale weakly supervised audio classification using gated convolutional neural network[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 121-125.

[11] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[12] 1 Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[13] Mesaros A, Heittola T, Virtanen T. A multi-device dataset for urban acoustic scene classification[J]. arXiv preprint arXiv:1807.09840, 2018.

[14] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.