ACOUSTIC SCENE CLASSIFICATION COMBINING LOG-MEL CNN MODEL AND END-TO-END MODEL

Technical Report

Xu Zheng

National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, China
zx980216@mail.ustc.edu.cn

ABSTRACT

This technical report describes the Zheng-USTC team's submissions for Task 1 - Subtask A (Acoustic Scene Classification, ASC) of the DCASE-2019 challenge. In this paper,two different models for Acoustic Scene Classification are provided.The first one is a common two-dimensional CNN model in which the log-mel energies spectrogram is treated as an image.The second one is an end-to-end model ,in which the features of a speech are extracted by a 3-layer CNN model with 64 filters. The experimental results on the fold1 validation set of 4185 samples and the leaderboard showed that the class-wise accuracy of the two models are complementary in some way.Finally we fused the softmax ouput scores of the two different systems by using a simple non-weighted average.

Index Terms— convolutional neural networks, deep learning,end-to-end model,specaugment ,random crop

1. INTRODUCTION

This report describes our submissions for Task 1 (Subtask A) — Acoustic Scene Classification (ASC) in the DCASE-2019 Challenge. The baseline of our model is based on the Cnn_9layers_AvgPooling model provided by Qiuqiang Kong[1][1]. We used random cropping and specaugment[2] for data augment,the best classification accuracy on the fold1 validation dataset was 74.20% with an improvement of 5% compared to the pure baseline without any data augments.Laterly,we also changed the input channel and the number of mel-bin and the details will be described in Section 2.

In addition,we designed an end-to-end system for ASC , in which the features of audio signal are extracted by a 3-layer cnn, and the feature size is 640×64,which is same to the size of log-mel energies spectrogram of Kong 's baseline model. The other part of the end-to-end system is the model of Cnn_9layers_AvgPooling.

In the end, we submitted the predictions of three systems: (1) a single system of Cnn_9layers_AvgPooling with data augments of random cropping and specaugment; (2) a single system of our end-to-end model with between-class learning［4］; (3) a combination of Cnn_9layers_AvgPooling model and end-to-end model obtained by averaging.We estimate the performance of our methods on the publicly available Kaggle-Leaderboard. The system(1)、(2)、(3) achieve the classification accuracy of 75.33% 、69.33% 、79.16%, respectively.

2. LOG-MEL CNN MODEL——Cnn_9layers_AvgPooling

In this section we describe the neural network architecture as well as the data preparation and augment strategies used for training our network.

2.1. Data Preparation and Data Augment

For data preprocessing,we have two approaches.In the first approach ,the audio signals are resampled to 32000 Hz and then the two -channel signals are transformed into a single channel by a simply averaging.We then calculated the Short Time Fourier Transform （STFT） using a 1024-sample window and a hop-size of 500 samples,and finally we applied 64-bin Mel filter bank to obtain the log-mel spectrogram with a size of 1×640×64.

In the second approach, the audio signals are resampled to 22050Hz,and both of the two channel audio signals are preserved. We then calculated the STFT using a 2048-sample window and a hop-size of 512 samples,and finally we applied 256-bin Mel filter bank to obtain the log-mel spectrogram with a size of 2×430×256.

For data augment ,we mainly have two methods of data augment: cropping and specaugment. In the method of cropping ,we firstly randomly cropped 2-s duration of audio signal from the 10-s audio ,and then we padded 1-s of zeros on each side of the firstly cropped audio,and finally we randomly cropped the audio from the padded audio of 4-s duration to obtain the 2 -s audio.

In the method of specaugment[2] ,we only used frequency masking and the parameter settings are showed in Table1.

2.2. Network Architecture

Our network architecture is depicted in table2,and this is a VGG style network[3] and the convolution-block consisting of convolution layer, batch normalization layer and activation of ReLu. After all the convolution layers, we used AvgPooling on the Frequency dimension and MaxPooling on the time dimension and we then obtained a 512-dimension vector .We Finally used dropout and fully connected layer to get the prediction score of each class .

[1] https://github.com/qiuqiangkong/dcase2019_task1

Table 1 The SpecAugment parameter for the two kinds of log-mel spectrogram

spectrogram size	Maximum frequency mask width	Mask number
1x640x64	4	16
2x430x256	8	32

Table 2: Our Cnn_9layers_AvgPooling. BN: Batch Normalization, ReLU: Rectified Linear Unit.

Name	Description	Output size
Input	Channel× Time× Frequency	1× 640 ×64
Data augment	Random cropping 2s Padding 2×1s Random cropping 2s SpecAugment	1× 128 ×64
ConvBlock1	Cov3×3 -64BN-ReLU Cov3×3 -64BN-ReLU AvgPooling2x2	64 × 64 ×32
ConvBlock2	Cov3×3 -64BN-ReLU Cov3×3 -64BN-ReLU AvgPooling2x2	128 × 32 × 16
ConvBlock3	Cov3×3 -64BN-ReLU Cov3×3 -64BN-ReLU AvgPooling2x2	256 × 16 × 8
ConvBlock4	Cov3×3 -64BN-ReLU Cov3×3 -64BN-ReLU AvgPooling1x1	512 × 16 ×8
Avgpooling	Avgpooling 8x1	512 × 16
Maxpooling	MaxPoing 16x1	512

fc	Dropout(0.5)
	Linear(512,10)
	Softmax
		10

3. END-TO-END MODEL

In this section we describe the end-to-end model we used in the ASC task.

3.1. Data Preparation

The audio signals are resampled to 16000 Hz and then the two -channel signals are transformed into a single channel by simply averaging.Thus each audio was represented in a vector of 160000-dimension.

We used random cropping ,which is described in Section 2,so We only used 2-s duration audio for training .In testing,we padded 1-s on both side of the test audio and regularly cropped the padded audio with a stride of 1-s to obtain 11 audio segments with length of 2-s.The final score is the average of 11 softmax scores.

3.2. Model Architecture

The model mainly consists of 2 parts.The first part is a 3-layer convolution block with 64 filters,which is used to extract the features of a audio.After 3-layer convolution, 2 -s audio is represented as a matrix of size of 128×64 . The second part of the model is the same to the architecture described in Section2.The model architecture is depicted in table3.

3.3. Between Class Learning

We used between-class learning[4] for the data augmentation method. In between-class learning ,the two different audios X_1, $X_2$are mixed with a random ratio λ :

$$X = \lambda X_1 + (1-\lambda)X_2$$
$$y = \lambda y_1 + (1-\lambda)y_2$$

Here , $\lambda \in [0,1]$,is acquired by beta distribution .And in between-class learning ,a new training sample (X,y) is created from the original training set.

Table 3: Our End-to-End model. BN: Batch Normalization. ReLU: Rectified Linear Unit.

Name	Description	Output size
Input	Channel× Time× Frequency	1× 160000 × 1
Data augment	Random cropping 2s Padding 2×1s Random cropping 2s	1× 32000 × 1
Feature extracting part	Cov7× 1 -64BN-ReLU AvgPooling5x1 Cov7× 1 -64BN-ReLU AvgPooling2x1 Cov7× 1 -64BN-ReLU AvgPooling25x1	64 × 128 × 1
Swap axes	——	1 × 128 ×64
Cnn 9layers AvgPooling	In Table 2	10

4. RESULTS

We used the proposed two systems and the class-wise accuracy are depicted in Table 4.Both the development and the leaderboard result show that the class-wise accuracy of the two models are complementary in some way.Finally we fused the softmax ouput scores of the two different systems by using a simple non-weighted average.The best accuracy of our fused system achieved 79.16% on the leaderboard .

Table 4: Our class-wise accuracy on different kind of models

Model	Network Input Size	Development Acc	Leaderboard Acc
Cnn 9layers AvgPooling Baseline	1x640x64	69.23%	69.16%
Cnn 9layers AvgPooling +cropping +specAugmet	1x128x64	74.22%	73.6%
Cnn 9layers AvgPooling +cropping +specAugment	2x86x256	78.5%	75.33%
End-to-End System +cropping +BClearning	1x32000x1	69.23%	69.33%
Fusion of Last two models	——	81.3%	79.16%

5. REFERENCES

[1] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yong Xu, Wenwu Wang, Mark D. Plumbley. Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems. arXiv preprint arXiv:1904.03476 (2019)

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[3] Daniel S. Park, William Chan ,et al.SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition.

[4] Yuji Tokozume,Yoshitaka Ushik,and Tatsuya Harada.LEARNING FROM BETWEEN-CLASS EXAMPLES FOR DEEP SOUND RECOGNITION. in ICLR, 2018