ACOUSTIC SCENE CLASSIFICATION MODEL BASED ON TWO PARALLEL RESIDUAL NETWORKS

Technical Report

Ziling Qiao, Hongxia Dong, Xichang Cai, Menglong Wu

North China University of Technology

ABSTRACT

This technical report describes our submission for task1a of dcase2021 challenge. We calculated 128 log-mel energies under the original sampling rate of 44.1KHz for each time slice by taking 2048 FFT points with 50% overlap. Additionally, deltas and delta-deltas were calculated from the log Mel spectrogram and stacked into the channel axis. The resulting spectrograms were of size 128 frequency bins, 423 time samples and 3 channels with each representing log-mel spectrograms, its delta features and its delta-delta features respectively. Then, the three channel feature map is divided into 0-64 and 64-128 Mel bins on the frequency axis, and the high and low frequency features are input into the two parallel residual networks with identical residual blocks and convolutional residual blocks for training, and then the two network models are concatenate on the channel axis. Finally, after 1×1 convolution and global average pooling, the classification results are obtained through softmax output.

Index Terms— Acoustic Scene Classification, Convolution Neural Network, Data Augmentation, ResNet

1. INTRODUCTION

Acoustic Scene Classification (ASC) is a task of classifying given data to a place where it was recorded. Each data corresponds to one class out of ten, and there is no data with multiple labels. The length of the data is ten seconds, but the useful information appears very rarely. This task is one of the major topics that has been covered every year in the DCASE challenge. This year, subtask A for Low-Complexity Acoustic Scene Classification with Multiple Devices [1,2].

The main issue of the subtask A is to design a classifier that works stably on various microphone types. However, the development dataset mostly includes the data collected from a specific microphone, and the evaluation data will include data recorded with a microphone that has not appeared in the development set. This task was treated in the previous year, and [3] was placed on top with spectrum correction method and Convolutional Neural Network (CNN) model.

The following sections include details of our model structure and training methods. Due to the model size limitation in subtask A, A model complexity limit of 128 KB is set for the non-zero parameters. 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 became impossible to solve both problems with a universal model design.

2. AUDIO DATASET

The development dataset for this task is TAU Urban Acoustic Scenes 2020 Mobile, development dataset. The dataset 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 [4].

Recordings were made using four devices that captured audio simultaneously. The main recording device consists in a Soundman OKM II Klassik/studio A3, electret binaural microphone and a Zoom F8 audio recorder using 48kHz sampling rate and 24-bit resolution, referred to as device A. The other devices are commonly available customer devices: device B is a Samsung Galaxy S7, device C is iPhone SE, and device D is a GoPro Hero5 Session.

3. SYSTEM ARCHITECTURE

3.1. data preprocessing

The data of subtask A are mono audio files with 44.1 kHz sample rate. We transformed them into power spectrogram by skipping every 1024 samples with 2048 length Hann window. A spectrum of 431 frames was yielded from 10 seconds audio file, and each spectrum was compressed into 128 bins of Mel frequency scale. Additionally, deltas and delta-deltas were calculated from the log Mel spectrogram and stacked into the channel axis. The number of frames of the input feature is cropped by the length of the delta-delta channel so that the final shape becomes [$128 \times 423 \times 3$].

3.2. data augmentation

Due to the limited data set provided by dcase, we propose a data augmentation method to increase the diversity of data distribution. Mixup augmentation [5] were applied. We did not use additional training datasets other than the official training dataset. Parameter of mixup $\alpha = 0.5$.

3.3. model design

Previous studies have verified the effectiveness of the ResNet [6] on the ASC [7, 8, 9]. Our model consists of two parallel residual networksthis parallel structure has been proposed in [8] and [3] to learn distinct features from different frequency bandsOur model consists of two paths for 0-64 and 64-128 Mel bins. After concatenating [10] the outputs from each network, one block of 1×1 convolution [11,12] and Global Average Pooling (GAP) [11] calculates the classification scores.

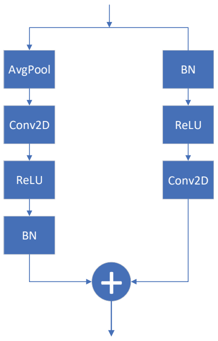

Figure 1: Convolution residual block

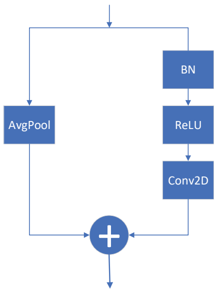

Figure 2: Identical residual block

The residual block in this paper uses two kinds of residual blocks: the identical residual block and the convolution residual fast block. The difference between convolution residual block and the traditional identical residual block is that the identical residual block is a short wire in the shortcut path, while the convolution residual block is a layer of conv2d in the shortcut path, It is used to adjust the number of input channels to the appropriate size to match the number of main path channels. In addition, the output of the shortcut path in the convolution residual block is added with the output of the identical residual block to increase the characteristic parameters. The convolution residual block used in this paper is shown in Figure 1. The identical residual block used in this paper is shown in Figure 2. The residual block used in this paper is shown in Figure 3. The overall structure of our model is shown in Figure 4.

3.4. Categorical Focal Loss

Focal loss attenuates the logloss generated by welltrained samples, so that the model can focus on the poorly trained samples [3]. The following equation describes focal loss with balancing parameter , focusing parameter and prediction score ,

$$FL\left(p_t\right) = -\alpha\left(1-p_t\right)^{\gamma}\log\left(p_t\right) \tag{1}$$

Increasing the value of γ increases the sensitivity of the model to misclassified samples, and scales the loss function linearly. Our setting was $\gamma = 1.0$ and $\alpha = 0.3$, respectively.

3.5. Training Setup

We trained our model using Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9. We use warm up and cosine annealing to plan the learning rate. Choosing warmup learning rate can make the learning rate of several epochs or some steps smaller. Under the warmup learning rate, the model can gradually become stable. After the model is relatively stable, we choose the preset learning rate for training to make the convergence speed of the model faster, The effect of the model is better. The preset learning rate is 0.001, warmup_learning is set to $4e-06$. The warm up phase lasts for 12341 steps. After the warm up phase, the learning rate remains unchanged until the end of the hold_rate_steps.

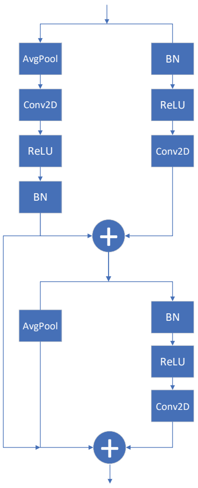

Figure 3: Residual block

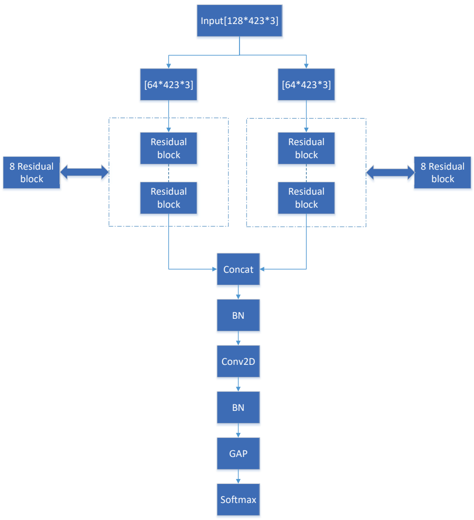

Figure 4: The overall structure of our model

4. EXPERIMENTAL RESULT

For baseline system: log-mel is used and CNN is used for classification network, total amount of non-zero parameters in the model is 46246, model_size is 90.3kb, the loss of baseline system was 1.461, and the classification accuracy was 46.9%. For our system, we trained the model using all the development data and the experimental results are shown in Table 1.

Table 1: Accuracy on the fold 1 evaluation set(class-wise)

Scene label	**Baseline**	**Our system**
Airport	31.1%	42.1%
Bus	40.1%	51.3%
Metro	48.1%	59.5%
Metro_station	29.6%	38.1%
Park	63.6%	71.5%
Public_square	36.0%	45.1%
Shopping_mall	61.3%	69.4%
Street_pedestrian	47.1%	54.5%
Street_traffic	68.0%	71.7%
Tram	44.3%	51.8%
Average	**46.9%**	**55.5%**
Loss	**1.461**	**0.847**
Model_size	$\mathbf{90.3}kb$	$\mathbf{124.4}kb$

5. CONCLUSION

In this technical report, we proposed a acoustic scene classification system. We use log-mel spectrograms, deltas and delta-deltas and two parallel residual networks with identical residual blocks and convolutional residual blocks to improve the performance of the system. We achieved a classification accuracy of 55.5%, which is 8.4% over than the baseline system.

6. REFERENCES

[1] http://dcase.community/challenge2021/.

[2] Irene Martín-Morató, Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Low-complexity acoustic scene classification for multi-device audio: analysis of dcase 2021 challenge systems. 2021. arXiv:2105.13734.

[3] Suh S, Park S, Jeong Y, et al. Designing acoustic scene classification models with CNN variants[R]. DCASE2020 Challenge, Tech. Rep, 2020.

[4] 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). 2020. Submitted. URL: https://arxiv.org/abs/2005.14623.

[5] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez -Paz, Mixup: Beyond empirical risk minimization, arXiv preprint arXiv:1710.09412, 2017.

[6] He, K., Zhang, X., Ren, S., and Sun, J. (2016), Deep residual learning for image recognition, in Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[7] Koutini, K., Eghbal-zadeh, H., and Widmer, G. (2019), CPJKU submissions to DCASE19: Acoustic Scene Classification and Audio Tagging with Receptive-FieldRegularized CNNs, in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019).

[8] Gao, W., and McDonnell, M. (2019), Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths, DCASE2019 Challenge, Tech. Rep.

[9] Liu, M., Wang, W., and Li, Y, THE SYSTEM FOR ACOUSTIC SCENE CLASSIFICATION USING RESNET, DCASE2019 Challenge, Tech. Rep.

[10] M.D.M.W. Gao, ACOUSTIC SCENE CLASSIFICATION USING DEEP RESIDUAL NETWORKS WITH LATE FUSION OF SEPARATED HIGH AND LOWFREQUENCY PATHS, IEEE International Confer-ence on Acoustics, Speech and Signal Processing (ICASSP), (2020).

[11] S.P. Sangwon Suh, Youngho Jeong, Taejin Lee, Designing Acoustic Scene Classification Models with CNN Variants, DCASE2020 Challenge, Tech. Rep, (2020).

[12] K. Koutini, H. Eghbal-zadeh, M. Dorfer, a.G. Widmer, The Receptive Field as a Regularizer in Deep Convolutional Neural Networks for Acoustic Scene Classification, 2019 27th European Signal Processing Conference (EUSIPCO), (2019).