ACOUSTIC SCENE CLASSIFICATION WITH RESIDUAL NETWORKS AND ATTENTION MECHANISM

Technical Report

Jie Liu

Maxvision, Wuhan, China

ABSTRACT

This technical report describes our submission for TASK1A of DCASE2020 challenge. We use log-mel spectrograms and a residual network. We follow the idea of McDonnell [1] in DCASE2019 and do not downsample in the frequency axis. Besides, we use attention mechanism to improve the performance of the system.

Index Terms— residual network; log-mel spectrograms ; attention mechanism

1. INTRODUCTION

Task 1a in the 2020 DCASE Acoustic Scene Classification challenge aims to classify sounds into one of predefined classes. The dataset of DCASE2020 is different from DCASE2019. The task 1a 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. Audio is provided in single channel 44.1kHz 24-bit format. The development dataset is provided with a training/test split and some devices appear only in the test subset, which increases the difficulty. By using proposed method, we achieved a classification accuracy of 72.1% on the officially provided fold 1 evaluation dataset.

2. PROPOSED SYSTEM

Since the introduction of AlexNet [2] in 2012, deep convolutional neural networks have become the dominating approach for image classification. Also, the results of the previous DCASE challenges suggest that CNNs are the most popular classifiers for acoustic scene classification [3]. We also use a CNN model like others and adopt different attention mechanisms. We do not use any additional data and train the model from scratch.

2.1. Acoustic Feature

A number of features, such as the constant-Q transform (CQT), mel frequency cepstral coefficients (MFCC), perceptual weighted power spectrogram, wavelet and etc., have been used in acoustic scene classification. We use log-mel energies, and additionally calculate deltas and delta-deltas of the spectrum. We also used other features in our experiments, such as CQT, gammatone and wavelet. We find that log-mel energies performs best of all.

For Task 1a, the data is single channel, which is different from DCASE2019. We calculated log-mel spectrograms and deltas and delta-deltas, and consequently, the overall input to our CNN had 3 channels.

2.2. Model Design

ResNet [4] has proved its power in image classification and many other networks make a progress based on it. We follow the idea of [1] and use the model as a baseline. Spectrograms are different from images, since features at different frequencies represent different meanings. So we want to adopt attention mechanism, which may help the model to focus on crucial features.

We tried different attention modules, such as Squeeze-and-Excitation (SE) [5], Convolutional Block Attention Module (CBAM) [6] and point-wise attention [7]. These three attention modules are show in Figure 1. The original SE module is for channel attention, and we modified it for frequency attention. Point-wise attention module gets the best performance in our experiments.

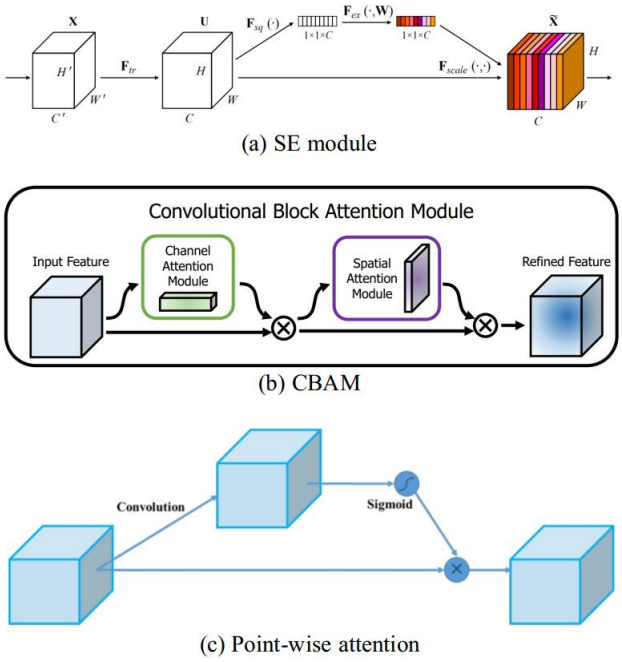
2.3. Data Augmentation

Data augmentation is a efficient way to avoid overfitting and enhance the model's generalization in deep neural networks. We use mixup [8] and crop for data augmentation. With cropping method, we can get more training data.

3. EXPERIMENTS AND RESULTS

3.1. Data Preprocessing

Audio of task 1a is provided in single channel 44.1kHz 24-bit format. To get the log-mel energies, we use 2048 FFT points, a hop-length of 1024 samples, frequencies from 0 to half of the sampling rate [1]. We use 128-bin mel filter bank and finally get log-mel spectrogram of size (128, 431). Then we calculate the log-mel deltas and delta-deltas. Finally we get the training data with mixup and crop.



(a) SE module

(b) CBAM

(c) Point-wise attention

Figure 1: Attention modules

3.2. Attention Mechanism

Attention mechanism has been widely adopted in image classification. Attention module can be integrated into CNN architectures seamlessly with negligible overheads. We tried three different attention modules, SE, CBAM and Point-wise attention.

The structure of an SE block is shown in Fig.1(a). SE block has two steps, squeeze and excitation. First, global average pooling is applied to the feature map. Then, we get the activation after a bottleneck with two fully-connected (FC) layers. The final output of the block is obtained by rescaling the original feature map with the activation. SE block can be regarded as a self-attention function on channels. In our experiment, we apply SE module on frequency axis.

CBAM has two sub-modules, the structure is shown in Fig.1(b). The channel sub-module utilizes both max-pooling outputs and average-pooling outputs with a shared network. It is different from SE block, which only use average pooling. The spatial sub-module of CBAM utilizes similar two outputs that are pooled along the channel axis. CBAM simultaneously applies channel attention and spacial attention.

Point-wise attention is proposed in YOLOv4 [7], the structure is shown in Fig.1(c). The attention value has the same shape with original feature map. Then a point-wise multiplication is applied. In out experiments, point-wise attention performs best. We finally adopt point-wise attention as our attention module.

3.3. Training

We used the official fold 1 procedure to evaluate our systems' performance. Stochastic gradient descent (SGD) optimizer was used with an initial learning rate of 0.01 and a mini-batch size of 32. Training epoch was set to 126.

Table 1: Class-wise accuracy for the development dataset

Class	Baseline(%)	Proposed(%)
airport	45.0	52.5
bus	62.9	86.9
metro	53.5	66.3
metro_station	53.0	77.1
park	71.3	89.9
public_square	44.9	52.9
shopping_mall	48.3	65.7
street_pedestrian	29.8	60.3
street_traffic	79.9	89.6
tram	52.2	80.5
average	54.1	72.1

3.4. Result

The experimental results obtained by our model over the validation dataset are shown in Table 1. We retrained the model using all the development data before running the model on evaluation data for submission.

4. CONCLUSIONS

In this technical report, we proposed a acoustic scene classification system. We use log-mel spectrograms and a residual network with attention mechanism to improve the performance of the system. We achieved a classification accuracy of 72.1% with a single model, which is 18% over than the baseline system.

5. REFERENCES

[1] Mark D. McDonnell and Wei Gao, "Acoustic Scene Classification Using Deep Residual Networks with Late Fusion of Separated High and Low Frequency Paths," DCASE2019 Challenge, Tech. Rep., June 2019.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[3] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification: an overview of DCASE 2017 challenge entries," in 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[5] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[6] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[7] Bochkovskiy, Alexey, Chien-Yao Wang, and Hong-Yuan Mark Liao. "YOLOv4: Optimal Speed and Accuracy of Object Detection." arXiv preprint arXiv:2004.10934 (2020).

[8] H. Zhang, M. Cisse, Y. N. Dauphin, and D. LopezPaz, “mixup: Beyond Empirical Risk Minimization,” in arXiv:1710.09412, 2017.