Multi-Scale Recalibrated Features Fusion for Acoustic Scene Classification

Technical Report

Chongqin Lei

Chongqing University
College of Optoelectronic Engineering, No.174 Shazhengjie
Shapingba, Chongqing, 400044, China
leichongqin@cqu.edu.cn

Zixu Wang

Chongqing University
College of Optoelectronic Engineering, No.174 Shazhengjie
Shapingba, Chongqing, 400044, China
201808021045@cqu.edu.cn

ABSTRACT

We investigate the effectiveness of multi-scale recalibrated features fusion for acoustic scene classification as contribution to the subtask of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2019). A general problem in acoustic scene classification task is audio signal segment contains less effective information. In order to further utilize features with less effective information to improve classification accuracy, we introduce the Squeeze-and-Excitation unit to embed the backbone structure of Xception to recalibrate the channel weights of feature maps in each block. In addition, the recalibrated features of multiscale are fused and finally fed into the full connection layer to get more useful information. Furthermore, we introduce Mixup method to augment the data in training stage to reduce the degree of over-fitting of network. The proposed method attains a recognition accuracy of 77.5%, which is 13% higher compared to the baseline system of the DCASE 2019 Acoustic Scenes Classification task.

Index Terms**—** acoustic scene classification, multi-scale features fusion, recalibrated features, mixup

1. INTRODUCTION

Acoustic scene carries exceedingly complex and diverse sounds information, and the background sounds of the target scene contain a large number of random and complex sounds, which makes the acoustic scene classification (ASC) task extremely challenging but quite meaningful. DCASE 2019 challenge organized by IEEE Audio and Acoustic Signal Processing (AASP) Technical Committee is one of the large-scale challenges for ASC research, which provides an excellent platform for ASC task to promote its development [1].

DCASE 2019 task 1 is fundamentally an extended version of the previous DCASE 2018 ASC task, providing a larger amount of data for the same scenes. Many of participants applied a deep learning approach such as convolutional neural networks (CNNs) [2, 3, 4, 5] and recurrent neural networks (RNNs) [6, 7], and top ranks were achieved by CNNs in DCASE 2017 and 2018’s submitted algorithms. On the contrary, top ranks were achieved by non-negative matrix factorization [8], which are comparatively traditional dictionary learning methods in the Challenge of DCSE 2016. And most of the submitted algorithms in challenge of DCASE 2018 used log-Mel spectrograms, one of the most popular handcrafted features. As we can see from the results of the DCASE task in the past, the deep learning approach has shown promising results.

With the rapid development of deep learning and continuous performance breakthroughs of CNNs, approaches based on CNNs which is widely being used for image processing has also been applied for ASC [9]. In the past three years, challenges of DCASE have received numerous approaches based on deep learning.

The convolutional neural network extracts abstract features by merging spatial information on a channel-by-channel basis using local receptive fields [10]. It is much difficult to train a performance-efficient network. On the one hand, from the perspective of spatial dimension, for example, the Inception structure [11] embeds multi-scale information and aggregates features of different receptive fields to improve performance. Based on Inception V3's improved network Xception [12][13], in this paper, we improve the classification performance of the network by combining feature maps of different scales as input to the classifier. On the other hand, according to the relationship of feature channels, Squeeze-and-Excitation Network (SENet) [14] selectively enhances the informatizable features and compresses useless features by using global information. On this basis, we introduce the Squeeze-and-Excitation unit to embed the residual structure of Xception, which can explicitly model the channel correlation between convolution layer features to improve the representation ability. In addition, deep neural networks have a large number of model parameters, so that for data with few samples or few effective information in the sample, which is extremely easy to produce a over-fitting phenomenon. In order to solve the problem, data augmentation is effective [15]. In this paper, the Mixup [16] method was introduced to augment the data, thereby reducing the degree of over-fitting of the model and improving the generalization ability of the model.

The remainder of this paper is organized as follows. Section 2 explains details of the proposed system. Section 3 discusses the experiments and results. Finally, the conclusion is provided in Section 4.

2. METHOD

2.1. Neural Network Architecture

The baseline network of this work is Xception, which is an improved network proposed by Google based on Inception V3.

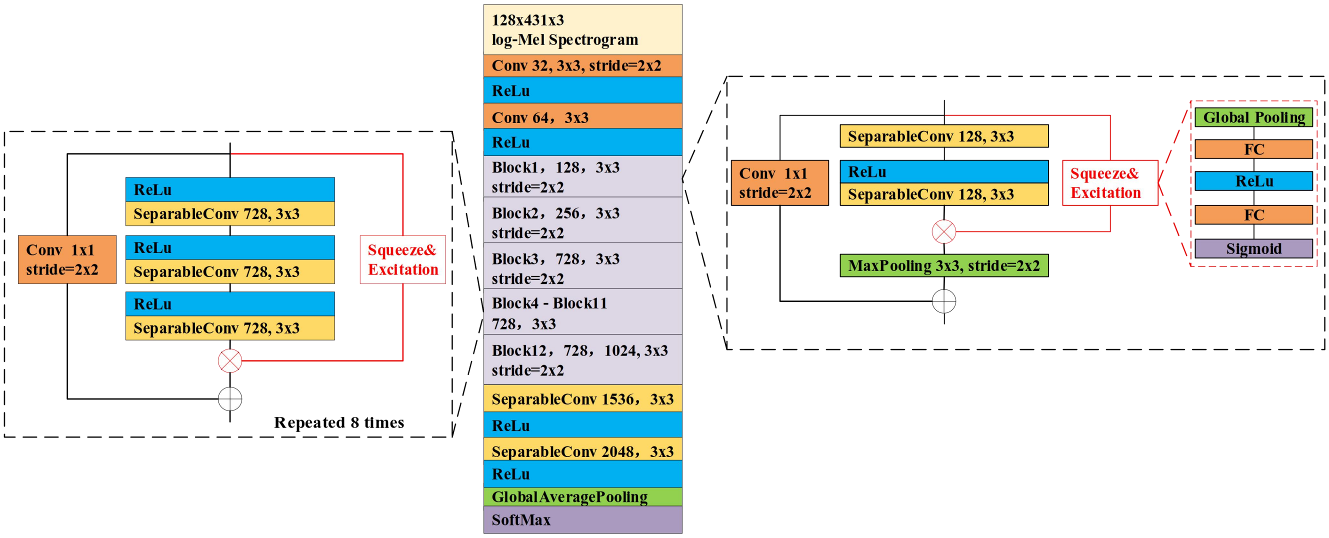


Figure 1：The overall architecture of the network. The intermediate structure is a backbone network containing 12 blocks, where Block4 to Block11 are the same Block, and the rest are similar Blocks. The structure on the left is the detail of Block4, and the structure on the right is the detail of Block1 (the green part is the detail of the SE module), the red connection part in the figure is the SE module introduced in this paper.

Its main improvement is to replace the convolution operation in the original Inception V3 with depthwise separable convolution, and add a residual connection mechanism to the model to speed up the convergence of the model.

In this paper, the SE module is added to Xception to enable the network to automatically acquire the importance of each feature channel through training. Figure 1 shows the overall network architecture. There are 12 blocks in the backbone network, each of which contains similar convolution and pooling operations. We describe the different Blocks in detail. The connection marked red in the enlarged Block structure is the SE module we introduced. The details of the SE module are further explained in the right enlarged Block structure. We add the SE module in front of the Maxpooling layer of each block in Xception, and the result is added to the output of the residual connection and then used as the input to the next block.

2.2. Spectrogram Extraction

Effective feature extraction is the basis of improving the accuracy of acoustic scene classification. Subtask 1A of DCASE 2019 Challenge provides data that is stereo, 48kHZ. The commonly used Python toolkit for feature processing, such as Librosa, defaults to using mono-channel data with a sampling rate of 22.05 kHz. In this paper, the sampling rate of the original audio signal is set to 44.1 kHz, and the log-Mel spectrograms are extracted from the two channels respectively, then the average values of the two channels are obtained. Finally, the three spectrograms are stacked into three-channel data. Figure 2 shows the feature extraction method used in this paper.

The sound in the acoustic scene is a random non-stationary signal. It is generally considered that the sound signal is a stationary signal within 20ms-30ms. Therefore, when the feature is extracted, the hop-length is set to 1024, the sampling rate is 44.1kHZ, the number of STFT points is 4096. and the Mel filters is 128, which results in a log-Mel Spectrogram size of (128, 443).

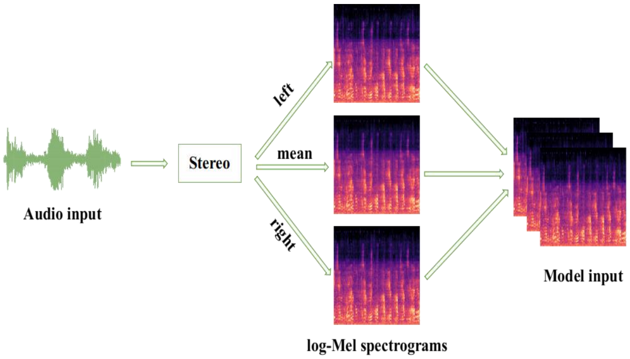


Figure 2: Illustration of log-Mel spectrogram for network input.

2.3. Multiple Scale Feature

In the acoustic scene classification task, the collected audio data is cut for equal duration, generally according to the standard of 5s or 10s per segment, and the audio segment of the dataset provided by subtask 1A is 10 seconds. The clipped audio clip may contain only a few target sounds, so the features extracted from the audio clip contain only a small amount of valid information. In order to make better use of the information in the features, we concat the features of several different scales and input them into the classifier through feature fusion, so that we can make more effective use of different layer features with different information to compensate the disadvantage that the input feature has less effective information.

Specifically, we connect the output of Block3, the output of Block11 and the input of MaxPooling layer, and then input the feature vectors to the full connection layers. This simple and effective operation improves the classification results of the model by combining the features of different scales, but it does not increase too much computational cost.

3. EXPERIMENTAL SETTINGS AND RESULTS

3.1. Dataset

The dataset for Subtask 1A is the TAU Urban Acoustic Scenes 2019 dataset, which extends the TUT Urban Acoustic Scenes 2018 dataset [17] with other 6 cities to a total of 12 large European cities. This subtask is concerned with the basic problem of acoustic scene classification, in which all available data (development and evaluation) are recorded with the same device, in this case device A. The dataset consists of 10-seconds audio segments from 10 acoustic scenes, each acoustic scene has 1440 10-second segments (48 kHz / 24bit / stereo, 240 minutes of audio).

The dataset was recorded in 12 large European cities. The development dataset contains audio material from 10 cities, whereas the evaluation dataset contains data from all 12 cities. The dataset is perfectly balanced at acoustic scene level, with very slight differences in the number of segments from each city.

3.2. Common Experimental Settings

The preprocessing and feature extraction of raw audio in this article relies on LIBROSA, a powerful audio processing python toolkit. And all experiments were completed in the PYTORCH environment of the Ubuntu 16.04 system.。

Our network is trained for 80 epochs in batches of 16 samples by optimizing the categorical cross-entropy and stochastic gradient descent (SGD) with Nesterov momentum, and we apply 40% dropout to the full connection layers. The learning rate, mini-batch size, and decay were set to 0.0001, 16, and 0.0001, respectively. The strategy of cosine annealing is used in training, which is one cycle for every 20 epochs with initial learning rate of 0.0001, and the learning rate decreases twice in each cycle. And we use Mixup method with α = 0.4 to augment the data.

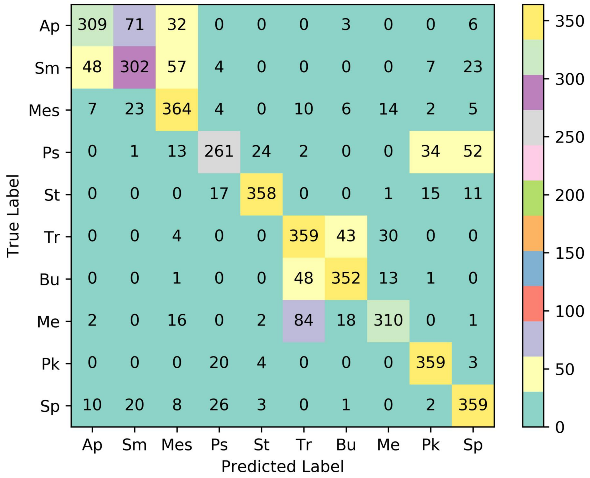
3.3. Experiment Results

The class-wise accuracy of the submitted method is summarized in Table 1. In the challenge of DCASE 2019, we only submitted the results of subtask A of task 1. As can be seen from the table, the classification results of our method are better than those of the official baseline model in all scenarios. The results of our method were obtained on the TAU Urban Acoustic Scenes 2019 development dataset, and the result on the official leaderboard dataset was 0.775.

The confusion matrix of submitted model result is presented in Fig. 4. The labels of the X-axis (or Y-axis) in the confusion matrix represent "Airport", "Shopping_mall", "Metro_station", "Public_square", "Street_traffic", "Tram", "Bus", "Metro", "Park" and "Street_pedestrian" from top to bottom (or left to right). And it can be observed that the confusion is relatively focused in the Metro and Tram, park, Airport and Shopping mall.

Table 1: The class-wise accuracy of the submitted method outperformed the baseline of the development set.

Scene class	Subtask A Accuracy(%)	
	Baseline	**Our method**
Airport	0.484	**0.734**
Bus	0.623	**0.8482**
Metro	0.651	**0.7159**
Metro_station	0.545	**0.8368**
Park	0.831	**0.93**
Public_square	0.407	**0.6744**
Shopping_mall	0.594	**0.6848**
Street_pedestrian	0.609	**0.8368**
Street_traffic	0.867	**0.8905**
Tram	0.64	**0.8234**
Average	**0.6251(+/- 0.14)**	**0.7964(+/- 0.09)**



True Label	Ap	Sm	Mes	Ps	St	Tr	Bu	Me	Pk	Sp
Ap	309	71	32	0	0	0	3	0	0	6
Sm	48	302	57	4	0	0	0	0	7	23
Mes	7	23	364	4	0	10	6	14	2	5
Ps	0	1	13	261	24	2	0	0	34	52
St	0	0	0	17	358	0	0	1	15	11
Tr	0	0	4	0	0	359	43	30	0	0
Bu	0	0	1	0	0	48	352	13	1	0
Me	2	0	16	0	2	84	18	310	0	1
Pk	0	0	0	20	4	0	0	0	359	3
Sp	10	20	8	26	3	0	1	0	2	359

Figure 4: Confusion matrix of the submitted model of the development set. X-axis indicates the predicted label and Y-axis indicates the true label.

The confusion matrix of submitted model result is presented in Fig. 4. The labels of the X-axis (or Y-axis) in the confusion matrix represent "Airport", "Shopping_mall", "Metro_station", "Public_square", "Street_traffic", "Tram", "Bus", "Metro", "Park" and "Street_pedestrian" from top to bottom (or left to right). And it can be observed that the confusion is relatively focused in the Metro and Tram, park, Airport and Shopping mall.

The same experiment settings from development set for the evaluation set are used. For the final submission, we submitted two different results. We used the model described in this paper for submission 1, and a voting rule is utilized for decision fusion of three different models for submission 2.

4. CONCLUSIONS

In this paper, we presented a useful model using Squeeze-and-Excitation unit and multi-scale feature fusion method developed for the DCASE Challenge 2019. We have addressed task 1A - Acoustic Scene Classification and have outperformed the baseline accuracy by 13% using our method. Multichannel log-Mel spectrograms are used as input of the model, and Mixup method is used for data augmentation in our work.

In future work, we try to test our method over a wide range of different acoustic classification tasks. We also want to collect further data from social multimedia to train the network with more real life audio recordings. Finally we intend to work on bioacoustic recognition, such as bird sound detection and recognition. Maybe our method is not suitable for these tasks, but the rapid development of deep neural network provides us with very good tools in these fields.

5. REFERENCES

[1] Han Y, Park J, Lee K. Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification[J]. the Detection and Classification of Acoustic Scenes and Events (DCASE), 2017: 1-5.

[2] E. Marchi, D. Tonelli, X. Xu, F. Ringeval, J. Deng, and B. Schuller, “The up system for the 2016 DCASE challenge using deep recurrent neural network and multiscale kernel subspace learning, ” DCASE2016 Challenge, Tech. Rep.,September 2016.

[3] S. H. Bae, I. Choi, and N. S. Kim, “ Acoustic scene classification using parallel combination of LSTM and CNN, ” DCASE2016 Challenge, Tech. Rep., September 2016.

[4] V. Bisot, R. Serizel, S. Essid, and G. Richard, “Supervised nonnegative matrix factorization for acoustic scene classification, ” DCASE2016 Challenge, Tech. Rep., September 2016.

[5] Xu K, Feng D, Mi H, et al. Mixup-based acoustic scene classification using multi-channel convolutional neural network[C]. Pacific Rim Conference on Multimedia. Springer, Cham, 2018: 14-23.

[6] Qian K, Ren Z, Pandit V, et al. Wavelets revisited for the classification of acoustic scenes[C]. Proc. DCASE Workshop, Munich, Germany. 2017: 108-112.

[7] J. Salamon and J. P. Bello, “Unsupervised feature learning for urban sound classification,” in IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Brisbane,Australia, Apr. 2015, pp. 171–175.

[8] V. Bisot, R. Serizel, S. Essid, and G. Richard, “ Feature learning with matrix factorization applied to acoustic scene classification, ” in IEEE Trans. Audio, Speech, Lang. Process., vol. 25, no. 6, pp. 1216-1228. June 2017Shanghai, China, Mar. 2016, pp. 6445–6449.

[9] Jimenez A, Elizalde B, Raj B. DCASE 2017 task 1: Acoustic scene classification using shift-invariant kernels and random features[J]. arXiv preprint arXiv:1801.02690, 2018.

[10] Jacobsen, Jörn-Henrik, Van Gemert J , Lou Z , et al. Structured Receptive Fields in CNNs[J]. 2016.

[11] Szegedy C , Liu W , Jia Y , et al. Going Deeper with Convolutions[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.

[12] Szegedy C , Vanhoucke V , Ioffe S , et al. Rethinking the Inception Architecture for Computer Vision[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016:2818-2826.

[13] Chollet, François. Xception: Deep Learning with Depthwise Separable Convolutions[J]. 2016.

[14] Hu J , Shen L , Albanie S , et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[15] Mikolajczyk A , Grochowski M . Data augmentation for improving deep learning in image classification problem[C]. 2018 International Interdisciplinary PhD Workshop (IIPhDW). IEEE, 2018:117-122.

[16] Zhang H , Cisse M , Dauphin Y N , et al. mixup: Beyond Empirical Risk Minimization[J]. 2017.

[17] Kong Q, Iqbal T, Xu Y, et al. DCASE 2018 Challenge Surrey Cross-task convolutional neural network baseline[J]. Parameters, 2018, 4: 4,691,274.