ACOUSTIC SCENE CLASSIFICATION BASED ON BINAURAL DEEP SCATTERING SPECTRA WITH CNN AND LSTM

Technical Report

Zhitong Li, Liqiang Zhang, Shixuan Du, Wei Liu

Beijing Institute of Technology
Laboratory of Modern Communication
No. 5 South Zhongguancun Street, Beijing , China
Lizhitong_bit@163.com, zlqbit@qq.com, 1377051501@qq.com, 1120143589@bit.edu.cn

ABSTRACT

This technical report presents the solutions proposed by the Beijing Institute of Technology Modern Communications Technology Laboratory for the acoustic scene classification of DCASE2018 task1a. Compared to previous years, the data is more diverse, making such tasks more difficult. In order to solve this problem, we use the Deep Scattering Spectra (DSS) features. The traditional features, such as Mel-frequency Cepstral Coefficients (MFCC), often lose information at high frequencies. DSS is a good way to preserve high frequency information. Based on this feature, we propose a network model of Convolutional Neural Network (CNN) and Long Short-term Memory (LSTM) to classify sound scenes. The experimental results show that the proposed feature extraction method and network structure have a good effect on this classification task. From the experimental data, the accuracy increased from 59% to 76%.

Index Terms— DCASE2018, acoustic scene classification, convolutional neural network, long short-term memory, Deep Scattering Spectra, binaural representations

1. INTRODUCTION

Sounds carry a large amount of information, in addition to speech information, the sound scene is also a key information point. For example, the sound of a coffee shop is often different from that of the subway. However, it is of great significance to automatically realize the classification of acoustic scenes by means of technology. For example, artificial intelligence can turn on automatic push based on the sound scene, or accurately identify the location of the elderly and children. At present, there are a large number of scholars in this area of research with good progress. DCASE's results from previous years also show potential.[1,2]

DCASE has designed a special task to classify acoustic scenes. The number of scenes classified this year has been reduced from 15 to 10, but the locations of the recordings have become richer and the dataset becomes more numerous. This makes the task more difficult, but also increases its generalizability in real-world applications.

Acoustic features play an important role in the classification of acoustic scenes. A good acoustic feature selection can be a better characterization of the characteristics of the sound, which can help to achieve better classification results. Conventional acoustic features, such as MFCC, Perceptual Linear Prediction (PLP), are all designed to be deformation stable, they remove important higher-order information from the speech signal. Deep scattering networks (DSN) have recently been introduced to solve this challenge. DSNs can generate a contractive representation of a raw signal, doing like this can preserves signal energy, while ensuring time-shift invariant and stability to time deformations. The representation generated by these networks id called Deep Scattering Spectra (DSS).

Depth neural network shows good potential in the field of ASC, based on the excellent scheme of DCASE2016 and 2017. In this paper, we investigate the possibility of integrating DSS features with Convolutional Neural Networks (CNN) for this acoustic scene classification (ASC) task.

According to the experimental results, the combination of DSS features and CNN network makes the classification task get good results, and the classification accuracy is improved by 17% compared with the baseline system.

2. DEEP SCATTERING SPECTRA

Audio feature should be time-invariant and stable to time deformation. The former means that that the audio segment always belongs to the same class even if it is shifted by a constant in time. Stability to time warping means that small deformation in the raw signal leads to small modification in audio feature. Mostly owing to its properties of group invariance and stability to deformations, DSS has shown to achieve state-of-the art results in the challenges of music genre recognition, image, texture classification, and fetal heart rate characterization. Its core feature relies on the construction of a scattering network, i.e. a stack of signal processing layers of increasing width. Each layer consists in the association of a linear filter bank with a non-linear operator, namely the complex modulus. The scattering transform of an input signal x is defined as the set of all paths that x might take from layer to layer. In this sense, the architecture of a scattering network closely resembles a convolutional deep network.[7,8]

2.1. Time Scattering

As shown in [6] , log-mel features can be approximated by convolving in time a signal x with a wavelet filterbank. This feature representation can be written as

$$F_1 = | x * \varphi_{\lambda_1} | * \phi(t) , \tag{1}$$

where φ_{λ_1} denotes a wavelet filterbank and $\phi(t)$ denotes a low-pass filter. While time averaging provides features which are locally invariant to small translations and distortions, it also leads to loss of higher-order information in the speech signal, such as attacks and bursts [6]. To recover this lost information another decomposition of the sub-band signals is performed using a second wavelet filter-bank, denoted by φ_{λ_2} , This second decomposition captures the information in the sub-band signal, $| x * \varphi_{\lambda_1} |$, left out by the averaging filter $\phi(t)$. The decomposed sub-band signals are denoted by

$$F_2 = | x * \varphi_{\lambda_1} | * \varphi_{\lambda_2} , \tag{2}$$

are once again passed through the low-pass filter $\phi(t)$ to extract stable features. The second order scatter is computed using a constant-Q filter-bank with Q = 1. Each of the decompositions can be written as

$$F_3 = || x * \varphi_{\lambda_1} | * \varphi_{\lambda_2} | * \phi(t) , \tag{3}$$

has a limited number of non-zero coefficients, due to the band-limited nature of the signals $| x * \varphi_{\lambda_1} |$. Typically, only first and second order scatter is used for speech [4, 10]. Again, following the terminology of [6], the second order scatter is referred to as S_2.

The above description is known as time-scatter, as the wavelet convolution is applied to the time domain signal only.

2.2. Frequency Scatter

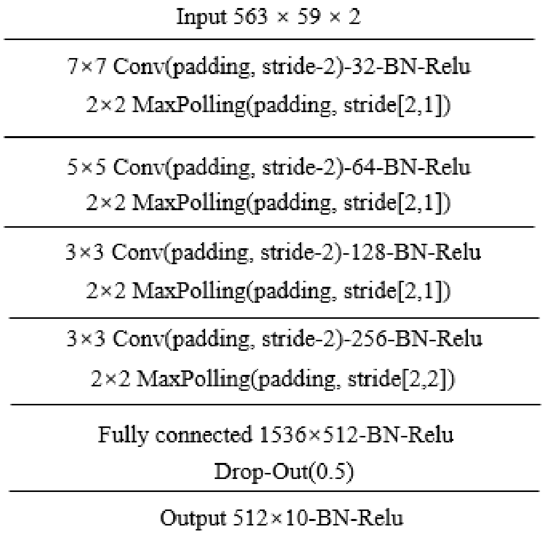
Frequency scatter can be seen as a way of removing variability in the frequency signal, for example due to translations of formants created from different speaking styles. A very simple type of frequency averaging is to apply a discrete cosine transform (DCT) to a log-mel representation and perform cepstral truncation, which is common when generating MFCCs. When applying frequency scatter in the DSS framework, the same time-scattering operation performed in time is now performed in the frequency domain on the S1 and S2 features. Specifically, frequency scattering features are created by iteratively applying wavelet transform and modulus operators, followed by a low-pass filter to the time-scatter features S_i , $| S_i * \varphi_{\lambda_1}^{f_r} | * \phi^{f_r}(t)$. All frequency-scattering features are produced using wavelets with Q = 1. Similar to [6], we only compute first-order frequency scatter.

2.3. Multi-Resolution Scatter

The first-order time-scattering operating described in Section 2.1, is performed using a wavelet with Q = 9. To capture different spectral and temporal dynamics, wavelets with different Q factors can be used, an operation known as multi-resolution time scatter. Frequency and second-order scatter are calculated on each first-order time scatter S_1 generated with filterbank Q.

3. PROPOSED SYSTEM

CNN (the convolution neural network), has the characteristics of local connection and weight sharing, which greatly reduces the number of parameters, improves the training speed, and reduces over-fitting. Because of its good processing ability to high dimensional array, CNN is widely used in speech recognition, image recognition and other fields. As for the acoustic scene classification proposed in this competition, we decided to use CNN structure to build our neural network. The structure is shown in figure 1. [3,4,5]



Input 563 × 59 × 2

7×7 Conv(padding, stride-2)-32-BN-Relu
2×2 MaxPolling(padding, stride[2,1])

5×5 Conv(padding, stride-2)-64-BN-Relu
2×2 MaxPolling(padding, stride[2,1])

3×3 Conv(padding, stride-2)-128-BN-Relu
2×2 MaxPolling(padding, stride[2,1])

3×3 Conv(padding, stride-2)-256-BN-Relu
2×2 MaxPolling(padding, stride[2,2])

Fully connected 1536×512-BN-Relu
Drop-Out(0.5)

Output 512×10-BN-Relu

Figure 1: Proposed Neural Network Structure

Our neural network consists of one input layer, four layers of convolution and maxpooling, one layer of full connected and one layer of output.

The input data is the 3d vector (563 × 59 × 2) obtained after the extraction of DSS features based on double channels.

The convolution layer mainly uses the convolution kernel and the input matrix to perform convolution operation, so as to obtain deeper features. Because of the weight sharing of the convolution layer, the information of the adjacent time dimension can be shared. At the same time, after the convolution, the size of the output matrix will also be reduced, thus reducing the number of parameters.

The pooling layer is sandwiched in the middle of the continuous convolutional layer. By selecting the largest element in the local matrix, the data is compressed, the useless information is discarded, the operation is accelerated, and the over-fitting is weakened. After 4 layers of neural network composed of convolutional layer and pooling layer, a new three-dimensional vector (3 × 2 × 256) is obtained. And then it is compressed into a one-dimensional vector and connected with the full connected layer.

The full connected layer adopts the 512-dimensional single hidden layer, and finally outputs a vector containing 10 dimensions. After softmax, the classification results are obtained. In addition, we also adopted the drop-out method at the full connected layer, which randomly omitted 50% of parameters during training to prevent over-fitting and improve generalization. Finally, we use cross-entropy as the loss function and update the parameters by using the gradient descent method with Adam optimizer, so that the loss function converges to the minimum value and the classification accuracy reaches the highest.

A Deep Neural Network (DNN) structure was also used to better utilize the frequency scattering features. The structure of the DNN is 1024*2018*1042. The output of the DNN is concatenated with the output of the CNN.

4. RESULTS AND CONCLUSIONS

DDSs features in the setting of different parameters, will get different dimensions and different effects of the characteristics. In order to get the best experimental results, we try different DSS features, The optimal characteristic parameters and network structure are determined. We submitted the best four groups of results as shown in the table1 below.

TABLE1: The submissions of the four final structures

ID	Features	Classifier	Acc
1	Wavelet-14-9	CNN	74.34%
2	Wavelet-13-9/Wavelet-14-9	CNN+DNN	75.21%
3	Wavelet-13-9/Wavelet-14-9/Wavelet-14-8	CNN+DNN	76.60%
4	Wavelet-13-9/Wavelet-14-9/Wavelet-14-8	CNN+DNN	76.44%

Among the four submissions, the forth one has the highest accuracy, and the confusion matrix is shown in table2.

TABLE2: The confusion matrix of the No.4 submission

173	35	30	16	2	8	1	0	0	0
36	217	7	12	7	0	0	0	0	0
4	8	230	3	0	6	2	1	5	0
6	8	0	173	39	17	0	0	0	4
2	0	1	14	108	54	0	0	0	37
0	0	0	19	5	220	0	1	0	1
0	4	4	0	0	4	195	11	43	0
0	0	0	0	0	0	43	191	7	1
0	0	28	0	0	1	41	8	183	0
0	0	1	0	5	1	0	0	0	235

5. ACKNOWLEDGMENT

This work is supported by National Nature Science Foundation of China (Grant No.61473041, No.11590772, No.61571044) and the joint research project between Beijing Institute of Technology and Ericsson AB.

6. REFERENCES

[1] http://www.cs.tut.fi/sgn/arg/dcase2016/index

[2] http://www.cs.tut.fi/sgn/arg/dcase2017/index

[3] Mun, S., Shon, S., Kim, W., & Ko, H. (2016). Deep Neural Network Bottleneck Features for Acoustic Event Recognition. INTERSPEECH(pp.2954-2957).

[4] Hyder, R., Ghaffarzadegan, S., Feng, Z., Hansen, J. H. L., Hasan, T., & Hyder, R., et al. (2017). Acoustic Scene Classification Using a CNN-SuperVector System Trained with Auditory and Spectrogram Image Features. INTERSPEECH (pp.3073-3077).

[5] Takahashi, N., Gygli, M., Pfister, B., & Gool, L. V. (2016). Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Recognition. INTERSPEECH (pp.2982-2986).

[6] Andén, J., & Mallat, S. (2014). Deep scattering spectrum. IEEE Transactions on Signal Processing, 62(16), 4114-4128.

[7] Joakim, A., Vincent, L., & Stephane, M. (2015). Joint time-frequency scattering for audio classification. 1-6.

[8] TN Sainath , V Peddinti , B Kingsbury , P Fousek , B Ramabhadran (2014). Deep scattering spectra with Deep Neural Networks for LVCSR tasks. INTERSPEECH(pp.900-904).