ACOUSTIC SCENE CLASSIFICATION BASED ON BINAURAL DEEP SCATTERING SPECTRA WITH NEURAL NETWORK

Technical Report

Sifan Ma, Wei Liu

Beijing Institute of Technology
Laboratory of Modern Communication
No. 5 South Zhongguancun Street, Beijing , China
masifan18@126.com, liuwei3589@163.com

ABSTRACT

This technical report presents our approach for the acoustic scene classification of DCASE2019 task1a. Compared to traditional audio features such as Mel-frequency Cepstral Coefficients (MFCC) and Constant-Q Transform (CQT), we choose Deep Scattering Spectra (DSS) features which are more suitable for characterizing acoustic scenes. DSS is a good way to preserve high frequency information. Based on DSS features, we choose a network model of Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU) to classify acoustic scenes. The experimental results show that our approach increase the classification accuracy from 62.5% (DCASE2019 baseline) to 85% .

Index Terms— DCASE2019, acoustic scene classification, Deep Scattering Spectra, convolutional neural network.

1. INTRODUCTION

Perceiving and understanding sound signals is an important research direction in the field of artificial intelligence. The sound signal carries a lot of information, such as the environment in which the sound is located. When we listen to a piece of audio, we often ignore its scene. So, to have a better understand of the acoustic scene is important. 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]

In DCASE2019, the goal is to classify a test recording into one of the provided predefined classes that characterizes the environment in which it was recorded.

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. The audio features used in the existing scene recognition methods are mostly based on the cepstrum domain MFCC, in addition to other frequency domain and time domain features. But they are statistical values either short-term features or of long-term features. The short-term feature cannot completely describe the audio scene, while the long-term statistical feature will lose the local structural information of the scene signal, which will ultimately affect the recognition effect of the audio scene. 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).

In this report, we introduce the framework we used. From the experimental results, we explored the possibility of combine DSS features with CNN for acoustic scene classification.

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.[3,4]

2.1. Time Scattering

As shown in [5], log-mel features can be approximated by convoluing in time a signal x with a wavelet filterbank. This feature representation can be written as

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

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 [5]. 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 bandlimited nature of the signals$|x^*\varphi_{\lambda_1}|$. Typically, only first and second order scatter is used for speech. Again, following the terminology of [5], 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 S_1 and S_2 features. Specifically, frequency scattering features are created by iteratively applying wavelet trans-form and modulus operators, followed by a low-pass filter to the time-scatter featuresS_i, $\left|S_i{}^*\varphi^{fr}_{\lambda_1}\right|^*\phi^{fr}(t)$. All frequency scattering features are produced using wavelets with Q = 1. Similar to [5], 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. SYSTEM STRUCTURE

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. [6,7,8]

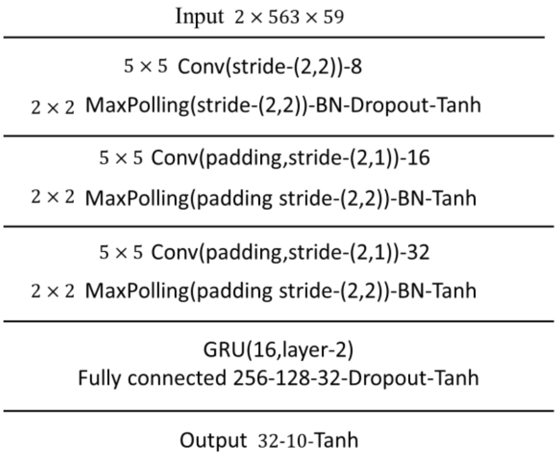

Input 2 × 563 × 59
5 × 5 Conv(stride-(2,2))-8
2 × 2 MaxPolling(stride-(2,2))-BN-Dropout-Tanh
5 × 5 Conv(padding,stride-(2,1))-16
2 × 2 MaxPolling(padding stride-(2,2))-BN-Tanh
5 × 5 Conv(padding,stride-(2,1))-32
2 × 2 MaxPolling(padding stride-(2,2))-BN-Tanh
GRU(16,layer-2)
Fully connected 256-128-32-Dropout-Tanh
Output 32-10-Tanh

Figure 1 System Structure

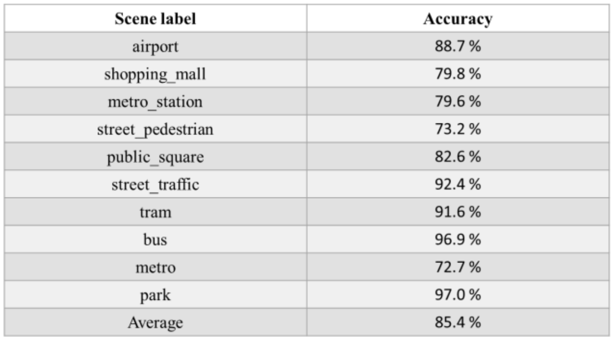
Our neural network consists of one input layer, three layers of convolution, one layer of gated recurrent unit, one layer of full connected and one layer of output. The input data is the 3d vector (2×563×59) obtained after the extraction of DSS features based on double channels.

Based on previous experience, we choose DSS features with dimension of 2×563×59. Use two channels input to cover binaural representations. Maxpolling was used to reduce the number of parameters. Next layer we use GRU with 10×2 cells. And then it is compressed into a one-dimensional vector and connected with the full connected layer. The full connected layer adopts the 256-128-31 hidden layer, and finally outputs a vector containing 10 dimensions.

4. RESULTS

We test our system on the test set containing 2880 pieces of audio, which are divided from the whole dataset. Mean and standard deviation of the performance from these 10 independent trials are shown in the results tables.

Table 1 Test Results



Scene label	Accuracy
airport	88.7 %
shopping_mall	79.8 %
metro_station	79.6 %
street_pedestrian	73.2 %
public_square	82.6 %
street_traffic	92.4 %
tram	91.6 %
bus	96.9 %
metro	72.7 %
park	97.0 %
Average	85.4 %

5. REFERENCES

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

[2] http://www.cs.tut.fi/sgn/arg/dcase2018/index

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

[4] 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).

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

[6] 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 Fea-tures. INTERSPEECH (pp.3073-3077).

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

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