THE SEIE-SCUT SYSTEMS FOR CHALLENGE ON DCASE 2018: DEEP LEARNING TECHNIQUES FOR AUDIO REPRESENTATION AND CLASSIFICATION

Yanxiong Li,Yuhan Zhang, Xianku Li

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
eeyxli@scut.edu.cn

ABSTRACT

In this report, we present our works about one task of challenge on DCASE 2018, i.e. task 1b:Acoustic Scene Classification with mismatched recording devices (ASC). We adopt deep learning techniques to extract Deep Audio Feature (DAF) and classify various acoustic scenes . Specifically, a Deep Neural Network (DNN) is first built for generating the DAF from Mel-Frequency Cepstral Coefficients (MFCCs), and then a Recurrent Neural Network (RNN) of Bidirectional Long Short Term Memory (BLSTM) fed by the DAF is built for ASC. Evaluated on the development datasets of DCASE 2018, our systems are superior to the corresponding baselines for tasks 1b.

Index Terms—DAF, BLSTM, Acoustic Scene Classification

1. INTRODUCTION

ASC is a process of determining a test audio recording belongs to which pre-given class of acoustic scenes, it can be regarded as the same task of audio representation and classification and tackled by using the same feature and classifier. It is useful for multimedia retrieval [1], audio-based surveillance and monitoring [2, 3]. What's more, they are under great attention of the research community with many evaluation campaigns [4-8], and are not effectively solved due to large variations of time-frequency characteristics within each class of sound events and acoustic scenes, non-stationary background noises, overlapping of sound events, and so forth [9].

The overall performance of audio classification system mainly depends on two stages: feature extraction and classifier building. Almost all of recent studies focused on these two stages for achieving better performance [10]. Many systems were submitted to the DCASE 2017 challenge for ASC and/or SED, and some of them achieved satisfactory results. They were based on the combinations of various features with different classifiers. The features include MFCCs, log Mel-band energy, spectrogram, Gabor filterbank, pitch, time difference of arrival, amplitude modulation filterbank, while the classifier mainly consists of Gaussian mixture model, Deep Convolutional Neural Network(DCNN), RNN, time-delay neural network, logistic regression, random forest, decision tree, gradient boosting, support vector machine, hidden Markov model. For example, Eghbal-Zadeh et al [11] proposed a novel I-vector extraction scheme for ASC using both left and right audio channels, and proposed a DCNN architecture trained on spectrograms of audio excerpts in end-to-end fashion.Their submissions achieved ranks first and second among 78 submissions in the ASC task of DCASE 2017 challenge. Adavanne et al [12] used spatial and harmonic features in combination with BLSTM RNN for SED. Their method improved the F-score by 3.75% while reducing the error rate by 6% compared with the baselines.

Although so many systems have been proposed for ASC and SED, to the best of our knowledge, there is no system by combining the DAF for audio representation with the BLSTM for audio classification. In our submissions for DCASE 2018, we propose to build a DNN for extracting the DAF based on MFCCs, and then feed the DAF into a classifier of BLSTM for ASC and SED. The rest of this report is organized as follows. Section 2 describes the proposed method and Section 3 presents experiments. Finally, conclusions are drawn in Section 4.

2. THE METHOD

The proposed framework for ASC is depicted in Figure 1, which mainly consists of two modules: DAF extraction and BLSTM classification. For task 1(i.e. ASC), the audio recordings of each acoustic scene are fed into the system and the labels of acoustic scene are output by the system.

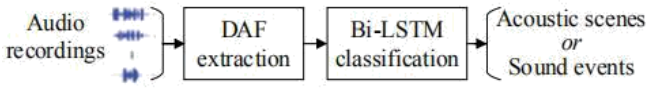

Figure 1: The proposed framework for ASC

2.1. DAF extraction

The proposed DAF is used for representing the properties of different acoustic scenes , whose extraction is illustrated in Figure 2. Each audio recording is split into frames for extracting MFCCs, and then a DNN feature extractor is built for extracting bottleneck feature (i.e. DAF) based on MFCCs. The DAF is output from the bottleneck layer of the DNN.

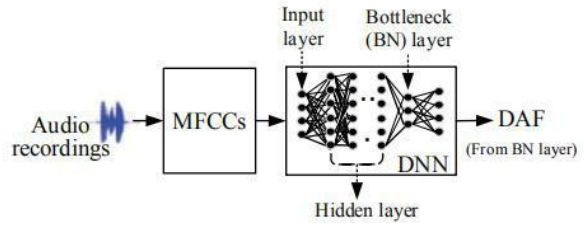

Figure 2: The DAF extraction

The MFCCs is the most popular feature for audio classification in the previous studies [7], and is used as a component for extracting the DAF here. The details of both the MFCCs extraction and the DNN building (including its training and parameters settings) are all discussed in our previous work [10].

2.2. BLSTM classification

A RNN has feedback connections and works efficiently and flexibly with time-series signals such as audio signal. Due to the exploding and vanishing gradient problem, a simple RNN is not easy to train, and not able to deal with long-range dependencies [13]. Hidden units of gated RNN are gate-based. Two common classes of Gated RNNs are LSTM and Gated Recurrent Units (GRUs), and the LSTM has been widely used. The introductions about LSTM and GRU are given in [14] and [15], respectively.

LSTM is very flexible in classifying sequential data in both cases of sequence-to-one classification and sequence-to-sequence classification. A BLSTM has a second hidden layer that learns input sequence in an inverse direction, which is expected to yield better prediction since information for prediction at each time- step is from both the backward and forward directions. Hence, we use the BLSTM as classifier for the ASC .

3. EXPERIMENTS

Our experiments are mainly performed on the TensorFlow [16]. We build a system for tasks 1b, respectively. The details about datasets, performance metrics and baseline systems are given in [8]. The predominant performance metrics for tasks 1b is classification accuracy. The configurations for the DAF extraction and the BLSTM building are listed in Table 1.

Table 1: The configurations for the DAF extraction and Bi-LSTM building.

DAF extraction	
MFCC	Dimension: 13, frame length/overlap: 40/20 ms.
DNN	DAF dimension: 50, learning rate: 0.001, maximum iterations: 3000, batch size: 256, context size: 7 frames, number of hidden layers: 5, weight decay: 0.1, dropout: 0.8, neurons of hidden layer: [200 100 50 100 200], output layer function: Sigmoid.
Bi-LSTM building	
Bi-LSTM	Cell number: 400, learning rate: 0.001, iterations: 300, batch size: 256, unrolled steps: 7, training algorithm: back-propagation through time, initial forget bias: 1.

3.1. Task 1b: acoustic scene classification

The goal of acoustic scene classification is to classify a test re- cording into one of the predefined classes that characterizes the environment in which it is recorded for example “Airport”, “Bus”, “Metro”. Table 2 shows average results obtained by our system and the baseline [8]. On Device B,C and (B,C) ,our system achieves an overall average classification accuracy of 54.4%,54.4%,53.9% which is higher than 45.1%,46.2%,45.6% obtained by the baseline respectively.

Table 2: Acoustic scene classification results on development

Acoustic scene （Device B）	Classification accuracy (%)	
	Baseline	Ours
Airport	68.9	38.9
Bus	70.6	55.6
Metro	23.9	61.1
Metro station	33.9	38.9
Park	67.2	94.4
Public square	22.8	38.9
Shopping mall	58.3	66.7
Street, pedestrian	16.7	38.9
Street, traffic	69.4	72.2
Tram	18.9	38.9
Overall	**45.1**	**54.4**

Acoustic scene （Device C）	Classification accuracy (%)	
	Baseline	Ours
Airport	76.1	27.8
Bus	86.1	94.4
Metro	17.2	44.4
Metro station	31.7	38.9
Park	51.1	61.1
Public square	26.7	38.9
Shopping mall	63.9	83.3
Street, pedestrian	25.0	44.4
Street, traffic	63.3	77.8
Tram	20.6	33.3
Overall	**46.2**	**54.4**

Acoustic scene （Average B,C）	Classification accuracy (%)	
	Baseline	Ours
Airport	72.5	38.9
Bus	78.3	75.0
Metro	20.6	44.4
Metro station	32.8	41.7
Park	59.2	80.6
Public square	24.7	41.7
Shopping mall	61.1	72.2
Street, pedestrian	20.8	38.9
Street, traffic	66.4	72.2
Tram	19.7	33.3
Overall	**45.6**	**53.9**

Acoustic scene （Device A）	Classification accuracy (%)	
	Baseline	Ours
Airport	73.4	53.6
Bus	56.7	78.1
Metro	46.6	39.1
Metro station	52.9	42.9
Park	80.8	80.6
Public square	37.9	38.4
Shopping mall	46.4	55.6
Street, pedestrian	55.5	56.3
Street, traffic	82.5	86.6
Tram	56.5	54.0
Overall	**58.9**	**58.3**

4. CONCLUSIONS

In this report, we have introduced our systems submitted to the challenge on DCASE 2018 and presented the systems performance on the development datasets of tasks 1b.In terms of the predominant performance metrics, the results have showed that our systems for tasks 1b outperform the corresponding baselines.

5. ACKNOWLEDGMENT

The work was supported by the NSFC (61101160), the Project of the Pearl River Young Talents of S&T, Guangzhou (2013J2200070), and Fundamental Research Funds for the Central Universities (2015ZZ102). We acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

1. REFERENCES

[1] Y. Li, Q. He, S. Kwong, T. Li, and J. Yang, “Characteris-tics-based effective applause detection for meeting speech,” Signal Processing, vol. 89, no. 8, pp. 1625-1633, 2009.

[2] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, “Audio surveillance of roads: a system for detecting anomalous sounds,” IEEE Transactions on Intelligent Transportation Systems, vol. 17, no. 1, pp. 279-288, Jan. 2016.

[3] M. Crocco, M. Cristani, A. Trucco, and V. Murino, “Audio surveillance: a systematic review,” ACM Computing Sur- veys, vol. 48, no. 4, pp. 1-46, 2016.

[4] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, “Clear evaluation of acoustic event detection and classification systems,” Lecture notes in computing sci- ence, vol. 4122, pp. 311-322, 2007.

[5] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M.D. Plumbley, “Detection and classification of acoustic scenes and events,” IEEE Transactions on Multimedia, vol. 17, no. 10, pp. 1733-1746, Oct. 2015.

[6] T. Virtanen, A. Mesaros, T. Heittola, M.D. Plumbley, P. Foster, E. Benetos, and M. Lagrange, “Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016),” 2016.

[7] J. Schröder, N. Moritz, J. Anemüller, S. Goetze, and B. Kollmeier, “Classifier architectures for acoustic scenes and events: implications for DNNs,TDNNs,and perceptual features from DCASE 2016”,IEEE/ACM Transactionson Audio Speech,and Language Processing,vol.25,no.6,pp.1304-1314,Jun.2017.

[8] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen. “DCASE 2017 chal- lenge setup: tasks, datasets and baseline system,” in Pro- ceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Nov. 2017. Submitted.

[9] H. Phan, M. Maaß, R. Mazur, A. Mertins, “Random regres- sion forests for acoustic event detection and classification,” IEEE Transactions on Audio, Speech, and Language Pro- cessing, vol. 23, no. 1, pp. 20-31, 2015.

[10] Y. Li, X. Zhang, H. Jin, X. Li Q. Wang, Q. He, and Q. Huang, “Using multi-stream hierarchical deep neural network to extract deep audio feature for acoustic event de- tection,” Multimedia Tools and Applications, doi: 10.1007/s11042-016-4332-z, pp. 1-20, Jan. 2017

[11] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, G. Widmer, “CP- JKU submissions for DCASE-2016: A hybrid approach us- ing binaural I-

vectors and deep convolutional neural net- works," in Proc. of Detection and Classification of Acoustic Scenes and Events 2016, Sep. 2016.

[12] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, T. Virtanen, "Sound event detection in multichannel audio us- ing spatial and harmonic features," in Proc. of Detection and Classification of Acoustic Scenes and Events 2016, Sep. 2016.

[13] R. Pacanu, T. Mikolov, and Y. Bengio, "On the difficulties of training recurrent neural networks," in Proceedings of the 30th International Conference on Machine Learning, no. 2, pp. 1310-1318, 2013.

[14] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recog-nition with deep recurrent neural networks," in International Conference on Acoustics, Speech and Signal Processing, pp. 6645-6649, 2013.

[15] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in Proceedings of the 2014 Confer- ence on Empirical Methods in Natural Language Pro- cessing, pp. 1724-1734, 2014.

[16] https://www.tensorflow.org/