THE SEIE-SCUT SYSTEMS FOR IEEE AASP CHALLENGE ON DCASE 2017: DEEP LEARNING TECHNIQUES FOR AUDIO REPRESENTATION AND CLASSIFICATION

Yanxiong Li, 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 three tasks of IEEE AASP challenge on DCASE 2017, i.e. task 1: Acoustic Scene Classification (ASC), task 2: detection of rare sound events in artificially created mixtures and task 3: sound event detection in real life recordings. Tasks 2 and 3 belong to the same problem, i.e. Sound Event Detection (SED). We adopt deep learning techniques to extract Deep Audio Feature (DAF) and classify various acoustic scenes or sound events. 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 Bi-directional Long Short Term Memory (Bi-LSTM) fed by the DAF is built for ASC and SED. Evaluated on the development datasets of DCASE 2017, our systems are superior to the corresponding baselines for tasks 1 and 2, and our system for task 3 performs as good as the baseline in terms of the predominant metrics.

Index Terms—DAF, Bi-LSTM, acoustic scene classification, sound event detection

1. INTRODUCTION

ASC is a process of determining a test audio recording belongs to which pre-given class of acoustic scenes, while SED is a process of labeling temporal regions within a test audio recording and resulting in a symbolic description such that each annotation gives the timestamps and sound event labels. Although ASC and SED are different in their specific steps, they can be regarded as the same task of audio representation and classification. Hence, they can be tackled by using the same feature and classifier. Both of them are 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 2016 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 49 submissions in the ASC task of DCASE 2016 challenge. Adavanne et al [12] used spatial and harmonic features in combination with LSTM 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 Bi-LSTM for audio classification. In our submissions for DCASE 2017, we propose to build a DNN for extracting the DAF based on MFCCs, and then feed the DAF into a classifier of Bi-LSTM 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 and SED is depicted in Figure 1, which mainly consists of two modules: DAF extraction and Bi-LSTM 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. For tasks 2 and 3 (i.e. SED in artificially-created and real-life recordings), the audio recordings containing the target sound events are fed into the system and the target sound events are detected by the system.

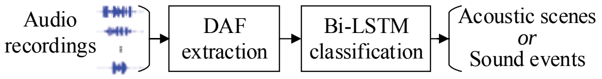

Figure 1: The proposed framework for ASC and SED.

2.1. DAF extraction

The proposed DAF is used for representing the properties of different acoustic scenes and sound events, 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.

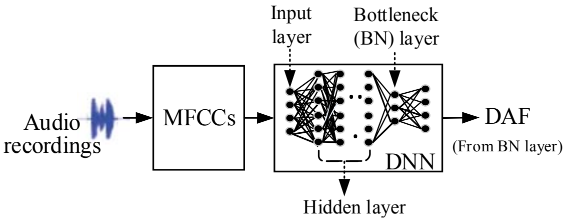

Figure 2: The DAF extraction.

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. Bi-LSTM 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 Bi-LSTM 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 Bi-LSTM as classifier for both ASC and SED.

3. EXPERIMENTS

Our experiments are mainly performed on the TensorFlow [16]. We build three systems for tasks 1 to 3, respectively. The details about datasets, performance metrics and baseline systems are given in [8]. The predominant performance metrics for tasks 1, 2 and 3 are classification accuracy, event-based and segment-based error rates, respectively. The configurations for the DAF extraction and the Bi-LSTM 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 1: acoustic scene classification

The goal of acoustic scene classification is to classify a test recording into one of the predefined classes that characterizes the environment in which it is recorded for example “park”, “home”, “office”. Table 2 shows average results over 4 folds obtained by our system and the baseline [8]. Our system achieves an overall average classification accuracy of 91.0% which is higher than 73.8% obtained by the baseline.

Table 2: Acoustic scene classification results on development dataset (average over 4 folds).

Acoustic scene	Classification accuracy (%)	
	Baseline	Ours
Beach	77.6	93.5
Bus	83.7	82.1
Cafe/Restaurant	55.1	91.5
Car	86.2	97.6
City center	88.5	94.9
Forest path	83.3	91.0
Grocery store	63.1	87.1
Home	74.5	97.4
Library	60.6	69.2
Metro station	88.5	97.4
Office	97.4	97.5
Park	64.4	90.0
Residential area	62.8	87.2
Train	38.1	89.7
Tram	82.7	98.7
Overall	**73.8**	**91.0**

3.2. Task 2: detection of rare sound events in artificially created mixtures

Task 2 focuses on the detection of rare sound events. The audio material used in this task consists of artificially created mixtures, allowing the creation of many examples at different event-to-background ratios. Here, “rare” refers to target sound events occurring at most once within a half-minute recording [8]. Table 3 shows average results obtained by our system and the baseline [8]. Our system obtains an overall average event-based error rate of 0.55 which is lower than 0.57 obtained by the baseline.

Table 3: Average results of detection of rare sound events in artificially created mixtures on development dataset.

Sound event	Event-based metrics			
	Baseline		Ours	
	Error rate	F-score	Error rate	F-score
Baby cry	0.79	68.1%	0.77	67.6%
Glass break	0.21	89.0%	0.35	82.8%
Gun shot	0.72	55.1%	0.54	67.2%
Average	**0.57**	**70.7%**	**0.55**	**72.5%**

3.3. Task 3: sound event detection in real life recordings

Task 3 evaluates the performance of sound event detection systems in multisource conditions similar to our everyday life, where the sound sources are rarely heard in isolation. Six predefined sound event classes are selected, and systems are expected to detect the presence of these sounds, providing labels and timestamps to segments of the test audio [8]. Table 4 shows average results obtained by our system and the baseline [8]. Our system achieves an overall average segment-based error rate of 0.69 which is equal to the counterpart obtained by the baseline.

Table 4: Results of sound event detection in real life audio on development dataset.

Sound event	Segment-based metrics			
	Baseline		Ours	
	Error rate	F-score	Error rate	F-score
Overall	**0.69**	**56.7%**	**0.69**	**54.5%**

4. CONCLUSIONS

In this report, we have introduced our systems submitted to the IEEE AASP challenge on DCASE 2017 and presented the systems performance on the development datasets of tasks 1, 2 and 3. In terms of the predominant performance metrics, the results have showed that our systems for tasks 1 and 2 outperform the corresponding baselines, and the performance of our system for task 3 is the same as that of the baseline.

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.

6. REFERENCES

[1] Y. Li, Q. He, S. Kwong, T. Li, and J. Yang, “Characteristics-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 Surveys*, 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 science*, 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 Transactions on 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 challenge setup: tasks, datasets and baseline system,” in *Proceedings 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 regression forests for acoustic event detection and classification,” *IEEE Transactions on Audio, Speech, and Language Processing*, 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 detection,” *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 using binaural I-vectors and deep convolutional neural networks,” 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 using 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 recognition 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 Conference on Empirical Methods in Natural Language Processing*, pp. 1724-1734, 2014.

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