THE SEIE-SCUT SYSTEMS FOR ACOUSTIC SCENE CLASSIFICATION USING CNN ENSEMBLE

Wucheng Wang, Mingle Liu, Yanxiong 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 concerning task 1b of DCASE 2019, i.e. acoustic scene classification (ASC) with mismatched recording devices. We propose a strategy of CNN ensemble for ASC. Specifically, an audio feature, such as Mel-frequency cepstral coefficients (MFCCs) and logarithmic filter-bank (LFB), is first extracted from audio recordings. Then a series of convolutional neural network (CNN) is built for obtaining CNN ensemble. Finally, classification result for each test sample is based on the voting of all CNNS contained in the CNN ensemble.

***Index Terms*—**convolutional neural network, acoustic scene classification, CNN ensemble

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 previous DCASE challenge for ASC, 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.

In our submissions for task 1b of DCASE 2019, we perform ASC using a strategy of CNN ensemble. 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 SYSTEM

The proposed framework for ASC is depicted in Fig. 1, which mainly consists of two modules: feature extraction and CNN 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.

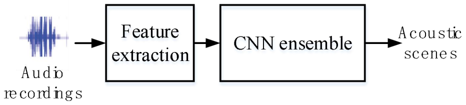

Fig. 1. The proposed system for ASC.

Many kinds of CNN structures, such as VGG, Inception, Resnet, have generated from image classification [11] and they obtained good results for image classification due to the CNN's strength on catching difference of various feature maps. These popular CNN structures were popularly used audio classification [12]. Only one type of CNN structure was adopted for audio classification instead of a combination of many CNN structures. Although each kind of CNN structure shows powerful ability of classification, they still have unique characteristics[13]. If they are fused in an effective way, we can obtain a stronger ASC system. Hence, we try to combine three CNNs to make classification decisions.

3. EXPERIMENTS

Our experiments are mainly performed on the TensorFlow and Keras. We extracted MFCC with 39 dimensions and LFB with 54 dimension from raw audio datasets. Frame length and overlapping are 40 ms and 20 ms, respectively. The configurations of VGG network, Inception network and Resnet network are presented in Fig. 2, 3 and 4, respectively.

The work was supported by the national natural science foundation of China (No. 61771200).

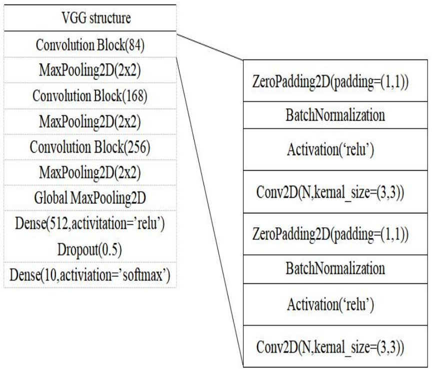

Fig. 2 VGG network.

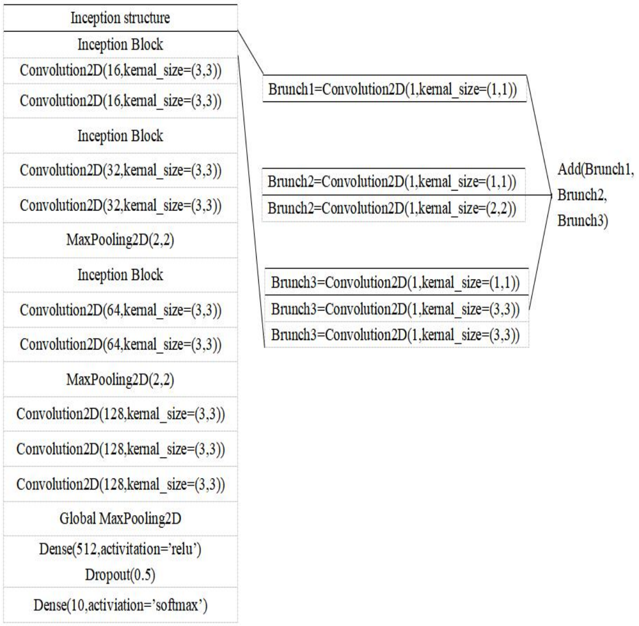
Inception structure
Inception Block
Convolution2D(16,kernal_size=(3,3))
Convolution2D(16,kernal_size=(3,3))
Inception Block
Convolution2D(32,kernal_size=(3,3))
Convolution2D(32,kernal_size=(3,3))
MaxPooling2D(2,2)
Inception Block
Convolution2D(64,kernal_size=(3,3))
Convolution2D(64,kernal_size=(3,3))
MaxPooling2D(2,2)
Convolution2D(128,kernal_size=(3,3))
Convolution2D(128,kernal_size=(3,3))
Convolution2D(128,kernal_size=(3,3))
Global MaxPooling2D
Dense(512,activitation='relu')
Dropout(0.5)
Dense(10,activiation='softmax')
Brunch1=Convolution2D(1,kernal_size=(1,1))
Brunch2=Convolution2D(1,kernal_size=(1,1))
Brunch2=Convolution2D(1,kernal_size=(2,2))
Brunch3=Convolution2D(1,kernal_size=(1,1))
Brunch3=Convolution2D(1,kernal_size=(3,3))
Brunch3=Convolution2D(1,kernal_size=(3,3))
Add(Brunch1,
Brunch2,
Brunch3)

Fig. 3 Inception network.

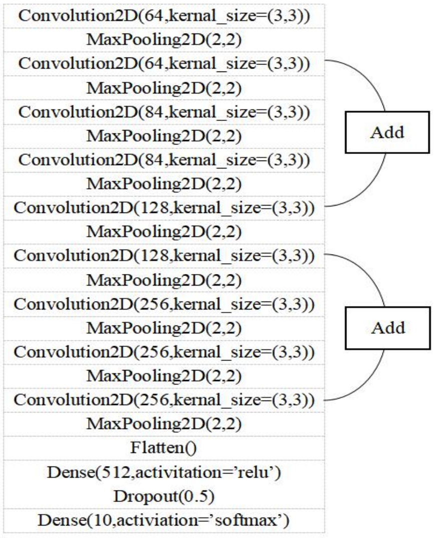
Convolution2D(64,kernal_size=(3,3))
MaxPooling2D(2,2)
Convolution2D(64,kernal_size=(3,3))
MaxPooling2D(2,2)
Convolution2D(84,kernal_size=(3,3))
MaxPooling2D(2,2)
Convolution2D(84,kernal_size=(3,3))
MaxPooling2D(2,2)
Convolution2D(128,kernal_size=(3,3))
MaxPooling2D(2,2)
Convolution2D(128,kernal_size=(3,3))
MaxPooling2D(2,2)
Convolution2D(256,kernal_size=(3,3))
MaxPooling2D(2,2)
Convolution2D(256,kernal_size=(3,3))
MaxPooling2D(2,2)
Convolution2D(256,kernal_size=(3,3))
MaxPooling2D(2,2)
Flatten()
Dense(512,activitation='relu')
Dropout(0.5)
Dense(10,activiation='softmax')
Add
Add

Fig. 4 Resnet network.

We divide the development datasets into five parts, and each training subset consists of 4 different parts. As a result, one type of CNN structure has five models trained by different training subsets. Two kinds of features (MFCC and LFB) and three CNNs (VGG, Inception and Resnet) are combined for obtaining different classification models, and 30 different models are finally built. At the first step, we get the prediction result from each model, and then make a vote based on the prediction results of each model. At the second step, we consider a result as a uncertain prediction which has the most votes and its votes are very close to the votes of the second place. On the contrary, if the votes of the first place are far more than that of the second place, the prediction is considered as a certain result. The test samples belonging to the certain part are used to retrain models. At the third step, we vote again with the retrained models, and renew the uncertain parts. The prediction process described above is illustrated in Fig. 5.

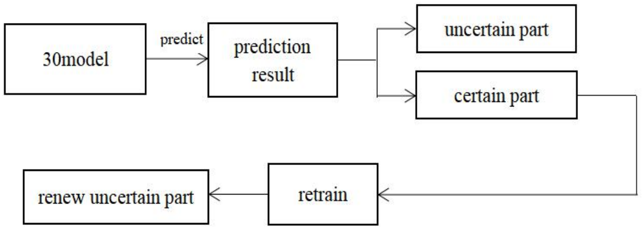

Fig. 5 Process of prediction.

Tables 1 and 2 show experimental results and classification accuracy of each scene, respectively. As shown in Table 1, our system outperforms baseline system.

Table 1 Classification results.

Models	Accuracy (%) (Development)	Accuracy (%) (Leaderboard)
Baseline	62.5	64.3
VGG	91.1	
Inception	89.5	
Resnet	88.7	
Vote	93.0	73.3
Retain and vote again		77.2

Table 2 Classification accuracy of each scene.

Scene label	Baseline(%)	Propose(%)
Airport	48.4	82.3
Bus	62.3	65.5
Metro	65.1	74.2
Metro station	54.5	62.5
Park	83.1	55.7
Public square	40.7	57.8
Shopping mall	59.4	66.2
Street, pedestrian	60.9	75.6
Street, traffic	86.7	82.3
Tram	64.0	85.3

4. CONCLUSIONS

we proposed a vote mechanism to divide vote result into certain part and uncertain part, and utilize certain prediction to retrain and finetune original model to adapt to test datasets. As a result, we obtain better accuracy on Leaderboard. Our method needs to train many networks and thus are time-consuming compared to the system using one network only.

5. REFERENCES

[1] Y. Li, Q. He, S. Kwong, T. Li, and J. Yang, “Characteris-ticsbased 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] 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.

[12] 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.

[13] 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.