Using different combination strategies to combine features for audio scene classification based on Convolutional Neural Networks

Chan Lin¹, Yan Leng¹,*, Rongyan Wang²,a and Zhaoli Lu¹

¹School of Physics and Electronics, Shandong Normal University, Ji’nan, 250014, China
²School of Information Management, Dezhou University, Dezhou, 253023, China

*Corresponding author e-mail: lengyan@sdnu.edu.cn, awangrongyan8232@163.com

Abstract. Audio scene classification (ASC) which can make good use of audio information for classification is currently a hot topic. In this paper, taking Convolutional Neural Networks (CNN) as the classification model, we have designed three combination strategies to combine different kinds of features for classification. In the first combination strategy, the multi-channel CNN is used to combine different features; in the second combination strategy, different kinds of features are concatenated, and the concatenated features are used as the input of the CNN; in the third combination strategy, new features are first extracted from each kind of feature through CNN, and then these new features are concatenated for classification. Experiments are done on the DCASE2017 datasets to test the effectiveness of the combination strategies, and the experimental results show that the second combination strategy performs best, also the experimental results show that on the whole, combining more features would get better classification results.

1. Introduction

Audio scene classification (ASC) is a technology which refers to determining the scene in which an audio recording has been produced. Currently, how to make good use of audio information is a hot topic [1-3], and ASC is a good way to make use of the audio information, it has great potential applications in industry, such as security surveillance [4], smart robot [5], and smart mobile device [6] etc.

Detection and Classification of Acoustic Scenes and Events (DCASE) challenge is a challenge which focuses on the tasks of acoustic scene classification, and sound event detection etc. DCASE has attracted more and more researchers to focus on ASC. Among the submissions of DCASE, Convolutional Neural Networks (CNN) is the most popular neural network for ASC, and it has shown good performance. For example, Soo et al. [7] combined CNN with Long Short-Term Memory (LSTM) to extract the sequential and spectro-temporal locality information. Hamid et al. [8] used a hybrid approach which uses binaural i-vectors and deep convolutional neural networks for classification, and their classification result ranked first in the DCASE2016 challenge. Currently, in the field of ASC, many researches choose to combine different kinds of audio features for classification, for example, Zheng weiping et al. [9] used deep convolutional neural network and multiple spectrograms fusion to perform ASC, Waldekar et al. [10] combined audio features in a
fusion-based framework, and Xu jinwei et al. [11] used convolutional neural network and two audio features to construct the fusion model for ASC.

Inspired by the above feature fusion systems, in this paper, we also combine different kinds of audio features for classification. Different from the above fusion systems, we propose three different combination strategies, and test which combination strategy is better for classification, also we do experiments to test if it is better to combine more features for classification.

2. The Proposed Method

In this paper, we propose three combination strategies, the first one is to use the multi-channel CNN to combine different features, the second one is to concatenate different features, and use the concatenated features as the input of the CNN, and the last one is to extract new features from each kind of feature through CNN, and then concatenate these new features for classification. We choose three different kinds of features which are very often used in the field of ASC for combination, that is, the Mel spectrogram, Short-Time Fourier Transform (STFT) feature and Constant Q-Transform (CQT) feature. In order to test if it is better to combine more features for classification, we respectively use the three proposed strategies to combine two or three of the above features for classification. We use CNN as the classification model. All the CNN models used in this paper has the same architecture which is described in section 2.1.

2.1. CNN Architecture

We choose the CNN model used in [12] as the classification model, its architecture is described in Table 1. The CNN model consists of three convolution layers., its input is a single-channel or multi-channel map with size BANDS×F, where BANDS indicates the feature dimension and F indicates the number of frames. For the first convolution layer, the filter size is set to be BANDS×F1, F1 is the number of frames, F1<F, the stride is set to be 1, and the number of filter units is set to be 100, the convolution layer is followed by batch normalization, Leaky ReLU is taken as the activation function (α=0.3), and the dropout parameter is set to be 0.25. For the second convolution layer, the filter size is set to be 1×1, and the other settings are the same as that of the first convolution layer. The final convolution layer consists of L filters with softmax as the activation function, where L indicates the number of acoustic scene classes. The final output is obtained by global average pooling.

Table 1. CNN Architecture

| Input | BANDS×F1 Conv-100-BN-LReLU Drop-Out (0.25) |
|-------|------------------------------------------|
|       | 1×1 Conv-100-BN-LReLU Drop-Out (0.25)     |
|       | 1×1 Conv-L-Sofmax Global-Average-Pooling  |

The network weights are initialized with the He uniform. The categorical cross-entropy is used as the loss function, and the Adam optimizer is used to train the network. The learning rate is set to be 0.0001. The training process is early stopped if the verification accuracy is not improved within 50 epochs, up to 500 epochs. The network is implemented using Keras (v2.1.3), with TensorFlow as the backend.

2.2. The first combination strategy

In the first strategy, each audio recording is first split into frames, and for each frame, the Mel spectrogram, STFT feature and CQT feature is extracted respectively, and then two or three of them are stacked together to form the multi-channel input of CNN. In Figure 1, we take the three-channel input for illustration.
2.3. The second combination strategy

In the second combination strategy, each audio recording is first split into frames, and then for each frame, the Mel spectrogram, STFT feature and CQT feature is extracted respectively, and then two or three of them are concatenated together to form a new longer representation for each frame, each audio recording as that described in formula (1) is taken as the input of CNN. In formula (1), we show the concatenation of three kinds of features for illustration, M, S, and C indicates Mel spectrogram, STFT feature, and CQT feature respectively.

\[ \begin{array}{c}
\text{M} \\
\text{S} \\
\text{C}
\end{array} \]

\text{audio frame}

\[ \begin{array}{c}
\text{one audio signal}
\end{array} \]

\[ (1) \]

2.4. The third combination strategy

In the third combination strategy, each audio recording is split into frames, and for each frame, the Mel spectrogram, STFT feature and CQT feature is extracted respectively, two or three of them are then used for combination. For the features which are used for combination, each kind of feature is taken as the input of a one-channel CNN to generate new feature, the new features are then concatenated together, this new concatenated feature is then taken as the input of the CNN model for classification. In Figure 2, we show the combination of all three kinds of features to illustrate the third combination strategy.

\[ \begin{array}{c}
\text{Generation of multiple spectrogram} \\
\text{Deep features concatenation}
\end{array} \]

\text{Figure 2. the block diagram of the third combination strategy}
3. Experiments and Results

3.1. Dataset and Data Processing

The dataset from the task 1 of the DCASE 2017 challenge is used to evaluate the proposed combination strategies. The dataset consists of two parts: the Development set and the Evaluation set. Each set contains 15 audio scene classes, and each audio recording is a stereo with a sampling rate of 44.1 kHz and a bit rate of 24 bits. In this paper, all audio recordings are converted into 16kHz and 16 bits for experiments. For the development set, each audio scene has 312 recordings of 10 seconds long, totally 4680 recordings. In the evaluation set, there are totally 1620 audio recordings. A 4-fold cross validation setting is provided with the dataset.

Librosa toolkit [13] is adopted to extract features. The band size is set to be 150 for Mel spectrograms, STFT and CQT. For Mel spectrograms, the window length and hop length are set to be 512 samples and 256 samples respectively; for STFT, the window length and hop length are set to be 298 samples and 256 samples respectively; for CQT, the hop length is set to be 256 samples.

3.2. Results

In the experimental stage, to test if it is better to combine more features for classification, we do three groups of experiments: 1) first, we use Mel spectrogram, STFT feature and CQT feature respectively as the one-channel input of CNN for classification; 2) then we respectively use the proposed three combination strategies to combine two kinds of features for classification; 3) finally, we respectively use the proposed three combination strategies to combine all three kinds of features for classification. For DCASE2017 Development set, the provided 4-fold cross validation setting is adopted, and the result of each fold as well as the average result of the 4 folds are shown in Table 2.; for DCASE2017 Evaluation set, all samples in development set are used as training samples, and the samples in the evaluation set are used as test samples, the classification result is shown in Table 2. Besides, experiments are done to compare our proposed classification system with some other systems proposed in references, that is, Multiple spectrograms fusion [9], Audio features fusion [10], Two features fusion [11] and DCASE2017 Baseline System[14], the comparison results are shown in Table 3. In Table 3, when using the development set for classification, we only show the average classification result of the 4 folds, and for our proposed strategies, we only use the best classification results shown in Table 2 for comparison. In Table 2, 'Mel', 'STFT', 'CQT' means to use Mel spectrogram, STFT feature and CQT feature respectively as the one-channel input of CNN for classification; 'twoC ' means in the first combination strategy, using the two-channel CNN for classification, and 'm+s', 'm+c', 's+c' means the combination of Mel spectrogram and STFT, the combination of Mel spectrogram and CQT, and the combination of STFT and CQT respectively; 'S' means using the second combination strategy, 'T' means using the third combination strategy; 'three-Channel' means in the first combination strategy, using the three-channel CNN for classification; 'm+s+c' means the combination of Mel spectrogram, STFT feature and CQT feature.

Table 2. The classification accuracy (%) of the three groups of classification experiments.

| System   | DCASE2017 Development Dataset | DCASE2017 Evaluation Dataset |
|----------|-------------------------------|-------------------------------|
|          | Fold1 | Fold2 | Fold3 | Fold4 | Ave  | Fold1 | Fold2 | Fold3 | Fold4 | Ave  |
| Mel      | 72.65 | 68.37 | 71.53 | 81.11 | 73.42 | 64.63 |
| STFT     | 72.48 | 74.00 | 71.61 | 78.63 | 74.18 | 59.01 |
| CQT      | 72.22 | 74.00 | 67.52 | 76.41 | 72.54 | 52.72 |
| twoC_m+s | 72.56 | 76.15 | 74.62 | 83.76 | 76.77 | 62.16 |
| twoC_m+c | 78.97 | 78.12 | 84.27 | 85.65 | 81.75 | 69.20 |
| twoC_s+c | 78.80 | 82.31 | 84.7  | 86.84 | 83.16 | 68.95 |
| S_m+s    | 73.58 | 75.80 | 77.08 | 80.33 | 76.70 | 61.57 |
| S_m+c    | 75.85 | 77.78 | 80.75 | 85.46 | 79.96 | 65.03 |
| S_s+c    | 79.83 | 80.51 | 84.87 | 86.67 | 82.97 | 68.70 |
| T_m+s    | 73.08 | 74.1  | 75.64 | 83.68 | 76.63 | 63.58 |
| T_m+c    | 77.18 | 75.47 | 82.99 | 85.98 | 80.41 | 65.31 |
| T_s+c    | 79.74 | 76.24 | 79.74 | 87.09 | 80.70 | 66.91 |
| three_Channel | 80.17 | 81.71 | 85.73 | 87.35 | 83.74 | 70.49 |
| S_m+s+c  | 79.83 | 81.71 | 85.04 | 88.97 | 83.89 | 70.74 |
| T_m+s+c  | 77.78 | 79.66 | 80.00 | 88.12 | 81.39 | 65.93 |
From Table 2, under each combination strategy, by comparing the classification result which is obtained by combining all three kinds of features with the classification results which are obtained by combining two kinds of features, it can be seen that on the whole, combing three kinds of features would get better results; by comparing the classification results which are obtained by combining two or three kinds of features with the classification results which are obtained by using Mel spectrogram or STFT feature or CQT feature alone, it can be seen that combining different kinds of features for classification would be better than using only one kind of features for classification, this is because that different kinds of features would provide information from different perspectives, such information would be complementary to each other, and then would be more favourable for classification. By comparing the proposed three different kinds of combination strategies, it can be seen that the second combination strategy which combines all three kinds of features has obtained the best results, this shows that concatenating the three kinds of features directly for classification would be better than concatenating their new features obtained through CNN for classification, and would be better than stacking them together to form the three-channel information for classification.

Table 3. The classification accuracy (%) of different systems.

| System                        | DCASE2017 Development Dataset | DCASE2017 Evaluation Dataset |
|-------------------------------|------------------------------|-----------------------------|
| Multiple spectrograms fusion  | 89.86                        | 77.70                       |
| Audio features fusion [10]    | 86.30                        | 67.00                       |
| Two features fusion [11]      | 85.30                        | 68.50                       |
| DCASE2017 Baseline System     | 74.80                        | 61.00                       |
| Our proposed system           | 83.89                        | 70.74                       |

From Table 3 it can be seen that the system proposed in [9] has achieved the best performance. Comparing our proposed system with that proposed in [10] and [11], though our proposed system performs worst on the development dataset, it performs best on the evaluation dataset, which means that our proposed system has better generalization ability. The reason why our proposed system performs worse on the development dataset may be that we has adopted a CNN model which has much simpler architecture for classification, in future work, we would try to increase the complexity of the network structure to see if it would get better classification results on the development dataset. Though our proposed system performs not better than the systems proposed in the references on the development dataset, it performs much better than the DCASE2017 baseline system, its classification accuracy is about 9% higher and 9.7% higher than the baseline system on the development dataset and the evaluation dataset respectively.

4. Conclusion
In this paper, we propose three combination strategies to combine different kinds of features for classification, and we do experiments to test which combination strategy would get better results, to test if combining more features would get better results. Also we do experiments to compare our proposed classification system with some other classification systems proposed in the references. The conclusion is that the second combination strategy, that is, concatenating different kinds of features directly for classification, has achieved the best results; on the whole, combining more features would get better results. Our proposed classification system performs much better than the DCASE2017 baseline system; although its performance is worse than two other systems on the development dataset, it performs much better on the evaluation dataset, which means that the proposed system has much better generalization ability.

5. Acknowledgments
This work was supported in part by the Project of National Natural Science Foundation of China under Grant 61401259 and Grant 61861033, China's Jiangxi Province Natural Science Foundation
under Grant 20181BAB202022, Research Fund for Young and Middle-aged Scientists of Shandong Province under Grant ZR2016FB25.

References
[1] Y. Leng, N. Zhou, C. L. Sun, X. Y. Xu, Q. Yuan, C. F. Cheng, Y. X. Liu, D. W. Li, Audio scene recognition based on audio events and topic model, Knowledge-Based Systems, 125 (2017) 1-12.
[2] Y. Leng, C. L. Sun, X. Y. Xu, Q. Yuan, S. N. Xing, H. L. Wan, J. J. Wang, D. W. Li, Employing unlabeled data to improve the classification performance of SVM, and its application in audio event classification, Knowledge-Based Systems, 98 (2016) 117-129.
[3] Y. Leng, C. L. Sun, C. F. Cheng, X. Y. Xu, S. Li, H. L. Wan, J. Fang, D. W. Li, Classification of overlapped audio events based on AT, PLSA, and the combination of them, Radioengineering, 24 (2) (2015) 593-603.
[4] R. Radhakrishnan, A. Divakaran, P. Smaragdis, Audio analysis for surveillance applications, in IEEE Workshop on Signal Processing to Audio and Acoustics (WASPAA’05), New Paltz, NY, USA, 2005.
[5] Johannes A. Stork, Luciano Spinello, Jens Silva, Kai O. Arras, Audio-Based Human Activity Recognition Using Non-Markovian Ensemble Voting, in Proc. IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN’12), Paris, France, 2012.
[6] J. Mäntyjärvi, P. Huuskonen, J. Himberg, Collaborative context determination to support mobile terminal applications, IEEE Wireless Commun., 9 (5) (2002) 39–45.
[7] S. H. Bae, I. Choi, N. S. Kim, Acoustic scene classification using parallel combination of LSTM and CNN, in Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), Budapest, Hungary, 2016.
[8] 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 the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), Budapest, Hungary, 2016.
[9] Z. Weiping, Y. Jiantao, X. Xiaotao, L. Xiangtao, P. Shaohu, Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion, in Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Munich, Germany, 2017.
[10] S. Waldekar, G. Saha, Audio features in a fusion-based framework for acoustic scene classification, in Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Munich, Germany, 2017.
[11] J. W. Xu, Y. Zhao, J. F. Jiang, Y. Dou, Z. Q. Liu, K. Chen, Fusion model based on convolutional neural networks with two features for acoustic scene classification, in Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Munich, Germany, 2017.
[12] K. J. Piczak, The details that matter: frequency resolution of spectrograms in acoustic scene classification, in Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Munich, Germany, 2017.
[13] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “Librosa: audio and music signal analysis in python,” in Proc. 14th Python in Science Conf., Austin, Texas, 2015.
[14] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, T. Virtanen, DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System, in Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Munich, Germany, 2017.