ACOUSTIC SCENE CLASSIFICATION USING DEEP CONVOLUTIONAL NEURAL NETWORK AND MULTIPLE SPECTROGRAMS FUSION

Zheng Weiping[1], Yi Jiantao[1], Xing Xiaotao[1] ,
Liu Xiangtao[2], Peng Shaohu[3]

Abstract

Making sense of the environment by sounds is an important research in machine learning community. In this work, a Deep Convolutional Neural Network (DCNN) model is presented to classify acoustic scenes along with a multiple spectrograms fusion method. Firstly, the generations of raw spectrogram and CQT spectrogram are introduced separately. Corresponding features can then be extracted by feeding these spectrogram data into the proposed DCNN model. To fuse these multiple spectrogram features, two fusing mechanisms, namely the voting and the SVM methods, are designed. By fusing DCNN features of the raw and CQT spectrograms, the accuracy is significantly improved in our experiments, comparing with the single spectrogram schemes. This proves the effectiveness of the proposed multi-spectrograms fusion method.

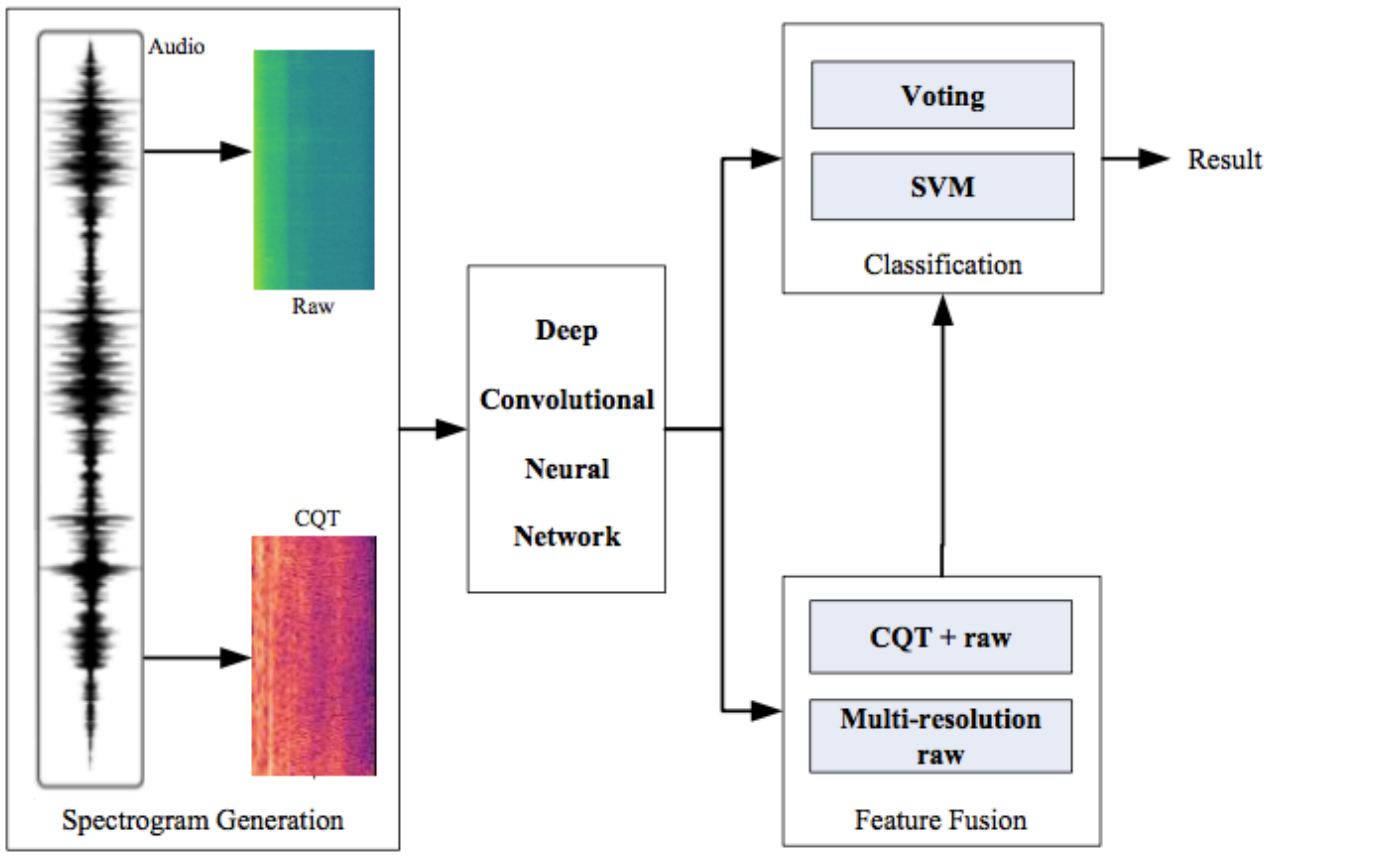
Introduction

In our DCASE2017 ASC submission, we also use a deep convolutional neural network (DCNN) based method to classify the acoustic scenes. Specifically, we produce multiple spectro- grams from audio files which are used to train a DCNN model. We have explored two different productions of spectrogram: raw spectrogram and Constant-Q-Transform (CQT) spectrogram[6]. According to the sliding window width and shift step length, multiple raw spectrograms with different resolutions are generat- ed. The classification performances of the DCNN model with multi-resolution raw spectrogram and CQT spectrograms are compared respectively. Next, we use the DCNN model to extract features, instead of classifying directly. A feature fusion method is applied in our submission. We have tried the fusion of features extracted from raw spectrograms with different resolutions, as well as the fusion of CQT spectrograms combined with raw spec- trograms. Among our experiments, the CQT plus raw spectro- grams fusion has achieved the best performance.

Table 1: Multi-Resolution parameters

Resolution Name	Sliding Window width	STFT			Bins × Freq	Num of samples
		NFFT	Pad	Over-lap		
R_{529}	24	529	1024	176	1249×512	12×2
R_{706}	32	706		276	1025×512	8×2
R_{882}	40	882		176	625×512	6×2

Methods and Materials



Results

As we can see, the four fusion solutions have achieved satisfactory results. All of the four accuracies are greater than 0.89. Actually, the highest one is 0.8986 and the lowest one is 0.8939. It is easy to find that the differences of accuracies among these four are very slight. However, compared to the best results of raw spectrogram and CQT spectrogram solutions (0.8536 and 0.8052 respectively), the improvements in accuracies of these fusion solutions are still significant, which proves the effectiveness of our multiple spectrograms fusion. Similarly, Table 5 shows the accuracy superiority of SVM meth- od over the voting in the fusion scenarios. To better understand the fusion performance, the class-wise accuracies of the best result, namely R706 +CQT84(SVM), are further given in Table 6.

Table 5: Accuracies of multiple spectrograms fusion solutions

	Folder 1	Folder 2	Folder 3	Folder 4	Average
R_{529} + CQT_{84} (Voting)	0.8769	0.9088	0.8406	0.8889	0.8788
R_{529} + CQT_{84} (SVM)	0.8684	0.919	0.8764	0.9162	0.895
R_{529} + CQT_{80} (Voting)	0.8752	0.902	0.8465	0.8949	0.8796
R_{529} + CQT_{80} (SVM)	0.8641	0.9173	0.8764	0.9265	0.896
R_{706} + CQT_{84} (Voting)	0.8547	0.8917	0.861	0.8983	0.8764
R_{706} + CQT_{84} (SVM)	0.865	0.9037	0.896	0.9299	**0.8986**
R_{706} + CQT_{80} (Voting)	0.8504	0.8832	0.8576	0.9043	0.8739
R_{706} + CQT_{80} (SVM)	0.8556	0.902	0.89	0.9282	0.8939

SUBMISSION RESULTS

All the development data are utilized to train the model, and the submitted results are tested on this final model. According to the fusion methods, two systems are included in our submission to the DCASE2017 challenge (task 1). The first one is DCNN based voting system, which fuses the raw (R706) and CQT84 spectro- grams by voting method (namely the R706+CQT84 (Voting) solu- tion). The second one is DCNN based SVM system, which fuses the same data by SVM method (namely the R706+CQT84 (SVM) solution).

Conclusions

The main contributions of this work lie in two aspects as follows. First, a deep CNN model is presented, which is originated from [5] and is improved to be more suitable for our problem. Second, a multi-spectrogram fusion method is proposed. Multiple spectrograms are fed into the same DCNN model and the corresponding features are fused to improve the accuracy of classification. In this work, the raw spectrogram and the CQT spectrogram are studied. The best accuracy of using the raw spectrograms is 0.8536; and the one of using CQT spectrograms is 0.8052. Although the accuracy of using CQT spectrograms is unsatisfactory, it can significantly improve the accuracy when fused with the raw spectrogram. The best result of the fusion scheme is 0.8986 and outperforms the best results of the single spectrogram schemes by more than 0.045.

Contact

Xing Xiaotao
South China Normal University
1299670261@qq.com

References

[1] Bae S H, Choi I, Kim N S. Acoustic Scene Classification using Parallel Combination of LSTM and CNN[J]. IEEE AASP Challenge on Detection and Classification of Acous- tic Scenes and Events (DCASE), 2016

[2] Valenti M, Diment A, Parascandolo G, et al. DCASE 2016 Acoustic Scene Classification using Convolutional Neural Networks[C]//Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2016), Budapest, Hungary. 2016.

[3] Dai Wei, Juncheng Li, Phuong Pham, et al. Acoustic Scene Recognition with Deep Neural Networks (DCASE challenge 2016)[R]. Robert Bosch Research and Technology Center, 3 September 2016.

[4] Rohit Patiyal, Padmanabhan Rajan. Acoustic Scene Classi- fication using Deep Learning[J]. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), 2016

[5] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, CP-JKU Submissions for DCASE-2016: A Hybrid Ap- proach using Binaural I-vectors and Deep Convolutional Neural Networks[C]//Workshop on Detection and Classifi- cation of Acoustic Scenes and Events (DCASE2016), Bu- dapest, Hungary. 2016.

[6] T. Lidy and A. Schindler. CQT-based Convolutional Neural Networks for Audio Scene Classification[C]//Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2016), September 2016, pp. 60-64.

[7] A Gorin, N Makhazhanov, N Shmyrev. DCASE 2016 Sound Event Detection System Based on Convolutional Neural Network[C]//Workshop on Detection and Classifica- tion of Acoustic Scenes and Events (DCASE2016), Buda- pest, Hungary. 2016.