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]

[1]School of Computer, South China Normal University
Guangzhou, China
250145025@qq.com, Yijiantao@hotmail.com, 1299670261@qq.com

[2]Shenzhen Chinasfan Information Technology Co., Ltd.
Shenzhen, China
liuxt@12366.net

[3]School of Mechanical and Electrical Engineering, Guangzhou University
Guangzhou, China
pengsh@gzhu.edu.cn

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 standard 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 standard 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.

Index Terms— Deep convolutional neural network, spectrogram, feature fusion, acoustic scene classification

1. INTRODUCTION

Environmental sound is a combination of sounds from many sources. It carries a lot of information that can help human to sense the surrounding environment. Acoustic scene classification (ASC) has been attracting the attention of researchers in machine learning communities and has been applied into surveillance, robotic navigation and context-aware services, etc.

Deep learning based solutions have been receiving great attentions from ASC researches. CNN[1][2], RNN[3], LSTM[1], DNN[4] and their combinations[1][3] have been applied to propose solutions. CNN has once again proved its powerful potential. In the DCASE2016 ASC challenge, a deep CNN solution[5] was proposed and won the rank first in the challenge task.

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 spectrograms from audio files which are used to train a DCNN model. We have explored two different productions of spectrogram: standard spectrogram and Constant-Q-Transform (CQT) spectrogram[6]. According to the sliding window width and shift step length, multiple standard spectrograms with different resolutions are generated. The classification performances of the DCNN model with multi-resolution standard 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 standard spectrograms with different resolutions, as well as the fusion of CQT spectrograms combined with standard spectrograms. Among our experiments, the CQT plus standard spectrograms fusion has achieved the best performance.

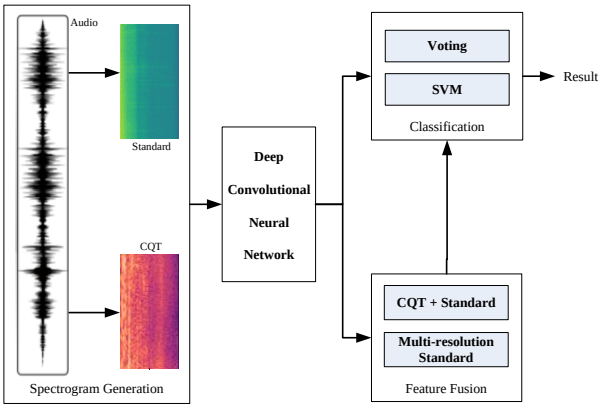

Figure 1: Flowchart of our method.

* Corresponding author

This work is partially supported by the Guangdong Provincial Scientific and Technological Projects (Grant Nos. 2016B010109005) and Characteristic Innovation Projects of the Educational Commission of Guangdong Province (Grant Nos. 2016KTSCX025)

The remainder of this paper is organized as follows. Section 2 introduce the generations of spectrograms. In Section 3, the detail of the DCNN model is given and the fusing algorithms used are described. Next, the experiment results are represented in Section 4. The submitted results are briefly explained in Section 5. Finally, we conclude the report in Section 6.

2. DATA PREPARATION

2.1. Standard Spectrogram

Instead of producing spectrograms from MFCC features[7], we generate spectrogram directly by performing Short Time Fourier Transform (STFT) on raw audio frames. Consequently, this spectrogram is referred to as "standard spectrogram". According to different sliding window widths and shift lengths, multi-resolution standard spectrograms can be generated; related setting parameters are shown in Table 1.

Table 1: Multi-Resolution parameters

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

Once the spectrogram has been generated, we split it into several smaller patches with fixed width and shift length. Finally, we resize every patches into 143×143. Then these patches are used as the training/test samples for the DCNN model.

2.2. CQT Spectrogram

The CQT spectrogram is generated on the CQT features which are computed from the raw audio frames by using the python library Librosa 0.5.0. When invoking the cqt function in the library, the sampling rate is set as 44100 and the other parameters are set as default, namely the number of bins per octave is 12 and the hop length is 512, etc. For each audio file, we generate two CQT spectrograms (size 832×143), each for a channel. Once again, we split the spectrogram into patches and feed them into a DCNN model as training/test samples. The patch width is 143 pixels and the shift step is 80 pixels. For each CQT spectrogram, 10 patches can be generated. As a result, we can generate 20 segments from a single audio file.

As mentioned in [6], as for CQT, its frequency resolution is better for low and mid-to-low frequencies. Hence, we generate two versions of CQT spectrograms: one uses all the 84 bands; the other uses 80 bands (the 4 bands related to high frequencies are discarded). For convenience of distinguish, they are mentioned as CQT_{84} and CQT_{80} respectively in the rest of this paper.

3. METHODOLOGY

3.1. Deep Convolutional Neural Network

Inspired by [5], we have adopted a DCNN model similar to the one proposed in [5]. The model follows a VGG style network for object recognition. As shown in Table 2, the input size of our model is 143×143. We have removed the global average pooling layer from the model, compared to the DCNN model in [5]. The removal of the global average pooling improves the performance in our experiments. The outputs of the next-to-last layer in Table 2 have fifteen feature maps of size 7×7. We flatten them into a 735-dimensions vector and feed it to the SoftMax layer.

The optimization setting is as follows. The batch size is 96. The initial learning rate is set as 0.1 and is decreased with 0.998 times every 10 epoches. We also use the L2-Regularization with a weight decay of 0.0001.

Table 2: DCNN model

Input 1 ×143 ×143
5 ×5 Conv(pad-2, stride-2)-32-BN-ReLu
3 ×3 Conv(pad-1, stride-1)-32-BN-ReLu
2 ×2 MaxPooling + DropOut(0.3)
3 ×3 Conv(pad-1, stride-1)-64-BN-ReLu
3 ×3 Conv(pad-1, stride-1)-64-BN-ReLu
2 ×2 MaxPooling + DropOut(0.3)
3 ×3 Conv(pad-1, stride-1)-128-BN-ReLu
3 ×3 Conv(pad-1, stride-1)-128-BN-ReLu
3 ×3 Conv(pad-1, stride-1)-128-BN-ReLu
3 ×3 Conv(pad-1, stride-1)-128-BN-ReLu
2 ×2 MaxPooling + DropOut(0.3)
3 ×3 Conv(pad-0, stride-1)-512-BN-ReLu
DropOut(0.5)
1 ×1 Conv(pad-0, stride-1)-512-BN-ReLu
DropOut(0.5)
1 ×1 Conv(pad-0, stride-1)-15-BN-ReLu
Flatten
15-way SoftMax

3.2. Fusing Methods

The DCNN can be used to classify acoustic scenes directly on an image sample. However, multiple samples have been generated from an audio file. To make good use of these samples, we further consider the fusing algorithms here.

3.2.1. Voting

Voting is a straightforward method in this situation. Each sample produces one vote and the class which wins the most votes is considered as the final result. For example, when standard spectrum is used and the revolution is R_{529} (as shown in Table 1), there are 24 samples for an audio file. They vote to decide the "correct" class.

By using voting, feature fusion can be easily implemented as well. If it is decided to use standard spectrograms (R_{529}, for

instance) and CQT spectrograms together for classification of the scenes, 44 votes are responsible for the result.

3.2.2. SVM

Instead of using the result given by DCNN directly, we can also use the next-to-last layer in the DCNN model to extract features for each sample. By concatenating all features of the samples from the same audio file sequentially, we can obtained a very long feature. Considering the risk of overfitting, a PCA dimensionality reduction operation is applied to the long features. As a result, a new feature has been generated which is encoded by all the samples. This new feature can be referred to as the aggregated feature.

By using the method described above, one aggregated feature can be generated for an audio file, according to a specific preparation of spectrograms. In other words, we can produce one CQT aggregated feature for an audio file, as well as another R_{529} aggregated feature, and so on (R_{706} etc.). When feature fusion is required, these aggregated features can be concatenated again into another feature. Note that PCA is not performed this time.

Finally, a SVM model is used to tell the final result by using these features as training/test samples. The linear kernel is applied in our experiments.

In our fusion experiments, SVM generally works better than voting mechanism. The reason for this is that the concatenating inputs of SVM provide sequential information which makes it possible for SVM to extract more comprehensive features for understanding the auditory scenes.

4. EXPERIMENTS AND RESULTS

In this section, we will demonstrate the experiments using the data and methods mentioned above. The experiments use the TUT Acoustic Scenes 2017 dataset (the part of acoustic scene classification). The results are conducted on the 4-fold cross validation set exactly the same to the baseline system in [8].

4.1. Classifying with Standard Spectrograms

Firstly, we try to explore the classification performances of standard spectrogram with different resolutions. The setting parameters about the resolutions involved here can refer to Table 1. For each resolution (R_{529}, for example), we will provide 3 types of accuracies: R_{529}(DCNN) is the accuracy computed by the DCNN model (see Table 2 in Section 3.1) on the patch sample as a unit; R_{529}(Voting) uses the voting algorithm to ensemble the baseline results from DCNN model; and R_{529}(SVM) uses SVM method instead. The accuracy results are shown as follows.

Table 3: Accuracies of standard spectrograms based solutions

	Folder 1	Folder 2	Folder 3	Folder 4	Average
R_{529}(DCNN)	0.7749	0.7779	0.6948	0.7557	0.7509
R_{529}(Voting)	0.8598	0.8789	0.7656	0.8632	0.8419
R_{529}(SVM)	0.8615	0.8721	0.7732	0.8684	0.8438
R_{706}(DCNN)	0.775	0.7892	0.7065	0.752	0.7557
R_{706}(Voting)	0.8496	0.873	0.7928	0.85	0.8451
R_{706}(SVM)	0.853	0.8679	0.8227	0.8709	**0.8536**
R_{882}(DCNN)	0.772	0.7836	0.6532	0.7489	0.7394
R_{882}(Voting)	0.8513	0.8508	0.7573	0.8602	0.8299
R_{882}(SVM)	0.8581	0.8687	0.7622	0.8635	0.8381

Looking at Table 3, we find that both voting and SVM can significantly improve the baseline accuracies of DCNN model for all the resolutions. Specifically, the SVM method is slightly better than voting algorithm. In this group of experiments, the best average accuracy is 0.8536 which is achieved by the R_{706} (SVM) solution.

4.2. Classifying with CQT Spectrograms

Using the same DCNN model, we conduct several CQT spectrogram based DCNN experiments. The accuracy results are presented in Table 4.

Table 4: Accuracies of CQT spectrograms based solutions

	Folder 1	Folder 2	Folder 3	Folder 4	Average
CQT_{84} (DCNN)	0.7278	0.6946	0.6958	0.7067	0.7062
CQT_{84} (Voting)	0.8154	0.7928	0.7937	0.8188	**0.8052**
CQT_{84} (SVM)	0.8231	0.7715	0.8005	0.8188	0.8035
CQT_{80} (DCNN)	0.6972	0.6878	0.6885	0.6896	0.6908
CQT_{80} (Voting)	0.7701	0.7715	0.7809	0.7872	0.7774
CQT_{80} (SVM)	0.7846	0.7519	0.7801	0.7889	0.7764

Generally, the accuracies of CQT spectrogram based solutions are unsatisfactory, compared with the standard spectrogram. Furthermore, the accuracies of CQT_{80} are worse than the ones of CQT_{84}, which is different with our original expectation [6].

4.3. Classifying with Standard and CQT Spectrograms

Although the accuracies of CQT spectrogram are not very competitive, significant improvements can be achieved when fused with standard spectrograms in our experiments. We have tried several feature combinations and have presented their results in Table 5.

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

As we can see, the four fusion solutions using SVM method 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 standard 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 method over the voting in the fusion scenarios. To better understand the fusion performance, the class-wise accuracies of the best result, namely R_{706} +CQT_{84}(SVM), are further given in Table 6.

Table 6: Class-wise accuracies of the best fusion solution

	Folder 1	**Folder 2**	**Folder 3**	**Folder 4**	**Average**	**Baseline**
beach	0.8718	0.7564	1.0	0.7949	0.8558	0.753
bus	0.9872	0.9615	0.8462	0.9615	0.9391	0.718
cafe/restaurant	0.3333	0.7051	0.7564	0.8462	0.6603	0.577
car	0.9744	0.9615	0.9744	1.0	0.9776	0.971
city_center	0.8718	0.8333	0.9231	0.9103	0.8846	0.907
forest_path	0.9615	1.0	0.9615	1.0	0.9808	0.795
grocery_store	1.0	1.0	0.8718	0.9359	0.9519	0.587
home	0.9744	0.8889	0.9753	0.8077	0.9116	0.686
library	0.6282	1.0	0.9359	0.9487	0.8782	0.571
metro_station	1.0	1.0	0.9872	1.0	0.9968	0.917
office	0.9872	1.0	0.9359	1.0	0.9808	0.997
park	0.6923	0.8462	0.6923	0.8974	0.7821	0.702
residetial_area	0.8846	0.9231	0.8718	0.8718	0.8878	0.641
train	0.8077	0.9487	0.7179	0.9744	0.8622	0.580
tram	1.0	0.7308	0.9872	1.0	0.9295	0.817
total	0.865	0.9037	0.896	0.9299	**0.8986**	0.748

The last column in Table 6 presents the performance of the baseline system provided along with the TUT Acoustic Scenes 2017 dataset in [8]. As we can see, the average accuracy of our best fusion system outperforms the one of baseline system by 20.13 percent.

5. 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 standard (R_{706}) and CQT_{84} spectrograms by voting method (namely the R_{706}+CQT_{84} (Voting) solution). The second one is DCNN based SVM system, which fuses the same data by SVM method (namely the R_{706}+CQT_{84} (SVM) solution).

6. CONCLUSION

In the ASC research domain, CNN is becoming more and more popular[1][2][5][6]. In this work, a DCNN solution is proposed for the acoustic scene classification. 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 the 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 standard spectrogram and the CQT spectrogram are studied. The best accuracy of using the standard 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 standard 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. We believe the performance can be further improved by using some other skills, such as fine tuning of parameters, normalization of spectrograms in the training of DCNN, utilizing the temporal characteristics, etc.

In our experiments, the fusion of multi-resolution standard spectrograms is also explored. The accuracy is also improved slightly, compared to the single resolution schemes. In summary, using the multiple spectrograms can greatly augment the size of training samples, which will result in a better DCNN performance.

When generating standard spectrograms, the width of sliding window as well as the overlap amount are important parameters. In our opinion, they both impact the accuracies of classification. Owing to the time limit, we have not performed grid searching for their values. In our future work, we will further explore the correlations between these parameters and the accuracies. It would be beneficial for finding out the best resolution for the DCNN model.

In [6], it is recommended to remove high frequency bins when preparing CQT inputs for the proposed CNN architecture. However, in our experiment, CQT_{84} works better than CQT_{80} in all cases, which differs with the results in [6]. In fact, the generation of CQT feature in our method is slightly different with the one proposed by [6], for example, we produce CQT samples for left and right channels separately. However, we don't think this contributes much to the difference of the conclusions. Actually, the main difference lies in the architectures of the two CNN model. The CNN structure in [6] is much simpler than the one in this paper. We suppose that the DCNN model in this paper can more effectively utilize the high frequencies bins. This should be validated in our future work.

7. 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 Acoustic 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 Classification 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 Approach using Binaural I-vectors and Deep Convolutional Neural Networks[C]//Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE2016), Budapest, 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 Classification of Acoustic Scenes and Events (DCASE2016), Budapest, Hungary. 2016.

[8] Acoustic Scene Classification[R/OL] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification