CIAIC-MODA SYSTEM FOR DCASE2018 CHALLENGE TASK5

Technical Report

Dexin Li[1]*,Mou Wang*[1]*,Di Li*[1]*, Xueyu Han*[1]*, Qian Wang*[1]*, Qing Liu*[1]*,*
Jisheng Bai[2]*, Ru Wu*[2]*, Bolun Wang*[3]*, Zhonghua Fu*[3]*,*

[1] Northwestern Polytechnical University, Center of Intelligent Acoustics and Immersive Communications, Xi'an, China, dexinli@mail.nwpu.edu.cn
[2] Northwestern Polytechnical University, Xi'an, China, baijs@mail.nwpu.edu.cn
[3] Northwestern Polytechnical University, Audio Speech Language Processing Group, Xi'an, China, berlloon@gmail.com

ABSTRACT

In this technical report, we present several systems for the task 5 of Detection and Classification of Acoustic Scenes and Events 2018 (DCASE2018) challenge. The task is to classify multi-channel audio segments into one of daily activities performed in a home environment. We develop three methods for the task. First, log mel-spectrogram is extracted from each segment and fed to CNN in baseline system with gated linear units (GLU). Then, we use VGGNet to improve the network. In addition, to exploit spatial information, we extract coherence features among all channels and use 1D-CNN with GLU to classify it. Finally, we make a fusion on posteriors from three subsystems to further improve the performances. The experimental results show the proposed systems can get at least 5% F1-score improvement compared to the baseline system.

***Index Terms*—** DCASE, domestic activities, multi-channel acoustics, convolutional neural network.

1. INTRODUCTION

Situational awareness in smart environments has received a increasing attention in recent years due to the rapid development of multimedia and artificial intelligence technology. It can enhance the quality of live for humans in terms of e.g. security, home care and etc. For example, some vocal assistants, such as Google Home and Amazon Echo, use the relevant smart functionality through acoustic information. In situational awareness, recognition of activities based on acoustics is a significant part. It has huge importance, particularly in several applications, such as monitoring of domestic activities, robots and etc. In the challenge on detection and classification of acoustic scenes and events (DCASE) [1] of this year, a new task is added, i.e. task 5 [3]. The goal of this task is to classify multi-channel audio segments into domestic activities performed in a home environment. For example, "Cooking", "Watching TV" and "Working"[1].

Many works on recognition of activities based on acoustics has been published. For example, in previous DCASE challenge, various approaches had been presented to cope with the similar task. Most of which are based on deep learning methods[4][5][6], such as CNN, RNN, generative adversarial networks, while the number of SVM, NMF and Gaussian mixture model based methods decreased. In addition, we can find that the acoustic models are typically based on single channel and single location recordings in previous studies. However, it is convinced that multi-channel acoustic recordings can give more information and improve the performance of the current works.

In this report, we propose several methods for task5 of DCASE2018. In the part of feature extraction, we extract log mel-spectrogram from each channel of each segment, and feed it to classifier as a image. Moreover, considering spatial information, we extracted correlation features from the multi-channel audio segments. And then, CNN with different activation function and architecture is used for classification. We used VGGNet to classify the spectrogram feature and try to use GLU to improve network. Since the correlation feature is one dimensional, 1D-CNN is adopted. Finally, we make 4 different fusion system with aforementioned 3 subsystems in our submissions.

The rest of this technical report is organized as follows. In the Section 2, we describe the datasets used in DCASE2018 task5 and feature briefly. The different models will be introduced in Section 3. In the Section 4, the experiments and results are presented. Lastly, Section 5 concludes this report.

2. DATASET AND FEATURES

2.1. Dataset

The dataset used in task5 of DCASE2018 is a derivative of the SINS dataset[7]. Where the activities recorded by 7 of 13 microphone arrays are used. Each audio segment contains 4 channels with 10 seconds length and 16kHz sampling rate. Each segment has corresponding activity class: absence, cooking, dishwashing, eating, other, social activity, vacuum cleaning, watching TV and working.

As development set, the data from 4 sensor nodes with the ground truth is given. While the data in evaluation set is provided from all the sensor nodes even not present in the development set.

2.2. Features

There are two sets of features used in our systems as follows:

(1) log-mel energy: Firstly, we calculate STFT with a hamming window of 40ms, and the hop size is set to 50%. Then, we apply 40-band mel filter bank from 50Hz to 8kHz to obtain the 40-dimension mel spectrograms. Finally, we apply logarithm to the results of last step and get the 40x501 sized feature matrix.

Table 1: Model description. Each column represents one model respectively. conv: convolution layer, maxpool: max-pooling layer, global-avgpool: global average pooling layer, FC: fully connected layer.

Convolution Neural Network Configurations.	
VGGNet10	1D-CNN
10 weight layers	
conv3-32	
conv3-32 conv3-32	
maxpool2	maxpool4
conv3-64 conv3-64	
maxpool2	maxpool4
conv3-128 conv3-128	
maxpool2	maxpool4
conv3-256 conv3-256	
global-avgpool	
FC-9	
soft-max	

(2) coherence of channels: To make good used of multi-channel audio segments, we tried to extract spatial features. Firstly, we apply hamming window and get STFT at a window length of 20ms. Secondly, we use the data of all channels in each TF point to calculate the auto-correlation matrix. Thirdly, we do mean along the frame axis. Last but not least, the feature vector at each frequency point is calculated by dividing cross-spectral density with the self-spectral density. Finally, we concatenate the real and imagine value at each frequency point to form the feature vector for one 10s segment.

3. PROPOSED METHODS

3.1. Models

We used three basic convolution neural network models in our systems, we will introduce each of them in details.

We rebuild the baseline system as model 1 with TensorFlow, but the batch normalization operation after convolution process was removed.

To build the second model, we replaced the network framework in baseline system with the changed VGGNet. Model details are described in Table 1. The input features are also log-mel energies.

As for using the coherence features, which is a column vector for one 10s segment, we build a 1D-CNN. The network framework also can be checked in Table 1, which is constructed by 10 weight layers the same as model 2, while the size of max pooling layer is different.

3.2. Changed learnable gated activation function

Using the learnable gated activation function for training the neural network in DCASE2017 task4 is presented in [8] by Xu. And the results shows the method can make great improvement for sound event detection sub-task. Here we tried a variant of the learnable gated activation function in our model.

Table 2: The results of F1-score on development set for D-CASE2018 task5.

Activity	F1-score	
	DCASE2018 baseline	Fusion
Absence	85.41%	89.40%
Cooking	95.14%	96.87%
Dishwashing	76.73%	87.50%
Eating	83.64%	94.42%
Other	44.76%	63.06
Social activity	93.92%	95.73%
Vacuum cleaning	99.31%	100%
Watching TV	99.59%	99.65%
Working	82.03%	88.46%
Macro-averaged F1-score	**84.50**%	**90.68**%

$$\mathbf{Z} = \mathbf{X} * \mathbf{W} + \mathbf{b} \quad (1)$$

$$\mathbf{Y1} = relu(\mathbf{Z}) \quad (2)$$

$$\mathbf{Y2} = sigmoid(\mathbf{Z}) \quad (3)$$

$$\mathbf{Y} = \mathbf{Y1} \bigotimes \mathbf{Y2} \quad (4)$$

Where **X** is the input feature, parameter **W** and **b** is the weight matrix and biases vector in each layer respectively. We applied ReLU and Sigmoid activation function to the output of product-sum process, finally the output **Y** is obtained by element-wise multiplication of **Y1** and **Y2**.

This method is used to training the baseline-based model and 1D-CNN model.

3.3. Fusion

In general, system fusion could get better performance. Here in DCASE2018 task5, with the models we have described before, we believe that the results of fusion system can be improved. In details, we do average to the posteriors of each subsystem.

4. EXPERIMENTS AND RESULTS

4.1. Experiment setting

We performed our system with development set provided by D-CASE2018 organizer to evaluate the performance of them. During developing phase, we use the official cross-validation setup which consists of four folds distributing the development dataset in order to make results uniform[1].

The VGGNet based system trains a single classifier model that takes a single channel as input the same as baseline. In the prediction stage a single outcome is computed for each node by averaging the 4 model outcomes (posteriors) that were computed by evaluating the trained classifier model on all 4 microphones.

Considering the optimization, we used an Adam optimizer with initial learning rate of 0.0001. There are different epoch number and mini-batch size setup for training each model according to the experiment experiences.

4.2. Experiment results

Table 2 presents the results of F1-score for monitoring of domestic activity task on development set. From which we can see that fusion of three models can get 90.68% F1-score on development set, while the official baseline system just obtained 84.50% F1-score.

The fusion results detailed in Table 2 is the highest score our system has obtained. In that fusion system, the posteriors of three models are considered. During training the baseline-based model and 1D-CNN model, we applied the learnable gated activation function.

5. CONCLUSIONS

In this technical report, we presents our system and models used for cope with task5 of DCASE2018 challenge. We used three deep CNN based model on log-mel energy and coherence features respectively. The GLU is also introduced to improve the performance. It can get 90.68% F1-score on official development set.

6. REFERENCES

[1] http://dcase.community/challenge2018/

[2] http://www.cs.tut.fi/sgn/arg/dcase2017/.

[3] http://dcase.community/

[4] H. E. 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," DCASE2016 Challenge, Tech. Rep., Sep. 2016.

[5] Y. Han, J. Park and K. Lee, "Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification," DCASE2017 Challenge, Tech. Rep., Nov. 2017.

[6] R. Lu and Z. Duan, "Bidirectional GRU for Sound Event Detection," DCASE2017 Challenge, Tech. Rep., Nov. 2017.

[7] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. V. Waterschoot, B. Vanrumste, M. Verhelst and P. Karsmakers, "The SINS database for detection of daily activities in a home environment using an acoustic sensor network," in *Proc. of Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, pp. 32-36, Nov. 2017.

[8] Y. Xu, Q. Kong, W. Wang and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Proc. IEEE ICASSP*, 2018.

[9] http://dcase.community/challenge2018/task-monitoring-domestic-activities