HOME ACTIVITY MONITORING BASED ON GATED CONVOLUTIONAL NEURAL NETWORKS AND SYSTEM FUSION

Yu-Han Shen[1], *Ke-Xin He*[1], *Wei-Qiang Zhang*[1]*,

[1] Department of Electronic Engineering, Tsinghua University, Bejing 100084, China
yhshen@hotmail.com, hekexinchn@163.com, wqzhang@tsinghua.edu.cn

ABSTRACT

In this paper, we propose a method for the task 5 of Detection and Classification of Acoustic Scenes and Events 2018 (DCASE2018) challenge. This task aims to classify multichannel audio segments into one of the provided predefined classes. All of these classes are daily activities performed in a home environment. This paper mainly adopts a model based on gated convolutional neural networks for domestic activity classification. We utilize multiple methods to improve the performance of our proposed system. Firstly, we use gated convolutional neural network to replace normal convolutional neural network and recurrent neural network in order to extract more temporal feature and improve working efficiency. Secondly, we mitigate the problem of data imbalance using a weighted loss function. Besides, we adopt model ensemble strategy to make our system stronger and more effective. Furthermore, we use a fusion of two systems to improve our performance. In a summary, we obtain 89.73% F1-score on the development dataset [1] while the official baseline system gets 84.50% F1-score.

Index Terms— Home activity monitoring, gated convolutional neural network, model ensemble, system fusion

1. INTRODUCTION

DCASE challenge is one of the most important international challenges in the field of acoustic event detection and classification and has been organized for several years. DCASE 2018 challenge consists of five tasks and we focus on task 5. This task evaluates systems for monitoring of domestic activities based on multi-channel acoustics. In this task, the audio segments can be classified into nine classes: absence, cooking, dishwashing, eating, other, social activity, vacuum cleaning, watching TV and working. All audio segments are derived from continuous recordings collected by seven microphone arrays and each segment contains four channels.

We can also refer to this task as acoustic activity classification. The main procedure of acoustic activity classification consists of two parts: extracting acoustic features and designing acoustic models as classifiers. Mel Frequency Cepstrum Coefficient (MFCC) is a common traditional acoustic feature and has been widely used [2]. But log Mel-scale Filter Bank energies (fbank) [3][4] are becoming more popular recently, and many works have been done based on fbank.

In recent years, Convolutional Neural Networks (CNNs) have achieved great success in many fields such as character recognition, image classification, speaker recognition. And many works [5][6] based on CNNs have been done in the field of acoustic event classification and detection. Besides, some researchers [3][4] have combined CNNs with Recurrent Neural Networks (RNNs) to capture temporal contexts of audio signals for further improvements. One of the biggest disadvantages of RNNs is that it takes a relatively long time to train a RNN-based mode because it cannot be calculated by parallelization over sequential tokens. In our work, we substitute recurrent neural networks with gated convolutional neural networks (GCNNs) proposed by Dauphin et al. in [7][8]. The rest of this paper is organized as follows. In Section 2, we introduce our methods in detail, mainly including acoustic features, classifier, weighted loss function, model ensemble and system fusion. The experiment setup, evaluation metric and our results are illustrated in Section 3. Finally, the conclusion of our work is presented in Section 4.

2. METHODS

2.1. System Overview

We mainly adopt two systems. The first one is based on gated convolutional neural network and the second one is based on GMM super vector followed by support vector machine. In the following part of this section, we will introduce the two systems respectively.

The input of both systems is two-dimensional acoustic features. We utilize deep learning model based on gated convolutional neural network as classifier in the first system. In the second system, we use GSV-SVM based model as classifier. The outputs of both classifiers are confidence probabilities for nine classes, and the class with the maximum probability is considered to be our prediction. In order to improve the robustness of our system, we adopt ensemble strategy. We fuse the confidence probabilities of multiple systems to get the final result.

2.2. Acoustic Feature

In our first system, we use fbank as the input of our system. Fbank is a two-dimensional time-frequency acoustic feature. It imitates the characteristics of humans ears and concentrates more on the low frequency components of audio signals. Compared with traditional MFCC feature, more original information can be kept in fbank and it has been widely used in deep learning. To extract fbank feature, each audio segment is divided into 40ms frames with 50% overlapping, and then 40 mel-scale filters are applied on the magnitude spectrum of each frame. Finally, we take logarithm on the amplitude and get fbank feature. Each feature is normalized to zero mean and unit standard deviation before it is input into the following classifier. As is mentioned in Section 1, the audio segments contain four channels, so our fbank feature contains four channels as well. In our work, four channels are fed into the system separately while

*Zhang is the corresponding author.

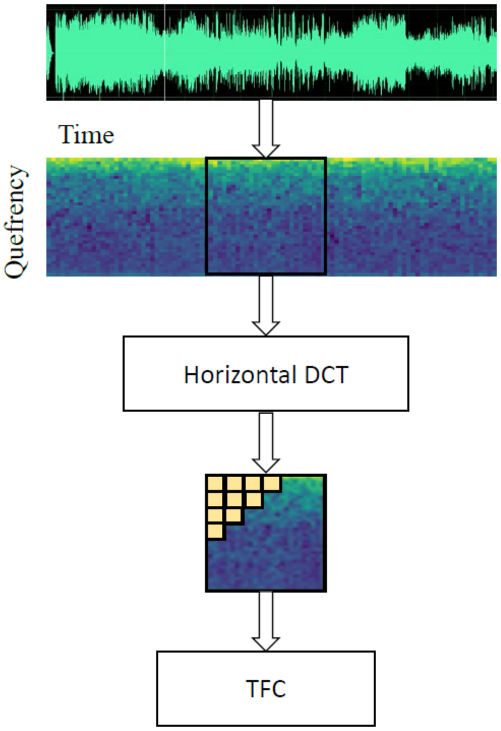


Figure 1: Procedures of TFC feature extraction.

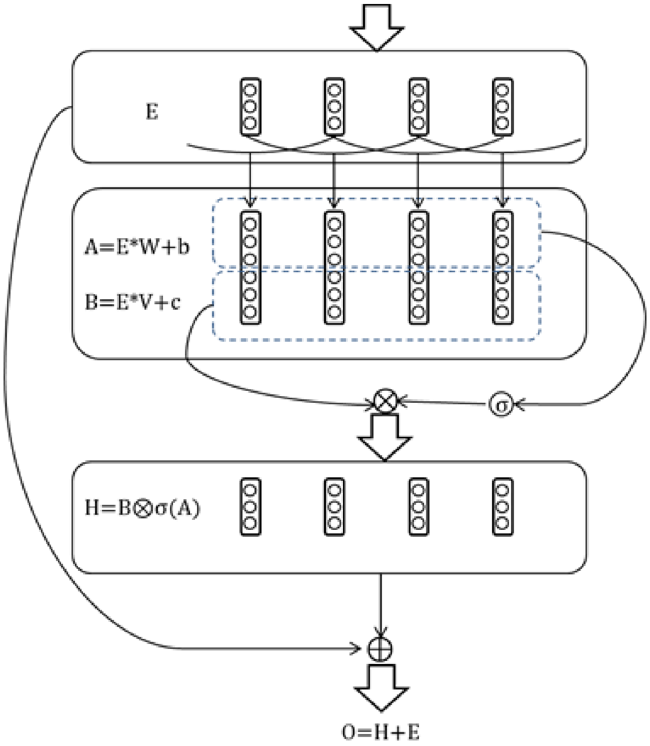


Figure 3: Gated convolutional neural network.

training. And the average output score of four channels is used for evaluation.

As for the second system, the input is TFC of synthetic mono channel audios. We adopt beamforming based on delay and sum to synthesize the four-channel audios. In our previous work, we have presented a time-frequency cepstral (TFC) feature [9], which is obtained by performing a temporal discrete cosine transform (DCT) on the cepstrum matrix and selecting the transformed elements in a zigzag scan order. And we increase discriminability through a heteroscedastic linear discriminant analysis (HLDA) on the full cepstrum matrix. As is shown in Figure 1, in the TFC feature extraction, the successive frames of basic feature vectors within a context width window are first extracted to form a cepstrum matrix. A temporal (in horizontal direction) DCT is then performed on the cepstrum matrix and the elements in the upper-left triangular area are in a zigzag scan order. We attempt to remove correlation in temporal direction and preserve the elements with greater variability [10]. The procedure of TFC is equivalent to perform a two-dimensional DCT on spectrum-time matrix, which can be interpreted as a compression of the information by a DCT truncation.

In this paper, we divide the audio segment into small segments of 20 frames. This leads the variances pattern of the cepstrum matrix after a horizontal DCT to be nearly an isosceles triangle, thus we can perform a zigzag scan to select elements in this area to form the TFC feature.

2.3. Classifier

The classifier used in our work consists of three main parts: 1) convolutional neural network (CNN), 2) gated convolutional neural network (GCNN), 3) feedforward neural network (FNN). And our overall architecture is shown in Figure 2.

Convolutional layers extract frequency features and connect features of adjacent frames. And the output of convolutional layer is followed by batch normalization [11], a ReLU activation unit and a dropout layer [12]. Then a max-pooling layer is applied to keep the most important features. The structure of gated convolutional neural network is illustrated in Figure 3.

In gated convolutional neural network, the output of convolutional layer is divided into two parts with the same size. The input of this structure is E = [e$_1$, e$_2$, , e$_n$], E passes through a convolutional layer and the output is divided into A and B. Then A passes through sigmoid activation function and multiplies with B by element-wise. In order to enable stronger work, we add residual connections from the input E to the output of this structure H. Residual network is introduced to avoid vanishing gradient problem. The specific formula is as follows:

$$A = E * W + b, \tag{1}$$

$$B = E * V + c, \tag{2}$$

$$H = B \otimes \sigma(A), \tag{3}$$

where W, V represent convolutional kernel values, and b, c mean biases. ⊗ represents element-wise production. The gated convolutional layer is also followed by batch normalization, a ReLU activation unit, a dropout layer and a global max-pooling layer.

After the gated convolutional neural network, the features on multiple channels are flattened into frequency axis.

Then two dense layers are used to combine extracted features and output nine scores. A softmax activation unit is added to normalize those scores into confidence probabilities. The second sys-

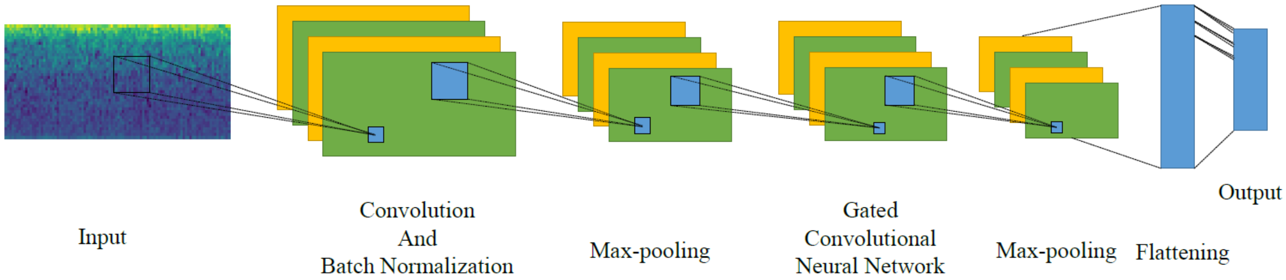


Figure 2: Overall architecture of our classifier.

tem is based on GMM super vector followed by support vector machine. The dimension of feature is 39, the number of GMM mixtures is 512, and the kernel of SVM is linear kernel. Specific technical details are explained in [9][10].

2.4. Weighted Loss Function

The daily activities for task 5 are shown in Table 1 along with the available 10s multi-channel segments in the development set and the amount of full sessions of a certain activity. It is clear that the amounts of 10s segments of different activities vary a lot, which causes a problem of data imbalance. Consequently, we adopted a weighted loss function to limit the negative effects of data imbalance. The loss function we use is computed as follows.

$$L = -\frac{1}{N}\sum_{n=1}^{N} w_n[y_n \log \widehat{y}_n + (1-y_n)\log(1-\widehat{y}_n)] \tag{4}$$

where N denotes the number of classes, w_n is the weight coefficient of the n-th class when computing the loss. We raise the weight coefficient for those classes with fewer segments.

Table 1: Amounts of audio segments and sessions

Activity	#10s segments	#sessions
Absence	18860	42
Cooking	5124	13
Dishwashing	1424	10
Eating	2308	13
Other	2060	118
Social activity	4944	21
Vacuum cleaning	972	9
Watching TV	18648	9
Working	18644	33
Total	72984	268

2.5. Ensemble

Model ensemble is a common strategy in machine learning [3][4]. In our work, we adopt various levels of model ensemble. During our experiments, we notice that absence and working are two sorts of activities that are often misclassfied with each other. So we train a model in particular to classify those two classes of activities. When our main system classifies an audio segment as either of the two classes, we will use the specially trained model for one more classification. The output probability scores of two systems will be fused according to the following formula:

$$p[0] = p_2[0] \cdot (p_1[0] + p_1[8]) \tag{5}$$

$$p[8] = p_2[1] \cdot (p_1[0] + p_1[8]) \tag{6}$$

where p_1 is a 9-dim vector, denoting the output score of the first model, and p_2 is a 2-dim vector, denoting the output score of the second model.

Besides, we also ensemble the iterations among the same model and the outcomes of models with different parameters.

2.6. System fusion

We transfer an existing speaker recognition system based on GMM super vector followed by support vector machine for fusion. We fuse the output scores of our two systems and get the final result.

3. EXPERIMENT, EVALUATION AND RESULTS

3.1. Experiment setup

Our model is trained using Adam [13] for gradient based optimization. Weighted cross-entropy is used as the loss function in order to mitigate the problem of data imbalance. And the structure of our main model is shown in Table 2 along with parameters. The initial learning rate is 0.001 and the iteration is stopped when the validation loss does not decrease for 10 epochs.

Table 2: Model structure and parameters

Input 40×501×1
Conv (padding: valid, kernel: [40, 5, 64]) BN-ReLU-Dropout(0.2) 1×5 Max-Pooling(padding: valid) Gated Conv (padding: same, kernel: [1, 3, 128]) BN-ReLU-Dropout(0.2) Global-Max-Pooling
Feature Flattening
Dense layer1(unit num: 64) -ReLU-Dropout(0.2) Dense layer2 (unit num: 9) -softmax

3.2. Evaluation

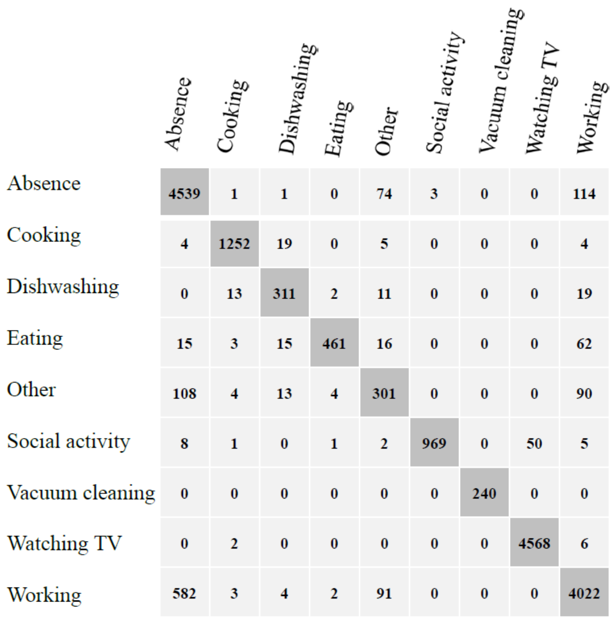
The official evaluation metric for DCASE2018 challenge task 5 is macor-averaged F1-score. F1-score is a measure of a test's accuracy and it is the harmonic average of precision and recall. Macro-averaged means that F1-score is calculated for each class separately and averaged over all classes. For this task, a full 10s multi-channel audio segment is considered to be one sample.

3.3. Results

The performances of the official baseline system [14] and our proposed system on the development set are shown in Table 3. Our proposed system includes the first GCNN-based system, the second GSV-SVM-based system and the fusion of the two systems. From the table, we can notice that the performance of fusion system is superior to both systems.

Table 3: Amounts of audio segments and sessions

Activity	System 1	System 2	Fusion	Baseline
Absence	89.03%	87.03%	90.89%	85.41%
Cooking	96.98%	79.38%	97.70%	95.14%
Dishwashing	83.50%	68.78%	86.50%	76.73%
Eating	85.85%	83.81%	88.48%	83.64%
Other	49.76%	31.40%	59.02%	44.76%
Social activity	96.99%	88.19%	96.51%	93.92%
Vacuum cleaning	99.99%	87.17%	99.99%	99.31%
Watching TV	99.51%	97.58%	99.37%	99.59%
Working	87.22%	83.51%	89.12%	82.03%
Average	87.65%	78.54%	89.73%	84.50%



	Absence	Cooking	Dishwashing	Eating	Other	Social activity	Vacuum cleaning	Watching TV	Working
Absence	4539	1	1	0	74	3	0	0	114
Cooking	4	1252	19	0	5	0	0	0	4
Dishwashing	0	13	311	2	11	0	0	0	19
Eating	15	3	15	461	16	0	0	0	62
Other	108	4	13	4	301	0	0	0	90
Social activity	8	1	0	1	2	969	0	50	5
Vacuum cleaning	0	0	0	0	0	0	240	0	0
Watching TV	0	2	0	0	0	0	0	4568	6
Working	582	3	4	2	91	0	0	0	4022

Figure 4: Prediction matrix of proposed system.

For almost all activities, our proposed system shows relatively better performance than baseline, except that the F1-score of watching TV of our system is a little lower than that of baseline since it is high enough. The prediction matrix of our proposed system is shown in Figure 4. The element in the *i*-th row and *j*-th column of this matrix represents the amount of audio segments that belong to class *i* and are classified as class *j*.

4. CONCLUSION

In this paper, gated convolutional neural network, model ensemble strategy and system fusion have been proposed for monitoring activities in home environment. We have introduced our work and the results show that the performance of our proposed system is superior to that of the official baseline. It has been shown that the performance can be further improved by model ensemble and system fusion.

5. REFERENCES

[1] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van 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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Munich, Germany, November 2017, pp. 32–36.

[2] R. Hyder, S. Ghaffarzadegan, Z. Feng, and T. Hasan, "Buet bosch consortium (b2c) acoustic scene classification systems for dcase 2017 challenge," *Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 Challenge*, 2017.

[3] E. Cakir and T. Virtanen, "Convolutional recurrent neural networks for rare sound event detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, pp. 803–806.

[4] H. Lim, J. Park, K. Lee, and Y. Han, "Rare sound event detection using 1d convolutional recurrent neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, pp. 80–84.

[5] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, pp. 46–50.

[6] W. Zheng, J. Yi, X. Xing, X. Liu, and S. Peng, "Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop*, 2017, pp. 133–137.

[7] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *arXiv preprint arXiv:1612.08083*, 2016.

[8] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," *arXiv preprint arXiv:1705.03122*, 2017.

[9] W.-Q. Zhang, L. He, Y. Deng, J. Liu, and M. T. Johnson, "Timecfrequency cepstral features and heteroscedastic linear discriminant analysis for language recognition," in *IEEE Transactions on Audio Speech & Language Processing*, vol. 19, no. 2, 2010, pp. 266–276.

[10] W.-Q. Zhang, Y. Deng, L. He, and J. Liu, “Variant time-frequency cepstral features for speaker recognition,” in *Interspeech*, 2010, pp. 2122–2125.

[11] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of The 32nd International Conference on Machine Learning*, 2015, pp. 448–456.

[12] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[13] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

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