Asymmetric Kernel Convolutional Neural Network for Acoustic Scenes Classification

Chien-Yao Wang and Jia-Ching Wang
Dept. of Computer Science & Information Engineering
National Central University
Jhongli City, Taiwan, ROC
jcw@csie.ncu.edu.tw

Yu-Chi Wu and Pao-Chi Chang
Dept. of Communication Engineering
National Central University
Jhongli City, Taiwan, ROC
pcchang@ce.ncu.edu.tw

***Abstract*—In this work, we propose an Asymmetric Kernel Convolutional Neural Network (AKCNN) for Acoustic Scenes Classification (ASC). Its kernel shape is not the traditional square but asymmetric in width and height. It also uses Weight Normalization (WN) to accelerate the training process because it can early converge the training loss and accuracy. The best of all, WN can help increase the accuracy of ASC. TUT Acoustic Scenes 2016 Dataset [1] is used for evaluation. The result shows that AKCNN achieves accuracy 86.7%. If we rank the score in DCASE2016 ASC Challenge, it shows that the system would have a higher score than the 5th place.**

Keywords—acoustic scenes classification; deep learning; convolutional neural network

I. INTRODUCTION

Audio plays an important role in environment recognition, so this makes the research of Acoustic Scenes Classification (ASC) important. However, there was no unified dataset for comparison before 2013. Due to the contributions to the first IEEE Audio and Acoustic Signal Processing (AASP) Challenge: Detection and Classification of Acoustic Scenes and Events (DCASE) in 2013 [2], there existed a benchmark to compare the performance of ASC. Furthermore, Tampere University of Technology reset a brand new dataset in DCASE2016 Challenge [1].

Nowadays, the applications of convolutional neural networks (CNNs) for acoustic correlation research become more and more popular. The most commonly used architecture in the top ten DCASE2016 Challenge was CNN. In this work, we propose an Asymmetric Kernel Convolutional Neural Network (AKCNN), which is designed based on LetNet [3] style networks. We specifically pay attention to the kernel shape to achieve the best performance.

II. PROPOSED METHOD

Figure 1 shows the proposed system for ASC. This system basically consists of three parts: feature pre-processing, CNN training stage, and CNN testing stage.

A. Feature Pre-processing

Before audio segments are sent into training stage or testing stage, feature pre-processing is applied. It includes feature extraction, feature normalization, and feature segmentation.

The audio feature the system used is log-mel spectrogram. To calculate log-mel spectrogram it applies STFT over windows of 40ms of audio with 50% overlap by using 2048 samples Hamming Window. Then, the absolute values of the STFT spectrogram are mapped to 40-band mel-scale filter bank. Finally, the logarithm is taken to generate 40 bins of the log-mel spectrogram, represented by 40-demension vectors.

The feature normalization is to normalize each bin by the mean and standard deviation of all training data. The final process of feature pre-processing is to segment the normalized log-mel spectrogram into 25 frames per segment without overlapping. Therefore, the input of CNN is a 40-bin 25-frame log-mel spectrogram.

B. CNN Training Stage

The architecture of AKCNN is shown in Figure 2. The parameters we set are chosen by various experiments. AKCNN includes two convolutional layers. The first convolutional layer performs a convolution over the input spectrogram with 128 kernels and the other layer is with 256 kernels. The weight and the height of kernels are *asymmetrical* which is 5x7 based on the audio feature processing setting, where 5 is the same axis as frequency domain and 7 represents the same axis as time domain. Also, these two convolutional layers include weight normalization (WN) [4] to accelerate training. The normalization function is shown as follow:

$$\mathbf{w} = \frac{g}{\|\mathbf{v}\|}\mathbf{v} \qquad (1)$$

where $\mathbf{v}$ means the weight as we used normally, g is a scalar representing the amplitude of $\mathbf{v}$. By this two variables, the normalized weight $\mathbf{w}$ is calculated which is the kernel of AKCNN. The activation function used for kernels in both convolutional layers is rectifier linear units (ReLU).

The two 5x5 non-overlapping max-pooling layers, which perform sub-sampling, are placed separately after two convolutional layers. Finally, because the classification includes 15 different labels, the flattened second max-pooling layer is connected to a WN softmax layer which is composed of 15 fully connective neurals.

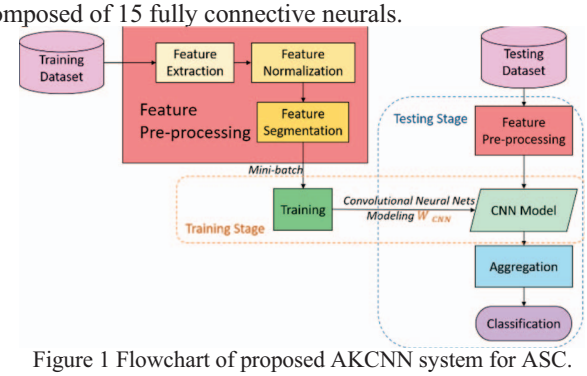

Figure 1 Flowchart of proposed AKCNN system for ASC.

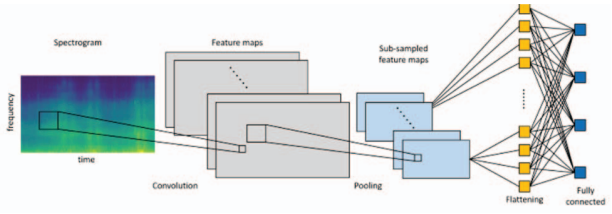

Figure 2 The architecture of proposed AKCNN.

TABLE I. ACCURACY FOR DIFFERENT SIZE OF KERNELS WITH 40 BANKS LOG-MEL SPECTROGRAM.

40banks log-mel-spec. Mel x Frame	Accuracy (%)
3x3	82.3
5x5	**85.6**
7x7	84.1
5x3	84.6
5x7	**86.7**
5x9	85.6
3x5	85.6
7x5	84.9
9x5	84.4

C. CNN Testing Stage

Since dividing a segment into smaller segments, the classification of the whole segment is obtained by averaging all the AKCNN prediction output. The AKCNN output $CNN_{out}^{j}[i]$ is a vector including all class-wise prediction probability for the j^{the} sequence. Therefore, the predicted class $class$ for the whole segment is calculated as follow:

$$class = \text{argmax}_i \left(\frac{\sum_{j=1}^{fr/d} CNN_{out}^{j}[i]}{fr/d} \right) \qquad (2)$$

where i is the number of labels and d is a number of the divided frames setting as 25 frames.

III. EVALUATION

The system has been implemented in Python with librosa library and Tensorflow library. Its training is performed with a Nvidia GTX TITAN X GPU showing as average training time of 20 seconds per epoch. The loss function we used is categorical cross-entropy.

A. Dataset

The dataset we used is TUT Acoustic Scenes 2016 Dataset [1]. Our training set is TUT Acoustic Scenes 2016 development dataset, and testing set is TUT Acoustic Scenes 2016 the evaluation dataset. The development set consists of 1170 audio segments where every label includes 78 audio segments of 30 seconds with total 15 labels. The 15 labels are beach, bus, café/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, park, residential area, train, and tram.

B. Neural Network Experiments

For our experiment, the accuracy to different kind of kernels is shown in Table I. The top three layers are symmetrical shapes of 3x3, 5x5, and 7x7.The best result is the 5x5 kernel, so we only show kernels that at least one of width or height is 5 in other experiments. The result shows that when frequency domain is 5 and the time domain is 7 (5x7), it achieves the best result of 86.7% accuracy. However, if the asymmetric kernel shape is 7x5, the performance will decrease 1.8%.

TABLE II. ACOUSTIC SCENE CLASSIFICATION ON DCASE2016 EVALUATION DATASET (%).

	Valenti [5]	DCASE2017 baseline [6]	AKCNN
Beach	84.6	80.8	92.3
Bus	100	100	92.3
Café/Restaurant	76.9	38.5	57.7
Car	100	96.2	100
City center	96.2	84.6	96.2
Forest path	100	100	100
Grocery store	92.3	65.4	88.5
Home	92.3	80.8	88.5
Library	92.3	46.2	46.2
Metro station	42.3	100	100
Office	96.2	100	100
Park	76.9	96.2	100
Residential area	76.9	88.5	84.6
Train	65.4	30.8	53.9
Tram	96.2	88.5	100
Overall Accuracy	**86.2**	**79.7**	**86.7**

Our best result is using 40-bank log-mel spectrogram into 5x7 kernel CNN, so these parameters are the setting of our AKCNN system. TABLE II shows the comparison of our system with Valenti [5] and DCASE2017 baseline [6]. We can discover that except for the label of café/restaurant, library, and train, the accuracy of other labels is over 80% in our system. Especially, car, forest path, metro station, office, park, and tram achieve 100% accuracy.

IV. CONCLUSION

In this paper, we proposed AKCNN which includes two WN convolutional layer with asymmetric 5x7 kernels with 40-bank log-mel spectrogram input. The accuracy of the system is 86.7% with 6 totally correct classes. If the system attended DCASE2016 Challenge, it would be ranked to the fifth place.

REFERENCES

[1] M., Annamaria, T. Heittola, and T. Virtanen, "TUT Database for Acoustic Scene Classification and Sound Event Detection," IEEE 2016 24th Euro-pean Signal Processing Conference, pp. 1128-1132, Aug. 2016.

[2] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events: An IEEE AASP challenge," in Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on. IEEE, 2013, pp. 1-4.

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998.

[4] T. Salimans and D. P. Kingma, "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks," in Advances in Neural Information Processing Systems, pp. 901-909, 2016.

[5] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, "DCASE 2016 Acoustic Scene Classification Using Convolutional Neural Networks," IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2016), Budapest, Hungary, Sep. 2016.

[6] DCASE2017 Challenge Baseline website, http://doi.org/10.5281/zenodo.400515, retrieved Mar. 17, 2017.