Convolutional-Recurrent Neural Network with the Tensor Fusion Mechanism for Acoustic Scene Classification

Pengxu Jiang, *Member, IAENG,* Ruxue Guo, Ruiyu Liang, Yue Xie, and Cairong Zou

***Abstract*—Acoustic scene classification (ASC) is one of the key fields of artificial intelligence. Due to the short duration of scene audio features, the existing deep learning network cannot fully capture information in short-term audio. In this regard, a convolutional-recurrent neural network with the tensor fusion mechanism (CRN-FM) is proposed for ASC. Each audio is divided into fixed-length segments, and the spectral features are extracted from the segment audio as the input. Then, a convolutional neural network (CNN) is used to obtain time-frequency related information, and long short-term memory (LSTM) is used to obtain time-related details. When receiving the output of the high-level features by the two modules, the designed tensor fusion attention layer fuses different tensors according to the difference in information saturation. Finally, a SoftMax classifier is used to classify scenes. Experimental results on DCASE 2018 and 2019 ASC datasets demonstrate the effectiveness of the proposed approach.**

***Index Terms*—Acoustic scene classification, Convolutional Neural Network, Long Short-Term Memory, Fusion attention layer.**

I. Introduction

People can perceive the scene from sound because the sound contains a lot of surrounding environment information. ASC automatically identifies specific scenes through simulating human perception, such as parks, airports, subway stations, etc. ASC has a variety of real-life applications such as hearing and navigation[1], [2].

Early ASC tasks used hand-designed features to classify different scenes. Although these features contributed to the development of ASC, these low-level features could not clearly express the audio environment. Most ASC-based researchers have focused on the detection and classification of acoustic scenes and events challenges (DCASE) in recent years [3]. As one of the hot tasks of DCASE, ASC has received significant attention. In ASC-based research, all teams with high system performance use the framework based on deep learning[4]. The deep learning algorithm extracts low-level features into high-level feature representation, eliminating the subjectivity of human selection of features. The generalization of automatic training of the deep learning model is better than that of hand-designed features.

The primary deep learning models include convolutional neural networks (CNN)[5] and long short-term memory (LSTM)[6]. Based on ASC research, convolutional neural networks account for most deep learning networks. LSTM-related models are rarely used in ASC-related studies because these models can not extract the information in audio effectively, mainly for the following reasons. First, the scene audio does not contain semantic information, making LSTM unable to combine the context information effectively. That is, scene audio does not need to contain human voices. Therefore, LSTM cannot use the speech-related information of the front and back associated frames in the scene audio features, which limits the utilization of LSTM in ASC-related research. LSTM has been widely used in sound event detection (SED) because SED-related research contains rich semantic information. Secondly, some scene audio may include many silent fragments, which may interfere with the discrimination of the system. In addition, the scene information in audio usually lasts for a short time, which may cause the LSTM to be unable to focus on it effectively. One solution is to put the LSTM module behind CNN, which is first used to reduce the dimension of features, but the LSTM module may lose the short-term information of input features.

To address these challenges, a convolutional-recurrent neural network with the tensor fusion mechanism (CRN-FM) is proposed for ASC, using the network structure presented in Fig. 1. Firstly, we extract spectral features from audio. Spectral features are the most commonly used model input features in ASC research based on neural networks. Similar to picture representation, spectral features contain time-frequency related information of audio. To make the model focus on the limited scene information in audio, we divide the spectral features into segment-level features as the input of the model. Then, the CNN module is used to learn the frequency information in audio, and the LSTM module is used to learn the time information in audio. To integrate these segment-level features, we designed a fusion attention layer to integrate these segmented features. The fusion attention layer fuses the outputs of different modules according to the difference in information weight between different segment-level features. Finally, a SoftMax classifier classifies different scenes.

Manuscript received May 19, 2022; revised October 22, 2022.

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2020YFC2004002 and 2020YFC2004003, the National Natural Science Foundation of China under Grant No. 62001215.

Pengxu Jiang is a PhD candidate of the School of Information Science and Engineering, Southeast University, P.R. China.(email:px20115c@163.com)

Ruxue Guo is a PhD candidate of the School of Information Science and Engineering, Southeast University, P.R. China.(email:grx0904@sina.com)

Ruiyu Liang is a professor of the School of Communication Engineering, Nanjing Institute of Technology, P.R. China.(email:liangry@njit.edu.com)

Yue Xie is a professor of the School of Communication Engineering, Nanjing Institute of Technology, P.R. China.(email:230169046@seu.edu.com)

Cairong Zou is a professor of the School of Information Science and Engineering, Southeast University, P.R. China.(email:cairong@seu.edu.cn)

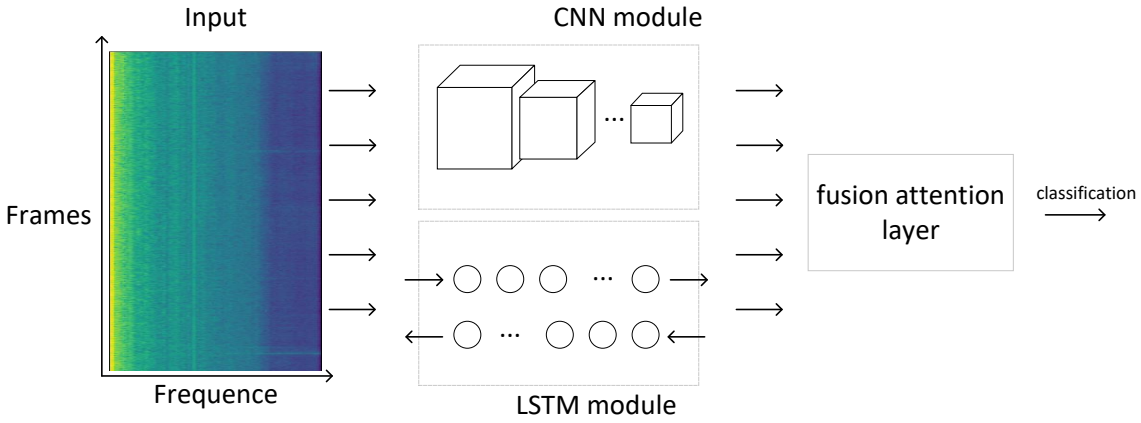

Fig. 1. The architecture of the proposed (CRN-FM) model.

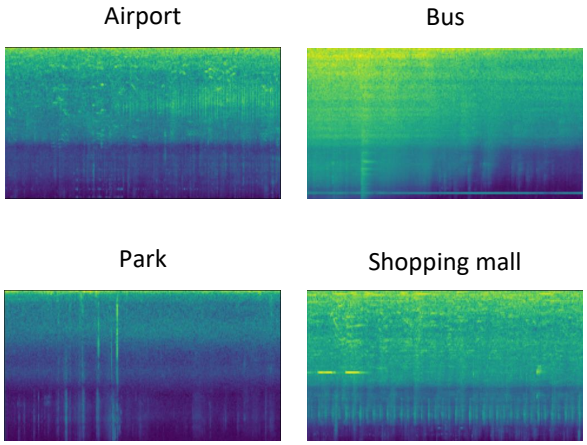

Fig. 2. The spectrum of the different scenes.

II. Model Structure

The network structure designed in this study is presented in Fig. 1. The segmented spectral features are used as the input of the model and learned in the CNN module and LSTM module, respectively. Then a fusion attention layer fuses different segment-level features. At last, a SoftMax classifier classifies different scenes. The details are introduced as follows.

A. Input design

In the related research of ASC, spectral features are usually used as the input features of CNN. Spectrum features are two-dimensional, and the two dimensions are the time dimension and frequency dimension, respectively. Modeling speech features as images can help CNN obtain rich scene details. We employ 128 Mel-filter banks for each audio file to obtain Mel-spectrum features, using Hamming windows with a frame size of 2048 samples and 1024 hop size. The spectrum features of different scenes are shown in Fig. 2, where the two axes represent time and frequency respectively.

Since the scene information in the recorded audio usually lasts for a short time, there may be a large amount of irrelevant information in the recorded audio. Hence, it is difficult for the model to fully obtain the scene information in the audio. In order to extract scene information effectively, each spectral feature is divided into ten segments of the same size on the time axis. Segment-level features are used as the input to help get short-term details.

B. Convolutional neural networks

The convolutional neural network we designed has three convolutional layers and one pooling layer. First, convolution helps the system extract target features. Convolution can be expressed as:

$$f(W^T x) = \sum_{i=1} W_i x_i + b, \tag{1}$$

where x represents the feature maps, W and b are weights and biases. The CNN module obtains time-frequency correlation information in spectral features in our system. In detail, our CNN module has three convolution layers and one pooling layer. The window size of each convolution layer is 5×5. The window step of the last convolution layer is $[3, 1]$, where 3 corresponds to the ordinate of the input tensor, 1 corresponds to the abscissa. The window moving step of all other convolution layers is 1. Each convolution layer is connected with a ReLu layer and a batch normalization (BN) layer. The size of the pooling window is [4, 4], and the step size of the pooling window is [2, 2]. Except that the core size of the first convolution layer is 64, the core size of all other convolution layers is 128. The parameter of the last fully connected layer is 128.

C. Long short-term memory

Compared with Recurrent Neural Network (RNN), LSTM has a more refined internal structure. The cooperation between different gate functions can effectively store and update context information. LSTM is mainly composed of five composite functions:

$$f_t = Sigmoid(W_f[h_{t-1}, x_t] + b_f), \tag{2}$$

$$i_t = Sigmoid(W_i[h_{t-1}, x_t] + b_i), \tag{3}$$

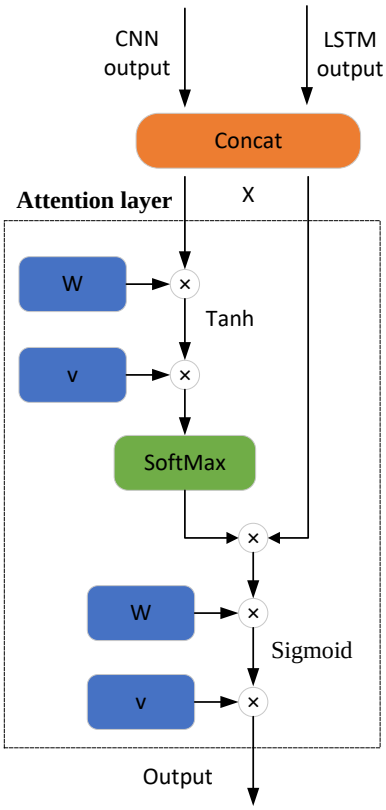

Fig. 3. The fusion attention layer.

$$O_t = Sigmoid(W_o[h_{t-1}, x_t] + b_o), \tag{4}$$

$$\widetilde{C}_t = Tanh(W_c[h_{t-1}, x_t] + b_C), \tag{5}$$

$$h_t = O_t + Tanh(f_t C_{t-1} + i_t \widetilde{C}_t), \tag{6}$$

Where W and b are weight matrix and offset, respectively. f_t. i_t and O_t are gate functions, C_t is cell activation. LSTM is less used in ASC than CNN. Because the audio recording scene information does not contain semantics and the context is not closely connected. Long-time recorded audio as the input of LSTM may affect the discrimination of the model. Therefore, segment-level features are used as the input of LSTM to reduce the interference of irrelevant parameters to the model. The number of internal units in LSTM is set to 128.

D. Fusion attention layer

The fusion attention layer contributes to the fusion of segmented high-level features. The fusion attention layer is presented in Fig. 3.

The outputs of CNN and LSTM are combined as the input $X = [x_1, x_2, ..x_n] \in \mathbb{R}^{n \times 256}$ of the fusion attention layer, where n is the number of segments of each audio segment. The attention probability of each feature α_i is first calculated:

$$s(x_i) = v^T \sigma(W x_i), \tag{7}$$

$$\alpha_i = \frac{exp(s(x_i))}{\sum_{j=1}^{N} exp(s(x_j))}, \tag{8}$$

TABLE I
PARAMETRIC SETUPS FOR THE PROPOSED CRN-FM MODEL.

Modules	Layer	Shapes
CNN	Input	$469 \times 128 \times 1$
	Conv	$5 \times 5 \times 64$
	Max Pooling	4×4
	Conv	$5 \times 5 \times 128$
	Conv	$5 \times 5 \times 128$
	GAP	
	Output	$n \times 128$
LSTM	Input	$469 \times 128 \times 1$
	# Hidden Units	128
	Output	$n \times 128$
FAL	Input	$n \times 256$
	Output	256
Dense	10	
Outputs	10	

where W and v are trainable weights, σ is the activation function *Tanh*. Calculating attention parameters between different segment-level features helps the model automatically select tensors with rich information. Then, calculate the weighted average of the input information: $h(X)$:

$$h(X) = \sum_{i=1}^{N} \alpha_i x_i. \tag{9}$$

The attention layer calculates the weighted average of different segment-level features. The attention calculation between segment-level features solves the uneven distribution of information, and then we need to consider the tensor fusion between different modules. We design a fusion layer to help the fusion of two high-level features:

$$fal(X) = \sigma(\sum_{i=1}^{N} \alpha_i x_i W) v^T, \tag{10}$$

where σ is the *Sigmoid* function, W, v are trainable parameters. The fusion attention layer uses attention parameter calculation and tensor fusion to improve the representation ability of features. Finally, a softmax classifier is cascaded for scene recognition.

III. EXPERIMENTS

A. Database and Training Setup

In the experiment, we used two data sets DCASE 2018 and DCASE 2019[7]. Both datasets were recorded in DCASE. DCASE 2018 is recorded in six major European cities, and DCASE 2019 expands to 12 major European cities. Both datasets contain 10 classes. For the recordings with a length of 10 seconds, DCASE 2018 has 8640 segments, and DCASE 2019 has 14400 segments. The ratio of training set to test set is 7:3.

momentum optimizer in training stage. The initial learning rate, batch size, and epochs are set to 0.01, 32, and 300 respectively. Considering the length of the original data, we set the segmentation number of spectral features to 10, equivalent to the audio input set to 1 second. The experimental evaluation standard is based on the classification accuracy recommended by DCASE. The proposed architecture is implemented using a Python platform with the TensorFlow framework. The elaborate setups of the parameters used in the model are shown in Table 1, “Conv” is the Convolution, ”FAL” is the Fusion Attention Layer, ”GAP” is the Global

TABLE II
THE AVERAGE ACCURACIES (%) OF DIFFERENT STRATEGIES ON THE TWO CORPORA.

Methods\Corpora	DCASE2018	DCASE2019
CNN	71.24	71.80
CNN(w/ SF)	70.09	70.80
LSTM	63.58	60.81
LSTM(w/ SF)	66.87	62.79
CRN-FM(w/o FAL)	71.52	72.52
CRN-FM	73.35	73.81

TABLE III
COMPARISON OF CLASSIFICATION ACCURACY (%) OF DCASE2018 AND DCASE2019 WITH OTHER WORK.

Methods\feature	DCASE 2018	DCASE 2019
Baseline	58.9	62.5
IITKGP ABSP Fusion18 [8]	66.2	/
ABCNN[9]	69.0	/
Fusion[10]	/	72.3
ASC by SFCC and DNN[11]	/	70.4
RW-CNN[12]	/	69.7
Atrous CNN[13]	72.7	/
Atrous CNN[14]	72.4	71.3
CNN	71.24	71.80
LSTM	63.58	60.81
CRN-FM	73.35	73.81

Average Pooling layer, and "n" represents the number of segments.

B. Experiment Results

To explore the performance of our proposed CRN-FM model, we set up several comparison models, including the model CNN with spectral features as input; CNN(w/ SF) model with segmented spectral features as model input; Model LSTM and model LSTM(w/ SF) with the same input settings; Model CRN-FM(w/o FAL) without fusion attention layer and our proposed model CRN-FM. The recognition rate in each experimental strategy is shown in Table 2.

Firstly, we compare the impact of using original audio and segmented audio as input on the system performance in the CNN model. We can see from Table 2 that in DCASE2018 and DCASE2019, the recognition rate of CNN(w/ SF) is lower than that of CNN. The recognition rates of the two corpora are reduced by 1.15% and 1%, respectively. Segment-level features as the input of the CNN model can not improve the performance of the system, which may be due to the local connection between CNN convolution layers, resulting in a low correlation of distant pixels. The dimensionality reduction operation in the convolution network eliminates the useless information of the tensor, so so the local information of the scene is retained in the global features. As the input, segment-level features can only learn local information, which may lead to misjudgment of the system.

Then we compare the impact of different input features on the performance of the LSTM model. The recognition rate of the LSTM model using segment-level features as input in the two databases has improved. Compared with the LSTM, the recognition rate of the LSTM(w/ SF) has increased by 3.29% and 1.98%, respectively, with significant performance improvement. The segment-level feature with a specific length can match the information required by the scene and exclude the influence of irrelevant factors in long-lasting audio on the model. The performance of the LSTM model is much lower than that of CNN. Because the scene audio does not contain semantics, it is more difficult for LSTM to obtain information than CNN.

Table 2 shows that the CRN-FM model we designed has achieved the best recognition effect. The recognition rates in the two databases are 73.35% and 73.81%, respectively, and the performance is higher than that of CRN-FM(w/o FAL). It is verified that the fusion attention layer plays a significant role in fusing different features.

Then, we compare the proposed model with the existing methods in Table 3. We employ DCASE Task 1 Baseline (Baseline). The baseline system implements a CNN-based approach, an attention-based atrous convolutional neural network (Atrous CNN). Then, we compare the proposed algorithm with the system submitted in DCASE, including "IITKGP ABSP Fusion18", "ABCNN", "Fusion", "ASC by SFCC and DNN", and "RW-CNN". Our proposed method has the best performance of the two databases. The performance of CRN-FM is significantly improved compared with the baseline system of the two datasets. The performance based on the two datasets is improved by 14.45% and 11.31%, respectively. In addition, the baseline system of DCASE task 1A adopts CNN as the framework, and the performance of the CNN module we designed is also better than the baseline. Furthermore, adding the LSTM module further improves the performance. [8] use the machine learning method, and the other comparative references use the deep learning method. From Table 3, the recognition performance of the methods based on traditional machine learning in ASC is poor, and the performance is lower than 67%. In the comparative experiment based on deep learning, only [11] used DNN, and the rest of the references used CNN. First, the performance of the CNN module we designed is higher than that of [11]. In addition, Most of the ASC-based deep learning models also use CNN, which can learn the time-frequency related information in the spectrum. Second, the performance of a single CNN module is limited, and the recognition effect of the CNN module we designed is lower than [10], [13-14]. On the other hand, the CRN-FM model has achieved the best performance. Compared with a single CNN network, CRN-FM supplements frame-level information, which means that our model can benefit from parallel structure and fusion attention layer.

IV. CONCLUSIONS

This letter proposes a convolutional-recurrent neural network with the tensor fusion mechanism for acoustic scene classification. Segment-level features are used as inputs, and the two neural networks learn together to fuse the outputs of different modules in the attention layer. This paper mainly has three contributions: first, segment-level feature matching the LSTM model improves the performance of the model. Then, CNN and LSTM form a parallel network to learn the time and time-frequency related information. Finally, the fusion attention layer helps to improve the fusion performance. Experiments verify the superiority of the proposed model.

REFERENCES

[1] V. Vivek, S. Vidhya, and P. Madhanmohan, "Acoustic scene classification in hearing aid using deep learning," in Proc. International Conference on Communication and Signal Processing (ICCSP). Chennai, India: IEEE, 2020, pp. 0695–069.

[2] S. Chu, S. S. Narayanan, C. Kuo, and M. J. Mataric, “Where am i? scene recognition for mobile robots using audio features,” in Proc. IEEE International Conference on Multimedia Expo (ICME), no. PP. Toronto, ON, Canada: IEEE, 2006, pp. 885–888.
[3] Abrol V , Sharma P . Learning Hierarchy Aware Embedding From Raw Audio for Acoustic Scene Classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, PP(99):1-1.
[4] Abeer J . A Review of Deep Learning Based Methods for Acoustic Scene Classification[J]. Applied Sciences, 2020, 10(6).
[5] Jing Xuan Yu, Kian Ming Lim, and Chin Poo Lee, ”MoVE-CNNs: Model aVeraging Ensemble of Convolutional Neural Networks for Facial Expression Recognition,” IAENG International Journal of Computer Science, vol. 48, no.3, pp519-523, 2021.
[6] Yuanbo Fang, Hongliang Fu, Huawei Tao, Xia Wang, and Li Zhao, ”Bidirectional LSTM with Multiple Input Multiple Fusion Strategy for Speech Emotion Recognition,” IAENG International Journal of Computer Science, vol. 48, no.3, pp613-618, 2021.
[7] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A multi-device dataset for urban acoustic scene classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), 9–13. November 2018. URL: https://arxiv.org/abs/1807.09840.
[8] Waldekar, Shefali and Saha, Goutam, ”Wavelet-Based Audio Features for Acoustic Scene Classification”, DCASE2018 Challenge, 2018.
[9] Ren, Zhao and Kong, Qiuqiang and Qian, Kun and Plumbley, Mark and Schuller, Björn, ”Attention-Based Convolutional Neural Networks for Acoustic Scene Classification”, DCASE2018 Challenge, 2018.
[10] Bilot, Valentin and Duong, Quang Khanh Ngoc, ”Acoustic Scene Classification with Multiple Instance Learning and Fusion”, DCASE2019 Challenge, 2019.
[11] Paseddula, Chandrasekhar and V.Gangashetty, Suryakanth, ”DCASE 2019 Task 1a: Acoustic Scene Classification by Sffcc and DNN”, DCASE2019 Challenge, 2019.
[12] Salvati, Daniele and Drioli, Carlo and Foresti, Gian Luca, ”Urban Acoustic Scene Classification Using Raw Waveform Convolutional Neural Networks”, DCASE2019 Challenge, 2019.
[13] Ren Z, Kong Q, J Han, et al. Attention-based Atrous Convolutional Neural Networks: Visualisation and Understanding Perspectives of Acoustic Scenes[C], IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
[14] Ren Z, Kong Q, Han J, et al. CAA-Net: Conditional Atrous CNNs with Attention for Explainable Device-robust Acoustic Scene Classification[J]. IEEE Transactions on Multimedia, 2020.