LETTER

An Integrated Convolutional Neural Network with a Fusion Attention Mechanism for Acoustic Scene Classification

Pengxu JIANG[†a)], *Nonmember*, Yue XIE[†], *Member*, Cairong ZOU[†], Li ZHAO[†], *and* Qingyun WANG[††], *Nonmembers*

SUMMARY In human-computer interaction, acoustic scene classification (ASC) is one of the relevant research domains. In real life, the recorded audio may include a lot of noise and quiet clips, making it hard for earlier ASC-based research to isolate the crucial scene information in sound. Furthermore, scene information may be scattered across numerous audio frames; hence, selecting scene-related frames is crucial for ASC. In this context, an integrated convolutional neural network with a fusion attention mechanism (ICNN-FA) is proposed for ASC. Firstly, segmented mel-spectrograms as the input of ICNN can assist the model in learning the short-term time-frequency correlation information. Then, the designed ICNN model is employed to learn these segment-level features. In addition, the proposed global attention layer may gather global information by integrating these segment features. Finally, the developed fusion attention layer is utilized to fuse all segment-level features while the classifier classifies various situations. Experimental findings using ASC datasets from DCASE 2018 and 2019 indicate the efficacy of the suggested method.

key words: *acoustic scene classification, ICNN-FA, CNN, attention mechanism, Mel-spectrograms*

1. Introduction

Sound offers diverse information about the surrounding environment, which can aid in machines' comprehension and perception of the world. Classifying a test recording into one of the specified acoustic scene classes is the objective of acoustic scene classification (ASC). Monitoring systems, personal archiving, robot navigation, and hearing aids are just a few of the numerous uses of ASC, which is a significant expanding field for identifying audio signals from an ambient backdrop.

Recent research on deep learning has provided ASC with deep models that outperform standard machine learning techniques. Principal deep learning models include Convolutional Neural Networks (CNNs) [1] and Long Short-Term Memory (LSTM) [2]. ASC techniques often employ CNN-based network topologies due to CNNs' superior ability to learn the abstract feature representation from spectrograms.

Sound data differs from picture data in that sound is generally sequential data of varying length, and sound often contains many silence periods and noisy, while the scene information may only be associated with a few frames. In addition, the recorded audio may comprise many segments, including information about the target scene. Thus, it is important to select scene-relevant frames for ASC. In recent years, several works [3], [4] have studied the effectiveness of the attention mechanism in ASC, attention method may score distinct frames.

In this letter, we propose an integrated convolutional neural network with a fusion attention mechanism (ICNN-FA) for ASC, using the network structure presented in Fig. 1. First, mel-spectrograms are extracted from the audio file. Since audio with a lengthy duration often involves audio segments with numerous label information, we split each Mel feature into a predetermined length as the model's input in order to make the constructed network more focused on these segment-level features with entire scene information. However, segment-level features as the network's input will cause the convolution network to lose global information, since the original input features contain complete time information, i.e. global information, whereas each segment of features after segmentation only contains local time information. Therefore, we propose a global attention layer in ICNN to assist the network in acquiring global auditory feature information. The fused attention layer is then utilized to combine segment-level information based on the distribution of attention weight and send them to the softmax classifier. The global attention layer combines time-related segmented data to produce the global attention parameter. The fusion attention layer incorporates all features by computing the attention weight distribution among segment-level features. The experiment validates the model's viability.

2. Model Structure

The network structure designed in this study is presented in Fig. 1. First, Mel-spectrograms are generated using sound as input. Then, these spectra are broken down into segment-level features of the same size as ICNN's input. The ICNN model captures high-level features from various segments, and the built fusion attention layer combines all extracted high-level features. Lastly, the SoftMax layer produces various scene probability values. Details are provided below.

2.1 Designed ICNN

CNNs are extensively used in ASC [5], [6]. In order to ac-

Manuscript received November 7, 2022.
Manuscript revised December 28, 2022.
Manuscript publicized February 6, 2023.

[†] The authors are with the School of Information Science and Engineering, Southeast University, P.R. China.
[††] The author is with School of Communication Engineering, Nanjing Institute of Technology, P.R. China.
a) E-mail: 230209051@seu.edu.cn

DOI: 10.1587/transfun.2022EAL2091

Copyright © 2023 The Institute of Electronics, Information and Communication Engineers

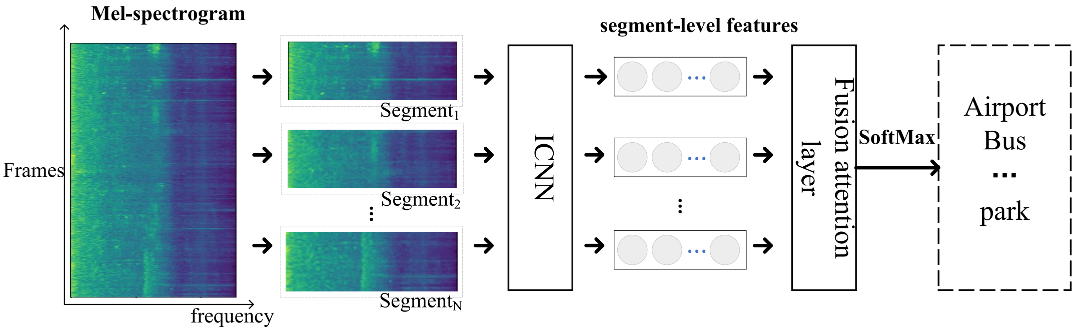


Fig. 1 Illustration of the proposed ICNN-FA architecture for ASC.

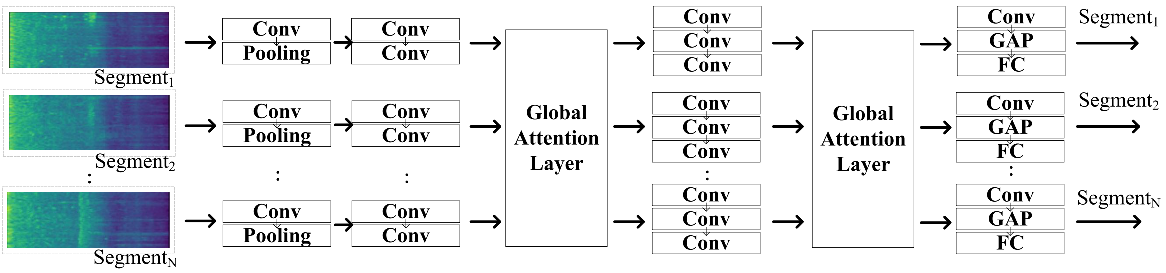


Fig. 2 Illustration of ICNN.

quire information on time-frequency correlation, we use a convolution network in our work. The CNN is shown in Fig. 2. The planned ICNN consists of seven convolution layers (Conv), one max pooling layer (Pooling), two global attention layers, one Global Average Pooling (GAP) layer, and one fully connected (FC) layer. The convolution layer is used to gather short-term time-frequency information, while the pooling layer is used to decrease parameter magnitude. The global attention layer is used to acquire global information, as detailed in the next section. 7 × 7 is the size of the first convolution layer, and the convolution window strides are 2. We only perform max-pooling after the first convolutional layer, the pooling size of 4 × 4, and strides of 2. In addition, the remaining convolution layers have a size of 5 × 5. Except for the convolution kernel of the first layer, which is 128, the kernels of the other convolution layers are 256.

2.2 Global Attention Layer

Due to the input of the ICNN involves segment-level features, the convolution layer may miss part of the crucial time-frequency information even if it may concentrate more on gaining short-term time-frequency information. In order to collect global information between segment-level features, we construct a global attention layer, motivated by the Convolutional Block Attention Module (CBAM) [7]. Figure 3 depicts the structure of the global attention layer. First, these segment-level features are joined along the frame axis to generate a global feature. The attention calculation module is then used to derive the parameters of attention from the global features.

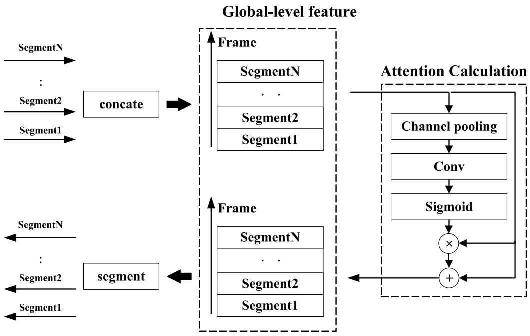


Fig. 3 Illustration of the global attention layer.

The primary components of the attention calculation module are a channel pooling layer, a convolution layer, and a sigmoid layer. The primary objective of the channel pooling layer is to minimize the input dimensions. The channel pooling layer uses the maximum pooling approach to limit the channel dimension of the input features to 1. The function of the convolution layer is to extract long-term time-frequency correlation information, where the movement step is set to 1, and the size of the convolution layer is 7 × 7. The sigmoid function is used to produce parameters for spatial attention. The global attention parameters are used to construct global features with a long-term time-frequency correlation. As the output of the global attention layer, these global features are subdivided once again into segment-level features.

2.3 Fusion Attention Layer

The ICNN model generates segment-level features that need

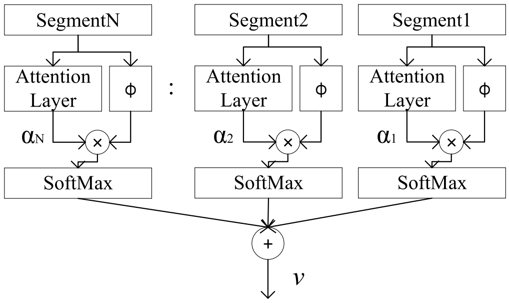


Fig. 4 Illustration of the fusion attention layer.

integration. Therefore, we developed a fusion attention layer to combine these time-related segment-level features. Figure 4 depicts the structure of the fusion attention layer. The input of the fusion attention layer may be written as:

$$\{segment_1, segment_2, \cdots, segment_N\} \in \mathbb{R}^{N\times C}, \tag{1}$$

where N is the number of feature segments, and C is the feature dimension of each segment. The attention layer is then used to calculate the weight distribution a_n between segment-level features:

$$u_n = v^T \sigma_t(segment_n W + b), \tag{2}$$

$$a_n = \text{SoftMax}(u_n), \tag{3}$$

where σ_t represents the $Tanh$ activation function, W and v^T are the weights and b denotes the bias of the attention layer. Then, a fully connected layer is employed to match the input dimensions to the number of outputs, and the SoftMax function is applied to each segment-weighted output:

$$p(segment_n) = \text{SoftMax}(\phi(segment_n)a_n), \tag{4}$$

where ϕ represents the fully connected layer. Then, all segment-level features are fused to obtain the classification probability of each feature for different scenes:

$$v = \sum_n p(segment_n), \tag{5}$$

v is the output tensor of each audio. Finally, a softmax classifier is cascaded for ASC.

3. Experiments

3.1 Datasets and Training Setup

We utilize two datasets of DCASE 2018 Challenge Task 1a and DCASE 2019 Challenge Task 1a [8] in the experiments to show the performance of our proposed model. Both contain 10 seconds of audio length with a sampling rate of 48 kHz from 10 classes, including airport, bus, metro, metro station, park, public square, shopping mall, street pedestrian, street traffic, and tram. The DCASE 2018 has 6122 files for training and 2518 for testing, and the DCASE 2019 has 9185 files for training and 4185 for testing.

Table 1 Comparison of the recognition rate (%) of different methods.

Methods	DCASE 2018	DCASE 2019
CNN(w/o sf)	74.78	75.91
CNN(w/ sf)	71.64	72.23
GA-CNN	74.78	74.95
F-CNN	76.44	76.51
ICNN-FA	77.48	77.80

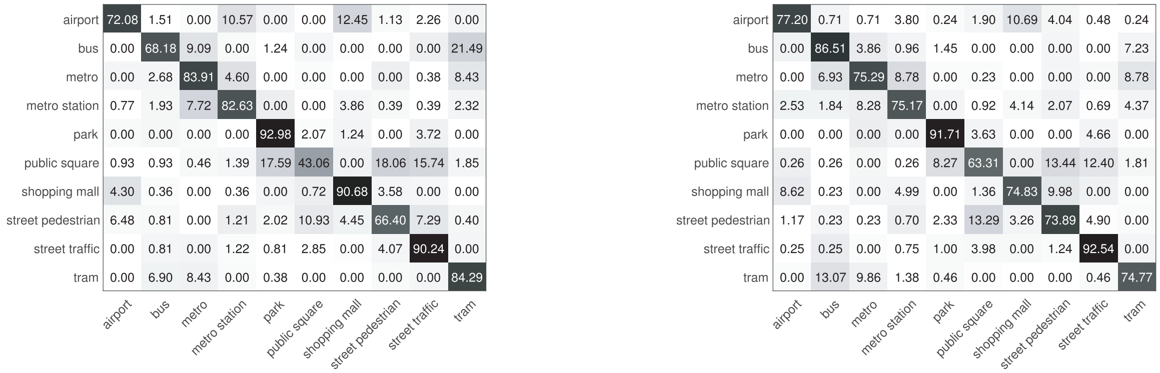
Mel-spectrograms are widely used and are the most effective features for many audio deep learning tasks. Therefore, we extracted the Mel-spectrograms of all audio samples from the above datasets. We employ 128 Mel-filter banks for each audio file to obtain Mel-spectrum features, using Hamming windows with a frame size of 2048 samples and 1024 hop size. The sampling frequency is set to 48 kHz.

Our experiments use a momentum optimizer and set its initial learning rate to 0.01 and the batch size to 128. The model's parameters are optimized by minimizing a cross-entropy objective function, while the maximum number of the epochs is set to 300. Since the audio duration provided by the DCASE dataset is 10 seconds, we divide each audio into five segments as the input of the model, that is, two seconds per segment. The proposed ICNN-FA architecture is implemented using a Python platform with the TensorFlow framework.

3.2 Experiment Results

The experimental strategies include CNN (w/o sf), CNN (w/ sf), GA-CNN, F-CNN, and ICNN-FA. CNN represents the convolution network without any attention mechanism. The input of CNN (w/ sf) is segment-level features, and the input of CNN (w/o sf) is non-segmented Mel-spectrograms features. GA-CNN and F-CNN are convolution networks with only global attention or fusion attention layers, respectively. GA-CNN and CNN (w/ sf) add all segment-level features in the final fusion stage. ICNN-FA contains all designed modules. The performance of each experimental strategy is shown in Table 1.

First, the test performance of CNN (w/o sf) on two datasets is greater than that of CNN (w/ sf): 74.78% and 75.91%, respectively. Segment-level features as input to CNN may not increase the model's performance. The greater recognition rate of GA-CNN compared to CNN (w/ sf) demonstrates that the acquisition of global information enhances the model's performance when segment-level features are used as input. However, the performance of GA-CNN remains inferior than that of CNN (w/o sf). This may be due to the fact that the global information gained in the global attention layer is not completely exploited in the feature fusion stage. F-CNN has a more significant recognition rate than CNN and GA-CNN, indicating that the weight computation of segment-level features in the fusion attention layer may enhance the model's performance. According to the Table 1, the performance of the proposed ICNN-FA is overwhelmingly impressive. Specifically, ICNN-FA



	airport	bus	metro	metro station	park	public square	shopping mall	street pedestrian	street traffic	tram
airport	72.08	1.51	0.00	10.57	0.00	0.00	12.45	1.13	2.26	0.00
bus	0.00	68.18	9.09	0.00	1.24	0.00	0.00	0.00	0.00	21.49
metro	0.00	2.68	83.91	4.60	0.00	0.00	0.00	0.00	0.38	8.43
metro station	0.77	1.93	7.72	82.63	0.00	0.00	3.86	0.39	0.39	2.32
park	0.00	0.00	0.00	0.00	92.98	2.07	1.24	0.00	3.72	0.00
public square	0.93	0.93	0.46	1.39	17.59	43.06	0.00	18.06	15.74	1.85
shopping mall	4.30	0.36	0.00	0.36	0.00	0.72	90.68	3.58	0.00	0.00
street pedestrian	6.48	0.81	0.00	1.21	2.02	10.93	4.45	66.40	7.29	0.40
street traffic	0.00	0.81	0.00	1.22	0.81	2.85	0.00	4.07	90.24	0.00
tram	0.00	6.90	8.43	0.00	0.38	0.00	0.00	0.00	0.00	84.29

(a)

	airport	bus	metro	metro station	park	public square	shopping mall	street pedestrian	street traffic	tram
airport	77.20	0.71	0.71	3.80	0.24	1.90	10.69	4.04	0.48	0.24
bus	0.00	86.51	3.86	0.96	1.45	0.00	0.00	0.00	0.00	7.23
metro	0.00	6.93	75.29	8.78	0.00	0.23	0.00	0.00	0.00	8.78
metro station	2.53	1.84	8.28	75.17	0.00	0.92	4.14	2.07	0.69	4.37
park	0.00	0.00	0.00	0.00	91.71	3.63	0.00	0.00	4.66	0.00
public square	0.26	0.26	0.00	0.26	8.27	63.31	0.00	13.44	12.40	1.81
shopping mall	8.62	0.23	0.00	4.99	0.00	1.36	74.83	9.98	0.00	0.00
street pedestrian	1.17	0.23	0.23	0.70	2.33	13.29	3.26	73.89	4.90	0.00
street traffic	0.25	0.25	0.00	0.75	1.00	3.98	0.00	1.24	92.54	0.00
tram	0.00	13.07	9.86	1.38	0.46	0.00	0.00	0.00	0.46	74.77

(b)

Fig. 5 (a) Confusion matrices of the results on DCASE 2018. (b) Confusion matrices of the results on DCASE 2019.

Table 2 Comparison of classification accuracy (%) of DCASE2018 and DCASE2019 with other work.

Methods	DCASE 2018	DCASE 2019
Atrous CNN [9]	72.70	/
MCTA-CNN [10]	72.40	75.71
HRAN-ASM [11]	70.50	/
scalogram-LMS [12]	/	76.70
SubSpectralNet [13]	74.08	73.44
wavelet [14]	66.20	/
ICNN-FA	77.48	77.80

outperformed CNN(w/o sf) and CNN(w sf) on both datasets by at least 2.7% and 5.84%, respectively. The approach of merging and segmenting various segment-level features in the model is viable, suggesting that the global attention and fusion attention layer plays a crucial role in information integration and fusion.

We present here the experimental comparison between the proposed model and existing approaches. Table 2 shows the performance of our proposed model and other state-of-the-art methods, including some recent models on the DCASE2018 and DCASE2019 datasets. Atrous Convolutional Neural Networks with global attention pooling (Atrous CNN) [9], an effective convolutional neural network structure with a multi-channel temporal attention block (MCTA-CNN) [10], high-resolution attention network with an acoustic segment model (HRAN-ASM) [11], a feature decomposition method based on temporal median filtering (scalogram-LMS) [12], an approach of using spectrograms in Convolutional Neural Networks (SubSpectralNet) [13], and wavelet-based audio features for acoustic scene classification (wavelet) [14]. [14] employing the machine learning approach, the performance of the deep learning model we developed outperforms the machine learning method by a wide margin. In addition, our ICNN-FA model outperforms other CNN-based models, which implies that our model structure based on attention mechanisms can significantly improve the performance of the ASC system.

Figure 5 represents the confusion matrices used to evaluate the performance of our suggested model. According to the data, for the proposed ICNN-FA, park and street traffic often correlate to high levels of accuracy, but public square, street pedestrian, and airport do not. One possible explanation is that these surroundings are noisy.

4. Conclusion

This paper described an integrated convolutional neural network with a fusion attention mechanism (ICNN-FA) for ASC. This model included a convolution network capable of splitting and combining tensors. Multiple attention methods utilizing global and fusion attention layers were presented for the ICNN-FA model. Compared to other state-of-the-art methodologies, the experimental findings on the two databases show that our model may effectively enhance the performance of an ASC system.

Acknowledgments

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

References

[1] H. Chen, Z. Liu, Z. Liu, and P. Zhang, "Long-term scalogram integrated with an iterative data augmentation scheme for acoustic scene classification," J. Acoust. Soc. Am., vol.149, no.6, pp.4198–4213, 2021.

[2] M.M. Morgan, I. Bhattacharya, R.J. Radke, and J. Braasch, "Classifying the emotional speech content of participants in group meetings using convolutional long short-term memory network," J. Acoust. Soc. Am., vol.149, no.2, pp.885–894, 2021.

[3] H. Liang and Y. Ma, "Acoustic scene classification using attention-based convolutional neural network," Technical Report, DCASE2019 Challenge, 2019.

[4] J.-W. Jung, H.-S. Heo, H.-J. Shim, and H.-J. Yu, "DNN based multi-level features ensemble for acoustic scene classification," Technical Report, DCASE2018 Challenge, 2018.

[5] M.D. McDonnell and W. Gao, "Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.141–145, 2020.

[6] T. Nguyen, F. Pernkopf, and M. Kosmider, "Acoustic scene classification for mismatched recording devices using heated-up softmax and spectrum correction," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.126–130, 2020.

[7] S. Woo, J. Park, J.Y. Lee, and I.S. Kweon, "CBAM: Convolutional block attention module," Proc. European Conference on Computer Vision (ECCV), Munich, Germany, pp.3–19, 2018.

[8] T. Heittola, A. Mesaros, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," Technical Report, DCASE2018 Challenge, 2018.

[9] Z. Ren, Q. Kong, J. Han, M.D. Plumbley, and B.W. Schuller, "CAA-Net: Conditional atrous CNNS with attention for explainable device-robust acoustic scene classification," IEEE Trans. Multimedia, vol.23, pp.4131–4142, 2021.

[10] Y. Wang, C. Feng, and D.V. Anderson, "A multi-channel temporal attention convolutional neural network model for environmental sound classification," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.930–934, 2021.

[11] X. Bai, J. Du, J. Pan, H.-S. Zhou, Y.-H. Tu, and C.-H. Lee, "High-resolution attention network with acoustic segment model for acoustic scene classification," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.656–660, 2020.

[12] Y. Wu and T. Lee, "Time-frequency feature decomposition based on sound duration for acoustic scene classification," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.716–720, 2020.

[13] S. Phaye, E. Benetos, and Y. Wang, "SubSpectralNet — Using sub-spectrogram based convolutional neural networks for acoustic scene classification," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.825–829, 2019.

[14] S. Waldekar and G. Saha, "Wavelet-based audio features for acoustic scene classification," Technical Report, DCASE2018 Challenge, 2018.