IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON INTELLIGENT INFORMATION SERVICES

Received June 7, 2020, accepted July 1, 2020, date of publication July 8, 2020, date of current version July 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007906

Environment Sound Event Classification With a Two-Stream Convolutional Neural Network

XIFENG DONG[1], BO YIN[1,2], YANPING CONG[1], ZEHUA DU[1], AND XIANQING HUANG[1]
[1]School of Information Science and Engineering, Ocean University of China, Qingdao 266100, China
[2]Pilot National Laboratory for Marine Science and Technology, Qingdao 266237, China

Corresponding author: Bo Yin (ybfirst@126.com)

This work was supported in part by the Key Research and Development Projects of Shandong Province under Grant 2019JMRH0109, and in part by the National Natural Science Foundation of China under Grant 61972367.

ABSTRACT In recent years, with the construction of intelligent cities, the importance of environmental sound classification (ESC) research has become increasingly prominent. However, due to the non-stationary nature of environment sound and the strong interference of ambient noise, the recognition accuracy of ESC is not high enough. Even with deep learning methods, it is difficult to fully extract features from models with a single input. Aiming to improve the recognition accuracy of ESC, this paper proposes a two-stream convolutional neural network (CNN) based on raw audio CNN (RACNN) and logmel CNN (LMCNN). In this method, a pre-emphasis module is first constructed to deal with raw audio signal. The processed audio data and logmel data are imported into RACNN and LMCNN, respectively to obtain both of time and frequency features of audio. In addition, a random-padding method is proposed to patch shorter data sequences. In such a way, the available data for experiment are greatly increased. Finally, the effectiveness of the methods has been verified based on UrbanSound8K dataset in experimental part.

INDEX TERMS Environmental sound classification, sound recognition, convolutional neural networks, data processing, pre-emphasis, two stream model.

I. INTRODUCTION

Speech recognition technology, as one of the representatives of the new generation of information technology, has become more and more mature. As one of the branches, the accuracy of speech classification and music classification has reached a considerable level, even exceeding the ability of human auditory perception [1], [2]. However, as another branch of speech recognition, environmental sound classification (ESC) still faces many difficulties in various aspects, such as non-stationary nature of environment sound and the strong interference of ambient noise [3]. On the other side, ESC research has an effect on the construction of smart cities [4]. For example, it could be used to automatically identify the specific types of sound in environment, such as children crying [5], animal sound [6] and siren [7]. Hence, it has caught lots of research attentions.

In early studies, research objects of ESC are mainly features extracted manually, such as Mel-frequency cepstral coefficient (MFCC), linear predicted cepstral coefficient (LPCC), short-term energy and zero-crossing rate [8]. These features are then classified by machine learning methods, such as Support Vector Machine (SVM), k-nearest neighbors (k-NN) and Gaussian mixture model (GMM) [9]–[11].

As deep learning method has been adopted in more and more fields, it is also introduced in ESC research. In this paper, a two-stream convolutional neural network model (CNN) is proposed based on deep learning to improve accuracy of ESC. In this method, both time domain and frequency domain features of audio signal are introduced as input signal, and a pre-emphasis module is constructed at input layer to improve signal-to-noise ratio (SNR). In addition, in terms of data pre-processing, this paper proposes a random-padding method to patch shorter data sequences.

The structure of this paper is arranged as follows. A brief introduction of related work of ESC based on CNN is given in Section II. The two-steam CNN method and the data preprocessing method called random-padding are elaborated in Section III. The detailed experimental process and results based on UrbanSound8K dataset are shown in Section IV.

II. RELATED WORK

Deep learning has been widely used in various fields, some researchers have also begun to introduce this technology into

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

research on ESC [11]. As a feedforward neural network with convolutional computations and deep structures, CNN have been used in image recognition research [12]. In recent years, CNN has been frequently used in ESC research [13]–[15]. And the studies can be divided into three categories.

In first category method, the network is trained by raw audio signal. Dai *et al.* [16] proposed a 1D-CNN with 34 weight layers. Compared with shallow neural networks, deeper networks can achieve better results due to the expansion of the receptive field. Abdoli *et al.* [17] proposed an end-to-end approach for ESC based on a 1D-CNN. The advantage of this method is that the process to manually extract features is cancelled. However, 1D-CNN extracts features at the global level without considering the temporal structure and frequency feature of environmental sounds [3]. In second category method, the network is trained by features extracted from raw signal, such as spectrogram and MFCC. In lots of studies, MFCC is used as input data to train classification model. However, due to discrete cosine transform (DCT), adopted to extract coefficient features in MFCC, will lead to a lack of structural information of audio signal, MFCC does not perform well for deep learning models [15]–[17]. On the contrary, logmel-CNN (LMCNN) model adopted logmel spectrogram feature is used well. Piczak [18] proposed a type of CNN model with logmel feature extracted from raw audio signal. Zhang *et al.* [15] used Mixup method combined logmel and gammatone spectrogram features to improve classification performance. In third category method, the network is trained by multiple input data. Tran and Tsai [7] used raw waveform and a combined feature formed by MFCC and logmel spectrogram as input data and proposed a SirenNet for siren-sound-based emergency vehicle detection. Li *et al.* [19] proposed an ensemble model, in which RawNet and MelNet are used individually. And the Dempster-Shafer (DS) method was then adopted to combine the training results. Su *et al.* [20] proposed a TSCNN-DS model and the performance of the model is pretty good on the UrbanSound8K dataset, but the input data of the model is too complex, which requires many features to be combined and the two stream network is fused by DS evidence theory. Furthermore [21] described a multi-stream CNN with temporal attention and decision fusion for ESC. However, the multi-stream CNN not only has complex structure, but also combines the original signal and the short-time Fourier transform, which leads to a large amount of data and requires a high level of hardware. Hence, the advantage of this method is that it can combine time domain and frequency domain features of audio signal, thereby compensating for the shortcomings of the single-input model. It should be mentioned that the method proposed in this paper belongs to the third category.

III. METHOD

The deep learning models represented by the CNN and Long Short-Term Memory (LSTM) have been widely used in the field of audio processing [22]–[24]. However, we only chose the CNN to construct the basic model of ESC, because the CNN has a many advantages over LSTM in ESC tasks. First, the ESC task emphasizes the types of sounds in the current environment and does not need to pay special attention to the sounds in the past period, the most obvious advantage of LSTM technology is not applicable here. Second, the CNN can obtain the time-frequency characteristics of sound signals by using data such as the acoustic spectrum as the input, which is difficult for LSTM to achieve.

TABLE 1. The recognition accuracy comparison of CNN based on different features.

Feature	Accuracy
Waveform	56.93%
CQT	82.21%
MFCC	93.71%
logmel	94.10%

It has been demonstrated that a model combining types of features of data has better performance [19]–[21]. Hence, a two-stream CNN method combining RACNN and LMCNN is proposed in this paper. In such a way, both the time domain and the frequency domain features of signal are considered. In addition, a random-padding method is also proposed to solve the inconsistent sample length problem of UrbanSound8K dataset.

A. SELECTING THE APPROPRIATE INPUT DATA

Compared with speech, environment sound event (ESE) is a kind of background sound, which is often mixed with various background noises, so it is more difficult to be identified. As known, the MFCC is used to solve automatic speech recognition (ASR) problem, and it do has a better performance for artificial audio recognition [25]. So, we decided to introduce MFCC to deal with ESC problem. Further, to solve the coordination problem between MFCC and deep learning network, a conversion step is needed to transform MFCC into logmel. In addition, a simple comparison experiment is carried out with the logmel and other popular audio features. And the experimental results are shown in Table 1. The waveform is the raw audio wave saved as a greyscale picture, and the constant Q transform (CQT) is a time-frequency transform algorithm often used in music signals. Obviously, the logmel is superior to other common feature algorithms.

Obviously, we can not only rely on the logmel feature for learning but also need to use other information to make up for the deficiency of the logmel feature information, which may be necessary for further development of the ESC task. The fast Fourier transform (FFT) is used in the process of extracting logmel features. In it, the time-domain signal is converted into a frequency-domain signal. Therefore, the logmel feature inevitably lacks relevant important features such as the time domain. The most direct method is to analyse the original signal. Finally, we use a two-stream CNN with the raw audio signal and logmel as the input data.

Algorithm 1 Random-Padding

Input: y_n:raw audio; $freq$:sampling frequency;
Output: Y_{pad}:after the random-padding of the raw audio;

$n \leftarrow len(y_n)$; $n_{all} \leftarrow 4 * freq$;
$n_{lack} \leftarrow n_{all} - y_n$; $t \leftarrow y_n/freq$;
if $t > 0$ **and** $t \leq 2$ **then**
    $k \leftarrow ceil(n_{lack}/n)$;
    $Y_{pad} \leftarrow copy\ the\ y_n\ k\ times$;
return $Y_{pad}[: n_{all}]$
else
    $point \leftarrow random.choice(1 : (y_n - n_{lack}))$;
    $Y_{pad} \leftarrow y_n[point : point + n_{lack}]$;
return Y_{pad}
end if

B. RANDOM-PADDING METHOD

As a public dataset, the UrbanSound8K [26] is commonly used in ESC researches. This dataset contains 8732 labelled sound excerpts (< = 4s) of urban sounds from 10 classes. 1798 of them are less than 4s, accounting for 20.59% of the total number of samples. It is a large waste to directly exclude these samples when preprocessing the data. In fact, there are already some methods, such as cubic spline interpolation, zero-padding, are designed to patch data. However, the duration of some samples are less than 1s or even less than 0.2s. Obviously, cubic spline interpolation is not applicable to such samples. Hence, a simple and effective data patching strategy called random-padding method is proposed, described as follows. (Its pseudocode is shown in Algorithm 1):

1. For the sample duration, where $0 < t \leq 2s$, it copies the entire sample to patch sample data until the sample length reaches 4 seconds. The copied sound will eventually be truncated at a random point.

2. For the sample duration, where $2s < t < 4s$, it selects a random data segment which can patch the duration of original simple data to make it reach 4 seconds at once.

There is a random point cut off in the sound in the above two padding situations, so we call this method random-padding. The advantage of this method is that it simultaneously retains 20.59% of the sample, and ensures the timing of the completed data. Figure 1. is the comparison diagram between the zero-padding and the random-padding: (a) is the raw data diagram after the zero-padding method, (b) is the raw data diagram after the random-padding method, (c) is the logmel diagram after the zero-padding method and (d) is the logmel diagram after the random-padding method.

C. PRE-EMPHASIS MODULE

Pre-emphasis is a widely used method for audio preprocessing [28]. The formula for the pre-emphasis is represented as $y(n) = x(n) - \alpha * x(n-1)$, where x is the original signal, y is the signal after pre-emphasis, and α is the pre-emphasis coefficient ($0.9 < \alpha < 1$). This paper sets

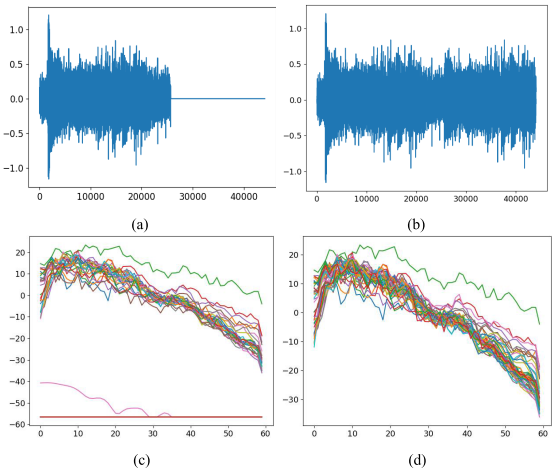


FIGURE 1. The comparison of the input data between zero-padding and random-padding. (a) and (c) are the raw data and logmel graph after zero-padding respectively, and (b) and (d) are the raw data and logmel graph after random-padding, respectively.

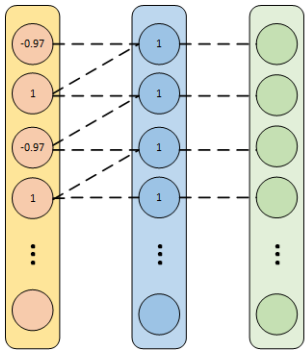


FIGURE 2. Pre-emphasis module. The raw audio data pass through the first two convolution layers that are initialized with weights. These two layers jointly constitute the pre-emphasis module and participate in the tuning with the whole network.

α to 0.97. In fact, there is no uniform selection criteria for this coefficient. In order to compare the effect among the proposed method and the other methods, the value of α is set according to [27], [28]. As described in many studies, 0.94 and 0.97 are commonly used. After a simple comparison experiment, 0.97 is selected for α according to the results, shown in Table 2 and the coefficient of 0.97 is better than 0.94 in the CNN+BN. In [27], a pre-emphasis layer is proposed for speaker verification. And in this paper, the pre-emphasis layer is also be used in the neural network. The first part of the proposed pre-emphasis module is a convolutional layer with a kernel length of 2, and the initial weights of the layer are set to -0.97 and 1. Compared to pre-emphasis layer, a convolutional layer with a kernel length of 2 and initial weight of 1 is added, as shown in Figure 2, the two convolutional layers together form a pre-emphasis module. The pre-emphasis module will participate in the tuning of the entire network as the first part of the hidden layer in the RACNN.

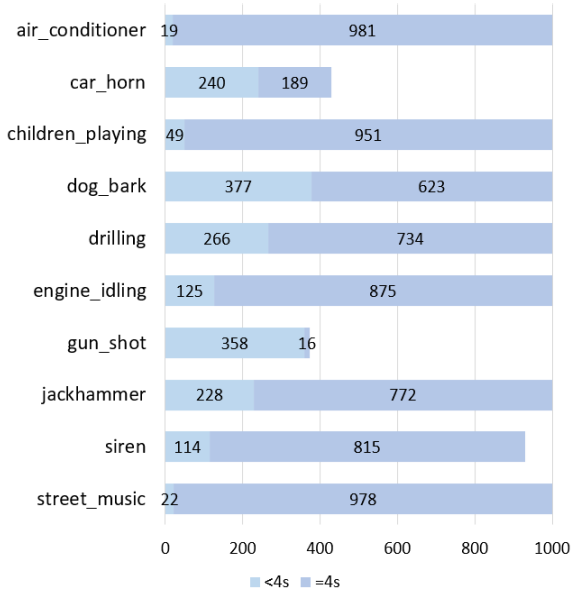


FIGURE 3. All samples of UrbanSound8K dataset (the light blue bars represent the samples <4s, the dark blue bars represent the samples = 4s).

TABLE 2. Initial weights (pre-emphasis coefficient) comparison of the pre-emphasis module.

Coefficient	1D-CNN	1D-CNN+BN
0.94	84.24%	88.20%
0.97	84.00%	89.14%

D. TWO-STREAM CNN

The network structure of the method proposed in this paper (Figure 4) is a two-stream CNN combining the RACNN and LMCNN. In the RACNN part, the input data are the raw audio signal. After the pre-emphasis module, the kernel length of the first layer convolution is set to 60 for corresponding to the logmel dimension of the sound signal. In addition, in order to increase the receptive field of the RACNN, 8 convolutional layers with kernel sizes of 3 and strides of 1 are added. The convolution layers and pooling layers are combined in a manner similar to the VGG, and a total of 11 layers are used in the RACNN. In the LMCNN part, 4 convolutional layers and a pooling layer are set after each layer convolution. The input data is logmel matrix, and the number of convolution kernels is sequentially increased, with a kernel size is 3*3 and stride is 1. And the fully connected layer is used to unify the format of network feature vector of two-stream CNN. Finally, the model uses ''addition'' to fuse the two stream feature maps to the ''softmax'' classifier to output the classification result. In addition, batch normalization (BN) and global average pooling are used in front of the fully connected layer of each stream CNN to reduce the number of parameters, which are not marked in the figure.

IV. EXPERIMENT

This part focuses on the analysis of the experimental results of the random-padding method, the pre-emphasis module and the two stream CNN model. Figure 5 shows the final flowchart of the experiment. Due to hardware limitations, we down-sampled the raw audio data are firstly down-sampled and then used to generate corresponding one-dimensional audio data and artificial logmel features by random-padding strategy. The one-dimensional audio data is input data for RACNN in combination with the pre-emphasis module, and the logmel feature is input data for LMCNN. In such a way, the fully connected layers of the two are added to output the classification result.

TABLE 3. Comparison of random-padding and zero-padding among different models.

Model	Baseline	Sample size	Zero-padding	Random-padding
1D-CNN	79.70%	1708	76.14%	78.21%
		8732	81.00%	82.62%
1D-CNN+BN	88.11%	1708	88.94%	88.60%
		8732	89.38%	89.42%
2D-CNN	94.04%	1708	91.29%	91.90%
		8732	92.63%	92.84%
2D-CNN+BN	94.79%	1708	89.49%	92.01%
		8732	94.10%	94.39%

TABLE 4. Comparison of two pre-emphasis methods.

Dateset	Method	1D-CNN	1D-CNN+BN
ESC10	Pre-emp layer	71.0%	76.75%
	This paper	72.5%	77.75%
UrbanSound8K	Pre-emp layer	84.30%	88.01%
	This paper	85.11%	89.14%

A. DATA PRE-PROCESSING

The hardware platform in this paper uses an Intel Core i5 9400F CPU, an NVIDIA GTX 1660 GPU and 16GB of RAM, and Keras2.2 as the development environment. The UrbanSound8K dataset is divided into 10 sound classes, which are air conditioner (AC), car horn (CH), children playing (CP), dog bark (DB), drilling (Dr), engine idling (EI), gunshot (GS), jackhammer (Ja), siren (Si) and street music (SM). The Librosa audio processing library was used to read the original sample with a sample rate of 11025 Hz, and then UrbanSound8K dataset was converted into 8732 sampled data with a length of 4s with a total of approximately 9.7 h via the random-padding method. The number of channels of the logmel spectrogram is 60, the length of the FFT window is 2048, and the frame shift is 1024. The final extracted logmel matrix size is 60*44.

B. EXPERIMENT AND RESULTS

1) COMPARISON BETWEEN RANDOM-PADDING AND ZERO-PADDING

To verify the availability of the random-padding method proposed in this paper, the experiment will be compared

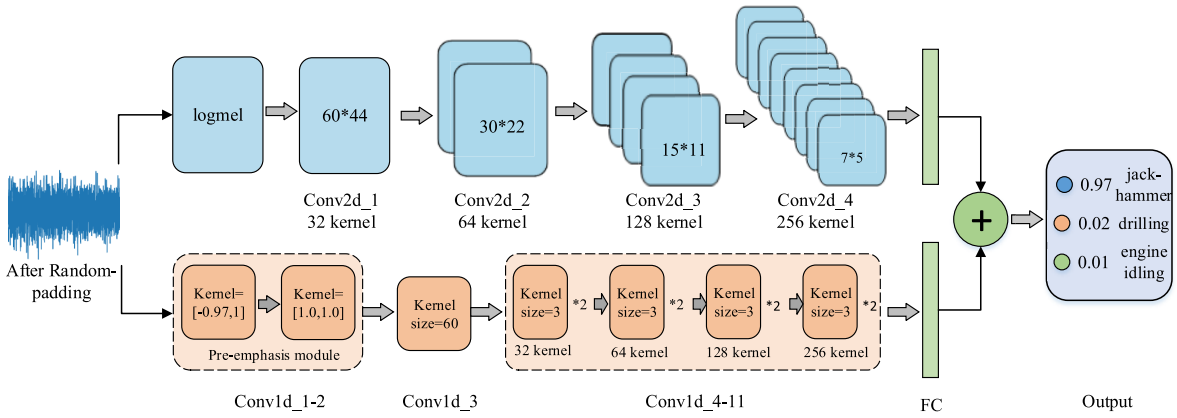


FIGURE 4. The model of a two-stream CNN.

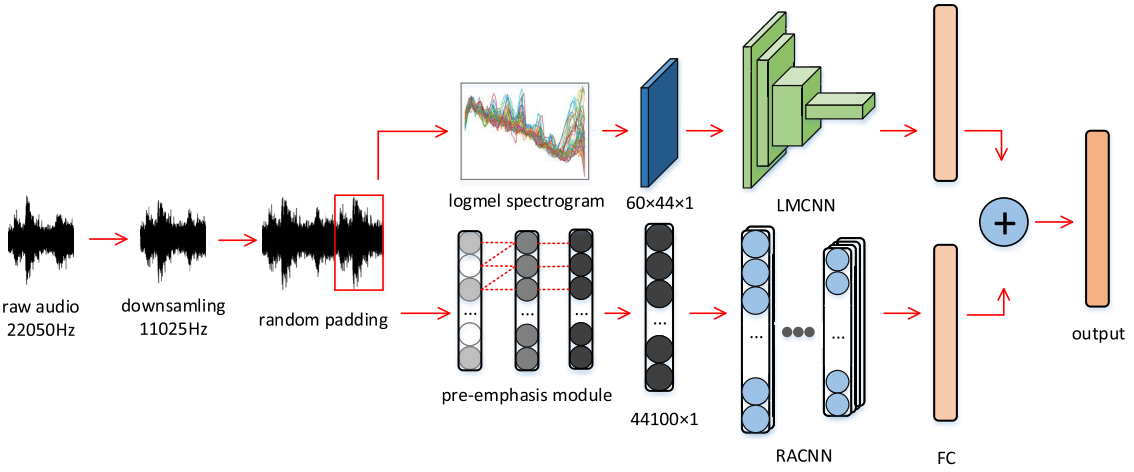


FIGURE 5. Diagram for the whole design and signal processing.

TABLE 5. Ablation experiments of the pre-emphasis modules (TS-CNN is a two-stream CNN without a pre-emphasis module).

Dateset	TS-CNN	This paper
ESC10	86.50%	87.25%
UrbanSoun8K	94.39%	94.97%

TABLE 6. Comparison between the proposed two-stream CNN and other models.

Model	Accuracy
PiczakCNN [18]	72.7%
Envnet-v2 [25]	78.3%
MixupCNN [15]	83.7%
DS-CNN [19]	92.2%
GoogLeNet [4]	93%
Proposed	95.7%
Proposed (best)	96.07%

to the most commonly used zero-padding method. In this experiment, the ratio of the training set to the test set is close to 8:2. To further prove the random-padding method, we also used the samples less than 4s as an independent dataset for experiments. It should be noted here that the samples less than 4s are extremely unevenly distributed in each category (the sample size is shown in the Figure 3). To avoid the impact of sample imbalance as much as possible, we have not selected the three categories air conditioner, children playing and street music, and the remaining 7 categories have a total of 1709 samples. Table 3 shows the comparison results. The 1D-CNN is a simple 7-layer one dimensional CNN based on the raw signal, and the 2D-CNN is a 4-layer two dimensional CNN based on the logmel. It can be seen that the accuracy of the random-padding method proposed in this paper is improved to different degrees compared with the zero-padding method. In particular, the random-padding method results in a large improvement in the 1D-CNN, but in the 2D-CNN, the accuracy improvement is not as obvious. The reason may be that the use of the zero-padding method destroys the time-order character of the signal and the 1D-CNN is sensitive to it. Therefore, random-padding results in an obvious improvement for the 1D-CNN, especially in the 1D-CNN model on a less than 4s dataset, which is approximately 2.07% better than the zero-padding and that's enough to see that our method works.

TABLE 7. The recognition accuracy of each category and the analysis of their audio features. The values for these features come from the average of all samples equal to 4s.

	Air condi-tioner	Car horn	Children playing	Dog bark	Drilling	Engine idling	Gun shot	Jack hammer	Siren	Street music
Zero-crossing	0.125	0.259	0.182	0.158	0.326	0.101	0.166	0.237	0.178	0.140
Auto-correlate	0.014	0.007	0.006	0.009	0.006	0.015	2.452	0.004	0.166	0.009
RMS	1.595	3.03	0.914	4.437	2.299	5.660	13.278	1.604	3.020	3.165
Accuracy	0.972	0.946	0.946	0.928	0.957	0.977	0.981	0.964	0.978	0.930

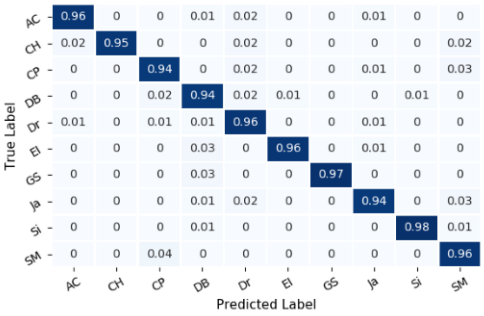


FIGURE 6. Confusion matrix of the proposed two-stream CNN on the UrbanSound8K dataset.

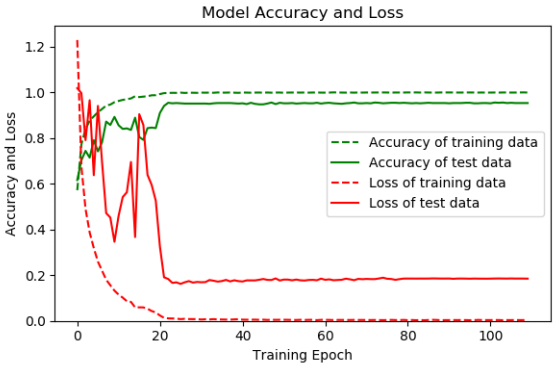


FIGURE 7. The training curves of the accuracy and loss of the proposed two-stream CNN on the UrbanSound8K dataset.

2) PRE-EMPHASIS MODULE EMDEDDED IN THE CNN

In this section, the pre-emphasis module is compared with the pre-emphasis layer proposed in [27], and the ESC10 dataset is also used in the experiment. Both the ESC10 and UrbanSound8K dataseta have 10 classes, but the sample size is much smaller in the former than in the latter, at only 400. The division of the training set and test set is the same as in the previous section. The comparative results of the experiment are shown in Table 4. It can be seen that the pre-emphasis module proposed in this paper is superior to the pre-emphasis layer in the ESC10 and UrbanSound8K datasets. This shows that our work can improve the performance of the model. Adding a convolutional layer with a kernel length of 1 and an initial value of 1 can better regulate the network than only one pre-emphasis layer. Meanwhile, the parameters of the network model hardly increase. Finally, ablation experiments were conducted on the pre-emphasis module and the performance of the model after the introduction of the pre-emphasis module was improved on both datasets (Table 5).

3) EXPERIMENT OF THE WHOLE NETWORK MODEL

In this part the performance of two-stream CNN is verified. The initial learning rate is set to 0.01, and the learning rate attenuation strategy is adopted. The learning rate is reduced to 0.1 times what it was in the 20th and 80th epochs, and the learning rate is reduced to 0.5 times what it was in the 50th epoch. A total of 110 epochs are conducted. The optimization function adopts the stochastic gradient descent method with momentum of 0.9. In this experiment, the 10-fold cross-validation method was used. First, the 8732 sample data were scrambled, and then divided into nine folds with 875 samples and one fold with 857 samples. The mean value of the 10-fold cross-validation is 95.7% and the the accuracy of the optimal model reached 96.07%. Table 6 compares the two-stream CNN model to other models on the UrbanSound8K dataset. It can be seen that the model proposed in this paper is superior to the model proposed by most other studies due to the higher recognition accuracy compared to Boddapati *et al.* [4] using the Spectrogram, MFCC and CRP combined features on GoogLeNet. The following uses typical experimental data for analysis. As seen from Table 7, gunshot, engine idling and siren sounds have better performance, and their recognition accuracy is over 97.5%, while children playing and street music had lower recognition rates. Here, we can roughly determine that sound pairs with larger autocorrelation coefficients and RMSs (root mean squares) can achieve better recognition accuracy. It means that the environment sound events, such as gun shot, siren and engine idling, are more recognizable due to the characteristics of obvious periodicity and large amplitude fluctuation. Figures 6 and 7 show a typical confusion matrix and training curve, respectively. It can be seen that due to the large learning rate used in the first 20 epochs, the loss keeps oscillating and decreases rapidly. After the 20th epoch, the model reduces the learning rate, becomes stable and seeks the optimal solution.

V. CONCLUSION

In this paper, a two-stream CNN model is proposed, which combines the RACNN and LMCNN. The two stream CNN uses the raw audio data and logmel matrix as input data, respectively. In such a way, the time-frequency characteristics of environment sound signals can be fully extracted. In terms of data preprocessing, this paper proposes a random-padding method to patch the uneven data samples. Hence, the available data for experiment are greatly increased. According to the comparison results with the zero-padding method, the advantage of random-padding method is confirmed. In terms of network structure, the pre-emphasis module is added to the convolution part of the RACNN. Hence, the network can be improved due to a better SNR of the signal. Finally, according to the experiment results, a high recognition accuracy of 95.7% is achieved based on the 10-fold UrbanSound8K dataset. In our future work, other newly developed models, e.g. [24] and [29], will be considered to explore the better recognation method in the field of ESC.

REFERENCES

[1] Y. Chen, Q. Guo, X. Liang, J. Wang, and Y. Qian, "Environmental sound classification with dilated convolutions," *Appl. Acoust.*, vol. 148, pp. 123–132, May 2019, doi: 10.1016/j.apacoust.2018.12.019.

[2] F. Demir, D. A. Abdullah, and A. Sengur, "A new deep CNN model for environmental sound classification," *IEEE Access*, vol. 8, pp. 66529–66537, 2020, doi: 10.1109/ACCESS.2020.2984903.

[3] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Learning attentive representations for environmental sound classification," *IEEE Access*, vol. 7, pp. 130327–130339, 2019, doi: 10.1109/ACCESS.2019.2939495.

[4] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Procedia Comput. Sci.*, vol. 112, pp. 2048–2056, Jan. 2017.

[5] L. Liu, W. Li, X. Wu, and B. X. Zhou, "Infant cry language analysis and recognition: An experimental approach," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 3, pp. 778–788, May 2019, doi: 10.1109/JAS.2019.1911435.

[6] K. Ko, J. Park, D. K. Han, and H. Ko, "Channel and frequency attention module for diverse animal sound classification," *IEICE Trans. Inf. Syst.*, vol. E102.D, no. 12, pp. 2615–2618, Dec. 2019, doi: 10.1587/transinf.2019EDL8128.

[7] V.-T. Tran and W.-H. Tsai, "Acoustic-based emergency vehicle detection using convolutional neural networks," *IEEE Access*, vol. 8, pp. 75702–75713, 2020, doi: 10.1109/ACCESS.2020.2988986.

[8] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980, doi: 10.1109/TASSP.1980.1163420.

[9] T. Theodorou, I. Mporas, and N. Fakotakis, "Automatic sound recognition of urban environment events," in *Proc. Int. Conf. Speech Comput.*, 2015, pp. 129–136.

[10] P. Khunarsal, C. Lursinsap, and T. Raicharoen, "Very short time environmental sound classification based on spectrogram pattern matching," *Inf. Sci.*, vol. 243, pp. 57–74, Sep. 2013, doi: 10.1016/j.ins.2013.04.014.

[11] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, May 2019.

[12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[13] S. Chu, S. Narayanan, and C.-C.-J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1142–1158, Aug. 2009, doi: 10.1109/TASL.2009.2017438.

[14] M. Lim, D. Lee, H. Park, Y. Kang, J. Oh, J. Park, G. Jang, and J. Kim, "Convolutional neural network based audio event classification," *KSII Trans. Internet Inf. Syst.*, vol. 12, no. 6, pp. 2748–2760, 2018, doi: 10.3837/tiis.2018.06.017.

[15] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, Nov. 2018, pp. 356–367.

[16] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 421–425.

[17] S. Abdoli, P. Cardinal, and A. Lameiras Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Syst. Appl.*, vol. 136, pp. 252–263, Dec. 2019, doi: 10.1016/j.eswa.2019.06.040.

[18] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Boston, MA, USA, Sep. 2015, pp. 1–6.

[19] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao, and J. Hu, "An ensemble stacked convolutional neural network model for environmental event sound recognition," *Appl. Sci.*, vol. 8, no. 7, p. 1152, Jul. 2018.

[20] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream CNN based on decision-level fusion," *Sensors*, vol. 19, no. 7, p. 1733, Apr. 2019, doi: 10.3390/s19071733.

[21] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-stream network with temporal attention for environmental sound classification," in *Proc. Interspeech*, Sep. 2019, pp. 3604–3608, doi: 10.21437/Interspeech.2019-3019.

[22] X. Tian, J. Zhang, Z. Ma, Y. He, J. Wei, P. Wu, W. Situ, S. Li, and Y. Zhang, "Deep LSTM for large vocabulary continuous speech recognition," 2017, *arXiv:1703.07090*. [Online]. Available: http://arxiv.org/abs/1703.07090

[23] Y. Kim, J. Sa, Y. Chung, D. Park, and S. Lee, "Resource-efficient pet dog sound events classification using LSTM-FCN based on time-series data," *Sensors*, vol. 18, no. 11, p. 4019, Nov. 2018, doi: 10.3390/s18114019.

[24] Z. Ren, K. Qian, Z. Zhang, V. Pandit, A. Baird, and B. Schuller, "Deep scalogram representations for acoustic scene classification," *IEEE/CAA J. Automatica Sinica*, vol. 5, no. 3, pp. 662–669, May 2018, doi: 10.1109/JAS.2018.7511066.

[25] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 2721–2725.

[26] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. ACM Int. Conf. Multimedia (MM)*, Orlando, FL, USA, 2014, pp. 1041–1044.

[27] J.-W. Jung, H.-S. Heo, I.-H. Yang, H.-J. Shim, and H.-J. Yu, "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5349–5353.

[28] R. Vergin and D. O'Shaughnessy, "Pre-emphasis and speech recognition," in *Proc. Can. Conf. Electr. Comput. Eng.*, Montreal, QC, Canada, vol. 2, 1995, pp. 1062–1065.

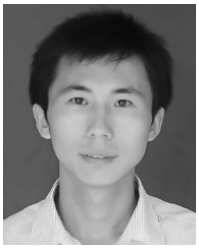
[29] S. Gao, M. Zhou, Y. Wang, J. Cheng, H. Yachi, and J. Wang, "Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 601–614, Feb. 2019, doi: 10.1109/TNNLS.2018.2846646.

XIFENG DONG received the B.E. degree from the Department of Information Science and Engineering, Shandong Agricultural University, China, in 2018. He is currently pursuing the professional master's degree with the Ocean University of China. His current research interests include signal processing and acoustic data analysis.

BO YIN received the Ph.D. degree from the Department of Computer Science and Technology, Ocean University of China, Qingdao, China, in 2006. He is currently an Associate Professor with the Ocean University of China, and a Visiting Scholar with Carnegie Mellon University and the University of Nantes, France. His research interests include acoustic systems, embedded system designing, and intelligent control technology.

YANPING CONG received the Ph.D. degree from the Department of Computer Science and Technology, Ocean University of China, Qingdao, China, in 2012. He is currently an Associate Professor with the Ocean University of China. He has participated in many provincial and ministerial level projects. His research interests include wireless networks and acoustic communication networks.

ZEHUA DU received the M.S. degree from the Department of Information Science and Technology, Shandong University of Technology, Zibo, China, in 2016. He is currently pursuing the Ph.D. degree with the Ocean University of China. His research interests include signal processing, and artificial intelligence and other related research fields.



XIANQING HUANG received the M.S. degree from the Department of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, China, in 2013. He is currently pursuing the Ph.D. degree with the Ocean University of China. His research interests include acoustic data analysis, complex signal processing, and neural network construction.