Applied Acoustics 164 (2020) 107243

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Urban sound classification based on 2-order dense convolutional network using dual features

Zilong Huang [a,c], Chen Liu [a,c], Hongbo Fei [a,c], Wei Li [b], Jinghu Yu [a,c], Yi Cao [a,c,*]

[a] School of Mechanical Engineering, Jiangnan University, Wuxi 214122, Jiangsu, China
[b] Suzhou Vocational Institute of Industrial Technology, Suzhou 215104, Jiangsu, China
[c] Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology, Wuxi 214122, Jiangsu, China

ARTICLE INFO

Article history:
Received 12 November 2019
Received in revised form 14 January 2020
Accepted 1 February 2020

Keywords:
Urban sound classification
2-DenseNet
Dual features fusion
D-2-DenseNet

ABSTRACT

Audio carry a large amount of life scenes and physical events in the city, therefore, developing deep learning approach to automatically extract this information has huge potential and application in building smart-city. In this paper, a novel urban sound event classification model based on 2-order dense convolutional network using dual features is proposed, which aims at the problems of insufficient classification accuracy and adaptability of current models. Firstly, the brief introduction of urban sound classification development and application is presented in Section 1. Then, the method of feature extraction and add noise environment is respectively introduced in Section 2. Moreover, a new network structure referred to as 2-order dense convolutional network (shorten as 2-DenseNet) and its algorithm are presented in Section 3. Meanwhile, an urban sound event classification model based on 2-DenseNet using dual features, i.e. D-2-DenseNet is proposed in this paper. Theoretically, D-2-DenseNet not only can accelerate the convergence speed when compared with DenseNet, but also can enhance the classification accuracy and guarantee a good generalization ability owing to the fact that dual features fusion is exploited in the proposed model. Finally, in order to validate advantages of the D-2-DenseNet, this new model is respectively exploited in the urban sound event classification based on UrbanSound8K and Dcase2016 datasets. The experimental result shows that the accuracy of the network is respectively 84.83% and 85.17%, which has increase up to 13.81% and 7.07% compared with baseline. Compared with single feature network, the classification accuracy of D-2-DenseNet has increased by 3.35% and 4.78% respectively in noise environment.

1. Introduction

Audio carry a large amount of life scenes and physical events in the city, the city designers identify each urban sound source through intelligent perception, human can perceive the sound scene they are within (city center, cafe, etc.), and recognize individual sound sources (car passing by, street, etc.) [1,2]. Therefore, developing an intelligent processing method which can automatically extracts the above information has huge potential and application in the construction of smart-city, such as: noise monitoring [3], urban security [4,5], soundscape assessment [6–8], multimedia retrieval [9], etc.

Currently, smart-city complex mainly uses visual acquisition and recognition, sensors, wireless positioning systems, bar code recognition, etc. to build visual Internet of Things (IOT) [10,11]. However, this IOT does not involve the way urban sound data acquired and specialize classification applications, only the visual IOT still has certain limitations in the urban intelligent sensing system [12]. It is great importance to conduct more comprehensive intelligent sensing [13], automatic data collection [14], and urban sound event acquisition and classification in the building smart-city in future [15]. Therefore, building an audio and visual IOT will further enhance intelliSense and make the smart-city complex more complete [16–18].

In the field of urban sound classification, to the best knowledge of the authors, many interstices have been reported for the topic of urban sound classification mainly based on support vector machine (SVM) [19], artificial neural network (ANN) [20] and convolutional neural network (CNN) [21–27]. Especially, CNN is the most widely used method, combining local and global acoustic information through conditional fusion scheme. Ref. [21] studied based on the deep convolutional neural network (DCNN). Basing on four datasets augmentation and using the convolutional neural

* Corresponding author.
E-mail address: caoyi@jiangnan.edu.cn (Y. Cao).

https://doi.org/10.1016/j.apacoust.2020.107243
0003-682X/

network [22] are further improved classification accuracy. Ref. [23] use dilated convolution for feature extraction in audio clips to improves classification accuracy. Compared to the original convolutions, dilated convolutions do not use max-pooling layers and achieve the state of art in urban sound classification. Especially Ref. [24] adjust the RF of variants of ResNet and DenseNet architectures to best fit the various audio processing tasks that use the spectrogram features as input. These architectures can be an effective tool that offer good generalization properties for various audio processing tasks. Ref. [25] adopt a data augmentation scheme employing generative adversary networks. After voting fusion, the final systems could achieve accuracies above 85.00%.

Above all, although research on urban sound events classification has been carried out by many scholars, on the one hand, there is still a challenge to effectively improve the accuracy and generalization of audio classification. On the other hand, there is still a lack of research methods on the adaptability of audio classification models in noise environment. In this paper, an urban sound event classification model based on 2-order dense convolutional network using dual features is proposed. The main contribution of this paper is threefold:

- This paper presents a novel urban sound classification model based on D-2-DenseNet, using dual features fusion in a multi-channel parallel 2-order dense convolutional network, which can be an effective method for urban sound classification.
- D-2-DenseNet takes both advantages of 2-DenseNet and dual features fusion for achieving better classification and generalization ability. This model can not only accelerate the convergence speed, but also achieve better comprehensive capability compared with current models.
- Based on UrbanSound8K and Dcase2016 datasets, urban sound classification experiments are conducted. The result shows that D-2-DenseNet has excellent classification accuracy and generalization ability. Moreover, experiments are carried out under noise environment, the robustness of D-2-DenseNet are also effectively verified.

This paper is organized as follows. The method of feature extraction and adding noise environment is respectively introduced in Section 2. A novel network referred to as 2-DenseNet is presented in Section 3. Meanwhile, an urban sound event classification model based on D-2-DenseNet is proposed in this paper. In Section 4, urban sound classification experiment is conducted in the UrbanSound8K and Dcase2016 datasets. The results verify the excellent performance of the D-2-DenseNet model. Finally, conclusions and future work are discussed in Section 5.

2. The method of feature extraction

Choosing some representative data to represent a piece of audio signal in processing audio signal often referred to as feature extraction [28]. It is not only the premise of the sound classification, but also paves an underling theoretical grounds for a better classification and generalization ability in the urban sound classification [29]. Therefore, the method of feature extraction should be introduced. Concerning feature extraction of audio can be classified into four different categories: Mel-scale frequency cepstral coefficients (MFCC), Gammatone frequency cepstral coefficients (GFCC), spectrogram, Filter Bank, etc. [30]. Among them, MFCC is the most widely used feature extraction schemes for speech recognition and audio classification, By the feature fusion of MFCC and GFCC, it can guarantee a good adaptability of network when noise is taken into consideration [31]. These are the reasons that why MFCC and GFCC are selected as the method of feature extraction in the present paper [32]. Meanwhile, since most of the audio datasets just provide clean samples (signal-to-noise ratios >60 dB), in order to further study the adaptability of urban sound classification in different environments, the method of add noise environment is proposed. White Gaussian noise (shorten as WGN) is the ideal model for analyzing channel noise [33,34]. Therefore, this paper use WGN with different signal-to-noise ratios to conducting classification in noise environments.

2.1. Feature extraction of MFCC and GFCC

As one of the most widely used feature extraction schemes for speech recognition and audio classification, based on Mel-scale filter shown in Fig. 1, MFCC uses the Hz spectral feature calculated by the nonlinear correspondence with the Mel-scale frequency using the auditory characteristics of the human ear [28]. However, one point to be noted is that GFCC has good recognition effect and better noise robustness, considering different factors such as background noise and signal-to-noise ratios, based on Gammatone filter shown in Fig. 2, GFCC uses Gammatone filter bank based on the human ear and cochlear auditory model instead of the Mel-scale filter bank [29]. The role of the different filter bank is mainly to smooth spectrum and eliminate effects of harmonics, and highlight the formant of the original sound [30].

The cepstral coefficient (MFCC, GFCC, etc.) feature extraction can be expressed by the following steps:

(1) Audio data preprocessing, which includes sampling and quantization, pre-emphasis processing, and windowing, the analog audio signal is converted into a sequence of audio frames.

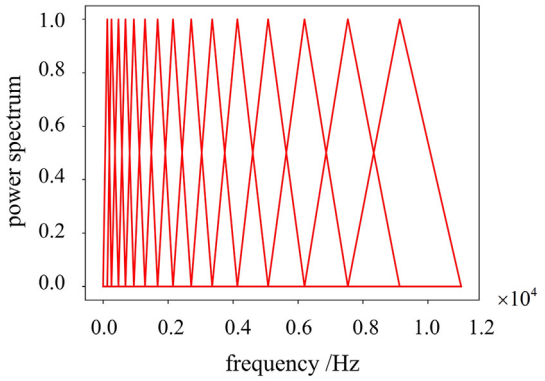


Fig. 1. Spectrogram of 15 Mel-scale filters.

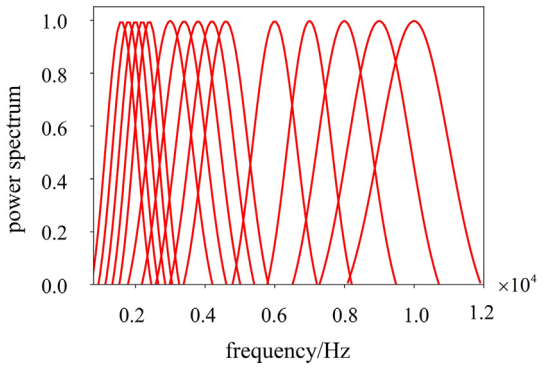


Fig. 2. Spectrogram of 15 Gammatone filters.

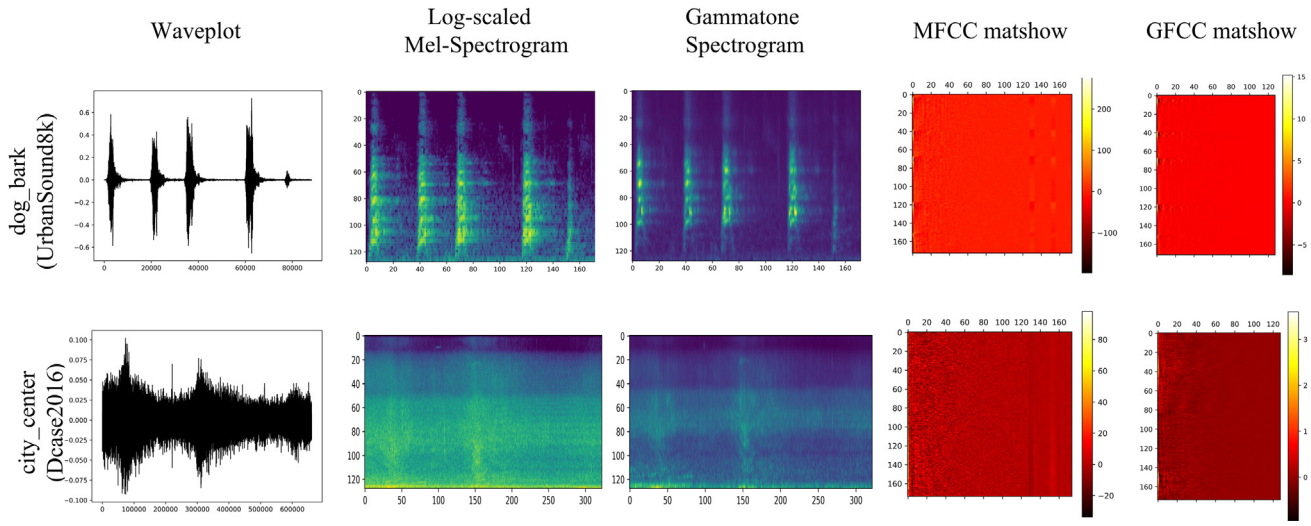


Fig. 3. Waveplot, Log-scaled Mel-spectrogram, Gammatone spectrogram, MFCC matshow and GFCC matshow of each audio segment.

(2) Audio frame sequence time and frequency domain analysis, including: fast Fourier transform (FFT), filter bank, log spectrum, discrete cosine transforms. Different features transformation is shown in Fig. 3.
(3) Audio signal time-varying information extraction, in order to obtain the time-varying information of audio frame, Delta and Delta_Delta characteristics are further extracted.
(4) Two-dimensional feature vector sequence [frame, dimension] output, which is used as an input to the network. The waveplot, spectrogram and cepstral coefficient matshow are shown in Fig. 3.

2.2. Adding noise environment

It is worth noting that most of the audio datasets just provide clean samples (signal-to-noise ratios >60 dB), however, the impact of noise on the classification model must be taken into account for audio classification in urban scenes. Therefore, in order to further study the a of urban sound model, especially the noise environments [32,33], a model of noise environment should be further constated. Based on the conclusion presented in [34–36]. WGN is an ideal model for analyzing channel noise, the WGN model is employed in the urban sound classification considering different signal-to-noise ratios (shorten as SNR), and conducting classification accuracy in noise environments.

Firstly, the definition of SNR is defined as the logarithm of the ratio of signal power to noise power.

$$\mathrm{SNR}_{dB} = 10\log_{10}(P_{signal}/P_{noise}) \tag{1}$$

Secondly, the digital signal of audio extracted by the sampling theorem is discrete and power of the signal can be directly calculated. Assuming that the discrete signal of the audio file is $S = \{s_1, s_2, \ldots, s_n\}$, the signal power P_{signal} is calculated as follow:

$$P_{signal} = \frac{1}{n}\sum_{k=1}^{n} s_k^2 \tag{2}$$

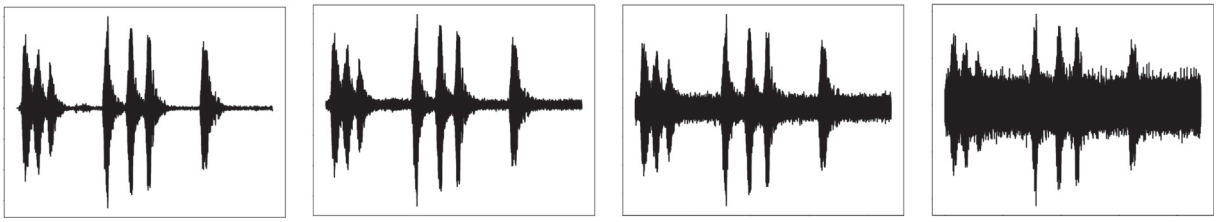
Then, for the current signal power P_{signal} and predetermined SNR (50 dB, 40 dB, 30 dB, etc.), the power of the noise P_{noise} can be calculated:

$$P_{noise} = P_{signal}/(10^{SNR/10}) \tag{3}$$

Finally, for the current power of noise with specified SNR, a noise sequence with a standard Gaussian distribution (the average is 0 while the standard deviation is 1) can be generated, and this noise sequence will be added to the original audio signal. A waveform of an audio sample of adding WGN with a specified SNR is shown in Fig. 4.

3. Network framework

In this Section, three major components of a novel network architecture are proposed. ①. A new network architecture, N-DenseNet was briefly introduced. ② Algorithm of forward propagation and back propagation (BP) of 2-DenseNet is theoretically introduced in detailed which paves theoretical grounds for advantages of the D-2-DenseNet. ③. Based on 2-DenseNet model, a novel network architecture, D-2-DenseNet is presented using dual features.



(a) audio file (dog_bark) **(b)** adding 50dB WGN **(c)** adding 30dB WGN **(d)** adding 10dB WGN

Fig. 4. The waveplot of an audio sample adding WGN with a specified SNR.

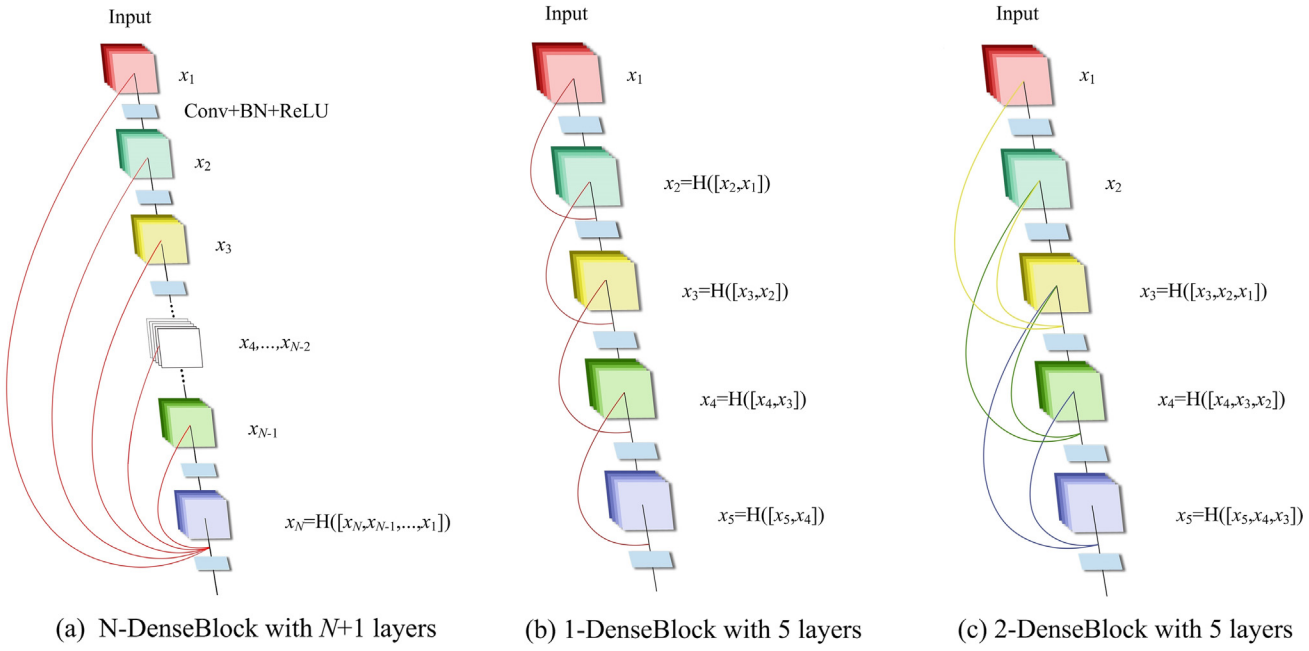


Fig. 5. Schematic diagram of DenseBlock with different orders.

3.1. N-Order dense convolutional network

It should be pointed that the DenseNet [37] network structure is providing an underlying theoretical ground for the proposition of the N-DenseNet, therefore, the main principle of DenseNet would be concisely introduced here. The input of each layer in the DenseNet comes from the output of all the previous layers as stated in [37]. In a l-layer DenseBlock structure, the input of each layer is defined as x_0, x_1, x_2, …, x_l, then:

$$x_l = \mathrm{H}([x_0, x_1, ..., x_{l-1}]) \tag{4}$$

where [x_0, x_1, …, x_{l-1}] refers to concatenation of the feature-map produced in layers 0, …, l-1. H$i(\cdot)$ represents the composite function of operations such as Batch Normalization (BN) [38], rectified linear units (ReLU) [39], or convolution (Conv) [40].

Combining characteristics of the DenseNet and N-Order Markov model in Ref. [41], a new network architecture, referred to as N-DenseNet shown in Fig. 5 is designed to further improve the classification accuracy and generalization ability of the domain in Refs. [42,43]. By the targeted and regulated tailoring of concatenation, the N + 1 layer in the N-DenseNet just only receives the feature-maps of the previous N layers based on the conclusion presented in Ref. [44].

It is worth noting that N-DenseNet is a theoretical ground instead of the keynote of the present paper, therefore, the complete and detailed interpretation of N-DenseNet will not presented here owing to space limitation, please refer to Refs. [42–44]. Based on the conclusion that 2-DenseNet as a sub-model of N-DenseNet, which has better classification accuracy and generalization ability based on the conclusion presented in Ref. [44]. Therefore, the 2-DenseNet and its working principle will be interpreted at length in the followings.

The 2-order state-dependent connection can be defined by the fact that the input of the l layer is just only related to the output of the previous 2 layers, and the connection is performed with concatenation, as shown in Fig. 5(c) the input of each layer is defined as: x_1, x_2, …, x_l. When the current layer would not be connected using the H$i(\cdot)$ function, as shown in Fig. 5(c) the forward propagation of x_1 and x_2 layer, it can be defined as:

$$\begin{aligned}\mathbf{X}^l(i,j) &= [\mathbf{X}^{l-1} \otimes \mathbf{w}](i,j) + \mathbf{b} \\ &= \sum_k \sum_m \sum_n [\mathbf{X}_k^{l-1}(i+m, j+n)\mathbf{w}_k(x,y)] + \mathbf{b}\end{aligned} \tag{5}$$

where $\sum$ denotes the forward propagation of the convolutional layer, $\mathbf{X}^{l-1}$ and $\mathbf{X}^l$ represent the input and output of the feature-map somber respectively; $\otimes$ denotes the convolution operation, $\mathbf{w}$ and $\mathbf{b}$ respectively denotes the kernel function and offset value, $\mathbf{X}$ (i, j) corresponds the pixel on the feature map; k is the number of channels of the feature map, and m and n are the size of the convolution kernel.

For a 2-order state-dependent connection, the current layer in a 2-DenseBlock model is the concatenation layer by targeted and regulated tailoring of concatenation, that is, the input of the current layer just comes from the output of the previous 2-layers, as shown in Fig. 5 (c) x_3 layer, which can be defined by:

$$x_l = \mathrm{H}([x_l, x_{l-1}, x_{l-2}]) \tag{6}$$

3.2. Forward propagation and back propagation of 2-Densenet

In a l-layers block structure, the number of state-dependent connection of DenseNet is $l(l-1)/2$, while in the 2-DenseNet, the number of state-dependent connection is $(2l$-4) owing to use of the regularity [44]. This decrease of number of the state-dependent connection will carfare goals to a faster convergence speed, a higher training efficiency [45,46], which will be theoretically demonstrated by the analysis of the forward propagation and back propagation of the 2-DenseNet in the following Subsection3.2 at length.

3.2.1. Forward propagation of 2-DenseNet

In the 2-DenseBlock structure, its forward propagation is shown in Fig. 6. One of a 1 × 1 and 3 × 3 convolutional layers is a set of non-linearly changing feature layer elements, shown in Fig. 6 $\boldsymbol{X}$, the input of each layer is defined by: $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, …, $\boldsymbol{X}_l$, in the 2-DenseBlock, the feature output $\boldsymbol{Uc}$ of the third layer starting network convolution transformation is defined by:

$$U_c = f(BN(W_{3\times3} \otimes f(BN(W_{1\times1} \otimes [X_l, X_{l-1}, X_{l-2}] + B)))) \tag{7}$$

where $[\boldsymbol{X}_l, X_{l-1}, X_{l-2}]$ indicates that the current layer passes the 2-order related connection mode and uses the feature mapping of two previous layers as inputs, $\boldsymbol{W}_{3\times3}$ and $\boldsymbol{W}_{1\times1}$ indicates that the convolution kernel size is 1×1 and 3×3 matrix respectively, BN $(\cdot)$ is the batch normalization, and $f(\cdot)$ is the activation function of ReLU.

Based on 2-order concatenation method, it is shown that 2-DenseNet is more targeted than the dense connection of DenseNet. Therefore, 2-DenseNet is just only connected with the feature information of two previous layers, that will reduce redundant feature information reuse, and more efficient and more targeted in feature information reuse.

3.2.2. Back propagation of 2-DenseNet

It should be pointed that in the 2-DenseBlock structure training process, the weight of each network layer is continuously updated

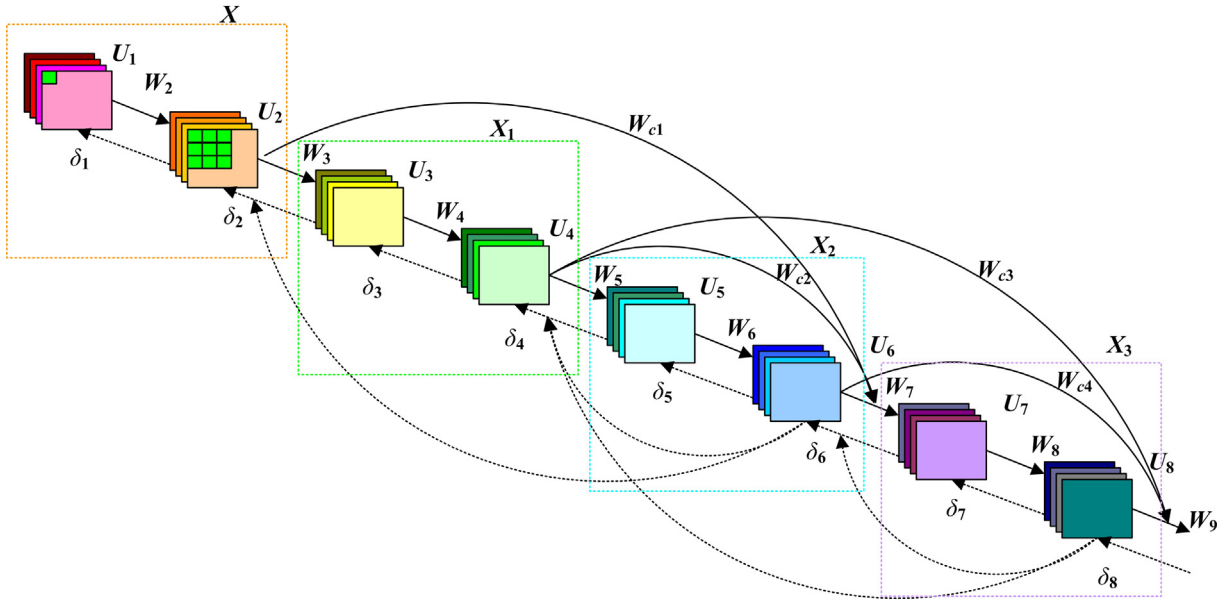


Fig. 6. Schematic diagram of the forward propagation and back propagation of the 2-DenseNet.

Table 1
Back propagation of 8-layers 2-DenseBlock.

Layer	Back Propagation	Layer	Back Propagation
Input	$\partial J/\partial W_1 = \delta_1 * W_1 \otimes X$	L_5	$\delta_5 = \delta_6 * W_6 \otimes (\partial U_6/\partial U_5)$
L_1	$\delta_1 = \delta_2 * W_2 \otimes (\partial U_2/\partial U_1)$	L_6	$\delta_6 = \delta_7 * W_7 + \delta_8 * W_{c4}$
L_2	$\delta_2 = \delta_3 * W_3 + \delta_6 * W_{c1}$	L_7	$\delta_7 = \delta_8 * W_8 \otimes (\partial U_8/\partial U_7)$
L_3	$\delta_3 = \delta_4 * W_4 \otimes (\partial U_4/\partial U_3)$	L_8	$\delta_8 = \delta_9 * W_9 \otimes (\partial U_9/\partial U_8)$
L_4	$\delta_4 = \delta_5 * W_5 + \delta_6 * W_{c2} + \delta_8 * W_{c4}$	Output	$\partial J/\partial U_9$

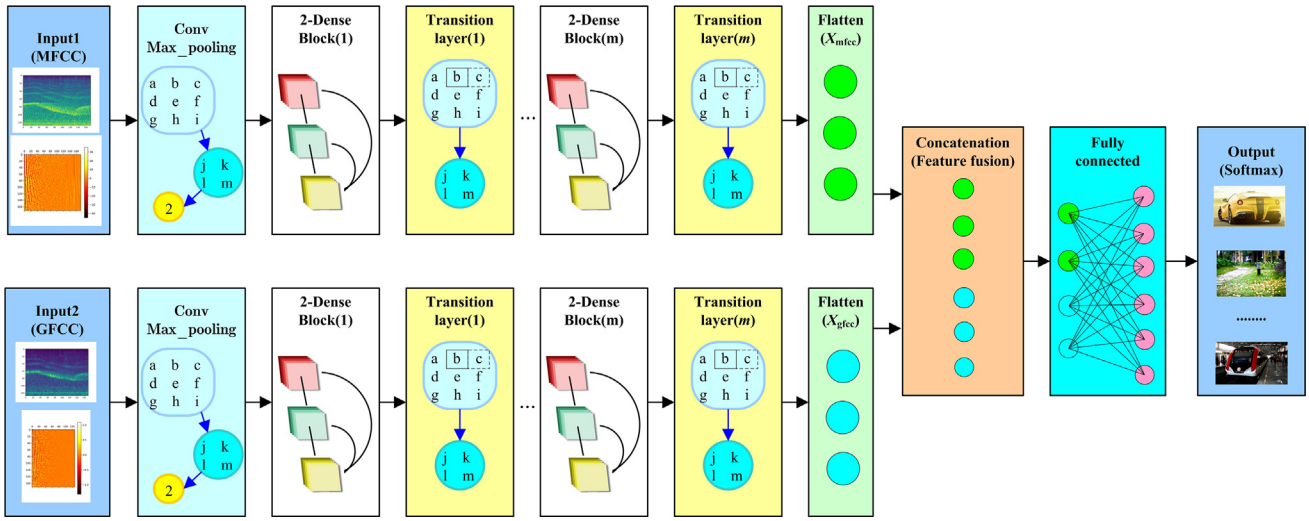


Fig. 7. A schematic diagram of 2-order dense convolutional network using dual features frame.

until the model is converged by the algorithm proposed in Refs. [47,48]. The back-propagation calculations in the 2-DenseBlock structure are shown in Table 1 and Fig. 6.

In Table 1 and Fig. 6, δ_l indicates the error of each layer, the output of each layer is defined as $\boldsymbol{U}_1, \boldsymbol{U}_2, \ldots, \boldsymbol{U}_l$, while $\boldsymbol{W}_l$ is the matrix of the convolution layer and $\boldsymbol{W}_{cl}$ is the matrix of the $Hi(\cdot)$ concatenation layer respectively, * is the convolution operation to flip.

Table 1 shown that the error term of the 2-order concatenation layer is propagated back to the two previous layers, who's the gradient value can be expressed by:

$$\partial J/\partial W_{l-2} = (\delta_{l-1} * W_{l-1} + \delta_l * W_{cl} + \delta_{l+1} * W_{c(l+1)}) \otimes X_{l-2} \tag{8}$$

However, in Denseblock structure, the error term of densely concatenation layer is back propagated to all the previous layers expressed by:

$$\partial J/\partial W_{l-2} = (\delta_{l-1} * W_{l-1} + \delta_l * W_{cl} + \delta_{l+1} * W_{c(l+1)} + \ldots + \delta_{l+n} * W_{c(l+n)}) \otimes X_{l-2} \tag{9}$$

From Table 1, Fig. 6, Eqs. (8) and (9), it can be concluded that 2-DenseNet back propagation only needs to calculate the influence of the back layer to the previous two layers owing to 2-order state-dependent connection, which is more conducive to the calculation of the gradient information and the overall convergence speed of the network comparing to densely connection of the DenseNet.

3.3. 2-Order dense convolutional network using dual features

As stated above that the 2-DenseNet model has many advantages, such as, a better convergence speed, a higher accuracy compared with the DenseNet, in order to further improve the comprehensive capability of the model including the generalization ability, especially the adaptability under noise environment, an urban sound event classification model based on 2-DenseNet using dual features, i.e. D-2-DenseNet, is proposed in Subsection 3.3. The network is characterized by dual features input, i.e. MFCC and GFCC, and a multichannel parallel 2-DenseNet for feature fusion. The construction of D-2-DenseNet classification model is mainly divided into four following steps:

The input of multichannel parallel network: Feature extraction of MFCC and GFCC is respectively carried out according to the method above proposed in Subsection 2.1. Two different features are input into a multichannel parallel network where a convolution operation is performed to extract more features, a pooling process is employed to compress layer size.

The construction of 2-DenseNet: Based on Fig. 6, the 2-DenseBlock structure is constructed, and a multichannel feature vector is input into the continuous modules of 2-Denseblock whose total number is m and Transition Layer as shown in Fig. 7. It should be pointed that this step is the key to the construction of D-2-DenseNet.

Feature fusion: Feature fusion is carried on the concatenation layer. The feature vector of MFCC and GFCC are flattened to one-dimensional data, and then concatenation layer merges the outputs of multichannel parallel into single channel, which is shown in Fig. 7 expressed by:

$$X = Concate([X_{\text{mfcc}}, X_{\text{gfcc}}]) \tag{10}$$

The classification result output: After the feature fusion processing, network input, i.e., the merged dual features information is input into the fully connected layer, and the information is adjusted and adaptive by dropout. Finally, the classification result is dealt with vector normalization and output.

To describe the above steps more clearly, a network framework construction of the D-2-DenseNet model can be represented by Fig. 7.

Based on the above research in Section 3, main advantages of the D-2-DenseNet model can be concluded as: ①. 2-DenseNet adopts the 2-order state-dependent connection method to effectively reduce the redundant connection of the feature layer, and convergence speed is faster than DenseNet; ②. D-2-DenseNet is a multichannel 2-DenseNet network and its convergence speed and effect should be similar to the 2-DenseNet performance; ③. D-2-DenseNet combining both advantages of dual features fusion and 2-DenseNet, theoretically, the classification accuracy, generalization ability, and especially the adaptability under noise should be better than DenseNet and 2-DenseNet.

4. Experiments and analysis

4.1. Dataset and experiment

In order to validate the comprehensive performance of the D-2-DenseNet model, urban sound event standard dataset UrbanSound8k [6] and IEEE AASP sound scene and event detection classification challenge dataset Dcase2016 [7] are used to conduct urban sound event classification in this paper. While GTX-1080Ti graphics card is used for training to ensure the smooth operation of the experiment. Keras+TensorFlow is employed as a deep learning framework. Three experiments are designed to respectively demonstrate the classification accuracy, the generalization ability, the adaptability of the D-2-DenseNet, that is:

(1) **Experiment 1:** The classification accuracy study will be conducted based on DenseNet, 2-DenseNet and D-2-DenseNet respectively using the same dataset UrbanSound8k.
(2) **Experiment 2:** The generalization ability study will be carried out based on DenseNet, 2-DenseNet and D-2-DenseNet respectively using the same dataset Dcase2016, and the classification accuracy will be exploited as a performance index for evaluation of the generalization of model.
(3) **Experiment 3:** The adaptability under noise study will be designed based on 2-DenseNet and D-2-DenseNet respectively using the Dcase2016 with different SNRs added to the dataset, while classification accuracy as a performance index for the evaluation of the adaptability.

4.2. Training

Based on DenseNet, 2-DenseNet, and D-2-DenseNet models, for a given number of convolution layers (22), convolution kernel size, channel number (32), optimizer Adam, Batch size = 32, etc., Figs. 8 and 9 respectively shows the loss values of the three above

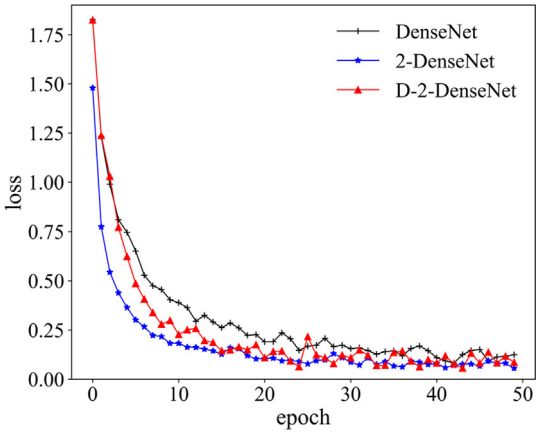


Fig. 8. UrbanSound8K _loss (50 epochs).

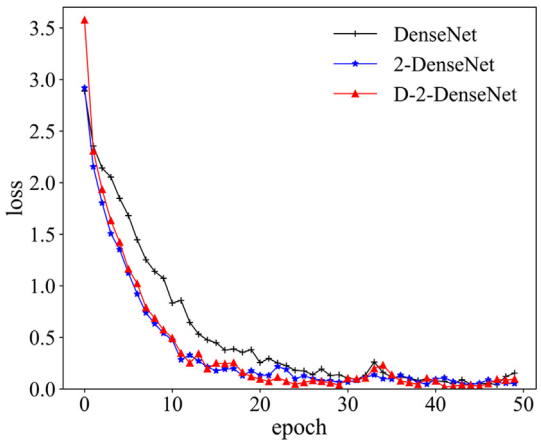


Fig. 9. Dcase2016_loss (50 epochs).

mentioned models under the given number of training epochs based on dataset UrbanSound8k and Dcase2016, respectively.

From Figs. 8 and 9, it can be concluded that: ①. The convergence speed of D-2-DenseNet is similar to that of 2-DenseNet. ②. Compared with DenseNet, the convergence speed of D-2-DenseNet is faster than that of DenseNet, and it trends to convergence with fewer rounds and a lower loss value. These conclusions also demonstrate that the D-2-DenseNet can guarantee a faster convergence speed, a lower loss value compared with that of DenseNet, while the convergence speed, loss value is similar to that of 2-DenseNet.

4.3. Results and analysis

4.3.1. Classification accuracy

In order to verify the classification accuracy of D-2-DenseNet, firstly, the classification accuracy is carried out based on D-2-DenseNet, 2-DenseNet and DenseNet using the UrbanSound8K dataset. Then, the results are compared with those based on SVM, DCNN, Dilated CNN, DNN and D-CNN-ESC, finally, classification accuracy under different models are is shown in Table 2.

From Table 2, it can be concluded that: ①. In the UrbanSound8K datasets, classification accuracy of D-2-DenseNet was 84.83%. ②. D-2-DenseNet is 2.93% higher than the current research results D-CNN-ESC. ③. Compared with DenseNet and 2-DenseNet, the accuracy of D-2-DenseNet has been improved up to 3.80% and 1.55% respectively. The above research proved that the D-2-DenseNet model has a better convergence effect, and the dual-feature fusion has higher classification accuracy than 2-DenseNet, DenseNet and current research results.

4.3.2. Generalization ability

In order to examine the generalization ability of the D-2-DenseNet, Dcase2016 dataset will be added and the classification accuracy will be exploited as a performance index for evaluation of the generalization ability of the D-2-DenseNet, which is shown in Table 3. In order to demonstrate the result more clearly, confusion matrix of the D-2-DenseNet using the UrbanSound8K and Dcase2016 is presented in Figs. 10 and 11 respectively.

Table 2
Different model accuracy results on UrbanSound8K dataset.

Model	Feature	Accuracy
SVM [6]	25mfcc	71.00%
DCNN [19]	Local & global features	77.36%
Dilated CNN [22]	60spectrograms	78.00%
DNN [20]	80FBANK	79.23%
CNN [21]	40FBANK	81.50%
D-CNN-ESC [23]	60mfcc+60mfcc_d	81.90%
DenseNet	174mfcc	81.03%
2-DenseNet	174mfcc	82.27%
D-2-DenseNet	[174mfcc,128gfcc]	84.83%

Table 3
Accuracy of each model with different datasets.

Model	Feature	UrbanSound8k	Dcase2016
Baseline	MFCC	71.00%	77.20%
DenseNet	MFCC	81.00%	80.28%
DenseNet	GFCC	78.27%	77.57%
2-DenseNet	MFCC	82.17%	81.03%
2-DenseNet	GFCC	79.57%	78.26%
2-DenseNet	MFCC+GFCC	82.75%	81.53%
D-2-DenseNet	[MFCC, GFCC]	84.83%	85.17%

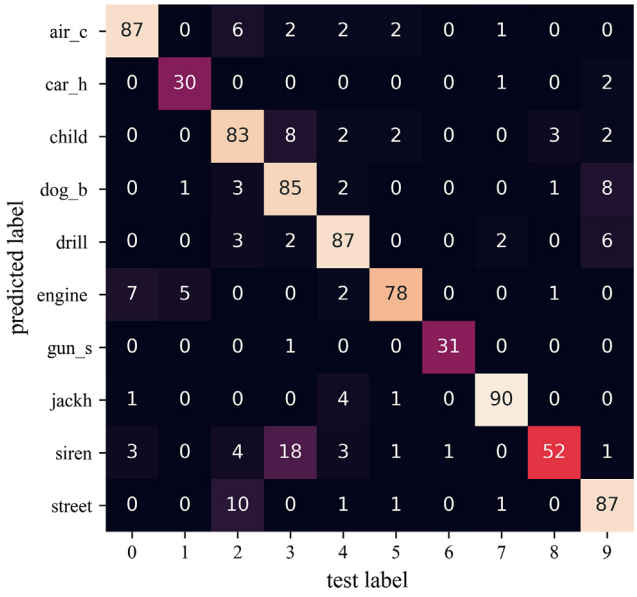


Fig. 10. Confusion matrix of the D-2-DenseNet with 84.83% accuracy.

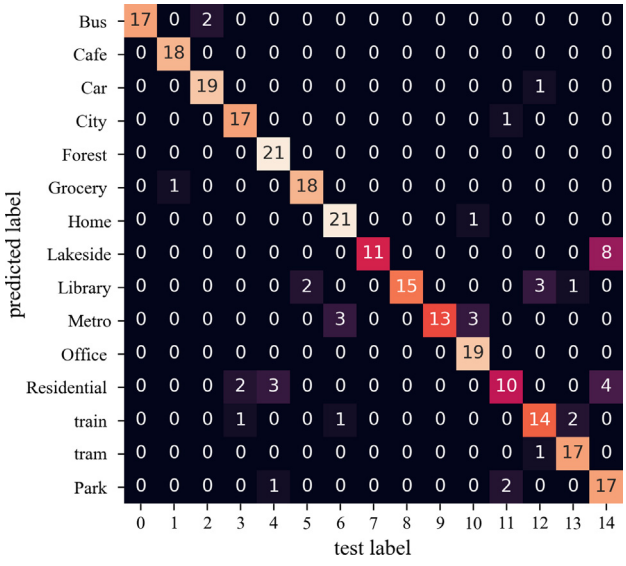


Fig. 11. Confusion matrix of the D-2-DenseNet with 85.17% accuracy.

Table 4
Classification accuracy of three model under different SNRs.

SNR	Clean(>60 dB)	50 dB	40 dB	30 dB	20 dB	10 dB
2-DenseNet (MFCC)	81.03%	73.79%	71.41%	69.66%	66.27%	63.83%
2-DenseNet (GFCC)	78.28%	71.03%	69.31%	68.03%	65.90%	63.57%
D-2-DenseNet	85.17%	77.24%	74.83%	72.41%	69.65%	68.29%

From Table 3, Figs. 10 and 11, it can be concluded that: ①. In the UrbanSound8K and Dcase2016 datasets, the best classification accuracy of D-2-DenseNet is 84.83% and 85.17%, respectively.②. Compared with the baseline, the accuracy of D-2-DenseNet has been increase up to 13.81% and 7.07%, respectively.③. Compared with the accuracy using single feature (MFCC or GFCC) based on DenseNet and 2-DenseNet, the average classification accuracy of D-2-DenseNet has been increased up to 4.36% and 3.40%, respectively.④ Compared with the accuracy using traditional feature fusion (MFCC+GFCC) based on 2-DenseNet, the average classification accuracy of D-2-DenseNet has been increase up to 2.08% and 3.64%, respectively.

In summary, the classification accuracy of D-2-DenseNet has been significantly improved and the classification accuracy in both datasets is about 85.00%, which also verifies the generalization ability of model. generalization ability of model.

4.3.3. Noise environment

In order to further study the adaptability of the model under different environment, especially the noise environment, experimental tests are carried out under the noise environment by adding WGN with different SNRs. The classification accuracy of noise environment is based on the Dcase2016 dataset, and results are shown in Table 4.

From Table 4, it can be concluded that: ①**.** The classification accuracy of each model decreases compared with the clean environment owing to the addition of WGN with different SNRs, and the smaller the SNR is, the lower classification accuracy it has. ②**.** Compared with the single feature (MFCC, GFCC) 2-DenseNet, D-2-DenseNet classification accuracy is higher under noise environment, and with the decrease of SNR, the accuracy decrease rate is relatively flat. ③**.** For the same SNR, compared with 2-DenseNet using MFCC or GFCC, the average classification accuracy of D-2-DenseNet has been improved up to 3.35% and 4.78% respectively. The above research proves that D-2-DenseNet has better classification accuracy in noise environment and adaptability to some extent.

5. Conclusion and future works

In this paper, a novel urban sound event classification model based on D-2-DenseNet is proposed, which aims at the problems of insufficient classification accuracy and adaptability of current models. A novel urban sound classification model is proposed based on D-2-DenseNet, using dual features fusion in a multichannel parallel 2-order dense convolutional network, which can be an effective method for urban sound classification. D-2-DenseNet takes both advantage of 2-DenseNet and dual features fusion for achieving better classification and adaptability performances. This model not only can accelerate the convergence speed, but also can achieve better comprehensive capability compare with current models. Based on UrbanSound8K and Dcase2016 datasets urban sound classification experiments are conducted. The experimental result shows that the accuracy of the network is respectively 84.83% and 85.17%, which has an increase up to 13.81% and 7.07% compared with baseline. Compared with single feature network the classification accuracy of D-2-DenseNet has been improved up to 3.35% and 4.78% in noise environment respectively. These conclusions verify D-2-DenseNet can effectively solves the problem of insufficient classification accuracy and adaptability, this model achieve the state of art in urban sound classification. Though the performance of D-2-DenseNet is studied in noise environment with WGN in this paper, there are still some limitations which should be pointed out: the classification and application in complex noise scene should be further carried out under this model. The authors will further consider different environment noise including standard noise dataset from NoiseX-92 and Thchs30 to verify the practicability of this network. All these above are the current aims of the authors and will be investigated in the future works.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work reported here was supported by the National Natural Science Foundation of China (Grant No. 51375209), 111 Project (Grant No. B18027), Research project (FMZ201901) from Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology. The Six Talent Peaks Project in Jiangsu Province (Grant No. ZBZZ-012), the Research and the Innovation Project for College Graduates of Jiangsu Province (Grant No. JNKY19_048, JNSJ19_005). Finally, the authors would like to thank for the support of UrabnSound8K and Dcase2016 datasets.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apacoust.2020.107243.

References

[1] Bogdanov D, Wack N, Salamon J, et al. An open-source library for sound and music analysis. In: 21st ACM international conference on multimedia. Spain: ACM Press; 2013. p. 855–8.
[2] Steele D, Krijnders JD, Guastavino C. A cognitive approach to soundscape research. J Acoust Soc Am 2004;56(2):214–8.
[3] Ramy H, Khaled BS, Ayman H. Robust feature extraction and classification of acoustic partial discharge signals corrupted with noise. IEEE Trans Instrum Meas 2017;99(1):1–9.
[4] Wang J, Li C, Xiong Z, et al. Survey of data-centric smart city. J Comput Res Dev 2014;51(2):239–59.
[5] Horwath D. MIT predicts 10 breakthrough technologies of 2018 sensing city. The MIT technology review, 2018.
[6] Sainath JP, Salamon J, Jacoby C. A dataset and taxonomy for urban sound research. 22nd ACM international conference on multimedia. USA: ACM Press; 2014.
[7] Mesaros A, Heittola T, Virtanen T. TUT database for acoustic scene classification and sound event detection. 24th European signal processing conference. Budapest, Hungary: IEEE Press; 2016.
[8] Antti J, Vesa T, Tuomi JT, et al. Audio-based context recognition. IEEE Trans Audio Speech Lang Process 2006;14(1):321–9.
[9] Kim D, Kwangyoun H, Hanseok K. Hierarchical approach for abnormal acoustic event classification in an elevator. Advanced video and signal-based surveillance (AVSS), 2011 8th IEEE international conference. Klagenfurt, Austria: IEEE; 2011.

[10] Pham C, Cousin P. Streaming the sound of smart cities: experimentations on the smart Santander test-bed. In: proceedings of the 2013 IEEE international the smart santander test-bed. In: proceedings of the 2013 IEEE international conference on green computing and communications and IEEE internet of '13. Washington, USA: IEEE Computer Society. 2013. p. 611–618.
[11] Valero X, Alas F. Gammatone wavelet features for sound classification in surveillance applications. In: 2012 proceedings of the 20th European signal processing conference (EUSIPCO). Bucharest, Romania. p. 1658–62.
[12] Woodland PC. Very deep convolutional neural networks for robust speech recognition. In: Spoken language technology workshop 2017. Tokyo, Japan: IEEE; 2017. p. 481–8.
[13] Guo J, Ma J. Trust recommendation algorithm for the virtual community based internet of things (IOT). J Xidian Univ 2015;42(2):52–7.
[14] Antonio J, Torija. A tool for urban soundscape evaluation applying support vector machines for developing a soundscape classification model. Sci Total Environ 2014;482483:440–51.
[15] Ahdanau D, Chorowski J, Serdyuk. End-to-end attention-based large vocabulary speech recognition. In: International conference on acoustics speech and signal processing. Barcelona (CCIB), Spain. p. 4945–9.
[16] Giannoulis D, Benetos E, Stowell D, et al. Detection and classification of acoustic scenes and events: an IEEE AASP Challenge. In: IEEE workshop on applications of signal processing to audio and acoustics (WASPAA). New Paltz, NY, United States. p. 1–4.
[17] Salamon J, Bello JP. Unsupervised feature learning for urban sound classification. In: IEEE international conference on acoustics, speech and signal processing. USA. p. 171–5.
[18] Piczak KJ. Environmental sound classification with convolutional neural networks. In: IEEE international workshop on machine learning for signal processing. Boston, USA. p. 1–6.
[19] Ye J, Kobayashi T, Masahiro M, et al. Urban sound event classification based on local and global features aggregation. Appl Acoust 2017;117:246–56.
[20] Lim M, Lee D, Donghyun K, et al. Audio event classification using deep neural networks. KSII Trans Internet Inf Syst 2015;12(10):27–33.
[21] Lim M, Lee D, Hosung P, et al. Convolutional neural network based audio event classification. KSII Trans Internet Inf Syst 2018;12(6):2748–59.
[22] Chen Y, Guo Q, Liang X. Environmental sound classification with dilated convolutions. Appl Acoust 2019;148:123–32.
[23] Zhang X, Zou Y, Shi W. Dilated convolution neural network with leaky ReLU for environmental sound classification acoustics. In: 22nd international conference on digital signal processing. USA: IEEE Press; 2017. p. 26–32.
[24] Chen H, Liu Z, Zo Liu, et al. Integrating the data augmentation scheme with various classifiers for acoustic scene modeling. 2019 IEEE AASP challenge on detection and classification of acoustic scenes and events. Western Finland, Finland, 2019.
[25] Koutini J, Eghbal H, Widmer G. Acoustic scene classification and audio tagging with receptive-field-regularized CNNs. 2019 IEEE AASP challenge on detection and classification of acoustic scenes and events. Western Finland, Finland, 2019.
[26] Peng Z, Zhu Z, Unoki M, et al. Speech emotion recognition using multichannel parallel convolutional recurrent neural networks based on gammatone auditory filterbank. In: 2017 Asia-pacific signal and information processing association annual summit and conference (APSIPA ASC). Kuala Lumpur, Malaysia: IEEE Press; 2017. p. 1750–5.
[27] Phaye HS, Benetos E, Wang Y. Subspectralnet-using sub-spectrogram based convolutional neural networks for acoustic scene classification. In: 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), Brighton, UK. p. 825–9.
[28] Zhong W, Fang X, Fan C, et al. Fusion of deep shallow feature and models for speaker recognition. Acta Acustica 2018;43(2):60–70.
[29] Li D, Zhang X, Duan S, et al. Dysarthria recognition combining speech fusion feature and random forest. J Xidian Univ 2018;45(3):149–55.
[30] Vergin R, O'Shaughnessy D, Farhat A. Generalized mel-frequency cepstral coefficients for large-vocabulary speaker independent continuous-speech recognition. IEEE Trans Speech Audio Process 1999;7(5):525–32.
[31] Hung JW, Heishe HJ, Chen B. Robust speech recognition viaenhancing the complex-valued acoustic spectrum in modulation domain. IEEE ACM Trans Audio Speech Language Process 2015;17(8):171–5.
[32] Rishabh NT, Dharmesh MA, Hemant AP, et al. Novel TEO-based gammatone features for environmental sound classification. In: 2017 25th european signal processing conference (EUSIPCO). Greek island of Kos. p. 318–25.
[33] Tan T, Qian YM. Adaptive very deep convolutional residual network for noise robust speech recognition. IEEE ACM Trans Audio Speech Language Process 2018;26(8):1393–405.
[34] Geiger JT, Helwani K. Improving event detection for audio surveillance using gabor filter bank features. In: European signal processing conf. (EUSIPCO), Nice, France. p. 714–8.
[35] Yi J, Tao J, Liu B, et al. Transfer learning for acoustic modeling of noise robust speech recognition. J Tsinghua Univ (Sci Technol) 2018;58(1):55–60.
[36] Cai S, Jin X, Gao S, et al. Noise robust speech recognition based on sub-band energy warping perception linear prediction coefficient. Acta Acustica 2012;6:667–72.
[37] Huang G, Liu Z, Weinberger KQ. Densely connected convolutional networks. In: 30th IEEE conference on computer vision and pattern recognition CVPR 2017, Honolulu, USA. p. 2261–9.
[38] Wu R, Li J, Qu J. Flight delay prediction model based on deep SE-DenseNet. J Comput Appl 2018. https://doi.org/10.11999/JEIT180644.
[39] Joffe S, Christian S. Batch normalization: accelerating deep network training by reducing internal covariate shift. 2015.arXiv:1502.03167v3 [cs. LG].
[40] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the 14th international conference on artificial intelligence and statistics. Sardinia, Italy. p. 315–8.
[41] Munkhammar J, Joakim W. An N-state Markov-chain mixture distribution model of the clear-sky index. Sol Energy 2018;173(1):487–95.
[42] Cao Y, Huang Z, Zhang W, et al. Urban sound event classification based on N-DenseNet and high-dimensional MFCC characteristics, 201910066 335.6. China, 2019-06-28.
[43] Cao Y, Huang Z, Liu C, et al. Urban sound event classification based on 2-order dense convolutional network using dual features, 201910539 745.8. China, 2019-10-29.
[44] Cao Y, Huang Z, Liu C, et al. Urban sound event classification with N-order dense convolutional network. J Xidian Univ 2019;12(6):9–17.
[45] Ye J, Kobayashi T, Murakawa M, et al. Robust acoustic feature extraction for sound classification based on noise reduction. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). Florence, Italy. p. 5944–8.
[46] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323(9):533–6.
[47] Bottou L, Bengio Y, Haffner P, et al. Gradient based learning applied to document recognition. Proc IEEE 1998;86(11):2278–98.
[48] Srivastava N, Hinton GE, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.