Applied Acoustics 148 (2019) 123–132

Contents lists available at ScienceDirect

Applied Acoustics

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

# Environmental sound classification with dilated convolutions

Yan Chen [a,c,1], Qian Guo [a,b,c,1], Xinyan Liang [a,b,c], Jiang Wang [a,c], Yuhua Qian [a,b,c,*]

[a] Institute of Big Data Science and Industry, Shanxi University, Taiyuan, 030006 Shanxi, China
[b] Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, 030006 Shanxi, China
[c] School of Computer and Information Technology, Shanxi University, Taiyuan, 030006 Shanxi, China

ARTICLE INFO

*Article history:*
Received 27 July 2018
Received in revised form 17 October 2018
Accepted 8 December 2018

*Keywords:*
Sound information retrieval
Environmental sound classification
Dilated convolutions

ABSTRACT

In sound information retrieval (SIR) area, environmental sound classification (ESC) emerges as a new issue, which aims at classifying environments by analysing the complex features extracted from the various sound data. As one of the most efficient feature extraction methods, convolution neural networks (CNN) has made its success in speech and music signal processing, and in particular, CNN with pooling has worked effectively in classifying environmental and urban sound sources. However, pooling causes information loss. In this paper, dilated CNN, being introduced to ESC problem, achieves better results than that of CNN with max-pooling and other state-of-the-art approaches. At the same time, we explore the effect of different dilation rate and the number of layers of dilated convolution to the experimental results, and find that expanding the number of covered frames or enlarging the dilation rate will make the accuracy reduce. That may be the sound signal has short-term stability, the size of the overlay frame seriously affects the feature extraction of the sound signal, and there is an inherent "gridding" in the dilation model conjunction defect.



## 1. Introduction

Sound signal retrieval (SIR) as a hot issue has been widely discussed that people in many application areas. For example, in the classification of marine mammalian sounds, a marine mammal classification calculation model was proposed [1] to extract and classify the data out of the online marine animal sound database such that scientists are able to more accurately detect, identify and locate different endangered species and high-intensity anthropogenic sources that may cause damage to marine ecosystems; for identifying the aircrafts, researchers analyze the noises of their take-off [2]. The interested reader is referred to [3–5]. In the city, various noise sources such as vehicle, swarms of people, blend in the urban soundscape, the structure of which is a complex. Separation or classification of these sources [6] are crucial to the understanding of urban sound and the controlling on urban noises. Based on MBLMS and BSS, an environmental acoustic analysis on recorded audio mixture was proposed for the separation [7]; and Predominant classification technique successfully identifies main noise sources [8].

Environmental sound classification (ESC) refers to the task of associating a semantic label to an audio stream that identifies the environment in which it has been produced, with classical classifiers such as Gaussian mixture models [9], support vector machines, hidden Markov models for manually extracting features like melfrequency cepstral coefficients [10]. In the past two years, deep networks have been introduced to this research area. Surveys on this subject [6] detailed the most frequently-used methods which however restricts to the analysis of highly preconditioned acoustic features [11–14].

Being one of the most renown, CNN emerged in 1980s [15], developed in 90s [16] and served various object classification or pattern recognition tasks [17–19] for almost three decades. CNN has made its way to speech processing [20–22] and music analysis [23,24] which highlighted data locality in sonic problem solving.

In 2015, Karol J. Piczak and etc. evaluated the potential of CNN classifying short environmental audio clips [25]. On a par with other feature learning methods, CNN model have been shown functional in ESC even with limited data sets. However the accuracy was confined by information loss in max-pooling operations which sacrifice the sample size for the increase of receptive field. We here use dilated convolution for feature extraction in audio clips to improves ESC accuracy. Compared to the original convolutions, dilated convolutions do not use max-pooling layers and achieve the state of art in ESC.

* Corresponding author at: Institute of Big Data Science and Industry, Shanxi University, Taiyuan, 030006 Shanxi, China.China
*E-mail address:* jinchengqyh@126.com (Y. Qian).
[1] Co-first author.

https://doi.org/10.1016/j.apacoust.2018.12.019
0003-682X/

## 2. Related methods

### 2.1. Convolutional neural networks

CNN, specializes in processing data of grid structure. For example, the time series data (one-dimensional grid with regular sampling on the time axis) and image data (a two-dimensional pixel grid). Convolution is a class of linear operation. The convolutional networks are neural networks that use convolutional operations instead of matrix multiplication operation. For example, it is defined as:

$$s(i,j) = (X * W)(i,j) + b = \sum_{k=1}^{n_{in}} (X_k * W_k)(i,j) + b \quad (1)$$

where $n_{in}$ is the number of input matrices or the dimension of the last dimension of the tensor. $X_k$ represents the $k$th input matrix. $W_k$ represents the $k$th sub-convolution kernel matrix of the convolution kernel. $s(i,j)$ is the value of the corresponding position element of the output matrix corresponding to the convolution kernel $W$. For example, input a two-dimensional matrix of $4 \times 4$, and the convolution kernel is a $2 \times 2$ matrix, the stride size is set to (1,1), the calculation is presented in Fig. 1.

For the output after convolution, the ReLU The Rectified Linear Unit activation function is generally used to change the element value corresponding to the position less than 0 in the output tensor to 0. The Relu activation function expression is:

$$f(x) = max(0, x) \quad (2)$$

The function image is represented as Fig. 2.

Generally CNN is component layers stacking in a deep architecture: an input layer, a set of convolutional layers that can be combined in various ways, a limited number of fully connected hidden layers, and an output layer. A typical convolution layer has three stages: first computes multiple convolutions in parallel to produce a set of linear activation responses; second, each linear activation response will pass through a nonlinear activation function, for example, a rectification linear activation function. it's called detector stage; lastly, pooling function adjusts the output of this layer. (depicted in Fig. 3).

A convolutional kernel (filter) in the convolution layer captures the local structure (mainly, but not limited to images) that presents in the two-dimensional input data. Each convolution kernel, instead of connects to all inputs from the previous layer, restricts to a small region of the entire input space (for example, a small $3 \times 3$ pixel block) called receptive field. The weighted convolution kernel is applied (tiled over) to the entire input space, and generates a feature map. In this way, a set of weights can be reused throughout the whole input space, under the assumption that locally useful features are also useful elsewhere in the input space. This assumption engenders both reduction in the number of parameters and robustness of data transformation. A typical convolutional layer will consist of many filters (feature maps).

Further dimensionality reduction can be done by merging adjacent cells of a feature map through pooling layers. The pooling function replaces the output of the network in this location by using the overall statistical features of adjacent outputs at a certain location. For example, the max pooling function (Zhou and Chellappa, 1988) [26] gives the maximum value in adjacent rectangular regions. Other commonly used pooling functions include average value in the adjacent rectangular regions, L2 norm, and weighted average function based on the center pixel distance. No matter what kind of pooling function it takes, the pooling can help the input approximation invariant when the input makes a small shift. The invariance of this translation is that when we make a small amount of shift on the input, most of the output after the pooling function does not change. Pooling can be seen as adding an infinitely strong priori: the function that is learned in this layer must be invariant to a small amount of translation. When this assumption is established, pooling can greatly increase the statistical efficiency of the network and reduce the storage requirements for parameters.

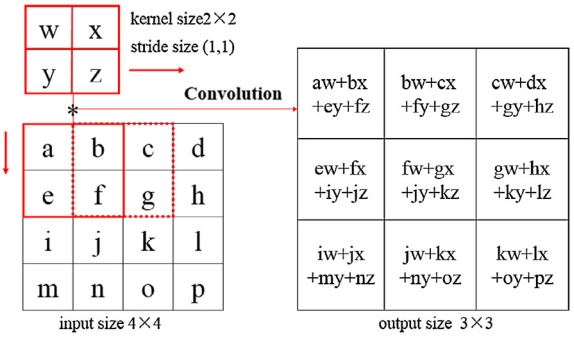


**Fig. 1.** Convolution calculation process.

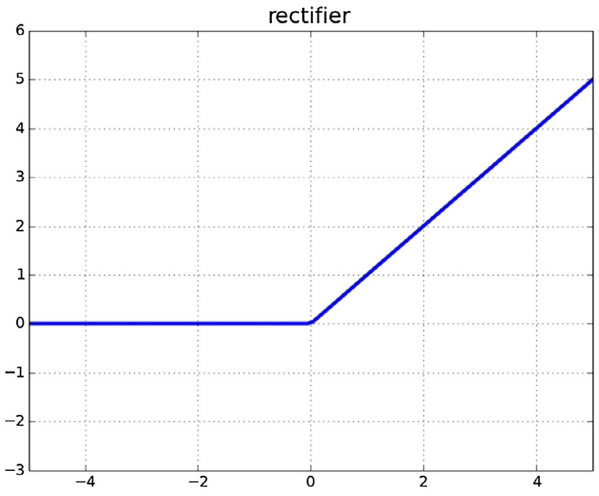


**Fig. 2.** The Rectified Linear Unit.

### 2.2. Dilated convolution

As an important means to extract features and reduce the amount of calculations, pooling plays essential role in CNN. For example, when FCN receives images [27] for a segmentation task, it runs the convolution and then pooling reduce the pixel size hence lower the resolution. However, since pooling the pixels is to simply truncate pixels and output pixel-wisely and the size of the pixel block is usually much smaller than the size of the entire image, pixel block size limits the size of the receptive field and thus only few local features can be detected,which hinder the network's performance. That is why pixel-wise networks (such as SegNet [28]) upsample the smaller image size after pooling to the original image size for prediction by deconvolution to relearn the missing pixels. Therefore, there are two key points in the image segmentation FCN, convolution-pooling to reduce the image size to dig abstract features, upsampling to expand the image size.

Although this downsampling is very successful in categorizing digits or iconic views of objects, information loss during the process has negtive effact on classification results, and heavily weakens the transferability of details in data. What's worse, if the object's signal is lost due to downsampling, there is little hope to reconstruct it during training. On the contrary, if high spatial

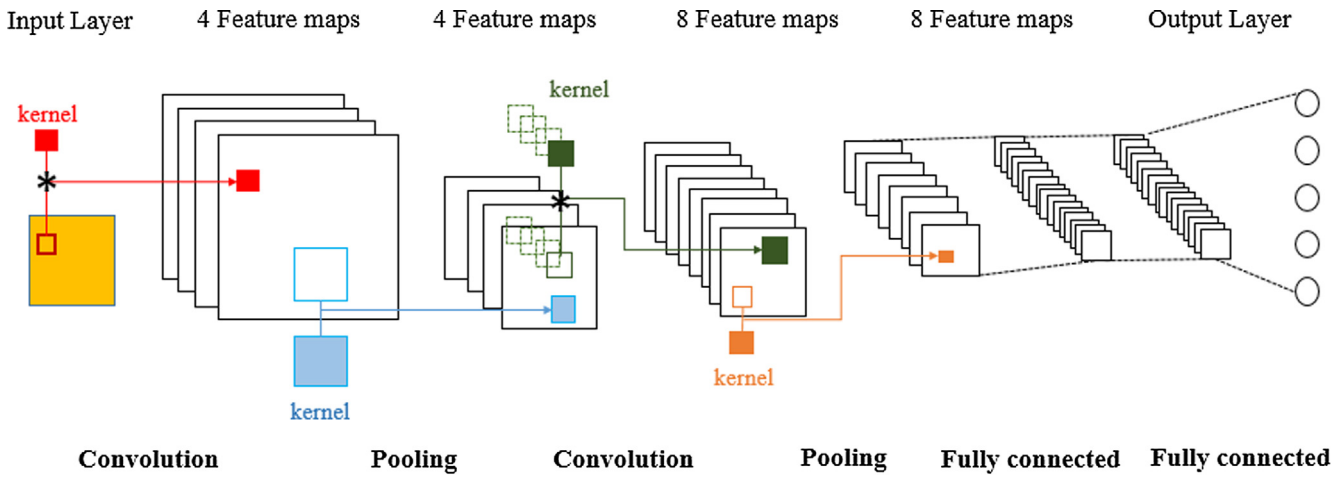


**Fig. 3.** Convolutional neural network.

resolution are preserved throughout the model and output signals densely cover the input space, back propagation can learn to nail down important information about smaller or less salient objects. So people question if there is an operation other than pooling that induces larger receptive fields to detect more information.

One answer is the dilated convolution [29]. It increases the receptive field of the higher layers, compensating for the reduction of the receptive field caused by the removal of sub-sampling. It is a rectangular prism of convolutional layers with no pool or subsampling. This module is based on dilated convolutions that supports exponential expansion of receptive fields without loss of resolution or coverage. The units in the dilated layers have the same receptive field as the corresponding units in original model. It also reduces the number of weights, thereby saves the computational cost.

Dilated Convolution(or Atrous convolution) was originally developed in algorithme *à* trous for wavelet decomposition [29]. It inserts a 0 in the convolution kernel to maintain the resolution of the network or to obtain a larger receptive field than the traditional convolution, thus avoid a downsampling (via the pooling or strided convolution) operation of deep CNNs. Dilated convolution has a hyper-parameter called dilation rate, which refers to the number of kernel intervals (eg, dilatation rate of normal convolution is 1).

Let F: $Z^2 \rightarrow R$ be a discrete function. Let $\Omega_r = [-r, r]^2 \cap Z^2$ and let k: $\Omega_r \rightarrow R$ be a discrete filter of size $(2r + 1)^2$. The discrete convolution operator $*$ can be defined as

$$(F * k)(p) = \sum_{s+t=p} F(s)k(t) \tag{3}$$

We now generalize this operator. Let l be a dilation factor and let $*_l$ be defined as

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)k(t) \tag{4}$$

We will refer to $*_l$ as a dilated convolution with dilation rate l, or an l-dilated convolution. The familiar discrete convolution $*$ is simply the l-dilated convolution.

The dilated convolution operator can apply the same filter to different ranges using different dilation factors, allowing us to skip the input value of a certain step in the calculation and apply the filter over an area larger than its length. This is equivalent to convolving with a larger filter obtained by dilating it with zeros from the original filter, but with significantly higher efficiency. A simple example is depicted in Fig. 4.

The number of parameters associated with each layer is the same. The number of parameters increases regularly with the growth of the receptive field. The system dilation supports expansion of the receptive fields without loss of resolution or coverage.

Dilation enlarges the receptive field or at any rate insures receptive field as in deep network like FCN, without loss of information in pooling or striding, such that the feature map includes adequate information since the size of each convolution output will guarantee to be no smaller than average convolutional structure. Dilated convolution, such as signal processing [29], object detection [30], audio generation WaveNet [31] and machine translation ByteNet [32], can be friendly applied to images that need global information or speech texts that require a long sequence information dependency.

### 2.3. Softmax distributions

Softmax Regression model is used in multiple classification tasks. It has the advantage of modelling conditional distributions, even when the data is implicitly continuous (such as the case for image pixel intensities or audio sample values) for a categorical distribution is flexible to simulate arbitrary distributions so that no assumptions about distributions are necessary. Literally, the output of multiple neurons is mapped to the (0,1) interval by the Softmax function. The sum of these output values is 1. The output node with the highest probability (that is, the value corresponding to the maximum) is picked as our prediction target in final selection. Suppose we have an array, $V_i$ represents the i-th element in V, then the element of the Softmax value is

$$S_i = \frac{e^{V_i}}{\sum_j e^{V_j}} \tag{5}$$

That is the ratio of the index to the sum of all the elements.

In order to train the model, we employ cross-entropy as the loss function when solving multi-classification problems. Defined as

$$L_i = -log\frac{e^{f_{y_i}}}{\sum_j e^j} \tag{6}$$

the value in the log is the Softmax value of the correct classification of this set of data. The larger the proportion, the smaller the Loss of this sample. This definition meets our requirements.

## 3. Model design for audio scene classification

In this subsection, we design a novel network architecture. The main idea is to replace the combination of pooling operation and traditional convolution with dilated convolution. As we all know,

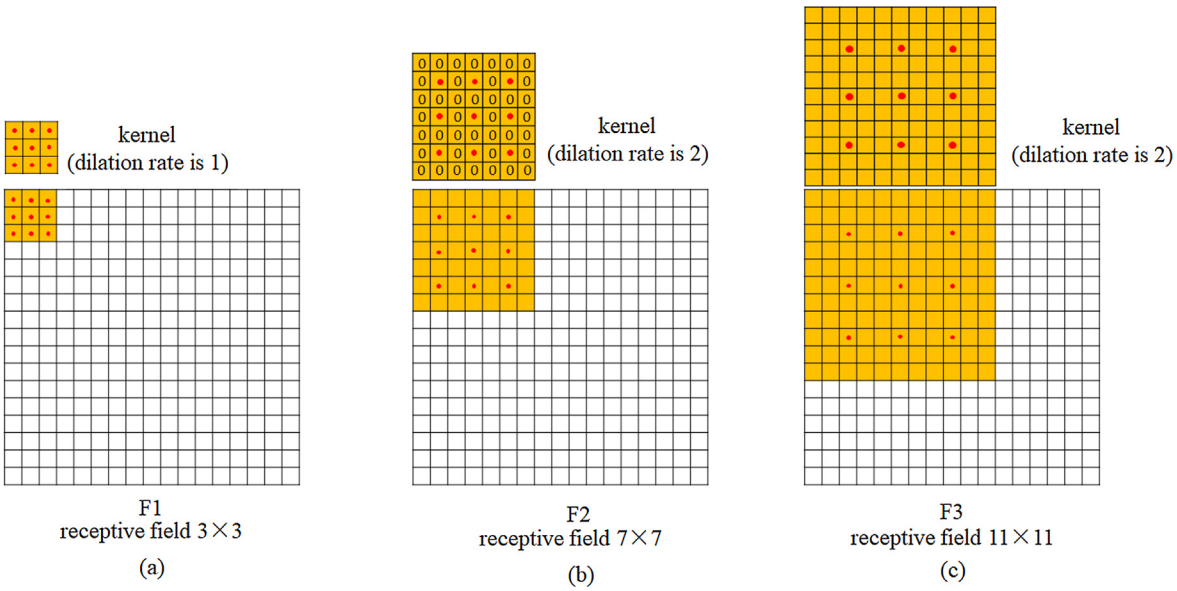


**Fig. 4.** Dilated convolution. (a): This is a regular 3 × 3 convolution. 1-dilated convolution yields F1 (the dilation rate is one). The receptive field for each position of F1 is 3 × 3. (b): Based on F1, perform a 2-dilated convolution (the dilation rate is 2), pay attention to its point-multiply position, not adjacent 3 × 3, get F2. The receptive field for each position of F2 is 7 × 7. (c): Based on F2, perform a 3-dilated convolution and get F3 (the dilation rate is three). The receptive field for each position of F3 is 11 × 11.

CNN causes to lost information due to pooling operation. Dilated convolution with dilation rate 2 can achieve the similar function to pooling with stride 2. Compared to pooling operation, after processed using dilated convolution, the data have the same size of feature map as before but the range of receptive fields increases. The detailed introduction is shown in the Fig. 5 and Fig. 6. Assume that the initial 5 feature size is 10 × 10. After pooling, the resulting feature map is greatly reduced to 5 × 5. And then during a convolution operation with a stride size of 1 and padding is same, although the obtained feature map remains unchanged (5 × 5), the size of the receptive field becomes 6 × 6. Therefore, through this traditional operation, the entire initial feature is reduced seriously, the receptive field is not big, and the information loss is serious. However, as shown in Fig. 4, if the original feature map directly passes through a dilated convolution with an dilation rate is 2, the receptive field will be 7 × 7 and the resulting feature map will be 10 × 10. That is, the size of the original features will not be reduced and the receptive fields will not be reduced. Therefore, after the dilated convolution is added to the multi-layer convolutional network, the number of parameters of the model does not change, the information will not have too much loss, and the receptive field will also become larger with the different dilation rate. So the dilated convolution can cover a wider range of information.

## 4. Experiment and analysis

Dilated convolution effectively allows the network to operate at a coarser scale than normal convolutions. This is similar to pooling and strided convolution, but here the output has the same size as the input.

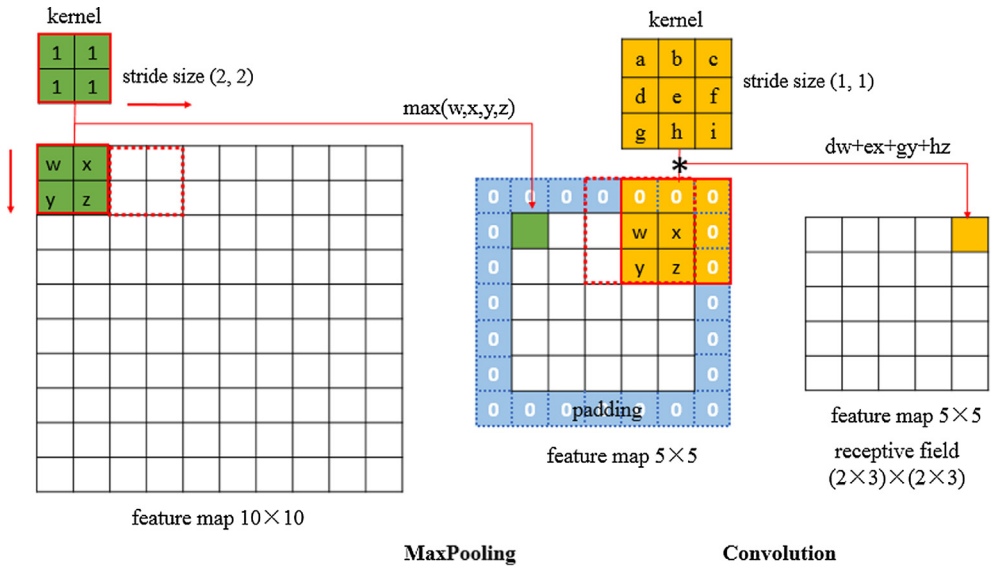


**Fig. 5.** Model architecture with pooling and convolution operations.

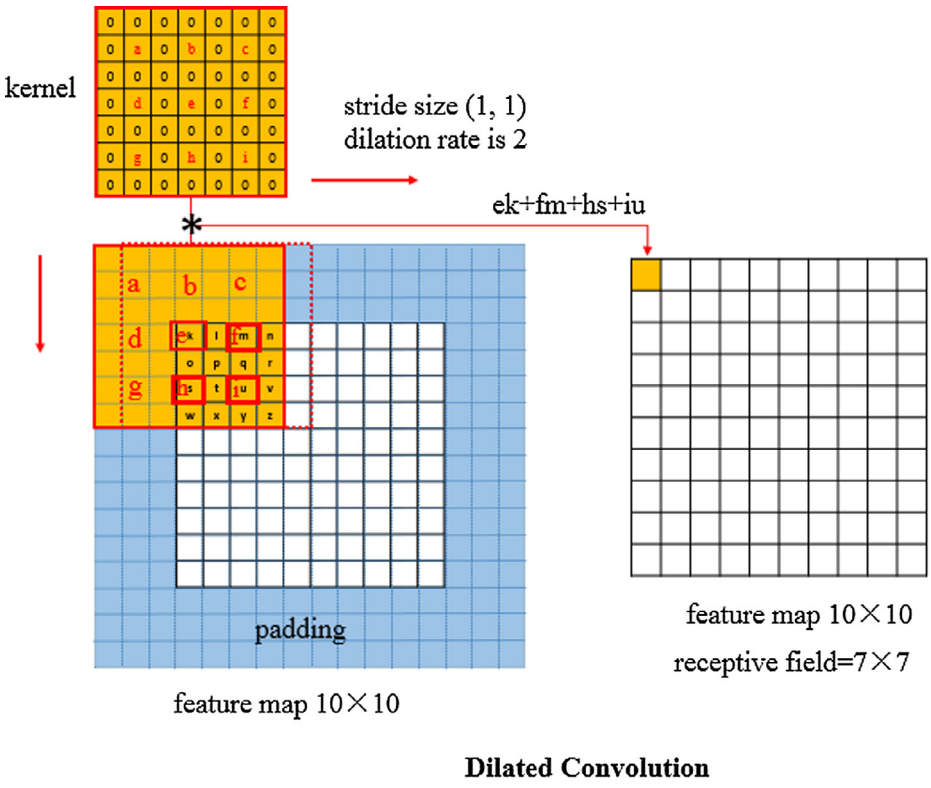


**Fig. 6.** Model architecture with dilated convolution.

### 4.1. Datasets

In order to examine our conclusion, we used the UrbanSound8K data set. It is widely used to verify the quality of the solution to the problem of environmental sound classification [33–36]. The data set is comprised of 8732 short (less than 4 s) excerpts of various urban sound sources extracted from field recording crawled from the Free-sound online archive. Based on data provided by the City of New York City's 311 service (more than 37,000 complaints from 2010 to date), these sources were selected from the Urban Sound Taxonomy [37], based on the high frequency with which they appear in noise complaints. Since these are real field-recordings,

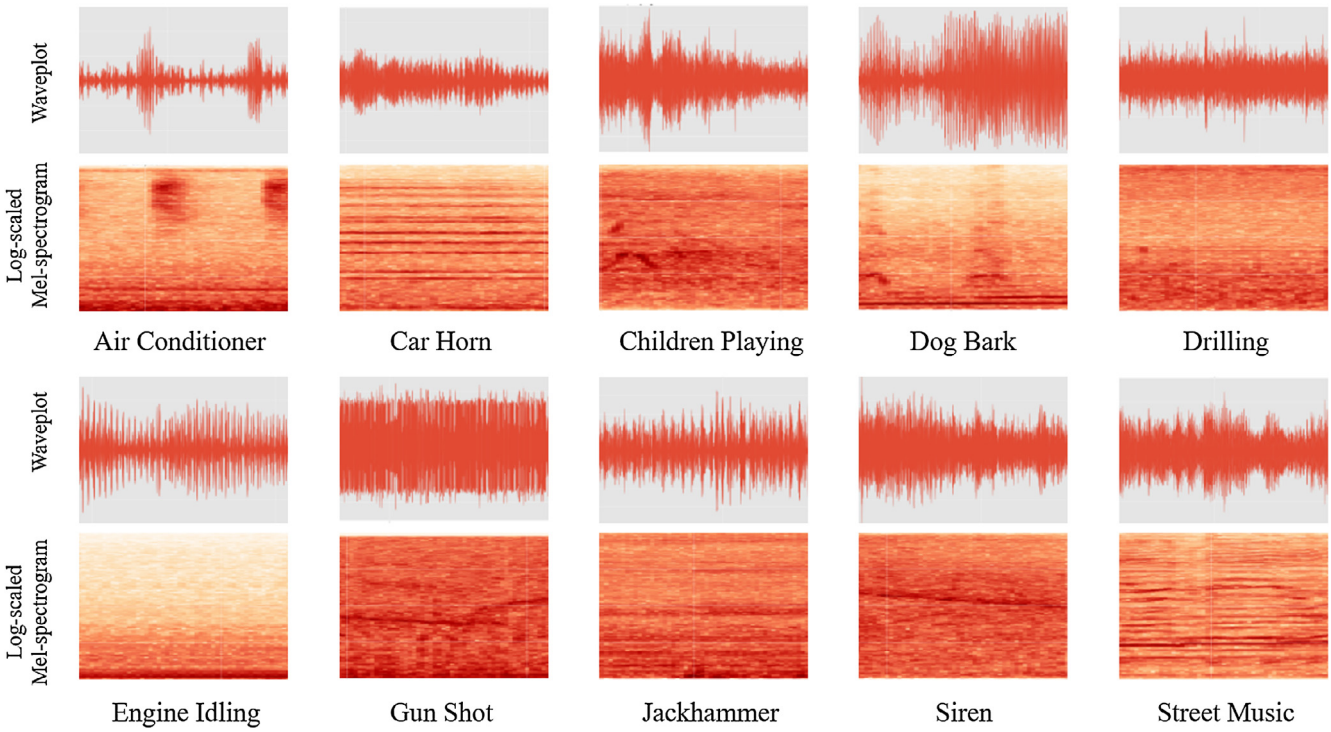


**Fig. 7.** Waveplot and Log-scaled Mel-spectrogram of each audio segment.

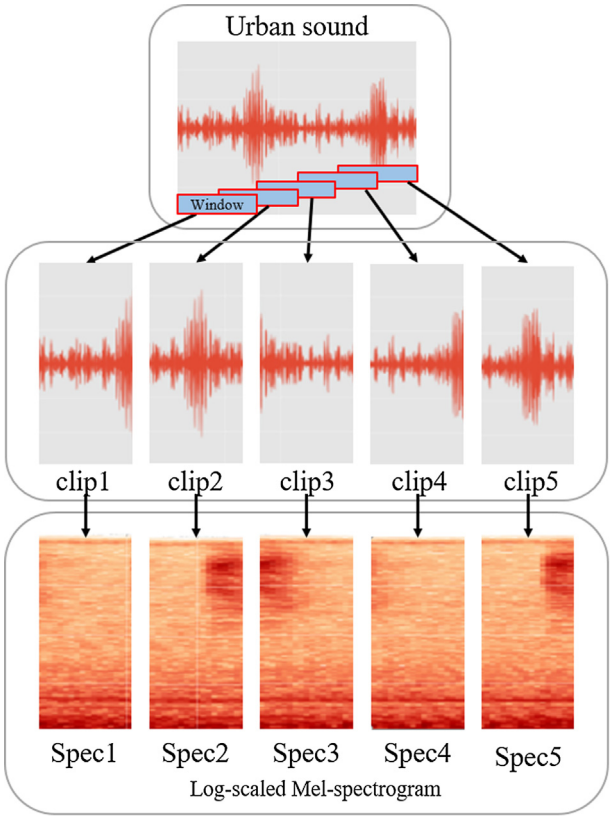


**Fig. 8.** The sound is divided by using a window function and log-scaled ratio of the mel-spectrograms being extracted.

it is possible (and generally) that there are other sources in the slice besides the labeled source.

All slices have been manually annotated with the source ID, and subjectively judged whether the source is in the foreground or background [38]. Each slice prearranged into one of 10 possible sound folds: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, street music. By looking at the plots shown in Fig. 7, we can see apparent difference between sound clips of different classes.

### 4.2. Experiment set up

We calculate the log-scaled mel-spectrograms and its corresponding deltas from the sound clip. First, all the urban sound are divided into frames, and the different length of the sound is divided into audio clips of the same size by using a window function (the default sampling rate is 22,050 Hz and normalized), with a log-scaled ratio of the mel-spectrograms being extracted. This process can be described in Fig. 8.

In this experiment we have chosen a window size of 1024, hop length of 512 and 64 n-mels, using the librosa implementation. Since the learning on whole clips was limited by the number of examples available for training, the spectrograms were split into 50% overlapping segments of 41 frames. So for fixed-size input, we split each sound clip into segments of $64 \times 41$ (64 rows and 41 columns). Also the labels (10 folds) were convert into one hot vector using one-hot-encode method. We use a method to extract the features and labels and save them in corresponding variables. The segments were provided together with their deltas (computed with default librosa settings) as a two-channel and will be input to different models. In all model designs, we use ReLUs activation function. Next, the output is flattened out for the Global Average Pooling layer input. Lastly, the Softmax layer is defined to output probabilities of the class labels. Because this experiment only compares the effects of different models, we only use the 50% drop-out layers to improve accuracy before the Softmax layer. This way big groups of neurons become helpful not only in the context of other neurons. Architecture averaging introduced by dropout tries to ensure that each hidden unit learns feature representations that are generally favorable in producing the correct classification answer.

The final system that was evaluated in detail can be described through the following process depicted in Fig. 9.

Our model is still in the form of convolution. For example, Cov1 represents the first conventional convolution structure, while DilaCov1(2) represents the dilation rate of the first dilated convolution is (2, 2). Convolution neural network training involves many decisions both the architecture (the format of the input data, the number and size of layers, the filter dimension) and learning hyperparameters (learning rate, momentum, and batch size). In

**Table 1**
Different convolution network structure comparison.

| | | Model structure |
|---|---|---|
| Group1 | Test1 | Cov1 + Cov2 + Cov3 + DilaCov1 |
| | Test2 | Cov1 + Cov2 + Cov3 + Max-pooling + Cov4 |
| Group2 | Test3 | Cov1 + Cov2 + DilaCov1 + DilaCov2 |
| | Test4 | Cov1 + Cov2 + Max-pooling + Cov3 + Cov4 |
| Group3 | Test5 | Cov1 + DilaCov1 + DilaCov2 + DilaCov3 |
| | Test6 | Cov1 + Max-pooling + Cov2 + Cov3 + Cov4 |

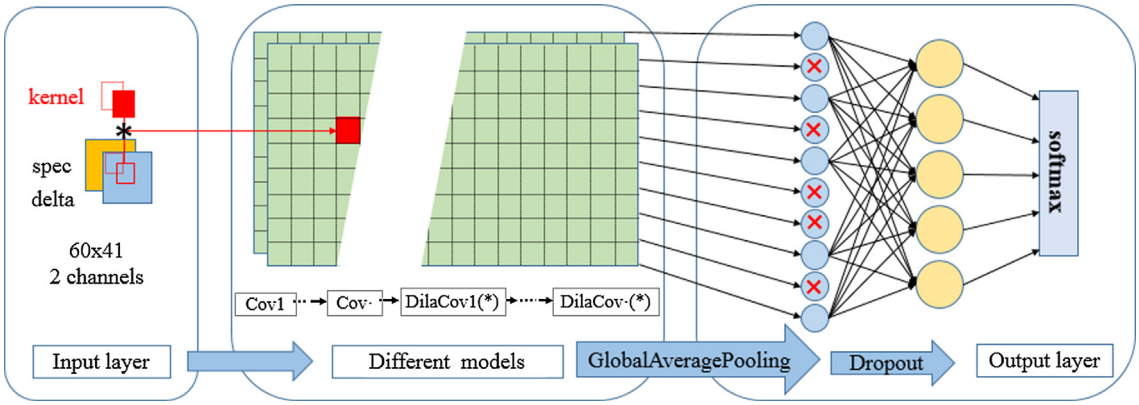


**Fig. 9.** The overall model structure of the experiment for the short sounds.

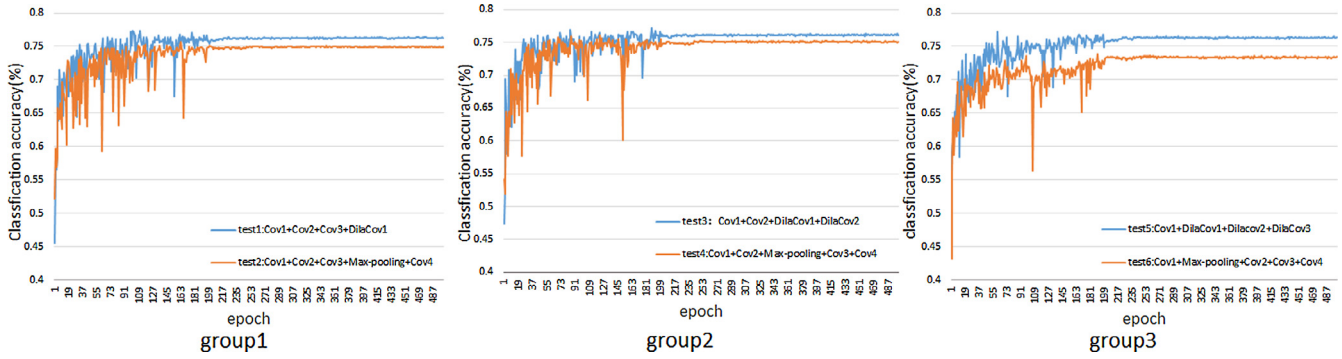


**Fig. 10.** Comparison of classification accuracy on the different convolution network structure. (the Cov1 means the first ordinary convolution, the DilaCov1 means the first dilated convolution).

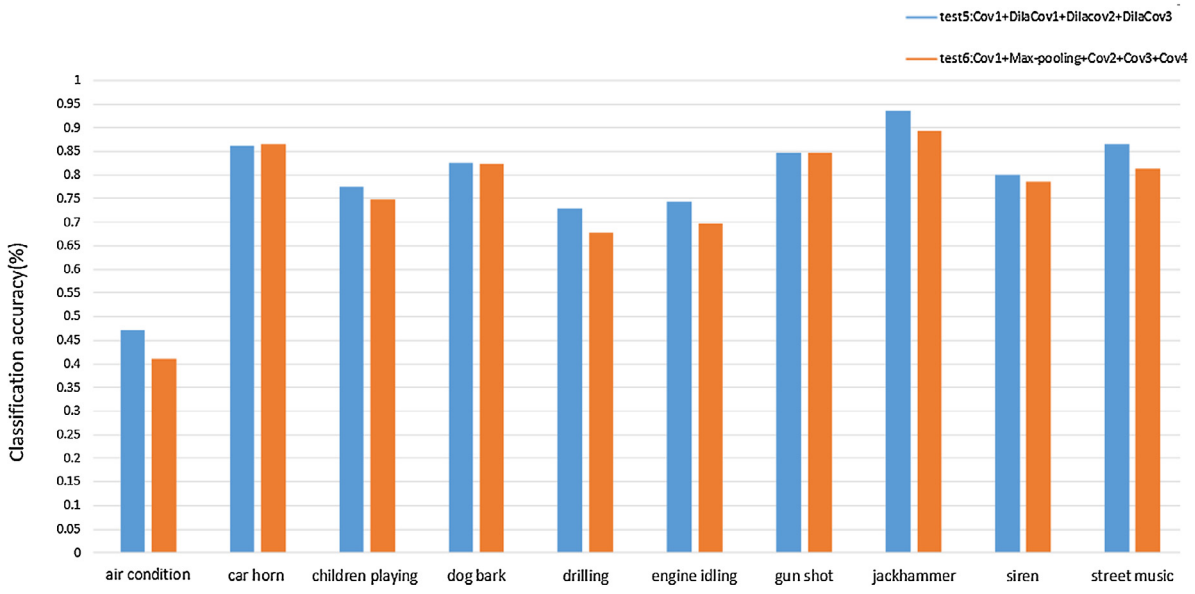


**Fig. 11.** The classification accuracy of the two models in group 3 on each folds.

this experiment, we use mini-batch stochastic gradient descent (SGD), with even the shuffled sequential batches (batch size is 32). Each layer of Nesterov momentum is 0.9 [39], the learning rate is 0.001. The training process was stopped after 500 epochs.

### 4.3. Experiment analysis

We designed three sets of experiments to examine our structure:

(1) Experiment 1: Different convolution network structure

The main purpose of this experiment is to compare conventional convolution structure equipped with pooling layer to dilated convolution. It contains three groups, each of which consists of two tests, test 1 with dilated convolution and test2 with combination of Max-pooling and convolution. All convolution kernels have a size of $3 \times 3$, stride is (1, 1), the Max-pooling size is $2 \times 2$, stride size is (2, 2). The dilation rate of all dilated convolutions is set to (2, 2). The model structures are listed in Table 1 and classification accuracy of the three groups are shown in Fig. 10.

In each group, we find that the model structure with dilated convolution is more accurate than the traditional convolution with max-pooling layer for urban sound, that is, 77% versus 75%. The classification accuracy of the two models in group3 on each fold folds can be seen in the Fig. 11. In fact, for feature extraction in ESC, dilated convolutions outperform the traditional convolution operation plus max-pooling operation, since urban sound is a complex collection of time-varying, short-term stationary signals carrying a large amount of effective information. When traditional convolution with max-pooling operation is exerted, the number of effective frames captured will be greatly reduced due to a heavy deterioration of the receptive field for original features caused by the pooling. In addition, it is difficult to sufficiently extracted sound signal features by adding window processing, because this process will include certain repetition. On the contrary, when dilated convolution is used, the size of the receptive field remains

**Table 2**
Dilated convolution with different dilation rate.

| | | Model structure |
|---|---|---|
| Group1 | Test1 | Cov1 + Cov2 + Cov3 + DilaCov1(2) |
| | Test2 | Cov1 + Cov2 + Cov3 + DilaCov1(3) |
| Group2 | Test3 | Cov1 + Cov2 + DilaCov1(2) + DilaCov2(2) |
| | Test4 | Cov1 + Cov2 + DilaCov1(3) + DilaCov2(3) |
| Group3 | Test5 | Cov1 + DilaCov1(2) + DilaCov2(2) + DilaCov3(2) |
| | Test6 | Cov1 + DilaCov1(3) + DilaCov2(3) + DilaCov3(3) |

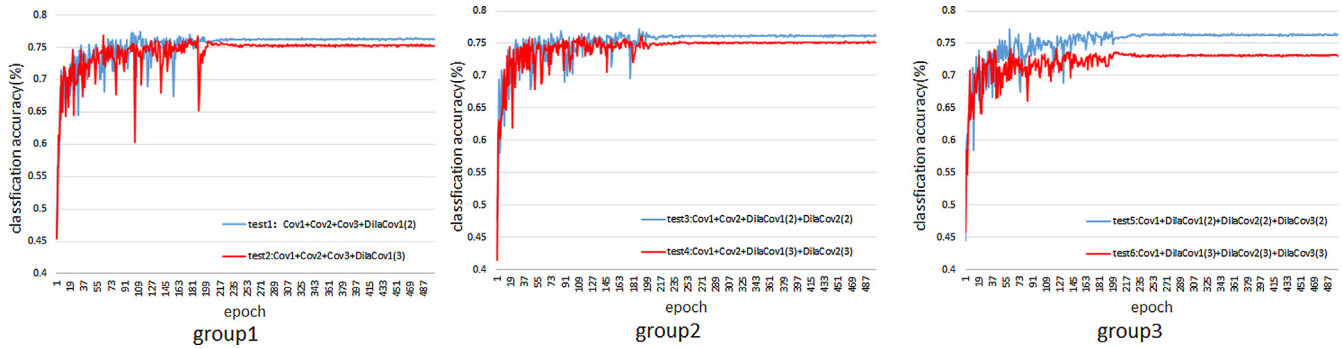


**Fig. 12.** Comparison of classification accuracy on the dilated convolution with different dilation rate.

unchanged or even increases, so that the operation can cover more frames, thereby fully extracting the sound features of the entire audio clip, and effectively improving the experimental effect.

(2) Experiment 2: Dilated convolution with different dilation rate

To explore the effect of changes in the dilation rate on the experimental results, based on the design of the dilated convolution structure in the experiment one, we designed three groups of comparative tests (in Table 2). The convolution structure used in all tests was dilated convolution. The size of the dilated convolution kernel is still $3 \times 3$ and the stride size is (1, 1). In each group, the two network structures are exactly the same (including the layers of traditional CNN and dilated CNN). However, in the two models designed in each group, the dilation rates of dilated convolutions are different. For example, DilaCov1(2) represents the dilation rate of the first dilated convolution is (2, 2), while DilaCov2(3) represents the dilation rate of the second dilated convolution is (3, 3). The classification accuracy can be seen in the Fig. 12.

From the comparative results of Experiment 2, we can find that the dilation rate of dilated convolution increases, the classification accuracy will decline rather than ascension. It is generally maintained at about 75%. It can be clearly seen from the experimental results that the dilated convolution can increase the receptive field, more effective frames can capture and more abundant sound signal features can be extracted. However, because the sound signal has time-varying features, its features change closely with time, it is difficult to effectively capture the precise features of sound signals over time. Therefore, when adopting dilated convolution, the type of "gridding" structure, enlarging dilation rate will make the whole network structure too sparse, not only a large amount of loss of information between adjacent frames, and will lead to the overall frame range too big. It is not good at extracting audio signal features change over time, resulting the classification effect worse.

(3) Experiment 3: Dilated convolution with different number of layers

In this experiment, we designed other third group comparative tests (Table 3) to explore the effect of the increase of the dilated convolutional layer on the accuracy of the urban sound classification. Also based on the model design of experiment 1, the internal structure design on each convolution in Experiment 3 is exactly the same (including convolution kernel size, stride size, and dilation rate). The only difference is the number of dilated convolution in each group.

By increasing the layers of dilated convolution, the effect of it on classification accuracy is observed. The result can be seen in the Fig. 13.

It can be clearly seen that along with the increase of the number of dilated convolution layers, the accuracy of the testing results decreased, from about 77% to less than 75%, same result as in Experiment 2. Too many layers lead to loose network structure and severe local information loss. In particular, in ESC, they destroy the precision of extracting the temporal features of the sound signal.

**Table 3**
Dilated convolution with different number of layers.

| | | Model structure |
|---|---|---|
| Group1 | Test1 | Cov1 + Cov2 + Cov3 + DilaCov1 |
| | Test2 | Cov1 + Cov2 + Cov3++DilaCov1 + DilaCov2 |
| Group2 | Test3 | Cov1 + Cov2 + DilaCov1 + DilaCov2 |
| | Test4 | Cov1 + Cov2 + DilaCov1 + DilaCov2 + DilaCov3 |
| Group3 | Test5 | Cov1 + DilaCov1 + DilaCov2 + DilaCov3 |
| | Test6 | Cov1 + DilaCov1 + DilaCov2 + DilaCov3 + DilaCov4 |

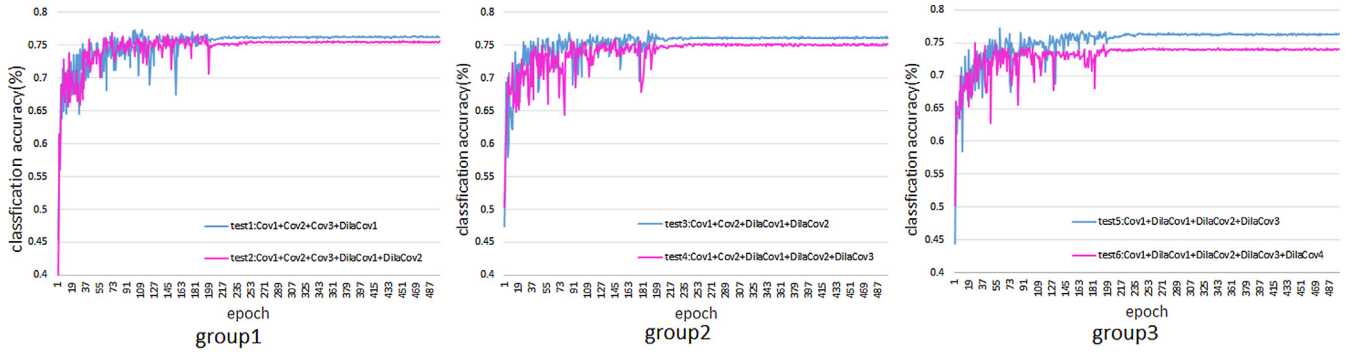


**Fig. 13.** Comparison of classification accuracy on the dilated convolution with different layers.

**Table 4**
Accuracy comparison with other methods.

| Method | Accuracy |
|---|---|
| Dilated convolution (ours) | 78% |
| Convolutional layers with max-pooling [25] | 74% |
| Unsupervised feature learning [38] | 73.6% |
| Long segments/majority voting [40] | 71.8% |
| Baseline system [37] | 68% |

### 4.4. Compared with other methods

Based on our data set, we compared our model with some other start-of-art methods for the environmental sound classification. For example, the work of Salamon and Bello [38] compared a baseline system to the unsupervised feature learning. The average classification accuracy of baseline obtained was 68% and that of the unsupervised feature learning is 73.6%. Karol J. Piczak et al. [25] also used a deep network consisting of 2 convolutional layers with max-pooling and 2 fully connected layers are trained on a low-level representation of audio data with deltas. And finally the model achieved an accuracy of around 74%. The result can be seen in Table 4. Through the results, we can find that our model is higher in classification accuracy than other advanced methods.

## 5. Conclusion and future work

We have tackled ESC problem with dilated CNN. Using this structure substitutes the traditional convolution operation with max-pooling results in higher accuracy of classification. At the same time, we explore the effect of different dilation rate and the number of layers of dilated convolution to the experimental results, because sound signal has short-term stability (immediate degeneration), carrying a lot of information and very complicated. Therefore, the frame is too large to obtain the features that the sound signal changes with time, and the frame is too small to extract the overall features of the sound signal in detailed. So, dilated CNN, being introduced to ESC problem, achieves better results than that of CNN with max-pooling, but expanding the number of covered frames or enlarging the dilation rate will make the accuracy reduce. We believe there is an inherent "gridding" defect in our dilation model: as the two pixels in the convolution kernel are padded with zeros, the receptive field of the kernel covers only the area with checkerboard patterns - samples only locations with non-zero values and loses some neighborhood information. The problem becomes worse as dilation rate increases, or when the dilated convolution is built on high layers that corresponds to bigger receptive field. Indeed, the convolutional kernel is too sparse to cover any local information because the non-zero values are too far apart. Because of the information that contributes to a fixed pixel always comes from its predefined gridding pattern, thus losing a huge portion of information. And excessive receptive field makes the frame is too large to obtain the characteristics that the sound signal changes with time. On the other hand, building too many dilated convolution layers causes the overall structure insufficiency for the training data since local information are seriously neglected. Our future work is to improve this model by carefully inspecting all parameters of dilated CNN and structural combination of in-depth inquiry.

## Acknowledgment

This study was supported by the National Natural Science Fund of China (Nos. 61672332, 61322211, 61432011, U1435212, 11671006, 61603173 and 61802238), the Program for New Century Excellent Talents in University (No. NCET-12-1031), the Program for the Outstanding Innovative Teams of Higher Learning Institutions of Shanxi, the Program for the Young San Jin Scholars of Shanxi, and the National Key Basic Research and Development Program of China (973) (Nos. 2013CB329404 and 2013CB329502).

## References

[1] González-Hernández Fernando Rubén, Sánchez-Fernández Luis Pastor, Suárez-Guerra Sergio, Sánchez-Pérez Luis Alejandro. Marine mammal sound classification based on a parallel recognition model and octave analysis. Appl Acoust 2017;119:17–28.
[2] Sánchez Luis P, Fernández Luis A, Pérez Sánchez, Carbajal Hernández José J, Ruiz Arturo Rojo. Aircraft classification and acoustic impact estimation based on real-time take-off noise measurements. Neural Process Lett 2013;38(2):239–59.
[3] Márquez-Molina Miguel, Suárez-Guerra Sergio. Aircraft take-off noises classification based on human auditory's matched features extraction. Appl Acoust 2014;84:83–90.
[4] Sánchez-Pérez Luis Alejandro, Sánchez-Fernández Luis Pastor, Suárez-Guerra Sergio, Carbajal-Hernández José Juan. Aircraft class identification based on take-off noise signal segmentation in time. Expert Syst Appl 2013;40(13):5148–59.
[5] Sánchez-Pérez LA, Sánchez-Fernández LP, Shaout A, Suárez-Guerra S. Airport take-off noise assessment aimed at identify responsible aircraft classes. Sci Total Environ 2016;542(Pt A):562–77.
[6] Barchiesi Daniele, Giannoulis Dimitrios, Dan Stowell, Plumbley Mark D. Acoustic scene classification: classifying environments from the sounds they produce. IEEE Signal Process Mag 2015;32(3):16–34.
[7] López-Pacheco MG, Sánchez-Fernández LP, Molina-Lozano H. A method for environmental acoustic analysis improvement based on individual evaluation of common sources in urban areas. Sci Total Environ 2014;468–469(468–469C):724–37.
[8] López-Pacheco María Guadalupe, Sánchez-Fernández Luis Pastor, Molina-Lozano Herón, Sánchez-Pérez Luis Alejandro. Predominant environmental noise classification over sound mixing based on source-specific dictionary. Appl Acoust 2016;112:171–80.
[9] So Stephen, Paliwal Kuldip K. Scalable distributed speech recognition using gaussian mixture model-based block quantization. Speech Commun 2006;48(6):746–58.
[10] Veisi Hadi, Sameti Hossein. Speech enhancement using hidden markov models in mel-frequency domain. Speech Commun 2013;55(2):205–20.
[11] Barkana Buket D, Uzkent Burak. Environmental noise classifier using a new set of feature parameters based on pitch range. Appl Acoust 2011;72(11):841–8.
[12] Ravanelli Mirco, Elizalde Benjamin, Ni Karl, Friedland Gerald. Audio concept classification with hierarchical deep neural networks. In: Signal Processing Conference. p. 606–10.
[13] Meysam Asgari, Izhak Shafran, Alireza Bayestehtashk. Inferring social contexts from audio recordings using deep neural networks. In: IEEE International Workshop on Machine Learning for Signal Processing; 2014. p. 1–6.
[14] Like Xue, Feng Su. Auditory scene classification with deep belief network. In Multimedia Modeling; 2015. p. 348–59.
[15] Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern 1980;36(4):193–202.
[16] Le Cun Y, Boser B, Denker JS, Howard RE, Habbard W, Jackel LD, Henderson D. Handwritten digit recognition with a back-propagation network. Adv Neural Inf Process Syst 1990;2(2):396–404.
[17] Chandrasekhar Vijay, Lin Jie, Morère Olivier, Goh Hanlin, Veillard Antoine. A practical guide to CNNS and fisher vectors for image instance retrieval. Signal Process 2016;128(C):426–39.
[18] Russakovsky Olga, Deng Jia, Hao Su, Krause Jonathan, Satheesh Sanjeev, Ma Sean, Huang Zhiheng, Karpathy Andrej, Khosla Aditya, Bernstein Michael. Imagenet large scale visual recognition challenge. Int J Comput Vision 2015;115(3):211–52.
[19] Dan Cireşan, Meier Ueli, Masci Jonathan, Schmidhuber Jürgen. Multi-column deep neural network for traffic sign classification. Neural Networks 2012;32(1):333–8.
[20] Deng Li, Abdel-Hamid Ossama, Dong Yu. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In: IEEE International Conference on Acoustics, Speech and Signal Processing. p. 6669–73.
[21] Cai Meng, Liu Jia. Maxout neurons for deep convolutional and LSTM neural networks in speech recognition. Elsevier Science Publishers B.V; 2016.
[22] Mitra Vikramjit, Sivaraman Ganesh, Nam Hosung, Espy-Wilson Carol, Saltzman Elliot, Tiede Mark. Hybrid convolutional neural networks for articulatory and acoustic information based speech recognition. Speech Commun 2017;89(C):103–12.
[23] Sander Dieleman, Philemon Brakel, Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In: International Society for Music Information Retrieval Conference, Ismir 2011, Miami, Florida, USA, October; 2011, p. 669–74.
[24] Dieleman Sander, Schrauwen Benjamin. Deep content-based music recommendation. Adv Neural Inf Processing Syst 2013;26:2643–51.

[25] Piczak Karol J. Environmental sound classification with convolutional neural networks. In: IEEE International Workshop on Machine Learning for Signal Processing; 2015. p. 1–6.
[26] Zhou YT, Chellappa R. Computation of optical flow using a neural network. In: IEEE International Conference on Neural Networks, vol 2; 2002. p. 71–78.
[27] Long Jonathan, Shelhamer Evan, Darrell Trevor. Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. p. 3431–40.
[28] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for scene segmentation. IEEE Trans Pattern Anal Mach Intell 2017;99:2481–95.
[29] Holschneider M, Kronland-Martinet R, Morlet J, Tchamitchian Ph. A real-time algorithm for signal analysis with the help of the wavelet transform. In: Wavelets. Time-Frequency Methods and Phase Space. p. 286–97.
[30] Dingqian Zhang, Hui Zhang, Haichang Li, Xiaohui Hu. RR-FCN: Rotational region-based fully convolutional networks for object detection. In: International Conference on Engineering Applications of Neural Networks; 2018. p. 58–70.
[31] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu. Wavenet: a generative model for raw audio. CoRR(abs/1609.03499); 2016.
[32] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron Van Den Oord, Alex Graves, Koray Kavukcuoglu. Neural machine translation in linear time. (arXiv:1610.10099); 2016.
[33] Piczak Karol J. ESC: Dataset for environmental sound classification. In: ACM International Conference on Multimedia; 2015. p. 1015–18.
[34] Ye Jiaxing, Kobayashi Takumi, Murakawa Masahiro. Urban sound event classification based on local and global features aggregation. Appl Acoust 2016;117.
[35] Salamon Justin, Bello Juan. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process Lett 2016(99). 1–1.
[36] Ting Wei Su, Jen Yu Liu, Yi Hsuan Yang. Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing; 2017. p. 791–95
[37] Salamon Justin, Jacoby Christopher, Bello Juan Pablo. A dataset and taxonomy for urban sound research. ACM; 2014. 1041–1044.
[38] Salamon Justin, Bello Juan Pablo. Unsupervised feature learning for urban sound classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing. p. 171–5.
[39] Bengio Yoshua, Boulanger-Lewandowski Nicolas, Pascanu Razvan. Advances in optimizing recurrent networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing. p. 8624–8.
[40] Nekruzjon Maxudov, Barş özcan, Furkan Kraç M. Scene recognition with majority voting among sub-section levels. In: Signal Processing and Communication Application Conference; 2016. p. 1637–640.