OPEN ACCESS Freely available online
X. Wang, Y. Xu, J. Shi, and X. Teng, "Acoustic Scene Classification Using Pixel-Based Attention"
J. Audio Eng. Soc., vol. 68, no. 11, pp. 843–855, (2020 November.).
DOI: https://doi.org/10.17743/jaes.2020.0052

Acoustic Scene Classification Using Pixel-Based Attention*

XINGMEI WANG[1],
(wangxingmei@hrbeu.edu.cn)

YICHAO XU[1],
(xuyichao@hrbeud.edu.cn)

JIAHAO SHI[1], AND XUYANG TENG[2]
(bjwgwgfy@hrbeu.edu.cn) (tengxuyang@hdu.edu.cn)

[1]*College of Computer Science and Technology, Harbin Engineering University, Harbin, 150001, People's Republic of China*
[2]*College of Communication Engineering, Hangzhou Dianzi University, Hangzhou, 310018, People's Republic of China*

In this paper, we propose a pixel-based attention (PBA) module for acoustic scene classification (ASC). By performing feature compression on the input spectrogram along the spatial dimension, PBA can obtain the global information of the spectrogram. Besides, PBA applies attention weights to each pixel of each channel through two convolutional layers combined with global information. In addition, the spectrogram applied after the attention weights is multiplied by the gamma coefficient and superimposed with the original spectrogram to obtain more effective spectrogram features for training the network model. Furthermore, this paper implements a convolutional neural network (CNN) based on PBA (PB-CNN) and compares its classification performance on task 1 of Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 Challenge with CNN based on time attention (TB-CNN), CNN based on frequency attention (FB-CNN), and pure CNN. The experimental results show that the proposed PB-CNN achieves the highest accuracy of 89.2% among the four CNNs, 1.9% higher than that of TB-CNN (87.3%), 2.2% higher than that of FB-CNN (86.6%), and 3% higher than that of pure CNN (86.2%). Compared with DCASE 2016's baseline system, the PB-CNN improved by 12%, and its 89.2% accuracy was the highest among all submitted single models.

0 INTRODUCTION

In recent years, as a subfield of acoustics, the problem of acoustic scene classification (ASC) has attracted the attention of many researchers. The task of ASC is to automatically analyze and recognize the environment in which acoustic signals are located, which is one of the most important and complex tasks in the field of computer auditory analysis [1]. Unlike traditional sound classification, such as speech recognition [2], ASC requires processing a wide range of frequency spectrum and solving the effects of various background noises [3].

In order to better complete the research on the problem of ASC, task 1 of Detection and Classification of Acoustic Scenes and Events (DCASE) 2016 Challenge provided a dataset and a well-designed test platform [4]. Task 1 of DCASE 2016 Challenge asks researchers to recognize the environment of each piece of audio. The dataset provided contains a total of 15 classes, such as cars, downtown, boulevards, and so on. The reason for choosing to study task 1 of DCASE 2016 Challenge is that the dataset provided is suitable in scale, and the results of the challenge and the relevant analysis of the year are disclosed, so that the performance of the proposed classification model can be evaluated under the same experimental conditions.

With the rise of deep learning, researchers try to complete the ASC tasks with the methods of deep learning, such as convolutional neural networks (CNNs) [5, 6], recurrent

*To whom correspondence should be addressed e-mail: wangxingmei@hrbeu.edu.cn

neural networks (RNNs) [7], and other popular deep learning models in recent years. For example, Mun et al. used the popular generative adversarial network (GAN) to expand the audio information to enhance the ASC effectiveness [8]. In 2017, CNN was used to complete ASC based on spatial feature extraction [9]. And a novel discriminative feature extraction for ASC using RNN-based source separation [10] was also proposed in 2017. In 2019, some obstacles under the concept drift framework were discussed, and the two fundamental adaptation approaches about active and passive conditions [11] were explored.

Although these deep learning methods show certain learning effects, the methods based on traditional acoustic models utilizing the traditional acoustic features, such as i-vector [12] and mel-frequency cepstral coefficient (MFCC) [13], still achieve the highest rankings in the challenge on DCASE in 2013. However, it is undeniable that the traditional acoustic model has poor anti-interference ability and that the model with autonomous learning ability based on deep learning, especially in resistance to intraclass noise and interclass noise, shows good performance.

Due to the development of deep learning and the expansion of datasets, the advantages of deep learning methods compared to traditional methods are obvious; more and more deep learning methods have been developed for ASC. Some of the most advanced ASC methods convert audio waveforms into time-frequency representations, which are called spectrograms, and the spectrograms are then used as input images to train models for the CNN [14]. In visual image classification, the target object is usually centered and occupies the main part of the image. But in the spectrogram classification, the acoustic event may only occur in a short time of audio recording and does not occupy a continuous image space (time period), so the event appears at intermittent times. For this feature of the ASC, many researchers have tried to classify acoustic scenes using attentional mechanisms.

In the 1980s, Treisman and Gelade proposed the ”characteristic integration theory,” which introduced the attention mechanism from biological research to focus attention on several vision features of images [15]. In 2014, Bahdanau proposed an attention-based ”Encode-Decode” machine translation model [16, 17] in the field of machine translation, which is used to solve the problem of source language alignment of different lengths in machine translation [18]. Since then, neural networks based on attention mechanisms have been applied to a variety of tasks, such as visual object classification [19], image caption [20], machine translation [18], speech identify [21], and other fields.

In recent years, the application of attention mechanism in image classification focuses on fine-grained classification tasks, which make the model more sensitive to some local feature data. For example, D. Wang et al. proposed a deep neural network model using hierarchical structure for fine-grained classification of images in 2015. The parallel unit construction in the model correspondingly matches the three classification criteria of the family and species of birds, and uses the attention mechanism in the area discovery module to make the model extract the feature information of different positions in the pixel matrix correspondingly, and output the category information of species after learning the fine-grained feature [22].

In 2017, J. Fu et al. proposed an attention mechanism applied on convolutional neural networks (RA-CNN). RA-CNN not only pays attention to global feature data but also strengthens the sensitivity of local features [23]. The squeeze-and-excitation (SE) network (SENet) structure proposed in 2018, using the SE-block-based attention mechanism, won the championship in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2017 competition [24].

However, due to the obvious difference between the spectrogram and the natural optical image (the attention mechanism based on the traditional vision on the ASC task with the sound spectrum as input), whether the final effect can be achieved still needs experimental verification. The attributes of spatial local feature information in real images cannot be reflected by the combination of time and frequency of the sound spectrum. Through the study of traditional acoustic models, the dominant acoustic signals that reflect the category characteristics in different scene events often have strong influence in a small range of one or a few frequency distributions. Obviously, the spectrogram also has a certain degree of local area characteristic information that has different effects on the task objectives, that is, the expected prediction classification results.

On the basis of considering the characteristic information in time and frequency dimension of the spectrogram, this paper proposes a pixel-based attention (PBA) module for ASC tasks. Taking into account the global information of each channel of the spectrogram, attention weights are applied to each pixel in order to extract more efficient spectroscopic features. Finally, experiments are carried out on the official dataset of DCASE 2016 Challenge, an international open-source dataset, to verify the effectiveness and adaptability of the CNN based on PBA (PB-CNN) proposed in this paper based on a PBA module.

The rest of this paper is organized as follows: Sec. 1 gives a detailed description of the attention mechanism, SE-block attention mechanism and the sub-spectrogram attention mechanism. Sec. 2 introduces the proposed PBA module and its calculation process in detail. Sec. 3 mainly shows the comparison results of the proposed PB-CNN, CNN based on frequency attention (FB-CNN), CNN based on time attention (TB-CNN), and CNN and gives the correlation analysis. Finally, Sec. 4 offers our conclusions.

1 ATTENTION MECHANISM

An attention mechanism can implement a focus operation based on data information. The attention vector is defined to estimate the degree of correlation between one element and other elements, and the weighted average of the attention vector is used as the approximation of the target. In the case of machine translation, traditional machine translation systems often rely on complex feature engineering based on textual statistical features and require a lot of engineering work to build the model. But the neural networks model

does not. In the neural networks model, the meaning of a sentence is mapped to a vector representation of fixed length, and the translation results are generated based on the vector representation. The neural networks model does not rely on n-gram and other traditional algorithms and attempts to extract higher layer features of the text. Hence, the neural networks model can be better generalized to new sentences, is more generalizable, and is easier to build and train, without any manual feature engineering. However, if a given complete sentence is very long, the effect of encoding complete information into a single vector for training and learning will be very poor. The pure neural networks model has defects in dealing with such long-range dependence.

Using an attention mechanism, it is possible to code a fixed length vector without using the complete original information. The model can selectively learn from the original information based on the historical input and features learned so far. Even what the model is learning can be interpreted and visualized. Inevitably, the attention mechanism sometimes has a certain price to pay. Compared with traditional models, attention mechanisms need to calculate the weight of attention between the specified element and other elements, which means that the number of parameters increases. Nevertheless, attentional mechanism greatly improve the performance of the model, especially in tasks that require better attention to information features at a local level.

$X = [x_1, \cdots, x_N]$ is defined to characterize the N input elements information. Considering resources such as the graphics processing unit and central processing unit, the model does not blindly process all N elements as input but instead picks out the elements associated with the specified task target as input to the model. The step of implementing the picking of related elements is done by specifying a query vector q, which is a vector associated with the task target that the current model needs to process. The attention variable $z \in [1, N]$ is defined to indicate the index position of the following elements of the selected element related to the current element. And $z = i$ is used to indicate that an element subscript i is currently selected as input information. After defining X, q, and z, next in the case of specifying X and q, the probability value α_i for selecting the element with the subscript i as the input is calculated by

$$\begin{aligned} \alpha_i &= p(z = i \mid X, q) \\ &= \frac{\exp(s(x_i, q))}{\sum_{j=1}^{N} \exp(s(x_i, q))}. \end{aligned} \tag{1}$$

where α_i is the attention distribution, $s(x_i,q)$ is the attention measurement function, selected by

$$s(x_i, q) = v^T \tanh(W x_i + U q) \tag{2}$$

$$s(x_i, q) = x_i^T q \tag{3}$$

$$s(x_i, q) = \frac{x_i^T q}{\sqrt{d}} \tag{4}$$

$$s(x_i, q) = x_i^T W q \tag{5}$$

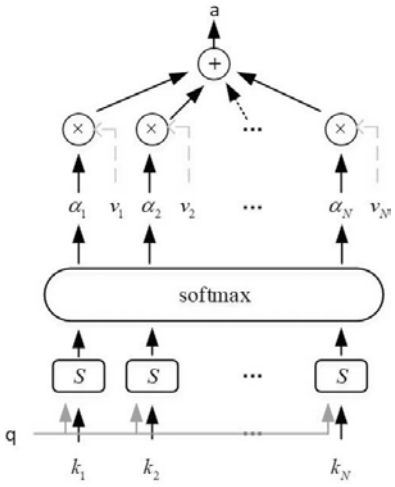


Fig. 1. Soft attention mechanism.

where W, U, and v represent the parameter values obtained by the model training, and d represents the input element dimension.

The attention distribution α_i is generally understood as the degree of association between the element with subscript i and the current element when querying the task target related to q during the process of model learning. The higher the degree of association, the more powerful the influence of the selected element on the current element has. The attentional distribution of the coded input elements given by

$$\begin{aligned} att(X, q) &= \sum_{i=1}^{N} \alpha_i x_i \\ &= E_{z \sim p(z \mid X, q)}[X] \end{aligned} \tag{6}$$

The attention mechanism represented by Eq. (6) is called soft attention mechanism, and its visualization is shown in Fig. 1.

With the further research on attentional mechanisms, more models based on attentional mechanisms have been proposed. The self-attention model proposed by the Google Machine Translation Team in 2017 has become a hot spot and has proven that it can effectively improve the effect of the model in various tasks [25].

CNN is commonly used in computer vision tasks. Although a convolution filter is good at exploring local information of image space, its receptive field may not be enough to cover a large pixel range. And if the size of the filter or the depth of the neural networks is increased, the model will be more difficult to train. Therefore, a self-attention mechanism can be applied to improve the convolution layer. Then, the convolution filter can extract the correlation features in the associated local region, and the model can improve the extraction performance of the subtle features.

$X = [x_1, \cdots x_N] \in \mathbb{R}^{d_1 \times N}$ is defined as the original element sequence, and $H = [h_1, \cdots h_N] \in \mathbb{R}^{d_2 \times N}$ is defined as the model output. And then we can use linear transformation to get the following formula.

$$Q = W_Q X \in \mathbb{R}^{d_3 \times N} \tag{7}$$

$$K = W_K X \in \mathbb{R}^{d_3 \times N} \tag{8}$$

$$V = W_V X \in \mathbb{R}^{d_2 \times N} \tag{9}$$

where Q, K, and V denote the query vector, the key vector, and the value vector, respectively. And W_Q,W_K, and W_V correspond to the parameter matrix of the iteration in the model training process.

The model output h_i is calculated by

$$\begin{aligned} h_i &= att\left((K, V), q_i\right) \\ &= \sum_{j=1}^{N} \alpha_{ij} v_j \\ &= \sum_{j=1}^{N} soft\max\left(s\left(k_j, q_i\right)\right) v_j \end{aligned} \tag{10}$$

where i,$j \in [1,N]$ represents the element subscripted i in X and the element subscripted j in H, and the connection weight α_{ij} is dynamically trained by model.

When the scaling dot product model is chosen to measure the attention distribution, the model output calculated by

$$H = V soft\max\left(\frac{K^T Q}{\sqrt{d_3}}\right) \tag{11}$$

where *soft* max characterizes the normalization function and performs column-by-column normalization operation on the input data.

The application of self-attention mechanism in practical tasks can replace the convolution operation or cyclic layer in the original model [25]. In this case, X is the output element of the network layer and can also be inserted into the model structure as a separate layer with convolution filtering operations or cyclic recursive operations. In addition, the attention weight α_{ij} generated by the self-attention model only depends on the degree of correlation between the two elements of q_i and k_j, and there is no record for the position subscript of the current element in the model input. Therefore, if self-attention mechanism need to be adopted separately, the position information based on subscript of elements is usually used in the model architecture to achieve the purpose of supplementation and correction [25].

1.1 SE-Block Attention Mechanism

The core idea of the attention mechanism based on SE-block [24] is to assign different attention weights to the feature map of each channel and train the models' ability to learn feature weights based on feature maps using back propagation of loss function values. Finally, the weights of feature maps with significant effect on classification tasks are larger, invalid, or ineffective, and the weights of feature maps with poor effect are smaller, which improves the training effect of the model.

The SENet structure uses SE-block to implement iterative training of feature weights. The SE-block structure is shown in Fig. 2.

The SE-block module can be applied to almost all current network structures. The schematic diagram of inserting the SE-block module into the Inception model structure and the ResNet model structure is shown in Fig. 3. However, due to the significant difference between natural optical images and spectrograms, the SE-block attentional mechanism proposed on the natural optical image dataset ImageNet is not able to extract the features of sounds in the spectrograms well in the field of ASC.

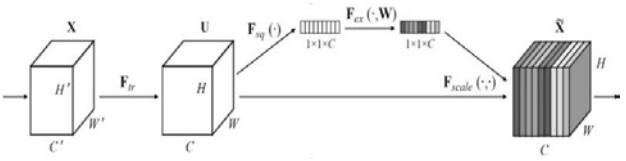


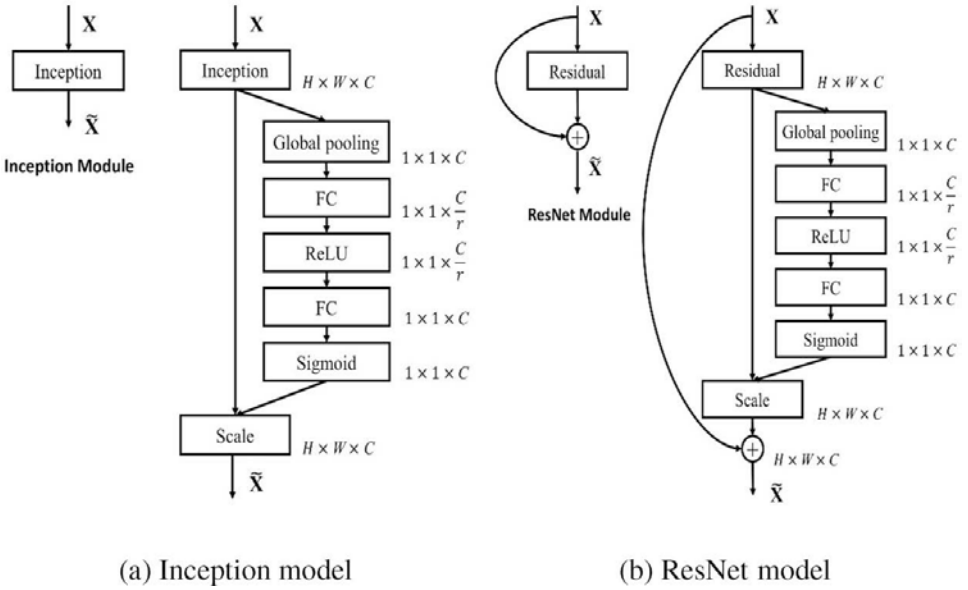
Fig. 2. Squeeze-and-excitation (SE)-block structure diagram.

1.2 Sub-Spectrogram Attention Mechanism

Natural optical images in real life have obvious local correlation characteristics in two spatial dimensions, but the spectrogram does not. The spectrogram object has a certain local relationship in the time dimension, and the correlation in the frequency dimension is fuzzy. For the sound of broadband spectrum similar to noise, there may be some local relationship of frequency dimension, while for the sound similar to harmonic, there is no local relationship to extract local features.

In the field of ASC, when the spectrogram is used as the input of the model, the model obtains the sound intensity information in the time-frequency range. In addition, through the study of traditional acoustic features, such as MFCC coefficient and normalized spectrum characteristics, if it is assumed that the background interference is poor, the category of homologous audio can be obtained, which has a strong correlation with the frequency distribution. Specific to the scene classification task, there is a specific frequency band that shows most activities, thus providing the main distinguishing function for this category. Combined with the idea of attention mechanism, attention distribution can be generated for each subfrequency band in the form of different weight coefficients, and the weighted average sum can be used as the influence factor to participate in the calculation of convolution filtering so as to represent the influence degree of a specific frequency band range on the predicted audio category. When it comes to the spectrogram of frequency intensity in the vertical direction, the high-dimensional characteristics of audio types can be found by extracting the features of the image in the vertical direction.

Based on this idea, in this paper, the spectrogram converted from the original audio was cut and clipped by means of the data processing method of marking the molecular frequency band on the spectrogram, the overlap percentage was adjusted to compensate for the features between the bands, and the frequency band level difference was introduced into the training process of the model. After the training, when the test data were used to verify the model effect, the molecular frequency band was also delimited. Finally, the category with the highest normalized probabil-



(a) Inception model (b) ResNet model

Fig. 3. Examples of networks structure application squeeze-and-excitation (SE)-block module. FC: fully connected layers, its function is to complete linear mapping. ReLU: Rectified Linear Unit, an activation function. ResNet: A convolution neural network model.

ity was selected as the final prediction result among all the prediction results.

The sub-spectrogram module can be chosen to insert after a convolutional layer. At this time, the input is the three-dimensional tensor matrix X of the feature map output after the convolution operation, assuming its dimension is $C \times H \times W$. Among them, C represents the number of feature maps, that is, the number of channels, and is also the number of convolution kernels of the previous layer convolution operation. H characterizes the height of a single feature map, namely, the frequency range of the audio in the high-dimensional space. W represents the width of a single feature map, which in the current task is the time scale of the audio in a high-dimensional space. And the overall structure of the model is shown in Fig. 4.

As shown in Fig. 4, the first transformation is as follows

$$G_{tr} : X \rightarrow F, X \in \mathbb{R}^{B\times C\times H\times W}, F \in \mathbb{R}^{B\times 1\times H\times 1} \quad (12)$$

This paper calculated the average value on H (it can also be understood as the global average pooling process δ_1 of size $1 \times W$), and the dimension of the output three-dimensional tensor U is $C\times H\times 1$.

Then, on the channel, a global average pooling process δ_2 of size $C \times 1$ is performed, and the dimension of the output three-dimensional tensor O is $1 \times H \times 1$. Define the batch size used in the model training process as B, and then, the dimension of the final output F is $B \times 1 \times H \times 1$. The calculation formula is defined by

$$F = \delta_2 (\delta_1 (O)) \quad (13)$$

Next, the obtained tensor F is transformed by the fully connected layer to generate an attention distribution α,$\alpha \in \mathbb{R}^{B\times 1\times H\times 1}$.

Finally, the original input X and α do the matrix bitwise multiplication operation to generate the output $\tilde{X}$ of the sub-spectrogram operation (the dimension is $B \times C \times H \times W$) and complete the weighted assignment operation of the attention thought, which achieves the frequency distribution of the high-dimensional feature map of the sound spectrum. The scope focuses on the purpose of attention.

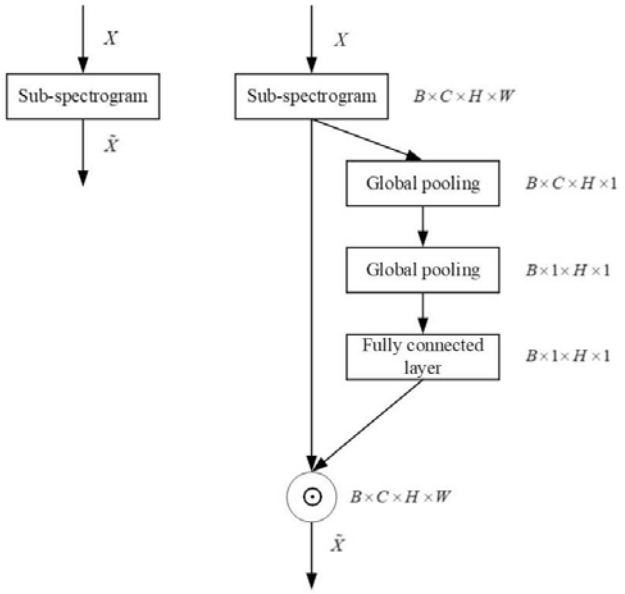


Fig. 4. Sub-spectrogram model structure.

Due to the frequency domain characteristics of audio signal that affect the classification of categories in the frequency domain, the sub-spectrogram attention mechanism can focus on the frequency-domain characteristics of the spectrogram, generate attention distribution, and make the frequency-domain attention vector act on the frequency dimension through transformation operation, achieving the purpose of paying active attention to the local characteristics of frequency. However, in the transformation operation based on the sub-spectrogram, after the global average pool-

ing δ_1 of the size of $1 \times W$ in the first step, the continuity features of the spectrogram on the time scale are directly discarded and not retained for further feature extraction in the subsequent steps. Moreover, the audio of environmental scene events has a certain degree of continuity in time scale. This part of the features is obviously of certain reference significance for ASC tasks, and it is not suitable to simply abandon them completely, resulting in waste of local features and weakening the model learning effect.

2 PROPOSED PBA MODULE

Based on the analysis of the attention mechanism of SE-block and sub-spectrogram, we are inspired and proposed PBA module. The method aims to focus on the feature distribution of pixels on each channel through the global information of each channel of the spectrogram, and combine with the original spectrogram to construct an ASC task network model based on the attention mechanism. The structure of the PBA module is shown in Fig. 5.

Suppose the module input is the feature map tensor X, C is the number of channels, H is the height of the single feature map, W is the width of the single feature map, and if the current batch size is B, the tensor X is $B \times C \times H \times W$.

First, a convolution operation of size c is performed on X. The number of convolution kernels is the number of channels C of the original input, and the output W_v has the same dimension as the input X, and the storage W_v is reserved.

Next, set the number of convolution kernels to 1, and recontract the size 1×1 on the input X. At this time, due to the change in the number of convolution kernels, the output is denoted as W_k, and its dimension is $B \times 1 \times H \times W$. This step implements the compression operation on the feature map channel dimension while preserving the continuity feature of the sound spectrum map in the high-dimensional transformation based on the time scale.

The nonlinear output operation is performed on the W_k of the second convolutional layer using a rectified linear unit (ReLU), and the final dimension is unchanged. This step adds nonlinear factors through the activation function ReLU, thereby achieving the purpose of improving the feature representation ability of the Pixed-ConvLayer module.

After calculating W_v and W_k separately, perform a bitwise multiplication of the matrix on both. Bitwise multiplication means that the elements at the same position of two matrices are multiplied by two to get a new matrix value. This step is calculated by

$$W_m = W_v \odot W_k \tag{14}$$

where the dimension of the output tensor W_m is $B \times C \times H \times W$. The original input tensor X is transformed using the SE-block module to achieve channel-based attention focusing, and the output tensor W_s has the same dimension as X. There are some operations required. First, perform a data transformation to convert the original image into a tensor in Eq. (15),

$$F_{tr} : X \to U, X \in \mathbb{R}^{W' \times H' \times C'}, U \in \mathbb{R}^{W \times H \times C} \tag{15}$$

where F_{tr} is the convolution operation, and the specific formula given by

$$u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * X^s \tag{16}$$

where v_c represents the cth convolution kernel and X^s represents the data of the sth input. After transformation, the three-dimensional tensor is obtained, which can be understood as C feature maps with scale $H \times W$, and u_c represents the first c two-dimensional matrix in the three-dimensional tensor U, that is, the feature map.

Subsequent squeeze operation can be understood as a global average pooling operation, given by

$$z_c = F_{sq}(u_c) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} u_c(i, j) \tag{17}$$

After the squeezing operation, the three-dimensional tensor U of size $W \times H \times C$ obtained in the previous step is transformed into z_c of size $1 \times 1 \times C$, which corresponds to the $F_{sq}()$ operation in Fig. 5, which obtains the numerical distribution of all the characteristic maps. After the squeezing operation is completed, it can be seen that the SE-block realizes the task of obtaining the numerical distribution of the characteristic features of the global feature map through the squeezing operation.

The following is the excitation operation, given by

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{18}$$

First, the result of the extrusion operation is through the fully connected layer W_1—the dimension of W_1 is $(C/r) \times C$, in which $r$represents the scaling factor, generally taking $r = 16$—and the number of channels can be reduced by the scaling factor, that is, the feature dimension is reduced, thereby greatly reducing the introduction of SE-block. The amount of calculation of the loss after the block module. Since the dimension of z is $1 \times 1 \times C$, the dimension of $W_1 z$ after the fully connected layer W_1 is $1 \times 1 \times (C/r)$. Then, the nonlinear transformation is performed by the linear rectification function to improve the fitting generalization ability of the data, and the dimension of the data is unchanged. Immediately after passing through another fully connected layer W_2, the dimension of W_2 is $C \times (C/r)$, and at this time, $W_2 \delta(W_1 z)$ is the output, and its dimension is $1 \times 1 \times C$. Next, the data normalization operation is performed by the sigmoid function σ to adjust the range of weight values, and the output of the operation is s.

It can be seen that the dimension of s is $1 \times 1 \times C$, where C represents the number of feature maps, that is, the number of channels. In the SENet architecture, s is used to characterize the weight distribution of C feature maps in the three-dimensional tensor U. Since the weight distribution s comes from the training of the fully connected layer and the nonlinear layer in the SE-block, end-to-end training can be achieved.

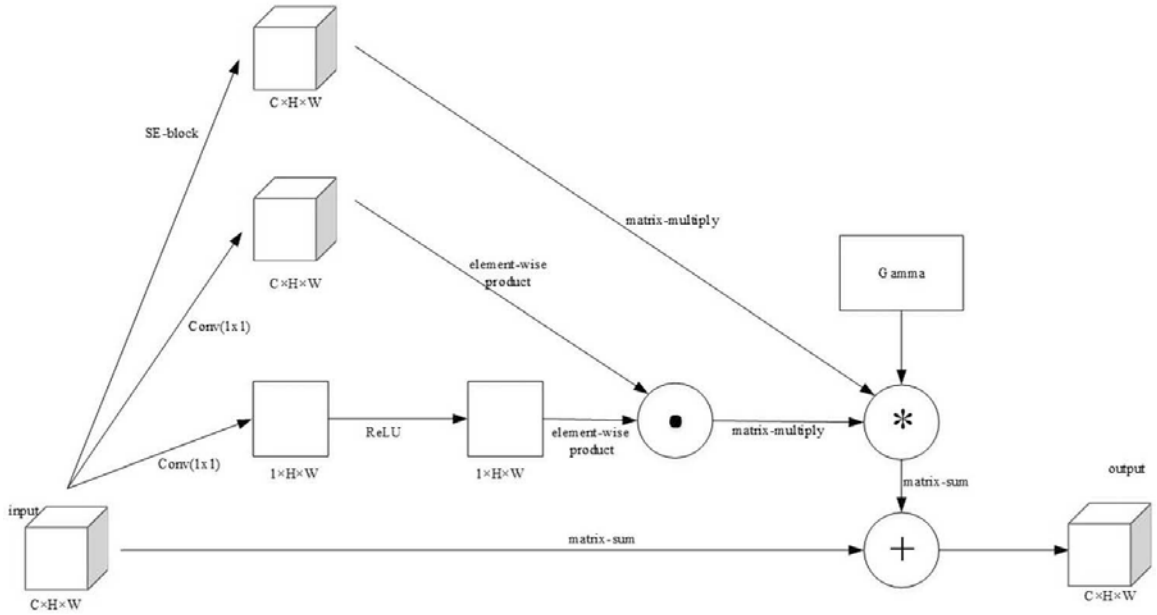


Fig. 5. The structure of the pixel-based attention (PBA) module. SE, squeeze-and-excitation. ReLU: Rectified Linear Unit, an activation function.

Next, the contraction coefficient Gamma, $Gamma \in [0, 1]$, is introduced to characterize the intensity of the influence distribution weight on the original model. The calculation is given by

$$\tilde{X} = X + Gamma * (W_v \odot f(W_k)) * W_s \tag{19}$$

where f is a nonlinear ReLU activation function. The operation of the activation function enhances the nonlinearity of the original structure and can better extract the task-related features in the training data. The generation process of output $\tilde{X}$ of PBA modules can be understood as the enhancement of feature extraction effects of spectral frequency domain distribution and time continuity on the basis of ReLU and combining the global information of each channel of the spectrum so that the output after convolution operation can more reasonably express the category characteristics of spectrogram. In addition, The application of the PBA module weights the generated attention distribution and merges it with the original input, taking into account the feature distribution of each dimension of the input data, thereby extracting as many effective features as possible of the spectrogram and reducing unnecessary information omissions. Compared with the traditional nonlinear activation function, it has more advantages in ASC tasks.

The basic classification model uses a traditional convolutional layer and a maximum pooling layer to construct a convolutional neural network for ASC tasks. The structure of the proposed PB-CNN is shown in Fig. 6. The rounded rectangle in the figure represents a set of operations, such as convolution, maximum pooling, etc. Each ellipse represents an operation, such as add, mul, each curve represents the flow direction of the data tensor, and each circle represents a constant.

As shown in Fig. 6, this paper uses four PBA modules. Before the data flows into each PBA module, the input needs to go through the convolutional layer and the batch normalize layer. Each PBA module performs global average pooling on each channel of the input, obtains global information for each channel, and then applies weights to each pixel on each channel based on this global information. The latest feature map is multiplied by the gamma coefficient and superimposed with the original feature map to flow into the ReLU activation function.

3 EXPERIMENT RESULTS

In this section, we first present the dataset we considered and describe the experimental setup. We then compare the effects of spectrogram generated by mel filters of different wavebands on the experimental results. We use the baseline convolutional neural network (Baseline), referring to the model proposed by Valenti et al., which won sixth place in DCASE 2016 Challenge, as the baseline network. Next, we compare the performance of the proposed PB-CNN, TB-CNN, FB-CNN, and CNN on ASC tasks. And finally, the performance of the proposed PB-CNN is analyzed in detail.

3.1 Experimental Setup

In this section, we describe the dataset used for analysis and detail the experimental setup.

3.1.1 Dataset

Our experiment was conducted on the DCASE 2016 dataset ASC [4]. The DCASE 2016 development dataset contains 1,170 audio files, including 15 different indoor and outdoor scene classes: beach, buses, cafes/restaurants, cars,

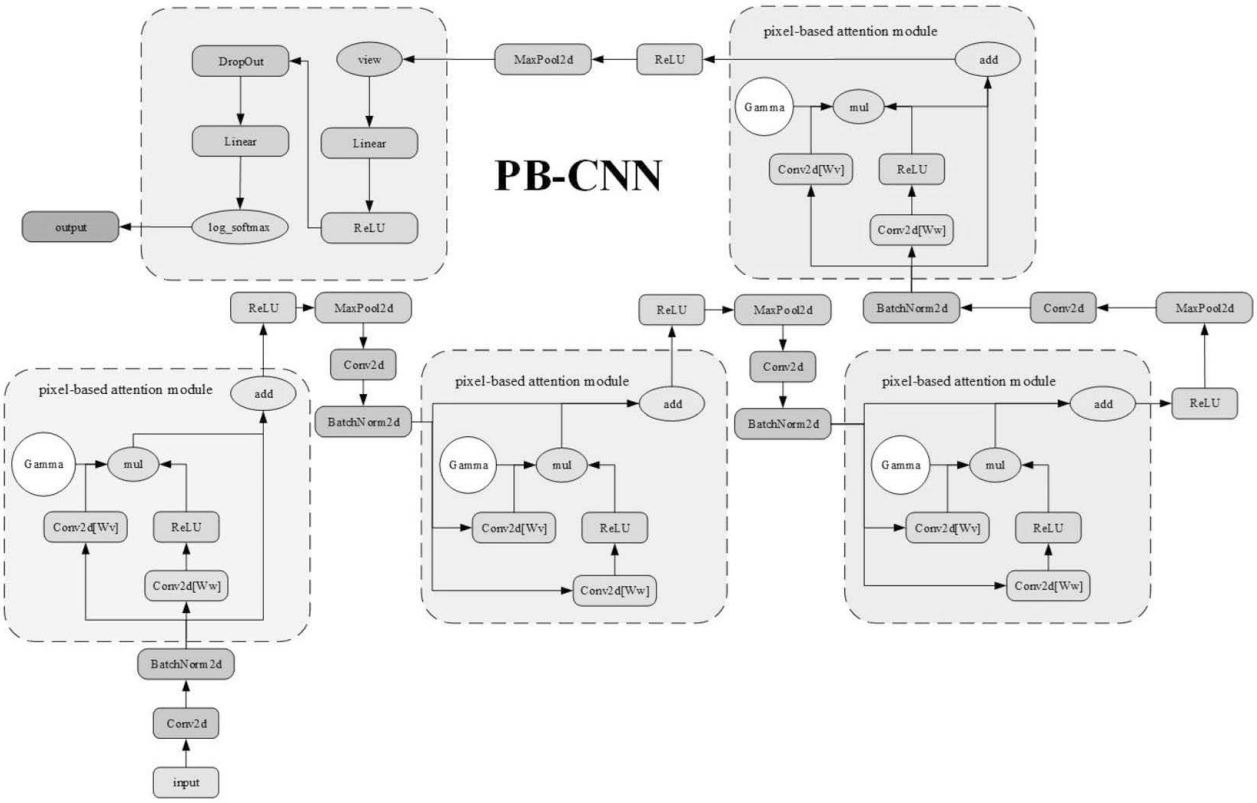


Fig. 6. The structure of the proposed convolutional neural network (CNN) based on pixel-based attention (PBA) (PB-CNN) ReLU: Rectified Linear Unit, an activation function.

downtown, boulevards, grocery stores, homes, libraries, subway stations, offices, city parks, residential areas, trains, and trams. Each field contains 78 samples, each of which was recorded for 30 s. For all sound files in the dataset, a sampling rate of 44.1 kHz and resolution of 24-bit depth are adopted for recording, and microphone earphones specially made to be worn on the ear are used for recording so as to achieve an effect similar to the human auditory system. After using the settings provided by DCASE 2016, the development set was further divided into four train and test sets.

For each fold, calculate the precision of each scene class I, where $I = 1,2...K$ and K are the number of classes. The calculation accuracy is the total number of correct scenarios divided by the total number of test scenarios. The precision of each fold is obtained by averaging the precision of 15 scene classes. Finally, the overall accuracy is evaluated by averaging the quadrature accuracy.

3.1.2 Feature Expression and Preprocessing

Firstly, the Hamming window (with 40 ms size and 50% frame shift) is used to slide on the audio signal. The short-time Fourier transform is applied to the audio signal. Then, the absolute value of each frequency bin is calculated and squared to get the power spectrum. Then, the four groups of processed datasets are obtained through mel scale filters of 60, 128, 256 and 512 bands, respectively. The logarithmic mel spectrogram is obtained by logarithmic transformation.

Preliminary manual features use Python as well as the librosa library for audio analysis. The original audio format is two-channel recording. In this paper, the two-channel audio is simply synthesized into a monophonic signal by taking the mean value of left and right two-channel for subsequent processing.

After the initial features are calculated, the features are normalized by subtracting the mean value and dividing by its standard deviation. The mean and standard deviation are calculated by pytorch throughout the train set. In the train set, the normalized logarithmic mel spectrogram is divided into shorter spectrogram segments.

In this experiment, a complete spectrogram is divided into about 3-s spectrogram, and there is 50% overlap between the spectrogram and the spectrogram, which provides about 22,000 training data for the model.

3.1.3 Aggregate Classification Result

Since the spectrogram generated by the original audio is divided into multiple spectrograms, each spectrogram will get a classification result through the classifier, so the classification result of the whole audio is obtained by aggregating all the classification results of the spectrogram generated by its segmentation. For the i-th spectral segment, the classification results are as follows

$$C_{segment} = \arg\max_{j} P\left(y^{(i)} = j\right) \tag{20}$$

Table 1. Classification accuracy (%) of using different mel filters bands for training baseline.

Mel filters bands	Fold 1 accuracy	Fold 2 accuracy	Fold 3 accuracy	Fold 4 accuracy	Average accuracy
60	85.862	85.172	84.828	87.241	85.776
128	86.897	86.207	85.172	87.931	86.552
256	86.207	85.172	85.862	87.241	86.121
512	86.897	85.172	84.828	87.931	86.207

Then, the final classification result of the whole audio C can be calculated as follows

$$C = \arg\max_{j} \sum_{i=1}^{M} \log\left(P\left(y^{(i)} = j\right)\right) \tag{21}$$

where M is the total number of sound spectrum segments divided into sound spectrum segments generated by the whole audio frequency.

3.1.4 Experimental Environment

All deep learning models in this paper were constructed and trained by pytorch1.0.1, and all models were trained by an Nvidia Tesla V100 graphics card. Compute Unified Device Architecture (CUDA) version 9.0 is adopted, cudnn version 7.4.2 is adopted, and all programs are implemented by Python 3.6, running in Ubuntu 16.04 Long Term Support (LTS). In addition, in this experiment, the librosa library based on Python is used for reading and preprocessing the original audio files in the dataset.

3.2 Experimental Results and Analysis

In order to verify that the proposed PB-CNN has a good effect on optimization of the ASC task, we first used the Baseline, which is a two-layer convolutional neural network structure, and achieved sixth place in the DCASE 2016 Challenge with an accuracy rate of 86.2%. The model is taken as the baseline network and the result as the basic evaluation standard.

We used the pytorch framework to build and train various deep learning models proposed. The adaptive momentum (Adam) algorithm was selected for the optimization algorithm, and the loss function was categorical cross-entropy. The average fourfold cross-validation accuracy of the reconstructed Valenti et al. CNN in generating datasets on 60, 128, 256, and 512 bands of mel filters is given by Table 1.

In Table 1, According to the experimental results of Baseline, we selected the 128-band mel filter with the best training effect to train the following network for the dataset. It can be seen that the result reproduced by this dataset is 86.55, which is similar to the result reported by Valenti et al. Then, we presented the classification results of the model proposed by several mentioned above, the final classification accuracy results are given by Table 2.

In Table 2, the accuracy of proposed PB-CNN reaches 89.23%, while the accuracy of the baseline method given by the DCASE 2016 Challenge official is 77.2%, 12.03 percentage points lower. Meanwhile, the accuracy of using CNN to classify acoustic scenes is 86.2%, 2.5 percentage points lower than that of the second place in DCASE 2016 Challenge (88.7%). In addition, this method is an integrated method, indicating that the use of attention mechanism can effectively improve the classification effect of acoustic scenes.

Table 2. Classification accuracy (%) of using different mel filters bands for training baseline.

Fold	CNN	TB-CNN	FB-CNN	PB-CNN
1	86.154	87.436	86.667	89.231
2	85.641	86.923	86.154	89.487
3	86.667	87.179	86.207	89.231
4	86.667	87.692	87.435	88.974
Average	86.282	87.308	86.616	89.231

CNN: Convolution neural network. TB-CNN: Convolution neural network based on time attention. FB-CNN: Convolution neural network based on frequency attention. PB-CNN: Convolution neural network based on Pixel-Based Attention.

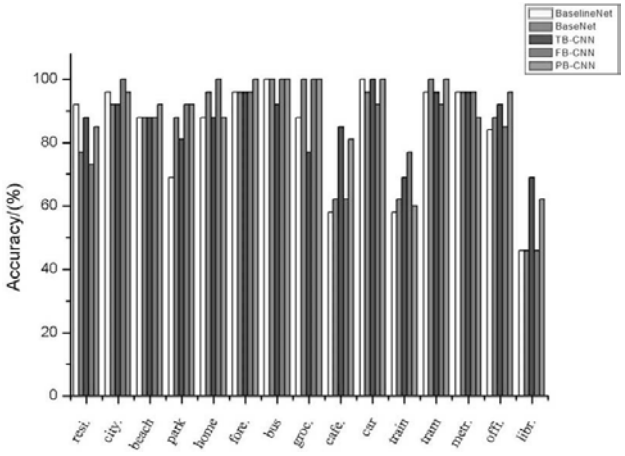


Fig. 7. The accuracy of each class in different models. FB-CNN, convolutional neural network (CNN) based on frequency attention; PB-CNN, CNN based on pixel-based attention; TB-CNN, CNN based on time attention.

In terms of the accuracy of each category, the proposed PB-CNN does not reach the highest accuracy in all categories, but it ensures that the recognition ability is not the worst in all categories and is ultimately superior to other desirable methods in terms of the overall average accuracy. The accuracy performance of each classification is shown in Fig. 7.

As shown in Fig. 7, the recognition effect of library, cafe/restaurant, and train is poor, but in the CNN with added attention mechanism, the recognition effect is improved compared with the Baseline without attention mechanism. Especially in beach, park, forest, grocery, cafe/restaurant, office, and library, the performance of the proposed PB-

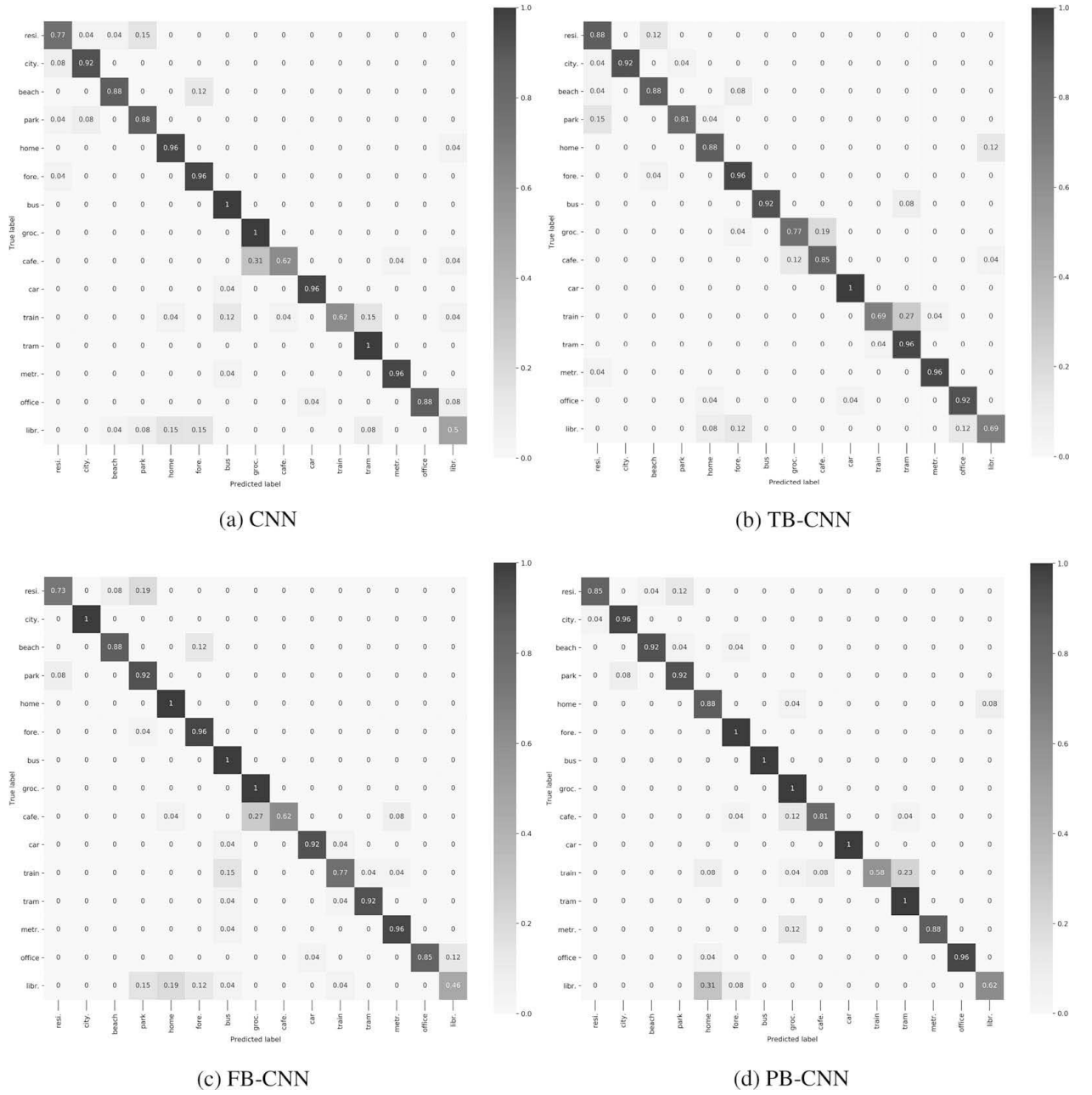


Fig. 8. The confusion matrix of convolutional neural network (CNN), CNN based on time attention (TB-CNN), CNN based on frequency attention (FB-CNN), and CNN based on pixel-based attention (PB-CNN).

CNN has been greatly improved. Therefore, overall, the recognition performance of the proposed PB-CNN has more advantages than that of CNN, TB-CNN, and FB-CNN.

In order to analyze the classification results of the model more intuitively, we drew the confusion matrix of classification results of CNN, TB-CNN, FB-CNN, and proposed PB-CNN, as shown in the Fig. 8. The value of each element in the confusion matrix is between 0 and 1, which represents the proportion of categories that are correctly classified. And we can see that most of the acoustic scene categories can be well separated, with confusion mainly occurring in residential areas, cafes/restaurants, trains, and libraries. All the models compared have poor recognition effects on library categories. Although the recognition accuracy of the proposed PB-CNN in these categories is not necessarily the highest, compared with the pure CNN, its recognition effect has been improved in these categories.

Through checking the dataset, it is found that most of the library scenes in the training center have continuous and loud voices, which are similar to buses and trains to some extent. Therefore, it is easy for confusion to occur in the training process.

In order to further and intuitively understand the classification results of the model of proposed PB-CNN, we did some visual analysis. We first selected three audio files at

Table 3. Classification accuracy (%) of different models in different datasets.

Model	No-split	Overlap	No-overlap
CNN	85.69	86.42	85.96
TB-CNN	86.64	87.22	86.42
FB-CNN	85.96	86.69	85.12
PB-CNN	89.23	89.23	88.21

CNN: Convolution neural network. TB-CNN: Convolution neural network based on time attention. FB-CNN: Convolution neural network based on frequency attention. PB-CNN: Convolution neural network based on Pixel-Based Attention.

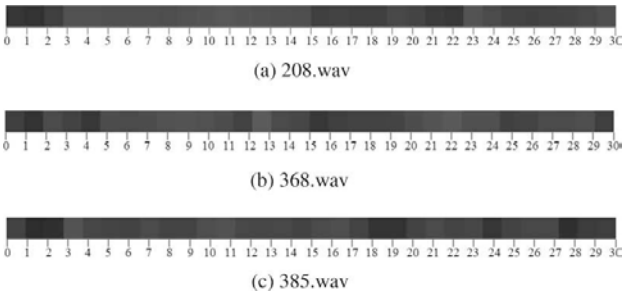


Fig. 9. Logarithmic mel spectrogram and comparison result of attention weight.

random in the DCASE 2016 dataset. The three audio files are 208.wav, 368.wav, and 385.wav. Their categories are metro_station, train, and bus. Then, we superimposed their logarithmic mel spectrogram with the attention weights. The comparison of the results is in Fig. 9.

From the comparison result of logarithmic mel spectrogram and attention weight of the 208.wav audio file, it can be seen that 1–3 s, 15–19 s, and 21–23 s are periods with large attention weight, and 1–3 s in 208.wav is the period of subway acceleration, 15–19 s is the period of subway emergency braking, and 21–23 s is the sound of subway stopping. And from the 368.wav file, the 1–5 s, 15–18 s, 25–26 s, and 29–30 s are periods with large attention weight. These are all periods of time when there are loud people on the train or when the train crashes. Finally, in 385.wav, the 1–3 s, 18–20 s, and 24 and 28 s are the periods with large attention weight. These are the periods when the bus stops, the bus closes, and the voice prompts. To some extent, it can be seen that the proposed PB-CNN does capture some important sound events.

Then, we found the segmentation method of spectrogram may have an impact on the classification results of the model. We conducted comparative experiments on the unsegmented original spectrogram, the segmentation with 50% overlap of audio segments, and the segmentation without overlap. The experimental comparison results are given by Table 3.

We can see from Table 3 that the proposed PB-CNN has the same classification performance in the datasets that are not segmented in the spectrogram and the datasets that overlap in the spectrogram segmentation, but the classification performance in the dataset that does not overlap when the spectrogram is segmented is reduced by about 1%. We think the full spectrograms divided into segments might give you a key event divided into two segments, the overlap of sliding window segmentation can be as much as possible intact, to some extent improve the classification performance.

However, on CNN, whether there is overlapping segmentation or not, the segmentation effect of spectrogram segments is better than that of unsegmented spectrogram segments. We believe that this is because the segmentation of spectrogram segments expands the dataset to some extent, and the increase of training datasets is helpful to improve the deep learning model.

On TB-CNN and FB-CNN, the training effect is best in the segmented spectrogram dataset with sliding window, followed by the unsegmented spectrogram dataset, and finally, the segmented spectrogram dataset without overlap. We believe that the existence of attention mechanism is the same reason, but the key events are divided into two ends, affecting the training performance.

4 CONCLUSION

This paper proposes a PBA module for an ASC task. The proposed PBA module mainly focuses on the attention weight distribution of each pixel on each channel of the spectrogram, obtains the initial weight distribution through the global information of each channel, and then trains through the loss function, which is combined with the original spectrogram, a more efficient spectrogram that improves the performance of acoustic scene classification. It has greater performance advantages in large datasets and anti-interference ability against interclass noise and intraclass noise, etc. In addition, since sound events may only occur in a short period of time recorded by audio, and instead of occupying a continuous image space, they appear intermittently at some moments.

Compared with common deep learning methods such as CNN and RNN, the PBA mechanism has a better ability to extract features of acoustic spectrogram objects. The experimental results show that the proposed PB-CNN achieves 89.23% classification accuracy in DCASE 2016, which exceeds 87.3% of TB-CNN, 86.6% of FB-CNN, and 86.2% of CNN. Compared with DCASE 2016's baseline system accuracy of 77.2%, the PB-CNN improved upon that number by 12% and is the highest among all submitted single models. From the experimental results, we can see that the application of PBA modules can indeed build models with better generalization and learning ability in ASC tasks.

From the experimental results, we can see that the application of attention mechanism in ASC tasks can really improve the performance. Furthermore, we can propose more relevant and adaptive attention mechanisms for ASC tasks according to the more detailed characteristics of ASC tasks and obtain more excellent performance models.

5 ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 41876110). This work was also supported by Fundamental Research Funds for the Central Universities (No. 3072019CFT0602) and the Zhejiang Provincial Natural Science Foundation (No. LQ19F020009). The authors are grateful to the guest editors and anonymous reviewers for their constructive comments, based on which the presentation of this paper has been greatly improved.

6 REFERENCES

[1] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic Scene Classification: Classifying Environments From the Sounds They Produce," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 16–34 (2015 May), https://doi.org/10.1109/MSP.2014.2326181.

[2] P. Belin, and V. B. Penhune R. J. Zatorre, "Structure and Function of Auditory Cortex: Music and Speech," *Trends Cogn. Sci.*, vol. 6, no. 1, pp. 37–46 (2002 Jan.), https://doi.org/10.1016/S1364-6613(00)01816-7.

[3] S. Chu, S. Narayanan, and C. Kuo, "Environmental Sound Recognition with Time–Frequency Audio Features," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 6, pp. 1142–1158 (2009 Aug.), https://doi.org/10.1109/TASL.2009.2017438.

[4] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 2, pp. 379–393 (2018 Feb.), https://doi.org/10.1109/TASLP.2017.2778423.

[5] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-Based Acoustic Scene Classification Using Multi-Channel Convolutional Neural Network," in *Proceedings of the 19th Pacific-Rim Conference on Multimedia (PCM)*, pp. 14–23 (2018 Sep.), https://doi.org/10.1007/978-3-030-00764-5_2.

[6] H. Tang and H. E. Chu, "SAR Image Scene Classification With Fully Convolutional Network and Modified Conditional Random Field-Recurrent Neural Network," *J. Computer Appl.*, vol. **36**, no. 12, pp. 3436–3441 (2016 Dec.), https://doi.org/10.11772/j.issn.1001-9081.2016.12.3436.

[7] W. J. Wang, Y. F. Liao, and S. H. Chen, "RNN-Based Prosodic Modeling for Mandarin Speech and Its Application to Speech-to-Text Conversion," *Speech Commun.*, vol. 36, no. 3–4, pp. 247–265 (2002 Mar.), https://doi.org/10.1016/S0167-6393(01)00006-1.

[8] S. Park, S. Munand H. Ko, D. Han "Generative Adversarial Network Based Acoustic Scene Training Set Augmentation and Selection Using SVM Hyper-Plane," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events*, pp. 93–102 (2017 Nov.).

[9] G. Takahashi, T. Yamada, and S. Makino, "Acoustic Scene Classification Based on Spatial Feature Extraction Using Convolutional Neural Networks," *J. Signal Process.*, vol. 22, no. 4, pp. 199–202 (2018 Jul.).

[10] S. Mun, S. Shon, W. Kim, D. K. Han, and H. Ko, "A Novel Discriminative Feature Extraction for Acoustic Scene Classification Using RNN Based Source Separation," *IEICE Trans. Inf. Syst.*, vol. E100.D, no. 12, pp. 3041–3044 (2017 Dec.), https://doi.org/10.1587/transinf.2017EDL8132.

[11] S. Ntalampiras, "Generalized Sound Recognition in Reverberant Environments," *J. Audio Eng. Soc.*, vol. 67, no. 10, pp. 772–781 (2019 Oct.), https://doi.org/10.17743/jaes.2019.0030.

[12] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "A Hybrid Approach with Multi-Channel I-Vectors and Convolutional Neural Networks for Acoustic Scene Classification," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pp. 2749–2753 (2017 Aug.–Sep.), https://doi.org/10.23919/EUSIPCO.2017.8081711.

[13] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, and J. Huopaniemi, "Audio-Based Context Recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 1, pp. 321–329 (2006 Jan.), https://doi.org/10.1109/TSA.2005.854103.

[14] W. Zheng, Z. Mo, X. Xing, and G. Zhao, "CNNs-Based Acoustic Scene Classification Using Multi-Spectrogram Fusion and Label Expansions," *arXiv e-prints*, submitted September 5, 2018. https://arxiv.org/abs/1809.01543v1.

[15] A. M. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136 (1980 Jan.), https://doi.org/10.1016/0010-0285(80)90005-5.

[16] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734 (2014 Oct.), https://doi.org/10.3115/v1/D14-1179.

[17] F. Meng, Z. Lu, Z. Tu, H. Li, and Q. Liu, "Neural Transformation Machine: A New Architecture for Sequence-to-Sequence Learning," *arXiv preprint*, submitted June 22, 2015. Last revised January 7, 2016. https://arxiv.org/abs/1506.06442v4.

[18] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv preprint*, submitted September 1, 2014. Last revised May 19, 2016. https://arxiv.org/abs/1409.0473v7.

[19] V. Mnih, N. Heess, and K. Kavukcuoglu A. Graves, "Recurrent Models of Visual Attention," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 2204–2212 (2014 Dec.).

[20] K. Fu, J. Jin, R. Cui, F. Sha, and C. Zhang, "Aligning Where to See and What to Tell: Image Caption With Region-Based Attention and Scene Factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2321–2334 (2017 Dec.), https://doi.org/10.1109/TPAMI.2016.2642953.

[21] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-End Attention-Based Large Vocabulary Speech Recognition," *arXiv preprint*, submitted August 18, 2015. Last revised March 14, 2016. https://arxiv.org/abs/1508.04395v2.

[22] D. Wang, Z. Shen, S. Jie, Z. Wei, and Z. Zheng, "Multiple Granularity Descriptors for Fine-Grained Categorization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015 Dec.), https://doi.org/10.1109/ICCV.2015.276.

[23] J. Fu, H. Zheng, and M. Tao, "Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017 Jul.), https://doi.org/10.1109/CVPR.2017.476.

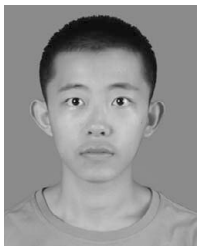
[24] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," *IEEE Trans.Pattern Anal. Mach. Intell.*, vol. **42**, no. 8, pp. 2011–2023 (2020 Aug.), https://doi.org/10.1109/TPAMI.2019.2913372.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv preprint*, submitted June 12, 2017. Last revised December 6, 2017. https://arxiv.org/abs/1706.03762v5.

THE AUTHORS


Xingmei Wang


Xuyi Chao


Jiahao Shi


Xuyang Teng

Xingmei Wang received a Ph.D. from the College of Automation, Harbin Engineering University, Harbin, China, in 2010. She was a Postdoctoral Researcher in computer science and technology with Harbin Engineering University, Harbin, China, in 2012. In 2016, she was a Visiting Scholar with the National University of Singapore, Singapore. She is currently an Associate Professor with the College of Computer Science and Technology, Harbin Engineering University, Harbin, China. Her research interests include artificial intelligence, sonar image processing, audio processing, computer vision, and pattern recognition.

•

Xuyi Chao received a B.S. degree from the College of Computer Science and Technology, Harbin Engineering University, Harbin, China, in 2018, where he is currently pursuing an M.S. degree in computer science and technology. His research interests include deep learning and audio classification.

•

Jiahao Shi received a B.S. degree from the College of Computer Science and Technology, Harbin Engineering University, Harbin, China, in 2019, where he is currently pursuing an M.S. degree in computer science and technology. His research interests include attention mechanism and audio classification.

•

Xuyang Teng received a Ph.D. from the College of Computer Science and Technology, Harbin Engineering University, Harbin, China, in 2017. He is currently a lecturer with the School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China. His research interests include artificial intelligence, evolutionary optimization algorithm, audio processing, and feature selection.