Multimedia Tools and Applications (2024) 83:10211–10231
https://doi.org/10.1007/s11042-023-16024-2

Parallel attention of representation global time–frequency correlation for music genre classification

Zhifang Wen[1] · Aibin Chen[1] · Guoxiong Zhou[1] · Jizheng Yi[1] · Weixiong Peng[2]

Received: 7 June 2022 / Revised: 6 May 2023 / Accepted: 11 June 2023 /
Published online: 20 June 2023

Abstract
Music genre classification (MGC) is an indispensable branch of music information retrieval. With the prevalence of end-to-end learning, the research on MGC has made some breakthroughs. However, the limited receptive field of convolutional neural network (CNN) cannot capture a correlation between temporal frames of sounding at any moment and sound frequencies of all vibrations in the song. Meanwhile, time–frequency information of channels is not equally important. In order to deal with the above problems, we apply dual parallel attention (DPA) in CNN-5 to focus on global dependencies. First, we propose parallel channel attention (PCA) to build global time–frequency dependencies in the song and study the influence of different weighting methods for PCA. Next, we design dual parallel attention, which focuses on global time–frequency dependencies in the song and adaptively calibrates contribution of different channels to feature map. Then, we analyzed the effect of applying different numbers and positions of DPA in CNN-5 for performance and compared DPA with multiple attention mechanisms. The results on GTZAN dataset demonstrated that the proposed method achieves a classification accuracy of 91.4%, and DPA has the highest performance.

Keywords Music genre classification · Attention mechanism · Convolutional neural network · Global time–frequency correlation · Mel-spectrogram

1 Introduction

With the rapid development of multimedia and communication technology, digital music is seen everywhere in life. The volume of music resources has become more extensive, and the retrieval of massive resources that rely on humans has become laborious. Therefore, music information retrieval (MIR) has become a challenging problem [3, 17]. The algorithm will play an essential role in MIR if it can

✉ Aibin Chen
hotaibin@163.com

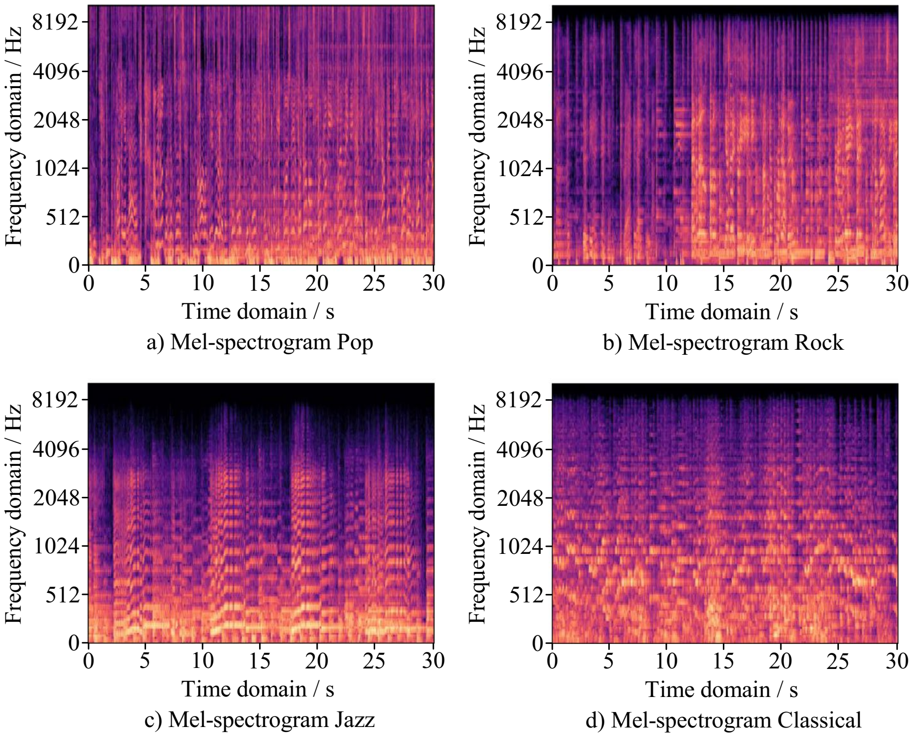
[1] Institute of Artificial Intelligence Application, Central South University of Forestry and Technology, Changsha 410004, China

[2] Hunan Zixing Artificial Intelligence Technology Group Co, Ltd, Beijing, China

automatically divide music into different genres according to the music content. Therefore, an accurate and effective music genre classification (MGC) algorithm is indispensable [18-20]. In the traditional method, MGC is mainly composed of two parts: (1) feature extraction; (2) classifier model. Feature extraction expresses the inherent properties of music as feature vectors, and the classifier model maps features vectors to different genres. Baniya and Lee [21] used two different types of features, tone texture, and rhythm content features, to represent music. They used Extreme Learning Machine (ELM) [7] with bagging as the classifier for classifying the genres. Arabi [22] proposed capturing the high-level concepts of music, harmonics, pitch, and rhythmic content feature combined with low-level features and Support Vector Machines (SVM) [23] as a classifier. Sarkar [24] used Empirical Mode Decomposition (EMD) to capture tonal characteristics in the mid-frequency range and used a multilayer perceptron (MLP) [6] as a classifier. Although these methods have contributed much to MGC, they all rely on hand-crafted features for classification. This requires researchers have professional music knowledge to design more worked features.

Deep learning has made breakthroughs in Natural Language Processing (NLP) and Computer Vision (CV) in recent years [8, 10, 25, 26]. The advantage of deep learning is that it provides an end-to-end learning mode, so it does not need to design features separately. Therefore, the works of [15, 27-29] try to apply Convolutional Neural Networks (CNN) in deep learning to audio classification. It is worth noting that Low-level audio features [4] of short-time Fourier transform spectrogram (STFT) [30] and Mel-spectrogram are particularly widely used. In MGC, Zhang [31] proposed a method based on Convolutional Neural Network combined with pooling and short connection [26] to apply to MGC. To capture the temporal dependence of audio, Choi [32] proposed a Convolutional Recurrent Neural Network (CRNN) for music classification. Yu [14] found that: spectra with different temporal steps have different importance. Therefore, they proposed a new model incorporating with attention mechanism based on Bidirectional Recurrent Neural Network [33] and discussed the influence of serial attention and parallel attention. The above methods based on CNN and attention mechanism consider such factors as the temporal dependence of audio and spectral importance in different time steps.

However, there is a strong correlation between temporal frames of sounding at any moment and sound frequencies of all vibrations in the Mel-spectrogram. This can be easily found by observing the Mel-spectrogram (as shown in Fig. 1), choosing a temporal frame randomly in the time domain, and there is a vibrating sound frequency in the vertical direction of the temporal frame. Similarly, choosing a sound frequency randomly in the frequency domain, and there are sounding temporal frames in the horizontal direction of the sound frequency. Therefore, we proposed parallel channel attention (PCA) to build a global time–frequency correlation. Specifically, PCA constructs a weight matrix to obtain global feature correlation, weight and sum the time–frequency information in the Mel-spectrogram and generate new features to build global time–frequency dependencies. From horizontal direction observed, Mel-spectrogram represents time domain figure, and from vertical direction observed, Mel-spectrogram represents frequency domain figure. Therefore, we discuss the influence of weighting methods based on time domain, frequency domain, and time–frequency domain building global time–frequency dependencies for attention mechanism. In addition, when CNN extracts feature, the importance of time–frequency information of each channel for feature map is different. We design dual parallel attention (DPA) composed of PCA and SE Attention [34], which focuses on global



a) Mel-spectrogram Pop

b) Mel-spectrogram Rock

c) Mel-spectrogram Jazz

d) Mel-spectrogram Classical

Fig. 1 Mel-spectrograms of four music genres, Pop, Rock, Jazz, and Classical, in the GTZAN dataset, where all temporal frames from left to right constitute the time domain, and all sound frequencies from top to bottom constitute the frequency domain

time–frequency dependencies in the song and adaptively distinguishes importance of different channels. The main contributions of this paper are summarized as follows:

(1) We propose parallel channel attention, which builds global time–frequency dependencies in the song by representing the correlation between temporal frames of sounding at any moment and sound frequencies of all vibrations.
(2) We discuss the influence of weighting methods based on time domain, frequency domain, and time–frequency domain building global time–frequency dependencies for attention mechanism.
(3) We design dual parallel attention focuses on global time–frequency dependencies in the song and adaptively calibrates contribution of different channels to feature map.

The rest of this paper is organized as follows: Sect. 2 introduces music genre classification and attention mechanism related works. Section 3 describes model of proposed parallel attention applied in CNN-5, including CNN-5, parallel channel attention, SE Attention, and dual parallel attention. Section 4 shows Dataset and experimental setup. Section 5 analysis experimental results of GTZAN dataset. Finally, Sect. 6 concludes the paper.

2 Related work

In audio and signal classification, deep learning is widely used. Yang [35] proposed duplicate convolutional layers whose output will be applied to different pooling layers and concatenated features after each pooling layer, providing more classification statistics. Chang [36] learned 2D representations from 1D raw waveform signals as input feature. Meanwhile, they proposed a new network architecture—MS-SincResNet, which can learn 1D and 2D convolutional kernels together. Choi [37] proposed a transfer learning method for MGC. They used pre-trained convolutional network features to perform music labeling. Then transfer [13] to classification task related to music. Cai [2] proposed a novel music classification framework incorporating the auditory image feature with traditional acoustic features and spectral feature. Srinivasu [38] uses the deep learning model based on finetuned AlexNet to classify the signals associated with glucose levels in the human body. Scalvenzi [11] proposed multiresolution analysis based on discrete wavelet-packet transform (DWPT) associated with a support vector machine (SVM) to classify music singals, such as major and minor chords.

In order to capture global dependencies between input and output across distances, Vaswani [25] proposed Transformer, a model architecture entirely relies on an attention mechanism. Inspired by the classical non-local means method in CV, Wang [39] proposed a non-local neural network (Non-local), which calculates the weighted sum of all position features as the response of a position. Wang [40] proposed parallel temporal-spectral attention based on the time–frequency domain properties of the spectrogram, which enhances the temporal and spectral features by capturing the importance of different time frames and frequency bands. Huang [41] proposed an end-to-end attention-based deep feature fusion (ADFF) approach for music emotion recognition to learn affect-salient features. Dosovitskiy [42] believes that the reason for the superior performance of the attention mechanism is that the Transformer structure plays a decisive role, so they put forward the Vision Transformer (ViT), which has made a new breakthrough in CV. Gong [43] applied Transformer structure in the direction of audio classification and proposed Audio Spectrum Transformer (AST).

Recently, the attention mechanism has been applied to MGC. Yang [44] continue to study the global dependencies of long audio sequences, employing parallel structures instead of recurrent architecture, multi-head attention as feature extractors, and SVM as classifiers. Their models show considerable generalization ability. In the MGC, few researchers focus on the time–frequency correlation in songs. We proposed parallel channel attention, which builds global time–frequency dependencies in the song and merges the parallel channel attention with SE Attention to form dual parallel attention.

3 Proposed method

In this section, we first introduce the basic network CNN-5, then describe parallel channel attention that builds global time–frequency dependencies, then show SE Attention, and finally present the dual parallel attention that fuses SE Attention with parallel channel attention.

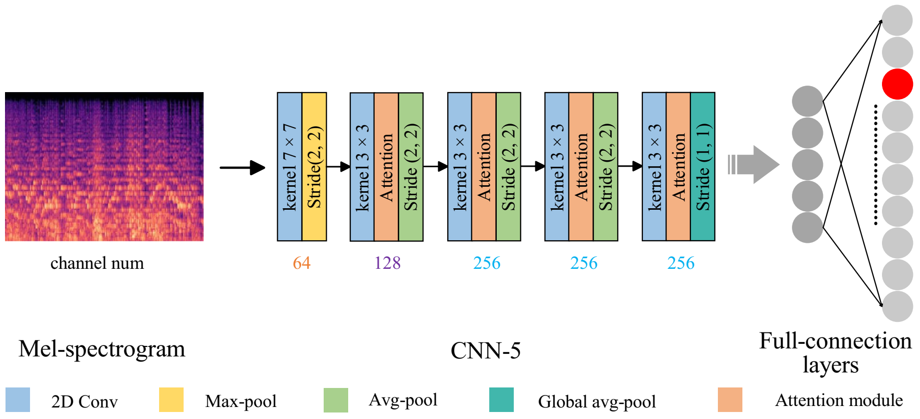


Fig. 2 The overall architecture of applying attention mechanism in CNN-5"

3.1 Overview

The general form of song is waveform signal, which is converted into Mel-spectrogram through short-time Fourier transform and Mel filter. There is a strong correlation between temporal frames of sounding at any moment and sound frequencies.

However, there is a strong correlation between temporal frames of sounding at any moment and sound frequencies of all vibrations in the Mel-spectrogram. Moreover, the importance of time–frequency information in each channel is different. Therefore, we propose dual parallel attention, which focuses on global time–frequency dependencies in the song and adaptively calibrates contribution of different channels.

The overall architecture of the proposed DPA approach is shown in Fig. 2: First, the input of the proposed model is Mel-spectrogram. Secondly, the backbone network of the model is CNN-5. The DPA proposed in this paper will be applied to the Attention module in the backbone network. CNN-5 captures the local features in the Mel-spectrogram, and DPA builds the global feature dependencies. As shown in Fig. 3, DPA is mainly divided into two parts. The upper part is parallel channel attention for building global time–frequency dependencies, and the lower part is SE Attention for constructing the global channel dependencies. Finally, the features captured by the backbone network are sent to the full-connection layers, which map the features into genre classes as the output results.

3.2 CNN-5

This section will introduce CNN-5, which mainly consists of five convolutional layers. First, the small channel convolutional layers capture low-level features such as texture and contour. With deeper layers, the number of channels is increased to gather deep semantic feature. The architecture parameters are as follows:

- Layer 1: The first convolutional layer, consisting of 64 kernels with a 77 respective field and stride of (1, 1), sets a large respective field to aggregate more spectrogram feature. Next, max-pooling with stride (2, 2) is used for down-sampling, capturing critical information in the pooling block.

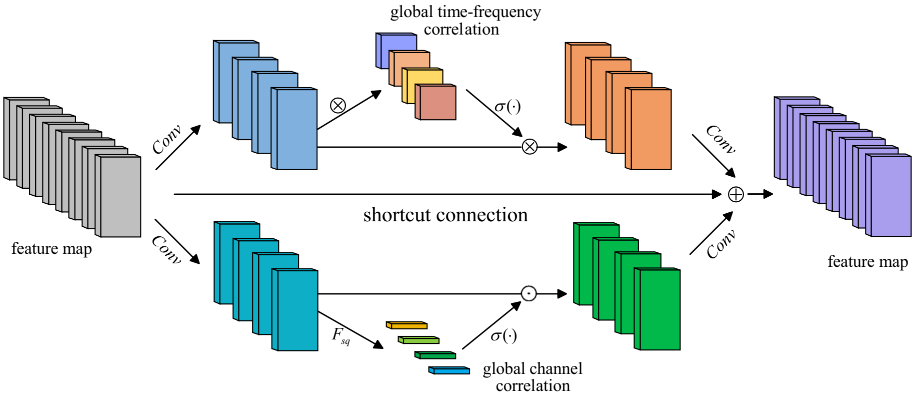


Fig. 3 Dual parallel attention, where the top part is PCA, and the bottom part is SE Attention. Here, $\sigma(\cdot)$ denotes sigmoid function, Conv denotes convolutional layer, $\otimes$ denotes matrix multiplication, $\oplus$ denotes matrix sum and $\odot$ denotes element-wise product

- Layer 2: The second convolutional layer consists of 128 kernels with a 33 respective field and stride (1, 1). Down-sampling is done by avg-pooling with stride (2, 2), which captures All information in the pooling block.
- Layer 3: The third convolutional layer consists of 256 kernels with a 33 respective field and stride of (1, 1). Down-sampling with avg-pooling which stride of (2, 2).
- Layer 4: The fourth convolutional layer consists of 256 kernels with a 33 respective field and stride of (1, 1). Down-sampling with avg-pooling which stride of (2, 2).
- Layer 5: The fifth convolutional layer of 256 kernels with a 33 respective field and stride of (1, 1). Finally, global average pooling [45] for down-sampling.

Each convolutional layer follows with Batch Normalization (BN) [46] to speed up the network's training. The activation function is Rectified linear units (ReLU) [47]. Detailed parameters are shown in Table 1.

3.3 Parallel channel attention

There is a strong correlation between temporal frames of sounding at any moment and sound frequencies of all vibrations. We propose parallel channel attention (Fig. 4) to build global time–frequency dependencies. PCA constructs a weight matrix for each channel in the feature map to obtain global feature correlation, weights and sums time–frequency information of each channel in parallel and generates new features to build global time–frequency dependencies. Given input feature map $\mathbf{X} \in \mathbb{R}^{c \times f \times t}$, c denotes the number of channels, f denotes the number of Mel filter banks in the frequency domain, and t denotes the number of temporal frame in the time domain(actually f and t denote height and width in feature map). The feature map $\mathbf{X}$ has c channels, denotes, $\mathbf{X} = \left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_c\right)$ where $\mathbf{x}_i \in \mathbb{R}^{f \times t}$, $i \in \{1, 2, \ldots, c\}$, $\mathbf{x}_i$ represents a channel and is also a feature set of Mel-spectrogram, composed of two dimensions: time domain and frequency domain. This section describes parallel channel attention based on time domain, frequency domain, and time–frequency domain.

Table 1 CNN-5 network structure

Layer	Type	Shape	Output size
L1	2D Conv	6477	64,128,313
L1	BN & ReLU	-	64,128,313
L1	Max pool	22	6,464,156
L2	2D Conv	12,833	12,864,156
L2	BN & ReLU	-	12,864,156
L2	Avg pool	22	1,283,278
L3	2D Conv	25,633	2,563,278
L3	BN & ReLU	-	2,563,278
L3	Avg pool	22	2,561,639
L4	2D Conv	25,633	2,561,639
L4	BN & ReLU	-	2,561,639
L4	Avg pool	22	256,819
L5	2D Conv	25,633	256,819
L5	BN & ReLU	-	256,819
L5	Global Avg pool	819	25,611

Parallel channel attention based on time domain weighting, as shown in Fig. 5: first, building a time domain weight matrix, then, weighting and summing all sounding temporal frames at each sound frequency of Mel-spectrogram in parallel. At last, aggregating time-domain features across distances to build global time–frequency dependencies, as follows:

$$\mathrm{F}_{PCA\_T}(\mathbf{X}) = g\left(\left(\sigma\left((\phi(\mathbf{X}))^{\mathrm{T}}\psi(\mathbf{X})\right)(\theta(\mathbf{X}))^{\mathrm{T}}\right)^{\mathrm{T}}\right) + \mathbf{X} \tag{1}$$

where three 1 × 1 2D convolutions include θ, ϕ and ψ, a 3 × 3 2D convolution g and Sigmoid function denotes σ. ϕ and ψ reduce the number of channels in feature map and the number of parameters during operation. Transpose $\phi(\mathbf{X})$ to get $(\phi(\mathbf{X}))^{\mathrm{T}}$ and multiply with $\psi(\mathbf{X})$ to get time domain weight matrix. We only select a channel $\mathbf{x}_i$ of $\psi(\mathbf{X})$ and $(\phi(\mathbf{X}))^{\mathrm{T}}$ to express this process, as follow:

$$\psi\left(\mathbf{x}_i\right) = \left(\mathbf{w}_{\psi_1}\mathbf{y}_1, \mathbf{w}_{\psi_2}\mathbf{y}_2, \ldots, \mathbf{w}_{\psi_t}\mathbf{y}_t\right) \tag{2}$$

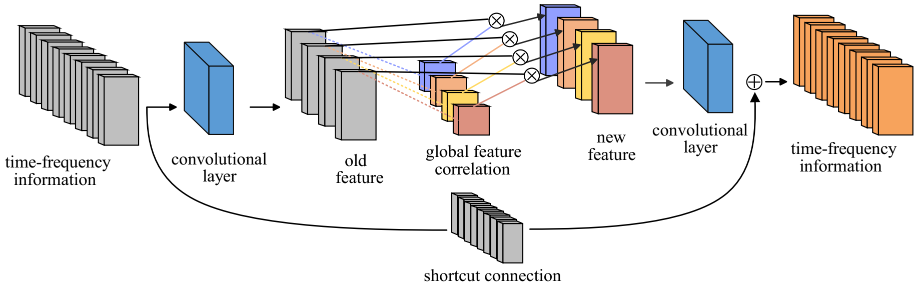


Fig. 4 Parallel channel attention, which constructs a weight matrix to obtain global feature correlation, weights and sums time–frequency information and generates new features to build global time–frequency dependencies

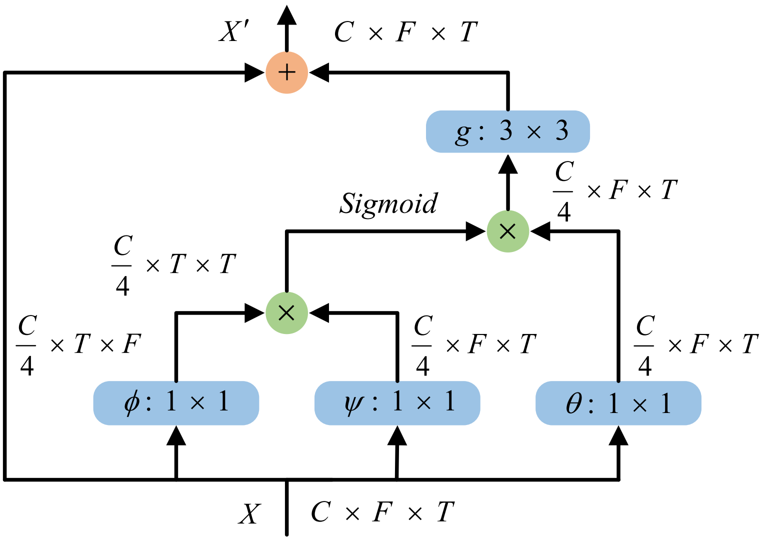

X'
$C \times F \times T$
$g: 3 \times 3$
$\frac{C}{4} \times F \times T$
$Sigmoid$
$\frac{C}{4} \times T \times T$
$\frac{C}{4} \times T \times F$
$\frac{C}{4} \times F \times T$
$\frac{C}{4} \times F \times T$
$\phi: 1 \times 1$
$\psi: 1 \times 1$
$\theta: 1 \times 1$
X
$C \times F \times T$

Fig. 5 PCA based on time domain. Sigmoid denotes activation for each element in the weight matrix. The blue circular box denotes 2D convolution, reducing the number of channels in the feature map

$$\left(\phi\left(\mathbf{x}_i\right)\right)^{\mathrm{T}} = \left(\mathbf{w}_{\phi_1}\mathbf{y}_1, \mathbf{w}_{\phi_2}\mathbf{y}_2, \dots, \mathbf{w}_{\phi_t}\mathbf{y}_t\right)^{\mathrm{T}} \tag{3}$$

where $\mathbf{w}_{\psi_j}\mathbf{y}_j \in \psi\left(\mathbf{x}_i^t\right)$, $\mathbf{w}_{\phi_j}\mathbf{y}_j \in \left(\phi\left(\mathbf{x}_i^t\right)\right)^{\mathrm{T}}$, $j \in (1, 2, \dots, t)$, $\mathbf{w}$ denotes 1D convolution kernel, $\mathbf{x}_i$ denotes a channel in the feature map $\mathbf{X}$, and $\mathbf{y}_j$ denotes a temporal frame on channel $\mathbf{x}_i$.

The weight matrix obtained by multiplying $\psi\left(\mathbf{x}_i\right)$ and $\left(\phi\left(\mathbf{x}_i\right)\right)^{\mathrm{T}}$ denotes the correlation between any two temporal frames on channel $\mathbf{x}_i$ as follows:

$$\left(\phi\left(\mathbf{x}_i\right)\right)^{\mathrm{T}}\psi\left(\mathbf{x}_i\right) = \begin{pmatrix} \mathbf{w}_{\phi_1}\mathbf{y}_1\mathbf{w}_{\psi_1}\mathbf{y}_1 & \cdots & \mathbf{w}_{\phi_1}\mathbf{y}_1\mathbf{w}_{\psi_t}\mathbf{y}_t \\ \cdots & \mathbf{w}_{\phi_j}\mathbf{y}_j\mathbf{w}_{\psi_j}\mathbf{y}_j & \cdots \\ \mathbf{w}_{\phi_t}\mathbf{y}_t\mathbf{w}_{\psi_1}\mathbf{y}_1 & \cdots & \mathbf{w}_{\phi_t}\mathbf{y}_t\mathbf{w}_{\psi_t}\mathbf{y}_t \end{pmatrix} \tag{4}$$

[16] proposed to employ Sigmoid as a scaling function in audio signals, which can avoid the concentration of attention on several temporal frames. Therefore, Sigmoid is the scaling function in this paper, as follows:

$$\sigma\left(\phi\left(\mathbf{x}_i\right)\right)^{\mathrm{T}}\psi\left(\mathbf{x}_i\right) = \sigma\begin{pmatrix} \mathbf{w}_{\phi_1}\mathbf{y}_1\mathbf{w}_{\psi_1}\mathbf{y}_1 & \cdots & \mathbf{w}_{\phi_1}\mathbf{y}_1\mathbf{w}_{\psi_t}\mathbf{y}_t \\ \cdots & \mathbf{w}_{\phi_j}\mathbf{y}_j\mathbf{w}_{\psi_j}\mathbf{y}_j & \cdots \\ \mathbf{w}_{\phi_t}\mathbf{y}_t\mathbf{w}_{\psi_1}\mathbf{y}_1 & \cdots & \mathbf{w}_{\phi_t}\mathbf{y}_t\mathbf{w}_{\psi_t}\mathbf{y}_t \end{pmatrix} \tag{5}$$

feature map $\mathbf{X}$ reduces number of channels through θ and also only selects a corresponding channel, $\theta\left(\mathbf{x}_i\right)$, as follows:

$$\theta\left(\mathbf{x}_i\right) = \left(\mathbf{w}_{\theta_1}\mathbf{y}_1, \mathbf{w}_{\theta_2}\mathbf{y}_2, \dots, \mathbf{w}_{\theta_t}\mathbf{y}_t\right) \tag{6}$$

$\mathbf{w}_{\theta_j}\mathbf{y}_j \in \theta\left(\mathbf{x}_i^t\right)$, transpose $\theta\left(\mathbf{x}_i\right)$ and multiply with $\sigma\left(\phi\left(\mathbf{x}_i\right)\right)^{\mathrm{T}}\psi\left(\mathbf{x}_i\right)$, which makes all temporal frames in $\theta\left(\mathbf{x}_i\right)$ multiply and add corresponding to each row of the weight matrix, aggregate into new features, to complete the operation of capturing time domain features across distances, to build global time–frequency dependencies, as follows:

$$\sigma\left(\left(\phi\left(\mathbf{x}_i\right)\right)^{\mathrm{T}}\psi\left(\mathbf{x}_i\right)\right)\left(\theta\left(\mathbf{x}_i\right)\right)^{\mathrm{T}} = \left(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t\right) \tag{7}$$

$\sigma\left(\left(\phi\left(\mathbf{x}_i\right)\right)^{\mathrm{T}}\psi\left(\mathbf{x}_i\right)\right)\left(\theta\left(\mathbf{x}_i\right)\right)^{\mathrm{T}} \in \mathbb{R}^c$, Then, the original channel shape is restored through g, and the input and output shapes are kept consistent. Finally, shortcut connections are added to the attention module to avoid losing the original information.

Parallel channel attention based on frequency domain weighting: first, build a frequency domain weight matrix, then weight and sum all sound frequencies in each temporal frame in parallel. And last, aggregate frequency domain features across distances to build global time–frequency dependencies, as follows:

$$\mathrm{F}_{PCA\_F}(\mathbf{X}) = g\left(\sigma\left(\phi(\mathbf{X})(\psi(\mathbf{X}))^{\mathrm{T}}\right)\theta(\mathbf{X})\right) + \mathbf{X} \tag{8}$$

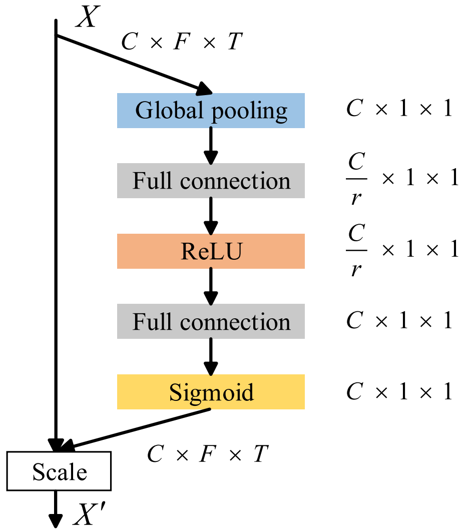
Parallel channel attention based on time–frequency domain is composed of the feature fusion of time domain weighting and frequency domain weighting, as follows:

$$\mathrm{F}_{PCA\_TF}(\mathbf{X}) = g\left(\sigma\left(\phi(\mathbf{X})(\psi(\mathbf{X}))^{\mathrm{T}}\right)\theta(\mathbf{X}) + \left(\sigma\left((\phi(\mathbf{X}))^{\mathrm{T}}\psi(\mathbf{X})\right)(\theta(\mathbf{X}))^{\mathrm{T}}\right)^{\mathrm{T}}\right) + \mathbf{X} \tag{9}$$

3.4 Squeeze-and-excitation attention

CNN extracts feature by fusing information in local receptive field, and each convolutional kernel independently completes the fusion process. However, not the time–frequency information of each channel in feature map is equally important. HU et al. proposed SENet (General attention module, Fig. 6), which adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. SE Attention is the

Fig. 6 Squeeze-and-Excitation Attention



core module of SENet, and we will introduce it. It is mainly divided into two parts: squeeze and excitation. As follows:

$$\mathrm{F}_{SEA} = \mathrm{F}_{\mathrm{scale}}\left(\sigma\left(\mathrm{W}_2\delta\left(\mathrm{W}_1\mathrm{F}_{sq}(\mathbf{X})\right)\right)\right) \odot \mathbf{X} \tag{10}$$

where F_{sq} denotes squeeze function, $\mathrm{W}_1 \in \mathrm{FC}^{\frac{m}{r}\times m}$ and $\mathrm{W}_2 \in \mathrm{FC}^{m\times\frac{m}{r}}$ denote two full connection layers, δ denotes ReLU function, $\mathrm{F}_{\mathrm{scale}}$ denotes scaling function.

First, introduce squeeze operation, using global average pooling to squeeze features on each channel in the feature map into an element. We only select one channel $\mathbf{x}_i$ to describe squeeze operation, as follows:

$$\mathrm{F}_{sq}\left(\mathbf{x}_i\right) = \frac{1}{f \times t}\sum_{\mathrm{p}=1}^{f}\sum_{\mathrm{q}=1}^{t} u_k(p, q) \tag{11}$$

Next, excitation operation, features are passed through W_1 and W_2. Two full connection layers learn the relationship between the channels and use different function to activate after each full connection layer. Finally, $\mathrm{F}_{\mathrm{scale}}$ is used to copy the output feature, make its shape consistent with the feature map $\mathbf{X}$, and multiply with $\mathbf{X}$ channel by channel to adaptively calibrate the relationship between channels. This paper adopts the general attention module squeeze and exception attention (SE Attention) in SENet.

3.5 Dual parallel attention

The limited receptive field of CNN cannot capture the correlation between temporal frames of sounding at any moment and sound frequencies of all vibrations. At the same time, when CNN extracts feature of the Mel-spectrogram, convolutional kernel captures time–frequency information of different levels and fuses them into channels. However, not the time–frequency information on each channel is equally important. We design dual parallel attention Fig. 3, which builds global time–frequency dependencies in the song and distinguishes the contribution of each channel to feature map. Dual parallel attention is composed of parallel channel attention and SE Attention fusion, as follows:

$$F_{DPA}(\mathbf{X}) = F_{PCA}(\mathbf{X}) + F_{SEA}(\mathbf{X}) + \mathbf{X} \tag{12}$$

After the feature map $\mathbf{X}$ is weighted by the parallel channel attention and SE Attention. next, the element-wise summation completes the feature fusion, and the shortcut connection is added to avoid losing the original information.

4 Dataset and experimental setup

4.1 Dataset and preprocessing

In this paper, dataset is GTZAN collected by Tzanetakis [12], widely applied in MGC. It includes 1000 songs, evenly distributed in each genre, 100 songs in each genre, of which ten music genres are Blues, Classical, Country, Hip-hop, Jazz, Metal, Pop, Reggae and Rock. Each song excerpt is about 30 s, stored as 22,050 Hz, 16 bits. To avoid repetitive information in multi-channel, we down-sample the song to 16,000 Hz, transform it to mono-channel processing.

We transform the song into Mel-spectrogram as input feature. The length of the FFT window is 512, hop length is 256, and the number of frequency bins is 128. We sliced the songs [14, 31, 35], and each song was divided into 11 music clips (lasting 5 s), and each clip overlaps by 50%, and the clip shape is 128×313 [5].In the experiment, train, validation, and test set ratio is divided into 8:1:1, and the balance between genres is maintained. In addition, the results of a single experiment in GTZAN fluctuate significantly, we use ten-fold cross-validation to ensure stability of results. In this paper, all test results are average values after ten runs.

4.2 Experimental setup

In this paper, Pytorch as deep learning platform, GPU is RTX 3090, Adam [48] as optimizer, batch size is 22, and loss function is Cross-Entropy. Each fold data training with 50 epochs and the tenfold cross-validation training for 500 epochs. The 0.0001 is the initial learning rate, which decays to one-tenth after 20 epochs. The convolutional kernel was initializing to Xavier Normal and Batch normalization initializing to constant. During training, we treat all song clips as independent samples for training. However, during verification and testing, we use a voting mechanism to select a genre with the highest probability from all the same song clips as final output result. For example, dividing a song into m clips and the song has k genres, prediction result of a song, as follows:

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ \cdots & \cdots & \cdots & \cdots \\ y_{i1} & y_{i2} & y_{ij} & y_{ik} \\ y_{m1} & y_{m2} & \cdots & y_{mk} \end{pmatrix} \tag{13}$$

where$i \in \{1,\ 2,\ ...,\ \mathrm{m}\}$,j $\in \{1,\ 2,\ ...,\ k\}$, y_{ij} denotes probability of genre j in the i song clip, calculate the average probability of genre j in all clips, as follow:

$$y_j = \frac{\sum_{i=1}^{m} \mathrm{y}_{ij}}{m} \tag{14}$$

Then, selecting a genre with the highest probability as final output, as follows:

$$y_{label} = \max(y_1, ...,\ y_j,\ ...,\ y_k) \tag{15}$$

Table 2 Comparison results of CNN-5 + DPA and existing methods on GTZAN dataset

Method	Feature	Accuracy
KCNN(k = 5) + SVM [49]	Mel-spectrum, SFM, etc	83.90%
nnet2 [31]	STFT	87.40%
Transform learning [37]	Convnet features	89.80%
Hybrid model [50]	MFCC, SSD, etc	90.00%
net1 [35]	Mel-spectrogram	90.70%
BRNN + PCNNA [14]	STFT	90.00%
CRNN with GLR [1]	Mel-spectrogram	87.79%
MhaNN-SVM [44]	Mel-spectrogram	88.40%
MS-SincResNet [36]	Raw waveform	91.49%
CNN-5	Mel-spectrogram	89.30%
CNN-5 + DPA (Ours)	Mel-spectrogram	**91.40%**

Table 3 Detailed comparison results between MS-SincResNet and CNN-5 + DPA

Method	Ten-fold cross-validation	Params size	Training time	Accuracy
MS-SincResNet [36]		43 MB	37 h 02 min	91.49% (Chang) 90.20% (ours)
CNN-5 + DPA		10.5 MB	04 h 15 min	91.40%

5 Experimental results and analysis

5.1 Experimental results on the GTZAN dataset

In this section, we compare the proposed method with many existing methods, and the results are summarized in Table 2. KCNN(k = 5) + SVM, nnet2, and net1 networks are based on convolutional neural networks, shortcut connection, pooling, and other operations, and the network structure is redesigned. Although the classification accuracy of net1 network is higher than that of the backbone network CNN-5, net1 network cannot apply attention mechanism, so this paper does not improve the net1 network. In addition, from the final results, the accuracy of net1 network is lower than the method proposed in this paper. The accuracy of Transform learning method based on hybrid feature and transfer learning training and Hybrid model method based on a two-stage hybrid classifier is lower than the methods proposed in this paper. BRNN + PCNA and MhaNN-SVM methods combine the attention mechanism. BRNN + PCNA method proposes a model based on a bidirectional recurrent neural network and parallel attention. The mhaNN-SVM method uses multi-head attention as a feature extractor and SVM as a classifier to recognize all classes. These two methods do not consider the application of attention mechanism in time–frequency dependency. MS-SincResNet innovatively applies the method of extracting features from waveform signals to music genre classification and then sends the features into the deep neural network ResNet for classification.

MS-SincResNet is slightly better than the proposed CNN-5 + DPA in classification accuracy. Therefore, we compare CNN-5 + DPA with MS-SincResNet in detail from Params size, training time, and accuracy. Params size and training time are not provided in the MS-SincResNet paper. We experimented again according to the code provided in the paper.[1] As shown in Table 3, 91.49% (Chang) represents the accuracy of [36], and 90.20% (ours) represents the accuracy of our experiment. The Params size of MS-SincResNet is 43 MB, while CNN-5 + DPA is only a quarter of it, 10 MB. MS-SincResNet lasts 37 h, while CNN-5 + DPA lasts about 4 h, only one-ninth of its time. Although the accuracy of CNN-5 + DPA is 0.09% lower than MS-SincResNet, the Params size and training speed are much better than theirs. Therefore, CNN-5 + DPA is quite competitive in the methods mentioned in Table 2.

[1] GTZAN is divided according to the proportion of 9:1 for the training set and test set in the paper [36], and ten-fold cross-validation is adopted. The average of the ten test results is the final result, which is consistent with the strategy of our paper. We treat the results of the test set as final when the model training is complete. Unfortunately, we were unable to reproduce the classification accuracy mentioned in their paper, which may be due to insufficient training details provided in the paper.

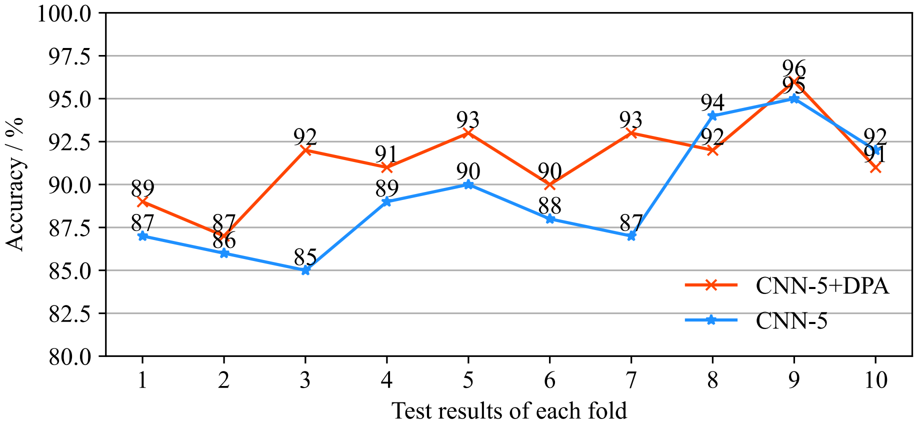
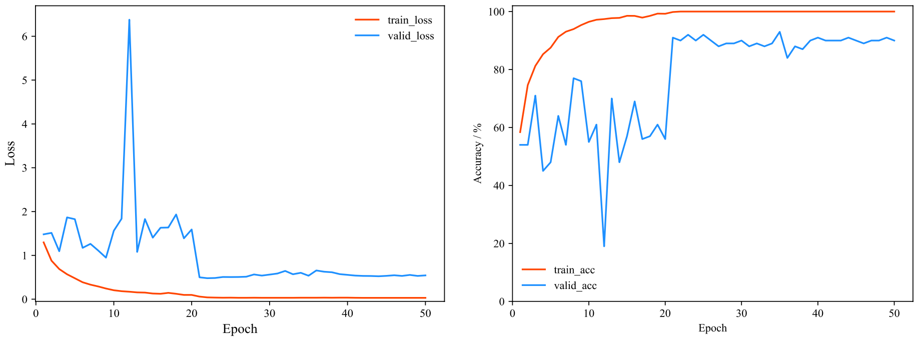


Fig. 7 Ten-fold cross-validation test results

The experimental results of the test set in tenfold cross-validation are shown in Fig. 7. The red line represents the proposed model CNN-5 + DPA, and the blue line represents the backbone network CNN-5. The results of the CNN-5 + DPA model are better than those of CNN-5 in most fold data. CNN-5 is slightly better than CNN-5 + DPA in eightfold and tenfold data. Overall, the DPA proposed in this paper can improve the model performance steadily and effectively.

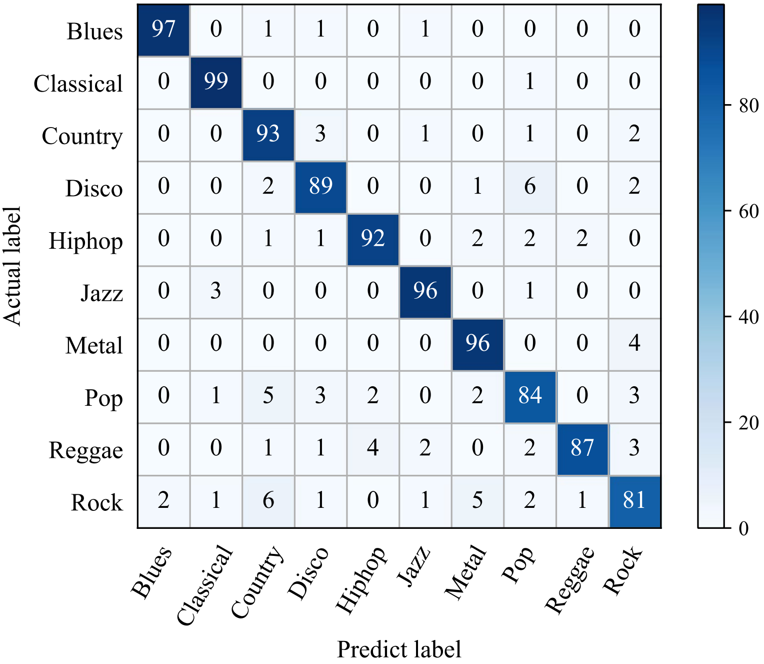
As shown in Fig. 8, taking the fourfold data in ten-fold cross-validation as an example, we provide the details of model training based on loss and accuracy. First, look at the loss figure. Training and validation loss in the first 20 epochs shows a downward trend. The training loss decreased steadily with the epoch increase, while the validation loss changed significantly, with the maximum value exceeding 6 and the minimum value only 1. After 20 epochs, the learning rate decreased to 0.00001. Between the 20th and 40th epoch, training and validation loss remained the same. The training loss was stable at around 0.03, and the validation loss was at about 0.5. After 40 epochs, the learning rate decreased to 0.000001. There was no significant change in training and validation loss, showing a convergence state, which remained around 0.03 and 0.5, respectively. Secondly, observing the



(a) Loss

(b) Accuracy

Fig. 8 Loss and accuracy of training and validation of CNN-5 + DPA model



Actual label \ Predict label	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
Blues	97	0	1	1	0	1	0	0	0	0
Classical	0	99	0	0	0	0	0	1	0	0
Country	0	0	93	3	0	1	0	1	0	2
Disco	0	0	2	89	0	0	1	6	0	2
Hiphop	0	0	1	1	92	0	2	2	2	0
Jazz	0	3	0	0	0	96	0	1	0	0
Metal	0	0	0	0	0	0	96	0	0	4
Pop	0	1	5	3	2	0	2	84	0	3
Reggae	0	0	1	1	4	2	0	2	87	3
Rock	2	1	6	1	0	1	5	2	1	81

Fig. 9 Confusion matrix of CNN-5 with DPA on GTZAN dataset

accuracy figure. In the first 20 epochs, the training accuracy increased steadily with the increase of epoch, while the validation accuracy showed an out-of-order state. Similarly, the learning rate decayed after 20 epochs. Between the 20th and 40th epoch, the training accuracy reaches 100%, and the validation accuracy fluctuates from 90%. After 40 epochs, the learning rate decreased again. The training accuracy remains at 100%, and the validation accuracy is at 90%. After the first decay of the learning rate, the proposed model's loss remains stable, indicating that the model converges fast. After two decays of the learning rate, the validation accuracy is basically stable, and the model has converged in combination with the loss figure.

As illustrated in Fig. 9, the confusion matrix figure represents the comparison of predicted and actual results of applied DPA in CNN-5 on the GTZAN dataset. The higher diagonal value and the darker color, the higher recognition rate of music genre. We found that the classification accuracy of Classic and Blue is relatively high, reaching 99% and 97%, Pop and Rock classification accuracy is relatively low, only 84% and 81%. From the perspective of music style and experimental methods: classic and blues music styles are relatively stable, and the melody and beat of the song itself change little. This paper adopts the song slice and voting mechanism (as shown in Sect. 4.2) to combine the prediction probability of each slice and ensemble the final prediction result. Therefore, songs with consistent melody and style tend to be more easily recognized by the model. It is worth noting that the precision value of classical in Table 4 is not the highest. This index evaluates the proportion of songs whose actual genre is classical in the songs whose predicted genre is classical by the model. Although Classical has a high recognition rate, Jazz, which is similar in style, is easily misclassified as classical. Fewer songs were misclassified as Blues. Therefore, Blues' Precision index is the highest.

Table 4 Precision, Recall, and F-score of each genre obtained on the GTZAN dataset

Genre	Precision	Recall	F-score
Blues	98.0%	97.0%	97.5%
Classical	95.2%	99.0%	97.1%
Country	85.3%	93.0%	89.0%
Disco	89.9%	89.0%	89.4%
Hiphop	93.9%	92.0%	92.9%
Jazz	95.0%	96.0%	95.5%
Metal	90.6%	96.0%	93.2%
Pop	84.8%	84.0%	84.4%
Reggae	96.7%	87.0%	91.6%
Rock	85.3%	81.0%	83.1%

Pop and Rock are generally marked by an inconstant rhythmic element, various styles, and a complex structure. Take Rock.23 of Rock genre in GTZAN dataset as an example. Rock.23 is a 30-s clip of the Bohemian Rhapsody Ballad part. Promane [9] argues that the Bohemian Rhapsody style fuses elements of Glam and progressive rock with those found in musical theatre, opera buffa, and vaudeville. The whole Rock.23 clip, vocal elements account for a prominent proportion, accompanied by a piano solo. Although the genre of Rock.23 is defined as Rock, the rock characteristics of the song are not obvious, and the content is not inclined to rock music. Under the method of song slice and voting mechanism, the more complex structure of a song, the more significant difference in the prediction results of each clip. It is unsuitable to employ a voting mechanism to ensemble the final results. Therefore, the various styles of songs and the method of this paper are fundamental reasons for the low recognition rate of Rock and Pop.

5.2 Ablation study for attention

In order to explore the effect of the weighting method based on time domain, frequency domain, and time–frequency domain for building global time–frequency dependencies in the spectrogram, we conducted experiments with different settings in Table 5. Similarly, to verify the function of two parts in DPA, we conducted experiments with different settings in Table 6.

As shown in Table 5, we found that applied PCA in CNN-5 brings a remarkable improvement in accuracy rate when comparing baseline CNN-5 with CNN-5 + PCA. Specifically, applied based on time domain, frequency domain, and time–frequency domain PCA in CNN-5 respectively improved 1.9%, 1.6%, 1.7% accuracy rate compared with baseline. Music is a 1D time-series signal, and it is the most effective to use based on

Table 5 Experimental results of constructing global time–frequency dependencies by PCA based on time domain, frequency domain and time–frequency domain

Method	Time domain	Frequency domain	Accuracy
CNN-5 (baseline)			89.30%
CNN-5 + PCA			91.20%
CNN-5 + PCA			90.90%
CNN-5 + PCA			91.00%

Table 6 Ablation experimental results of PCA and SE Attention in DPA

Method	DPA		Params size	Accuracy
	PCA (Time domain)	SE Attention		
CNN-5 (Baseline)			6.0 MB	89.30%
CNN-5			8.0 MB	90.50%
CNN-5			8.4 MB	91.20%
CNN-5			10.5 MB	91.40%

time-domain PCA to build time–frequency dependencies. Similarly, music was expressed as frequency signals after Fourier transform, based on frequency domain PCA also works. However, classification accuracy based on the time–frequency domain PCA is not fantastic. We argue that: building time–frequency dependencies with time domain and frequency domain weighting fusion, which time-series and frequency fuse in the feature. However, the mixture features could not represent an audio signal, and classification accuracy has not improved further.

Table 6 indicates that applied PCA and SE Attention in CNN-5, respectively, improved by 1.9% and 1.2% compared with baseline. We found that applied PCA in.

CNN-5 improved model performance more than SE attention. These results note that when Mel-spectrogram is the input feature, there is a significant similarity between channel feature information. Even if SE attention is applied in CNN-5, the performance improvement of model is limited. However, the fixed receptive field of traditional CNN cannot capture global time–frequency information in the song, then applying PCA in CNN-5 works excellently. Finally, we applied DPA in CNN-5, and accuracy improved by 2.1% compared with baseline, outperforming SE Attention and PCA. This result verifies that DPA focuses on global time–frequency dependencies in the song and adaptively calibrates contribution of different channels to the feature map. (PCA used time domain weighting). In addition, we can observe that with CNN-5 only, the params size is 6.0 MB. When SE Attention and PCA (Time domain) are applied separately, the params size are 8.0 MB and 8.4 MB, respectively. It can be seen that SE Attention and PCA (Time domain) have increased the params size by 2.0 MB and 2.4 MB. Finally, when DPA is applied, the params size is 10.4 MB. This shows that when the two attention mechanisms are applied in parallel, the params size of model is accumulated by adding.

5.3 Attention applied in CNN-5

In this section, we study the effect of different numbers and positions of DPA applied in CNN-5 for performance. Specifically, we performed experiments on applying different numbers of DPA in CNN-5. Next, fixed the number of DPA and experiments with DPA applied in different positions of CNN-5. As shown in Fig. 10, it improves performance most that applied DPA in the second, third, fourth, fifth layer of CNN-5, which is higher 2.1% than baseline. It brings the least improvement that applied DPA in second, third and third, fifth layer of CNN-5, which is only 1.0% higher than baseline. In Fig. 11, we found that the performance is sensitive to positional relationship when two DPA applied in CNN-5, the gap between Max and Min reach 1.0%, and average value is closer to Min. When

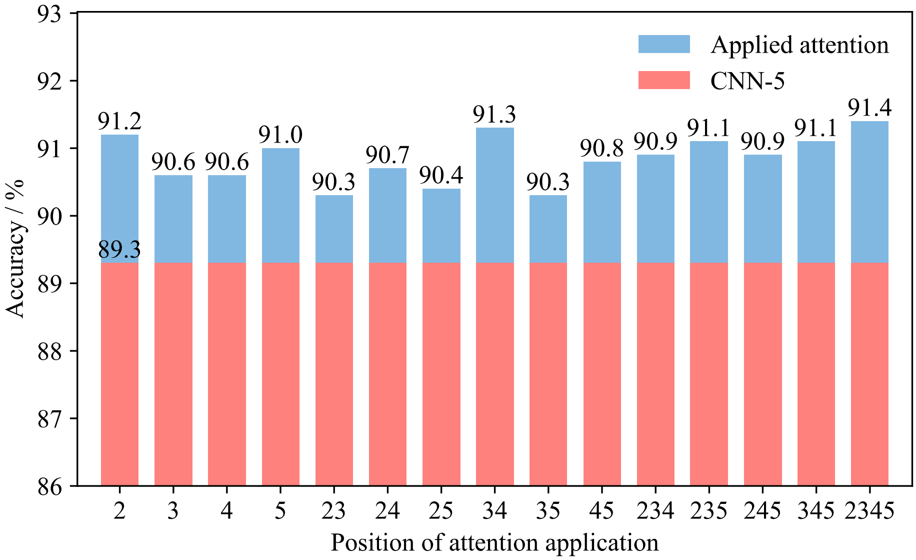


Fig. 10 The accuracy rate of different numbers and positions of attention applied in CNN-5. The horizontal axis denotes position and number of attention mechanism applied in CNN-5, and the vertical axis denotes accuracy rate. For example: "245" means that attention applied in the second, fourth, and fifth layers of CNN-5

three DPA applied in CNN-5, performance has a minor dependency in position, the gap between Max and Min only 0.2%, and average value in the middle of extreme values. (In this paper, applied attention to second, third, fourth, and fifth layers of CNN-5).

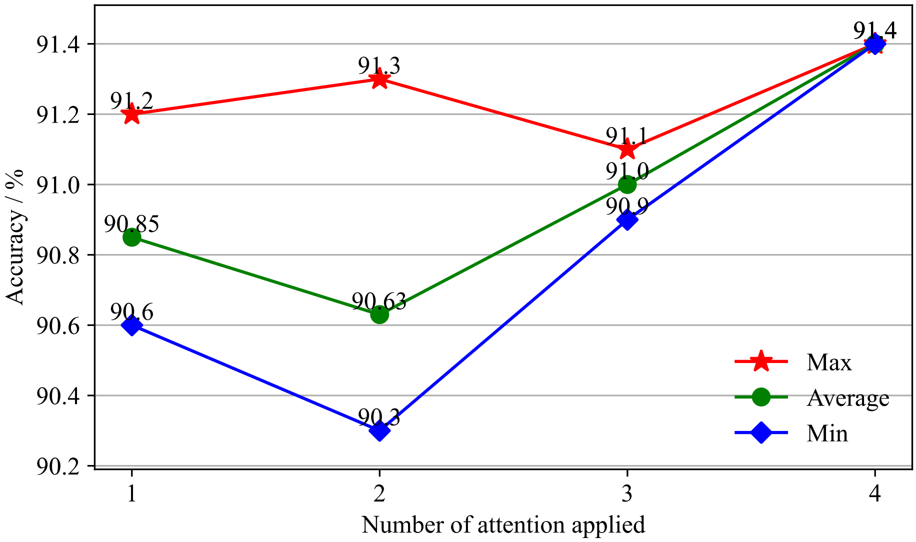


Fig. 11 The maximum value (Max), minimum value (Min), an average value of all positions (Average) denotes accuracy that applied the same number of attention mechanism in different positions of CNN-5

Table 7 Comparison results of applying multiple attentions in CNN-5

Method	Accuracy
CNN-5 (baseline)	89.30%
CNN-5 + Non-local	86.10%
CNN-5 + DANet	86.30%
CNN-5 + FLA	90.10%
CNN-5 + PTS-A	90.30%
CNN-5 + DPA (Ours)	**91.40%**

5.4 Contrast study for multiple attention mechanism

In this section, we compare DPA with Non-local, dual attention networks (DANet) [51], frame-level attention (FLA), and parallel time–frequency attention (PTS-A). Non- local and DANet are similar to DPA in architecture, which builds long-distance feature dependencies for features weighted summation through weight matrix. FLA and PTS-A are similar to DPA in function, which combines time-series or frequency characteristics of the audio signal.[2]

As shown in Table 7, we proposed the DPA has highest accuracy. The results show that the accuracy of applying Non-local and DANet in CNN-5 is 3% lower than the baseline. We argue that: squeezed the channel into a row, and then the features of all channels are multiplied together to obtain the weight matrix representing any two feature.dependencies in the feature map, and finally, the weighted summation of original feature to build global dependencies, this method, Non-local and DANet, is not suitable to a spectrogram. The spectrogram is composed of each temporal frame in the time domain or each sound frequency in the frequency domain arranged in parallel. Squeezing the spectrogram into a row, the time frame or sound frequency is connected end to end, violating the spectrogram's characteristic. Therefore, the accuracy rate decreases. In addition, the accuracy of applying FLA and PTS-A in CNN-5 is also lower than that of CNN-5 with DPA.

6 Conclusion

Automatic music genre classification is a research topic that classifies music (songs) into different genres according to their content. This can replace tedious manual labeling methods and provide a theoretical basis for commercial applications of music genre classification. In this paper, we proposed to apply dual parallel attention (DPA) in CNN-5 for music genre classification. DPA is composed of parallel channel attention (PCA) and SE Attention. PCA employs a weight matrix to construct new feature by weighted summation of time–frequency information to build global time–frequency dependencies in the song. This paper also has research on the effect of the weighting method based on time domain, frequency domain, and time–frequency domain for building global time–frequency dependencies. Among these methods, the based time domain method is the most effective. In

[2] It is worth noting that: Non-local and DANet are codes provided by the author. FLA and PTS-A are implemented based on the content of the paper. All attention is applied in CNN-5 for comparison in the same way.

addition, we analyzed the effect of different numbers and positions of DPA applied in CNN-5. The experimental results demonstrate that: the performance of applying DPA in the second, third, fourth, and fifth layers of CNN-5 is the most outstanding. Moreover, when two DPA are applied, the change of position relationship has a sensitive effect on the performance. When three DPA are applied, performance has a minor dependency on position relationships. Compared to the Non-local and DANet, etcetera attention mechanism, the classification accuracy of DPA is the highest in the optimal application position setting.

In commercial applications: This method can provide users with acceptable classification accuracy by retrieving music from different genres. More importantly, we propose building global dependencies of the song. It is worth thinking deeply and can provide new ideas for music genre classification. Similarly, there are still areas for improvement in this article. When DPA is used in the basic network CNN-5, it brings more parameters and computational overhead to the model. However, the classification accuracy of the model is not significantly improved. In future work, we will continue to focus on how to reduce the computational complexity of DPA. At the same time, recognizing music genres with rich styles based on hand-crafted features is worth researching.

Funding This work was supported by Postgraduate Scientific Research Innovation Project of Hunan Province (CX20210879), Postgraduate Scientific Research Innovation Project of Central South University of Forestry and Technology (CX202102059) and Hunan Key Laboratory of Intelligent Logistics Technology (2019TP1015).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ashraf M et al (2020) A Globally Regularized Joint Neural Architecture for Music Classification. IEEE Access 8:220980–220989
2. Cai X, Zhang H (2022) Music genre classification based on auditory image, spectral and acoustic features. Multimedia Syst 28(3):779–791
3. Downie JS (2003) Music information retrieval. Ann Rev Inf Sci Technol 37(1):295–340
4. Fu Z et al (2011) A Survey of Audio-Based Music Classification and Annotation. IEEE Trans Multimedia 13(2):303–319
5. Gao Y (2020) Research on Music Audio Classification Based on Deep Learning. South China University of Technology Guangzhou, China
6. Gardner MW, Dorling S (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ 32(14–15):2627–2636
7. Huang G-B, Zhu Q-Y, Siew C-K (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1–3):489–501
8. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
9. Promane BC (2009) Freddie mercury and queen: Technologies of genre and the poetics of innovation. University of Western Ontario, School of Graduate and Postdoctoral Studies
10. Sarikaya R, Hinton GE, Deoras A (2014) Application of deep belief networks for natural language understanding. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22(4):778–784
11. Scalvenzi RR, Guido RC, Marranghello N (2019) Wavelet-packets associated with support vector machine are effective for monophone sorting in music signals. Int. J. Semant. Comput. 13(03):415–425
12. Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Transactions on speech and audio processing 10(5):293–302
13. Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. Journal of Big data 3(1):1–40

14. Yu Y et al (2020) Deep attention based music genre classification. Neurocomputing 372:84–91
15. Zhang X et al (2019) Spectrogram-frame linear network and continuous frame sequence for bird sound classification. Eco Inform 54:101009
16. Zhang Z et al (2021) Attention based convolutional recurrent neural network for environmental sound classification. Neurocomputing 453:896–903
17. Schedl M, Gómez Gutiérrez E, and Urbano J (2014) Music information retrieval: Recent developments and applications. Foundations and Trends in Information Retrieval. 12; 8 (2–3): 127–261
18. Ndou N, Ajoodha R, Jadhav A (2021) Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches. in 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS). IEEE
19. Gupta R, Yadav J, and Kapoor C (2021) Music information retrieval and intelligent genre classification. in Proceedings of International Conference on Intelligent Computing, Information and Control Systems Springer
20. Pálmason H, et al (2017) Music genre classification revisited: An in-depth examination guided by music experts. in International Symposium on Computer Music Multidisciplinary Research 7 Springer
21. Baniya BK, Ghimire D, Lee J (2014) A novel approach of automatic music genre classification based on timbrai texture and rhythmic content features. in 16th International Conference on Advanced Communication Technology IEEE
22. Arabi, A.F. and G. Lu. Enhanced polyphonic music genre classification using high level features. in 2009 IEEE International Conference on Signal and Image Processing Applications. 2009. IEEE
23. Saunders C et al (1998) Support vector machine reference manual
24. Sarkar R, and Saha SK (2015) Music genre classification using EMD and pitch based feature. in 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR) IEEE
25. Vaswani A et al (2017) Attention is all you need. in Advances in neural information processing systems
26. He K et al (2016) Deep residual learning for image recognition. in Proceedings of the IEEE conference on computer vision and pattern recognition
27. Piczak KJ (2015) Environmental sound classification with convolutional neural networks. in 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP) IEEE
28. Himawan I, Towsey M, Roe (2018) P 3D convolution recurrent neural networks for bird sound detection. in Proceedings of the 3rd Workshop on Detection and Classification of Acoustic Scenes and Events. Detection and Classification of Acoustic Scenes and Events
29. Kahl S et al (2017) Large-Scale Bird Sound Classification using Convolutional Neural Networks, in CLEF (working notes)
30. Yang B (2008) A study of inverse short-time Fourier transform. in 2008 IEEE Int. Conf. Acoust. Speech Signal Process. IEEE
31. Zhang W et al (2016) Improved Music Genre Classification with Convolutional Neural Networks, in Interspeech 2016. 3304–3308
32. Choi K et al (2017) Convolutional recurrent neural networks for music classification. in 2017 IEEE Int. Conf. Acoust. Speech Signal Process (ICASSP) IEEE
33. Cho K et al (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078
34. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit
35. Yang H, Zhang W.-Q (2019) Music Genre Classification Using Duplicated Convolutional Layers in Neural Networks, in Interspeech 2019 3382–3386
36. Chang P-C, Chen Y-S, Lee C.-H (2021) MS-SincResNet: Joint Learning of 1D and 2D Kernels Using Multi-scale SincNet and ResNet for Music Genre Classification, in Proceedings of the 2021 Int. Conf Multimed. Retr.. 29–36
37. Choi K et al (2017) Transfer learning for music classification and regression tasks. arXiv preprint arXiv:1703.09179
38. Srinivasu PN et al (2022) Ambient Assistive Living for Monitoring the Physical Activity of Diabetic Adults through Body Area Networks. Mob. Inf. Syst **2022**
39. Wang X et al (2018) Non-local neural networks. in Proceedings of the IEEE Conf. Comput. Vis. Pattern Recognit
40. Wang H et al (2019) Environmental sound classification with parallel temporal-spectral attention. arXiv preprint arXiv:1912.06808
41. Huang Z et al (2022). ADFF: Attention Based Deep Feature Fusion Approach for Music Emotion Recognition. arXiv preprint arXiv:2204.05649
42. Dosovitskiy A et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

43. Gong Y, Chung Y-A, and Glass J (2021) Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778
44. Yang L, and Zhao H (2021) Sound Classification Based on Multihead Attention and Support Vector Machine. Math. Probl. Eng **2021**
45. Lin M, Chen Q, and Yan S (2013) Network in network. arXiv preprint arXiv:1312.4400
46. Ioffe S, and Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. in Int confe machine learning. PMLR
47. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. in Icml
48. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980
49. Zhang P et al (2015) A Deep Neural Network for Modeling Music, in Proceedings of the 5th ACM on International Conference on Multimedia Retrieval 379–386
50. Karunakaran N, Arya A (2018) A scalable hybrid classifier for music genre classification using machine learning concepts and spark. in 2018 Int Confe Intell Auton Syst (ICoIAS) IEEE
51. Fu J et al (2019) Dual attention network for scene segmentation. in Proceedings of the IEEE/CVF Conf. Comput. Vis. Pattern Recognit

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.