Hindawi
Mobile Information Systems
Volume 2021, Article ID 7629994, 10 pages
https://doi.org/10.1155/2021/7629994

Research Article

Music Feature Classification Based on Recurrent Neural Networks with Channel Attention Mechanism

Jie Gan

Huanghuai University, Zhumadian, Henan 463000, China

Correspondence should be addressed to Jie Gan; 20091153@huanghuai.edu.cn

Received 7 April 2021; Revised 11 May 2021; Accepted 18 May 2021; Published 11 June 2021

Academic Editor: Fazlullah Khan

With the advancement of multimedia and digital technologies, music resources are rapidly increasing over the Internet, which changed listeners' habits from hard drives to online music platforms. It has allowed the researchers to use classification technologies for efficient storage, organization, retrieval, and recommendation of music resources. The traditional music classification methods use many artificially designed acoustic features, which require knowledge in the music field. The features of different classification tasks are often not universal. This paper provides a solution to this problem by proposing a novel recurrent neural network method with a channel attention mechanism for music feature classification. The music classification method based on a convolutional neural network ignores the timing characteristics of the audio itself. Therefore, this paper combines convolution structure with the bidirectional recurrent neural network and uses the attention mechanism to assign different attention weights to the output of the recurrent neural network at different times; the weights are assigned for getting a better representation of the overall characteristics of the music. The classification accuracy of the model on the GTZAN data set has increased to 93.1%. The AUC on the multilabel labeling data set MagnaTagATune has reached 92.3%, surpassing other comparison methods. The labeling of different music labels has been analyzed. This method has good labeling ability for most of the labels of music genres. Also, it has good performance on some labels of musical instruments, singing, and emotion categories.

1. Introduction

With the popularity of computers and mobile phones, more and more people choose to listen to music on the Internet instead of using traditional tapes and CDs. The change in the way people listen to music has led to an explosive growth of digital music. More and more music websites that provide online music listening services have emerged. Facing a large amount of digital music on the Internet, quickly and accurately retrieving the music users' want has become more important. It has become an essential indicator of the user-friendliness of a music website [1]. Music genre classification [2] is an essential branch of music information retrieval. Correct music classification is of great significance for improving the efficiency of music information retrieval. At present, music classification mainly includes text classification and classification based on music content. Text classification is mainly based on music metadata information, such as singer, lyrics, songwriter, age, music name, and other labeled text information [3]. This classification method is easy to implement, simple to operate, and fast to retrieve, but the shortcomings are also evident. First of all, this method relies on manually labeled music data, which requires much workforce, and manual labeling is challenging to avoid incorrectly labeling music information. Secondly, this texting method does not involve the audio data of the music itself. Audio data includes many vital characteristics of music, such as pitch, timbre, melody, pitch, etc. These characteristics are almost impossible to label with the text [4]. Based on the classification of content is to extract the features of the original music data and use the extracted feature data to train the classifier to achieve the purpose of music classification. Therefore, music classification based on content has also become a research hotspot in recent years. Based on this, the research direction of this article is also based on content-based music classification [5].

At present, artificial intelligence [7–9] is in full swing, and deep learning [22, 24, 25] is the most significant credit. Deep learning is widely used in image processing and speech recognition. It has achieved much better results than traditional machine learning methods [23]. Thus, many researchers have also begun to introduce deep learning technology into the field of music information retrieval [16]. The correct classification of music genres helps improve the efficiency of music retrieval and improve music recommendations' accuracy. For example, suppose the user likes rock song A. In that case, the recommendation system can recommend song B under the rock category to the user based on the similarity calculation. Since songs A and B are both rock-style songs, the user feels that the recommendation is very accurate, invisible and improves the product's user experience.

Therefore, by stacking the convolutional layers [10, 11], the network can extract more abstract sound spectrum features layer by layer. However, the music signal is a kind of timing information. The music features at different moments may have a timing correlation that ignores the timing information inside the music [12]. In response to these problems in the previous methods, this paper combines the proposed convolutional neural network with the bidirectional recurrent neural network. It proposes a music classification model based on the convolutional recurrent neural network so that the model can learn the timing information in the music. Moreover, by assigning different attention weights to the output of the recurrent neural network at different moments, a better representation of the overall music features [13] can be obtained.

The following are the main contribution points of this paper:

(1) This paper combines the proposed convolutional neural network with the bidirectional recurrent neural network. It proposes a music classification model based on the convolutional recurrent neural network. The model can learn the time sequence information in the music. Through the neural network for different moments, different attention weights are assigned to the output to represent the overall characteristics of the music better.

(2) This paper converts the audio signal of music into a sound spectrum. The sound spectrum records the time domain and frequency domain information of the music signal. The scale of the data is reduced on the premise of preserving the music information to the greatest extent. The conversion methods of different music are unified, which avoids problems with manual feature selection.

(3) In this paper, comparison and ablation experiments were performed on the GTZAN and MagnaTagATune data sets. The experimental results prove the effectiveness and superiority of our algorithm.

The rest of the paper is organized as follows. The background study is given in Section 2, followed by the methodology in Section 3. Section 4 discusses the results, and the conclusion is given in Section 5.

2. Background

This section discusses the properties of sound signals and the element of sound signals in detail.

2.1. Basic Properties of the Sound Signals. Frequency is a physical quantity [14] that describes the number of vibrations of a sounding object per unit of time. The international system unit of frequency is Hz, representing the number of times the unit vibrates in 1 second. Music is mainly composed of human voice and instrument sound. The frequency, intensity, and duration of vibration of different instruments are not the same. The auditory system perceives a wide variety of tones and timbres. When the vibration wave propagates on the object, it will be reflected continuously. The reflected wave will form a standing wave when it encounters the superposition of the following waves. The standing wave is why the vibration of an object always has a fixed frequency and a fixed tone because only some frequency waves can form a standing wave in the object to continue to propagate [15]. In contrast, other frequency waves will quickly dissipate, thereby forming overtones. The sound with the strongest vibrational energy and the lowest frequency of an object is called the fundamental tone. Its frequency is called the fundamental frequency, and other sounds whose frequencies are integer multiples of the fundamental tone are homophonic [16]. The fundamental and overtone frequencies of several different musical instruments are shown in Table 1.

In mechanical vibration, the maximum value of the distance between the sound wave generated by the object's vibration and the equilibrium position of the object is called the amplitude. Amplitude measures the vibration energy and amplitude of the object. It is numerically equal to the maximum distance between the object and the equilibrium position during vibration. Loudness and amplitude are often discussed together. The main difference between the two is that the amplitude is a physical quantity and the loudness. The former can be obtained by analyzing sound waves [17]. The latter is a psychological quantity that characterizes the human ear's auditory perception of sound size, except for and in addition to amplitude, it is also related to frequency. Phase is an essential attribute of vibration waveform. Divide the difference between the two adjacent pole values in the horizontal axis projection of the coordinate system in two periodic motions of the same frequency and divide by the period size to get the ratio. Then, convert it into radians to obtain the phase of the waveform. It describes the state of a vibrating object at a particular moment. The phase of the simple harmonic motion is analogous to the angle of uniform circular motion. As time changes, the angle of motion will also change. The phase changes by 2π, which means that the uniform circular motion has performed a circle equivalent to the sounding object, and undergoes a vibration period. In this way, through a particular moment of phase, the object's position in the vibration can be obtained.

Table 1: Comparison of fundamental and overtone frequencies of different instruments.

Instrument type	Pitch frequency	Overtone frequency
Violin	196~1320 Hz	>12 kHz
Flute	247~2 kHz	>6 kHz
Piano	27.5~4186 Hz	10~15 kHz
Electric guitar	82~174 Hz	>10 kHz
Trumpet	165~1175 Hz	>15 kHz

2.2. Basic Elements of Music. As an abstract art form formed by a melody with rhythm, instrumental sound, and harmony, music can bring rich auditory enjoyment. Various musical instruments are used in different music, and the way the singer sings is also different. At the same time, there are differences in melody and rhythm. The human ear first hears a combination of different tones, timbres, or loudness. It then processes these music signals through the brain to produce high-level perceptions, such as genres, emotions, etc. [18]. From the above process of understanding music by the brain, it can be seen that the analysis of music data first needs to understand the characteristics of music as an audio signal. The three essential elements that make up music are introduced below.

Pitch describes the size of the vibration frequency of the musical instrument. The high-frequency musical instrument produces a higher pitch. The frequency of different musical instruments is related to their material and shape [19]. The human ear is compassionate to pitch, and the pitch is different. Male voices generally feel thick and complete, and female voices are generally bright and high-pitched. Female voices are generally an octave higher than male voices, so male midrange many female trebles. As shown in Table 1, different musical instruments have different frequencies of vibration and sound. For example, the fundamental frequency of a piano is between 27.5 Hz and 4186 Hz. The range is more comprehensive than other musical instruments, which enables the piano to produce rich sounds [20].

Sound intensity describes the volume of the human auditory system. Loudness is related to the amplitude described above. Generally speaking, the greater the amplitude, the louder the sound, but the relationship between amplitude and loudness is not linear. It is also related to the waveform and frequency of the sound. The human brain will feel that the sound between 1000 Hz and 4000 Hz has a higher loudness in the same sound intensity [21]. Outside this range, as the frequency increases or decreases, the human ear will feel less and less louden sensitive. When the frequency is outside the range of 20 Hz to 20000 Hz, the auditory system will not feel the sound. Therefore, the relationship between the human ear's perception of sound intensity and amplitude is not linear. The tone is often described as the color of the sound. Human ears can distinguish sounds with the same pitch and intensity through the difference in timbre. The number of homophonies and the intensity of the tone are essential factors that affect the sense of timbre [19]. According to Fourier's theory, any complex sound vibration process can be decomposed into many essential components. When these basic components are added together, the overall process of complex vibration can be described.

3. Methodology

Although the network can extract more abstract sound spectrum features layer by layer by stacking the convolutional layers, the music signal is a kind of timing information. Even if it is converted into a mel sound spectrum, the sound spectrum has a time dimension. Sequentially, simply using the convolution structure will ignore the timing information inside the music. One-dimensional convolution performs translation in the time dimension. While capturing the local sound spectrum characteristics, it also ignores the sequence relationship of the sound spectrum characteristics of different time frames. Only one-dimensional convolution cannot effectively model the music sequence relationship. In response to the above problems, this paper proposes a new type of convolutional recurrent neural network model, which can learn the timing information in music. Considering that the musical characteristics of a piece of music at different moments may have different effects on the overall category of music, the attention mechanism is used to assign different attention weights to the cyclic neural network's output at different moments sequence features are aggregated.

3.1. Music Sequence Modeling. In this section, we will provide details about recurrent neural networks and their importance in music modeling.

3.1.1. Basic Principles of Recurrent Neural Networks.

(1) *The Basic Structure of RNN.* In RNN, the output state of the current hidden layer is not only related to the input at the current moment but also depends on the state of the hidden layer at the last moment. This structure gives the network memory-like characteristics, and there are dependencies on contexts such as sequence prediction and classification. The problem is suitable to be solved by RNN. The basic network structure of RNN is shown in Figure 1. $X^{(i)}$ is the input of the ith step; $H^{(i)}$ is the state of the ith step of the hidden layer, which is a network unit with memory function in the RNN. $H^{(i)}$ is calculated from the input $X^{(i)}$ of the current ith input layer and the state $H^{(i-1)}$ of the previously hidden layer. The calculation equation is as follows:

$$H^{(i)} = f\left(WX^{(i)} + VH^{(i-1)} + b\right). \tag{1}$$

Here, f is a nonlinear activation function. In RNN, it is generally tanh b is the bias term and W is the connection matrix of the input layer. The weight matrix between the hidden layer at the previous time and the hidden layer at the next time is remarked as V said.

O_i is the output of the network at step i, where U is the connection matrix of the output layer, and c is the bias term:

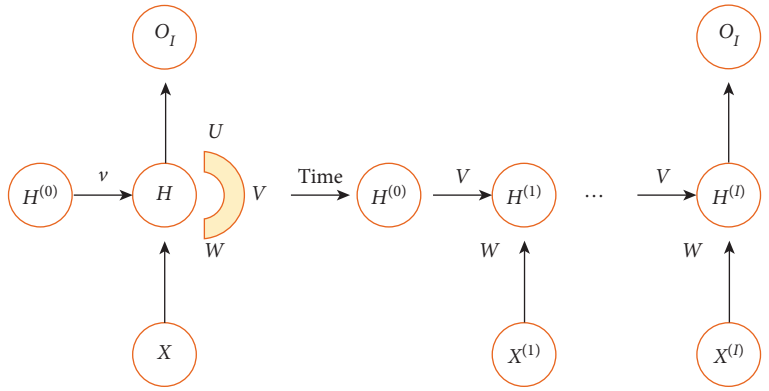

Figure 1: RNN network schematic diagram.

$$O_i = UH^{(i)} + c. \tag{2}$$

If equation (2) is looped into equation (1), ignoring the bias term can get

$$\begin{aligned} O_i &= UH^{(i)} \\ &= Uf\left(WX^{(i)} + VH^{(i-1)}\right) \\ &= Uf\left(WX^{(i)} + Vf\left(WX^{(i-1)} + VH^{(i-2)}\right)\right) \\ &= Uf\left(WX^{(i)} + Vf\left(WX^{(i-1)}\right.\right. \\ &\quad \left.\left. + Vf\left(WX^{(i-2)} + VH^{(i-3)}\right)\right)\right) \\ &= Uf\left(WX^{(i)} + Vf\left(WX^{(i-1)}\right.\right. \\ &\quad \left.\left. + Vf\left(WX^{(i-2)} + Vf\left(WX^{(i-3)} + \cdots\right)\right)\right)\right). \end{aligned} \tag{3}$$

It can be seen from the above equation that the output of the RNN not only depends on the current input but also is related to the previous historical input, which is why it has memory.

(2) *LSTM and GRU Memory Unit.* If a sequence is long enough, it is difficult for RNN to transfer information from a relatively early time step to a later time step. When the sequence is relatively long, it may lose information related to the task at the last time. In the process of backpropagation, RNN will encounter the problem of gradient disappearance. The existence of this problem makes it difficult for RNN to be widely used. In response to this problem, the academic community has proposed various memory unit variants; the most popular ones are LSTM and GRU.

The long short-term memory network (LSTM) has a special gate structure, enabling the network to realize the selective memory function of long sequences. The dependency between the sequence data before and after the sequence data can better make the learning process better. Three gate structures play a major role in LSTM: forget gate, input gate, and output gate, which correspond to g_f, g_i, and g_o in the figure. At the same time, to preserve the long-term state of the network, LSTM adds a state Ccalled the cell state. The forget gate g_f determines how much information about the cell state $C^{(i-1)}$ at the previous moment will be retained to the current moment $C^{(i)}$. The input gate g_i determines how much information about the network's input at the current moment will be retained into the cell state $C^{(i)}$ output gate g_o. How much of the control unit state $C^{(i)}$ is output to the output value $H^{(i)}$ of the current unit? The following formula can calculate the LSTM gate structure:

$$g_i = \sigma\left(W_i X^{(i)} + V_i H^{(i-1)}\right), \tag{4}$$

$$g_f = \sigma\left(W_f X^{(i)} + V_f H^{(i-1)}\right), \tag{5}$$

$$g_o = \sigma\left(W_o X^{(i)} + V_o H^{(i-1)}\right), \tag{6}$$

$$g_c = \tanh\left(W_c X^{(i)} + V_c H^{(i-1)}\right), \tag{7}$$

where σ represents the Sigmoid function, W_f and V_f are the parameter matrices of the forget gate, and W, V are the same. From this, the cell state $C^{(i)}$ and output $H^{(i)}$ in the LSTM can be obtained as follows:

$$C^{(i)} = g_c \circ g_i + g_f \circ C^{(i-1)}, \tag{8}$$

$$H^{(i)} = \tanh\left(C^{(i)}\right) \circ g_o. \tag{9}$$

The structure of the GRU is shown in Figure 2. GRU is similar to LSTM in that it uses a variety of gate structures. Unlike LSTM, there are only two gate structures in GRU: reset gate and update gate. These two gates together determine how to get from the hidden state HI in the previous step. The next hidden state, H1, discards the output gate in the LSTM memory unit. If the output of the reset gate is 1 and the output of the update gate is 0, then the GRU becomes a simple RNN.

The calculation method of the door structure in GRU is as follows:

$$g_u = \sigma\left(W_u X^{(i)} + V_u H^{(i-1)}\right), \tag{10}$$

$$g_r = \sigma\left(W_r X^{(i)} + V_r H^{(i-1)}\right), \tag{11}$$

$$g_c = \tanh\left(W_c X^{(i)} + g_r \circ V_c H^{(i-1)}\right). \tag{12}$$

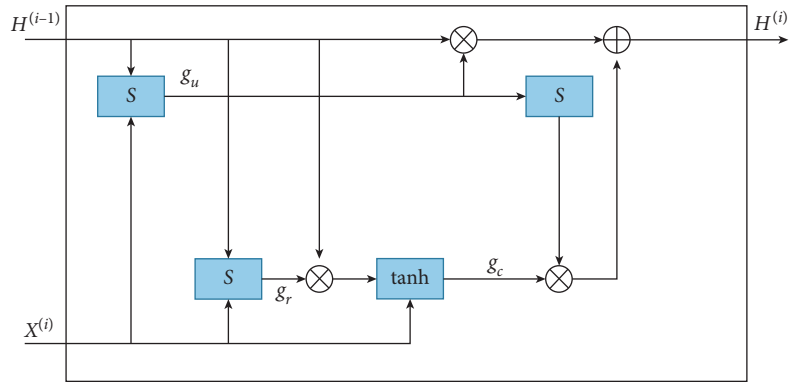

Figure 2: GRU network unit structure.

According to the above equation, the calculation method of the hidden layer state update is

$$H^{(i)} = g_u \circ H^{(i-1)} + (1 - g_u) \circ g_c. \tag{13}$$

3.1.2. Music Sequence Modeling Based on BRNN. The RNN described in the previous section can also be called a one-way RNN. The so-called one-way RNN means that the output of its next step is only affected by the input of all the previous steps. The two-way RNN believes that not only the previous input needs to be considered, but the latter input may also be affected. Modeling the data is helpful. Figure 3 shows the structure of BRNN. $\overrightarrow{H}^{(i)}$ is related to $\overrightarrow{H}^{(i-1)}$ forwarding calculation, $\overleftarrow{H}^{(i)}$ is related to $\overleftarrow{H}^{(i+1)}$ reverse calculation, and $\overrightarrow{H}^{(i)}$ represents the state of the hidden layer. The calculation equation $\overleftarrow{H}^{(i)}$ is as follows:

$$\overleftarrow{H}^{(i)} = f\left(W' X^{(i)} + V' \overleftarrow{H}^{(i+1)}\right). \tag{14}$$

Then, add the forward and reverse of each step of the network to get the final output of the network:

$$O_i = U\overrightarrow{H}^{(i)} + U'\overleftarrow{H}^{(i)}. \tag{15}$$

3.2. Sequence Feature Aggregation Based on Attention Mechanism. After BRNN is used to model the music feature sequence, each moment's high-level abstract feature representation is obtained. Before being passed into the fully connected layer, the abstract features of all moments are usually aggregated into an overall feature representation. The most common aggregation method is in the time dimension. The maximum pooling and average pooling of all-time features are performed. For music, the appearance of a specific sound spectrum feature at different moments of the music may have different contributions to the music category corresponding to the feature. For example, the same melody at the beginning or end of the music will give people different feelings. Certain audio features appearing at a particular moment may also be related to certain categories, such as emotion-related categories.

Figure 4 shows the attention module used in this section. Suppose the sound spectrum feature sequence output by the convolutional layer can be expressed as $X = [x_1, x_2, x_3, \cdots, x_L]$representing the time dimension of the convolution feature map, that is, the sequence length. The following formula can obtain the attention weight of each sequence feature:

$$a = \text{soft}\max\left(W_2 \varphi\left(W_1 X^T\right)\right), \tag{16}$$

where W_1 and W_2 are the weight matrix. The softmax function ensures that the attention weights of all feature sequences add up to 1 and φ represents the tanh activation function. It can be seen from the formula that attention weight learning is very similar to the forward neural network that removes the bias term. After obtaining the attention weight vector a, the overall feature representation v of the sequence feature is calculated as follows:

$$\begin{aligned} v &= \sum_{t=1}^{L} a_t x_t \\ &= aX. \end{aligned} \tag{17}$$

The overall feature representation is obtained by the weighted summation of the sequence features according to the corresponding attention weights. Through the attention mechanism, the abstract features at all moments are aggregated into an overall feature vector, which can be passed to the subsequent network to complete the music classification task.

3.3. Our Model. After the convolutional layer learns the sound spectrum, a feature map containing high-level abstract features can be obtained. The feature map is expanded in time to obtain a convolution feature sequence. The convolution feature sequence is input to BRNN to model the music sequence. Then, through the network, the learned attention weight performs a weighted summation on the feature sequence output by the BRNN and integrates the output of the BRNN at multiple moments into the overall feature expression of the music. Finally, it passes it to the fully connected layer for further learning to obtain the

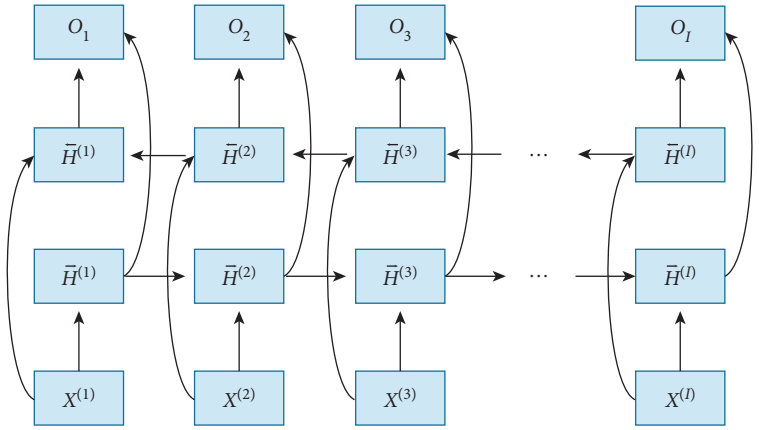

Figure 3: Schematic of BRNN network structure.

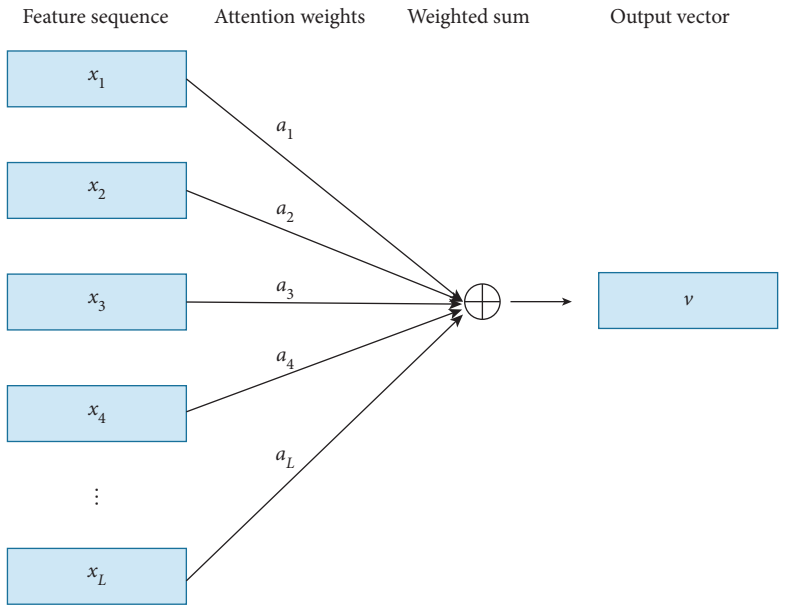

Figure 4: The calculation process of feature attention weight of music sequence.

classification result. According to the different functions of each part of the network, according to the direction of information transmission in the network, it can be divided into music representation learning layer, music sequence modeling and sequence feature aggregation layer, and fully connected layer. The network structure proposed in this section is shown in Figure 5.

4. Experiments and Results

This section discusses the experimental environment, data set, evaluation index, and performance of the proposed scheme.

4.1. Experimental Environment. This paper converts all audio samples in the two data sets into mono processing, sampling, or resampling at a sampling rate of 16 kHz. The Fourier transform window length used in the conversion to mel sound spectrum is 512; the window skip size is 256. The frequency of the number of bins is 128. For the GTZAN data set, the audio segmentation method is used. The segmentation time is 5 seconds, and the generated mel sound spectrum specification is ((313, 128). For the audio samples of the MagnaTagATune data set, no segmentation will be performed, and the mel sound will be generated. The spectrum specification is (1813, 128).

4.2. Data Set. The GTZAN data set contains 10 music genres: Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, and Rock. Each genre has 100 30s audio. The MagnaTagATune data set contains a total of 2, 5863 audios. Each audio is about 29 seconds long, and the sampling rate is 16000 Hz. The data set has 188 music tags, including genres, musical instruments, emotions, and other categories. These audio samples are selected from 5223 real songs and 445 albums, involving 230 creators. The organizer

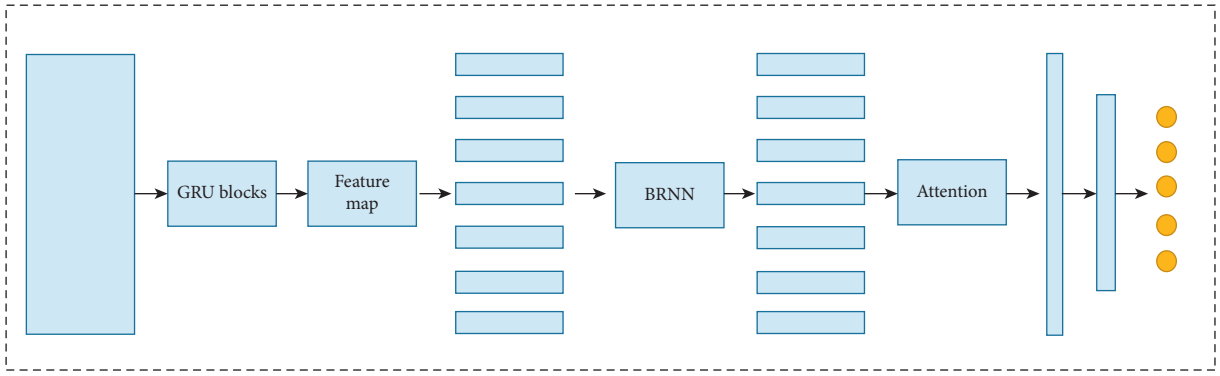

Figure 5: Schematic diagram of a music classification model based on convolutional recurrent neural network.

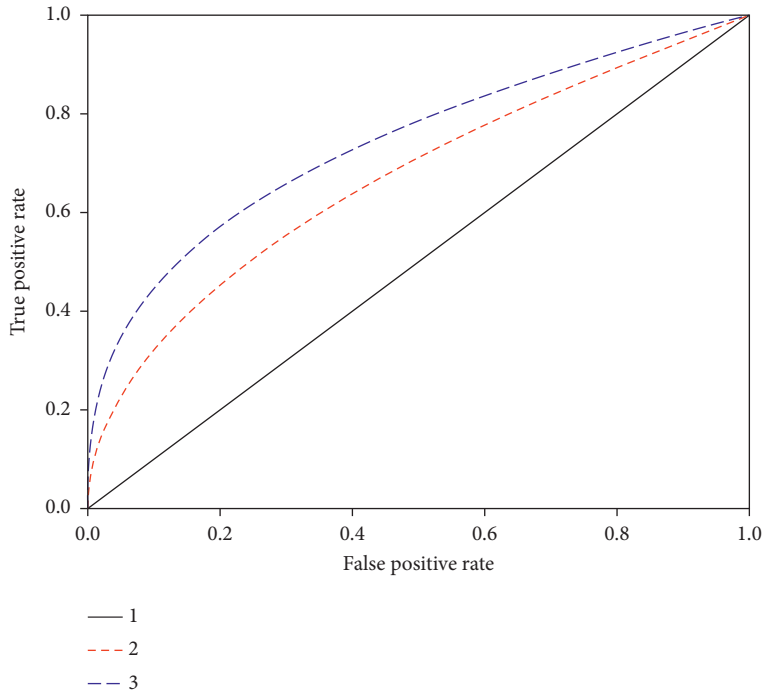

Figure 6: ROC curve diagram.

Table 2: AUC comparison of different RNN structures on GTZAN.

	GRU	LSTM	BGRU	BLSTM
RNN depth1	0.89	0.85	0.91	**0.92**
RNN depth2	0.91	0.89	0.90	**0.93**

collects the tags in the MagnaTagATune data set through an online game called TagATune. In the game, every two players will be asked to listen to the music clips provided by the game, and then the players need to give the category tags of the music clips heard. The last two players will discuss the category tags they give to determine the final music tag of the music clip.

4.3. Evaluation Index. AUC is obtained by calculating the area under the receiver operating characteristic curve (ROC). Figure 6 shows three different ROC curves. The abscissa of the coordinate axis represents the false positive rate (FPR) of the two-class model, and the ordinate represents the true positive rate (TPR). By setting different classification thresholds, different combinations of TPR and FPR will be obtained. Connecting these points in the coordinate axis constitutes a ROC curve. The AUC refers to the area enclosed by the corresponding ROC curve and the coordinate axis.

Table 3: Comparative experiment results on the GTZAN data set.

Methods	Acc (%)
CNN	81.1
LeNet	822
KCNN	83.6
RDNN	84.8
RGLU	89.2
Ours	93.1

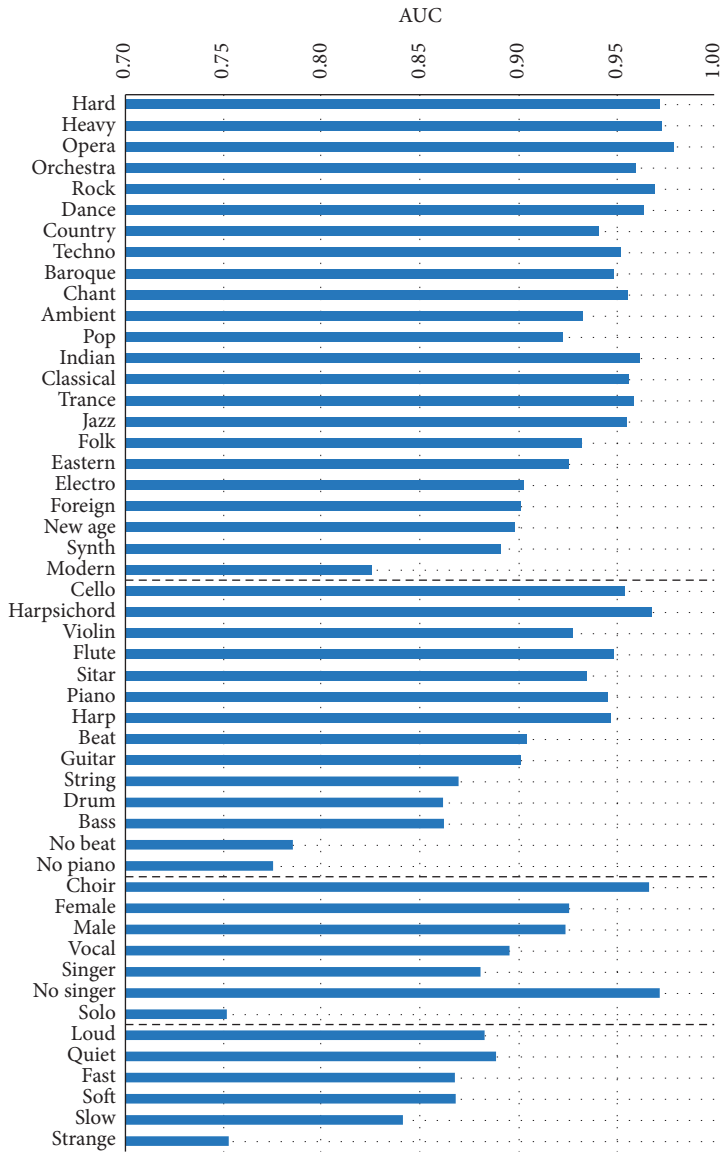

Figure 7: AUC value distribution of the top50 tags in the MagnaTagATune data set.

For multilabel classification data sets, the mutual exclusivity between labels of different categories is low. There may be a large number of labels for the same music at the same time. Therefore, the music classification on the MagnaTagATune data set can be regarded as a labeling problem. The labeling of the labels is more inclined to positively related labels. For example, the rock style of a song is evident, and it will be marked in the rock category with a high probability. However, the female or male label may be missing, which is a problem of labeling the strength of the music label, so if the model gives a certain. The label has a higher prediction probability. This label is not marked in the test sample, which is not necessarily the model prediction error. We should pay more attention to how many labeled labels are found by the model, which is the recall rate. To measure the performance of the model more comprehensively, this article will also use Recall@k as an evaluation indicator for multilabel classification. The calculation method is as follows:

$$\text{Recall@}k = \frac{\left|Y \cap R_{1:\,k}\right|}{|Y|}, \tag{18}$$

where Y represents the actual label set of the test sample and $R_{1:\,k}$ represents the set of the top k labels ranked from large to small according to the predicted probability of the model.

4.4. The Influence of Recurrent Neural Network Structure on Classification Performance. The experiment in this section will consider three main factors that affect RNN performance: the number of layers, the type of memory unit, and the direction (single/bidirectional). Different RNN structures will be used for experiments. The classification performance indicators of different RNN structures will be compared to know the cyclic nerves.

It can be seen from Table 2 that the classification effect of the BGRU and BLSTM network using the bidirectional

structure is better than that of the GRU and LSTM network of the unidirectional structure. For music, RNN's understanding of the characteristics of the sound spectrum at a certain moment is not only based on the previous musical moment but also on the way of expression of the subsequent music. Compared with understanding music in a single direction, it is in two directions. Perceiving the overall sequence information can help the model more intelligently understand the meaning of the features abstracted by the convolutional layer at a certain moment to the entire music. The network can better model the music sequence. This shows that music signals have similarities with sequence information such as speech and text. BRNN has been widely used in the related research fields of the latter two. It can also be seen from the comparison results that increasing the number of layers of the recurrent neural network does not improve the classification ability of the model. All four networks have more or less the phenomenon that the classification ability decreases as the number of RNN layers increases. Since this article uses a joint network model, there is a convolutional network layer in front. The gradient path is too long when backpropagating, which causes the difficulty of training the entire network to be increased, making the parameters of the previous network layer unable to be effectively updated. In this experiment, the single-layer BLSTM showed the best classification performance, indicating that the LSTM memory unit is more suitable for the model in this paper.

4.5. Comparison of Related Methods and Analysis of Results. Table 3 and Figure 7 show the comparison of classification performance on the GTZAN data set. The use of a two-way cyclic neural network for music sequence modeling and attention mechanism for sequence feature aggregation improves the model's classification performance on the GTZAN data set. It shows that using RNN to model short-term music sequence relationships within 5 seconds is still conducive to obtaining better music feature representation. The effect is better than maximum global pooling.

5. Conclusion

This paper proposed a novel recurrent neural network method with a channel attention mechanism for music feature classification. Because the music classification method based on convolutional neural network ignores the timing characteristics of the audio itself, therefore, this article combines the proposed convolution structure with the bidirectional recurrent neural network, proposes a music classification model based on the convolution recurrent neural network, and uses the attention mechanism to assign different attention to the output of the recurrent neural network at different times. The classification accuracy of the model on the GTZAN data set has increased to 93.1%. The AUC on the multilabel labeling data set MagnaTagATune has reached 92.3%, surpassing other comparison methods. The labeling of different music labels has been analyzed. This method has good labeling ability for most of the labels of music genres. Also, it has good performance on some labels of musical instruments, singing, and emotion categories.

Data Availability

The data used to support the findings of this study are included within the supplementary information file(s).

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

References

[1] Z. Li, "Design and implementation of fashion music resource website based on asp," *Journal of Physics: Conference Series*, vol. 1544, no. 1, Article ID 012194, 2020.

[2] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 4–21, 2018.

[3] J.-H. Su, C.-Y. Chin, T.-P. Hong, and J.-J. Su, "Content-based music classification by advanced features and progressive learning," in *Proceedings of the Asian Conference on Intelligent Information and Database Systems*, pp. 117–130, Springer, Yogyakarta, Indonesia, April 2019.

[4] Y. Mao, G. Zhong, H. Wang, and K. Huang, "MCRN: a new content-based music classification and recommendation network," in *Proceedings of the International Conference on Neural Information Processing*, pp. 771–779, Springer, Bangkok, Thailand, November 2020.

[5] F. Khan, A. U. Rehman, Y. Zhang et al., "A secured and reliable continuous transmission scheme in cognitive HARQ-aided internet of things," *IEEE Internet of Things Journal*, 2021.

[6] Y. V. S. Murthy and S. G. Koolagudi, "Content-based music information retrieval (CB-mir) and its applications toward the music industry," *ACM Computing Surveys*, vol. 51, no. 3, pp. 1–46, 2018.

[7] X. Ning, F. Nan, S. Xu, L. Yu, and L. Zhang, "Multi-view frontal face image generation: a survey," *Concurrency and Computation: Practice and Experience*, Article ID e6147, 2020.

[8] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[9] X. Ning, W. Li, and J. Xu, "The principle of homology continuity and geometrical covering learning for pattern recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 12, Article ID 1850042, 2018.

[10] W. Cai, B. Liu, Z. Wei, M. Li, and J. Kan, "TARDB-Net: triple-attention guided residual dense and BiLSTM networks for hyperspectral image classification," *Multimedia Tools and Applications*, vol. 80, no. 7, pp. 11291–11312, 2021.

[11] A. Hussain, S. Nazir, F. Khan et al., "A resource efficient hybrid proxy mobile IPv6 extension for next generation IoT networks," *IEEE Internet of Things Journal*, 2021.

[12] L. Zhang, X. Wang, X. Dong, L. Sun, W. Cai, and X. Ning, "Finger vein image enhancement based on guided tri-Gaussian filters," *ASP Transactions on Pattern Recognition and Intelligent Systems*, vol. 1, no. 1, pp. 17–23, 2021.

[13] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626, 2018.
[14] L. Yu, S. Tao, W. Gao, and L. Yu, "Self-monitoring method for improving health-related quality of life: data acquisition, monitoring, and analysis of vital signs and diet," *ASP Transactions on Pattern Recognition and Intelligent Systems*, vol. 1, no. 1, pp. 24–31, 2021.
[15] X. Ning, X. Wang, S. Xu et al., "A review of research on co-training," *Concurrency and Computation: Practice and Experience*, Article ID e6276, 2021.
[16] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "A tutorial on deep learning for music information retrieval," 2017, https://arxiv.org/abs/1709.04396.
[17] N. Usman, S. Usman, F. Khan et al., "Intelligent dynamic malware detection using machine learning in IP reputation for forensics data analytics," *Future Generation Computer Systems*, vol. 118, pp. 124–141, 2021.
[18] A. Holzapfel, B. L. Sturm, and M. Coeckelbergh, "Ethical dimensions of music information retrieval technology," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 44–55, 2018.
[19] M. A. Casey, "Music of the 7Ts: predicting and decoding multivoxel fMRI responses with acoustic, schematic, and categorical music features," *Frontiers in Psychology*, vol. 8, p. 1179, 2017.
[20] N. Sachdeva, K. Gupta, and V. Pudi, "Attentive neural architecture incorporating song features for music recommendation," in *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 417–421, Vancouver, Canada, September 2018.
[21] S. Oramas, O. Nieto, F. Barbieri, and X. Serra, "Multi-label music genre classification from audio, text, and images using deep features," 2017, https://arxiv.org/abs/1707.04916.
[22] J. Zhang, Y. Liu, H. Liu, and J. Wang, "Learning local-global multiple correlation filters for robust visual tracking with kalman filter redetection," *Sensors*, vol. 21, no. 4, p. 1129, 2021.
[23] J. Chen, C. Du, Y. Zhang, P. Han, and W. Wei, "A clustering-based coverage path planning method for autonomous heterogeneous UAVs," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2021.
[24] J. Zhang, J. Sun, J. Wang, and X.-G. Yue, "Visual object tracking based on residual network and cascaded correlation filters," *Journal of Ambient Intelligence and Humanized Computing*, 2020.
[25] Q. Liu, L. Cheng, A. L. Jia, and C. Liu, "Deep reinforcement learning for communication flow control in wireless mesh networks," *IEEE Network*, vol. 35, no. 2, pp. 112–119, 2021.