Hindawi
Scientific Programming
Volume 2021, Article ID 1651560, 9 pages
https://doi.org/10.1155/2021/1651560

Research Article

Music Feature Extraction and Classification Algorithm Based on Deep Learning

Jingwen Zhang

Music and Dance, Xi'an Peihua University, Xi'an Province 710199, China

Correspondence should be addressed to Jingwen Zhang; zhangjingwenpeihua@163.com

Received 5 April 2021; Revised 26 April 2021; Accepted 28 April 2021; Published 26 May 2021

Academic Editor: Shah Nazir

With the rapid development of information technology and communication, digital music has grown and exploded. Regarding how to quickly and accurately retrieve the music that users want from huge bulk of music repository, music feature extraction and classification are considered as an important part of music information retrieval and have become a research hotspot in recent years. Traditional music classification approaches use a large number of artificially designed acoustic features. The design of features requires knowledge and in-depth understanding in the domain of music. The features of different classification tasks are often not universal and comprehensive. The existing approach has two shortcomings as follows: ensuring the validity and accuracy of features by manually extracting features and the traditional machine learning classification approaches not performing well on multiclassification problems and not having the ability to be trained on large-scale data. Therefore, this paper converts the audio signal of music into a sound spectrum as a unified representation, avoiding the problem of manual feature selection. According to the characteristics of the sound spectrum, the research has combined 1D convolution, gating mechanism, residual connection, and attention mechanism and proposed a music feature extraction and classification model based on convolutional neural network, which can extract more relevant sound spectrum characteristics of the music category. Finally, this paper designs comparison and ablation experiments. The experimental results show that this approach is better than traditional manual models and machine learning-based approaches.

1. Introduction

With the rapid development of multimedia and digital technologies [1–3], there are more and more digital music resources on the Internet, and consumers' music consumption habits have shifted from physical music to online music platforms. Massive music resources and a huge online music library stimulate users to generate a variety of complex music retrieval needs. For example, at a certain moment, users are eager to listen to a certain genre or a song with a certain emotion. At this time, the music label is essential to the quality of music retrieval. In addition to music retrieval, many recommendation and subscription scenarios also require music category information to provide users with more accurate content [4, 5].

Music is diverse; it is made of different elements such as melody, rhythm, and harmony combinations according to certain rules of art forms. Understanding the music of different forms often requires some background knowledge, not as a music classification standard, so almost all music media platforms use text labels as the basis of the classification of music or retrieval. Music labels are text descriptors that express musical properties in high dimensions, such as "happy" and "sad" to express emotions, and "electronic" and "blues" to express musical styles [6, 7].

Music genre classification [8–10] is an important 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. The advantages of this classification method are easy to implement,

simple to operate, and fast to retrieve, but the shortcomings are also obvious. First of all, this method relies on manually labeled music data, which requires a lot of manpower, and manual labeling is difficult to avoid incorrectly labeling music information problems. Secondly, this text method does not involve the audio data of the music itself. Audio data includes many key characteristics of music, such as pitch, timbre, and melody. These characteristics are almost impossible to label with text; and based on the classification of the content, the features of the original music data are extracted, and the extracted feature data are used to train the classifier, so as 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 [11].

The emergence of deep learning has brought music classification technology into a new period of development. Deep learning has been widely used in image processing, speech recognition, and other fields, and its performance on many tasks surpasses traditional machine learning methods. Scholars have also begun to use deep learning technology to study related issues in the field of music information retrieval, so it is necessary to research music classification methods based on deep learning to improve the effect of music classification [12]. The following are the main innovation points of this paper:

(i) The paper aims to convert the audio signal of music into a sound spectrum as a unified representation, avoiding the problem of manual feature selection.

(ii) It aims to use 1D convolution, gating mechanism, residual connection, and attention mechanism, and it proposes a music feature extraction and classification model based on convolutional neural network, which can extract and correlate more closely related sound spectrum features.

(iii) Sufficient comparison and ablation experiments have been carried out. The experimental results have proved the effectiveness and superiority of our algorithm, surpassing several other well-known methods.

The organization of the paper is as given. Section 2 depicts the background knowledge of the proposed study. The methodology of the paper is shown in Section 3 with the details in the subsections. Experiments and results are presented in Section 4. The paper is concluded in Section 5.

2. Background

As a very important component in the field of music information retrieval, music feature extraction and classification recognition have been widely studied since the 1990s. In 1995, Benyamin Matityaho and Furst [13] proposed a method for frequency-domain analysis of music signals. First, fast Fourier transform is performed on the audio data, and then the logarithmic scale transformation is performed to use the obtained data as feature data. Training was done in a neural network [14–17] containing two hidden layers and two music genres were finally identified: classical and pop music. Tzanetakis and Cook [18] systematically proposed in 2002 the division of the characteristics of music into three feature data sets, namely, timbre texture characteristics, rhythm content characteristics, and tonal content characteristics; the authors adopted the Gaussian mixture model and *K*. The proximity method is used as a classifier. It is worth mentioning that, due to the numerous music genres, there was no relatively fixed classification standard in the academic circles before. Since the groundbreaking research results of Tzanetakis, the ten music genres contained in the GTZAN data set used by George Tzanetakis have become music information. The classification standard was generally recognized in the search field.

As George Tzanetakis' research results laid a lot of foundation for us, later scholars in the field of automatic recognition of music genres mainly focused on two aspects. On the one hand, they made corresponding improvements in the selection of music feature extraction and the dimension of feature vectors. On the other hand, they improved the choice of classification algorithm. The extraction of music features is a very critical part of music genre recognition. If the extracted features cannot represent the essential characteristics of music, then the music classification effect will undoubtedly be very bad. Scaringella et al. [19] divided the music signal characteristics into three categories: pitch, timbre, and rhythm. At present, the commonly used characteristics of music signals mainly include short-term zero-crossing rate, short-term energy, linear prediction coefficient, frequency spectrum, flux, Mel frequency inverse coefficient, spectral centroid, and spectral contrast. Since these characteristics are both in the time domain and in the frequency domain, they can reflect the musical perception characteristics of pitch, rhythm, timbre, and loudness to some extent. The process of music feature extraction is generally to first perform frame processing on the original audio signal, then perform related calculations based on the mathematical statistical significance of the features, and finally use the calculated results as the training data of the classifier in the form of vectors. Because music feature extraction is based on music signal analysis, the current audio-based music signal analysis techniques mainly include time-domain analysis methods and frequency-domain analysis methods. The so-called time-domain analysis method is to analyze and count the waveform state of the music signal from the time dimension. Frequency-domain analysis converts the music signal in the time domain into the frequency domain through Fourier transform, so many useful features in the frequency domain can be obtained, for example, Mel to general coefficient, spectral centroid, pitch frequency, subband energy, spectrogram, etc. Literature [20] cascades together the Mel-to-Pop coefficient and pitch frequency, spectral centroid, subband energy, and other perceptual characteristics to form a high-dimensional feature vector. In the music classification algorithm, traditional machine learning methods are mainly used, such as support vector machines, Gaussian mixture models, decision trees, nearest neighbors, hidden Markov, and artificial neural

networks [21–24]. In addition, there are some improvements to the above algorithms. For example, literature [25] adds a genetic algorithm to the Gaussian mixture model, which improves the accuracy of classification from the experimental results.

3. Methodology

3.1. Music Signal Features

3.1.1. Spectral Centroid. The spectral centroid is a metric used to characterize the frequency spectrum in digital signal processing. It indicates where the "centroid" of the frequency spectrum is located. It feels that it has a close relationship with the brightness of the sound. Generally speaking, the smaller the value is, the more energy is concentrated in the low frequency range. Since the spectral centroid can better reflect the brightness of the sound, it is widely used in digital audio and music signal processing. It is used as a measure of the timbre of music. Its mathematical definition is as follows:

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]}, \tag{1}$$

where $M_t[n]$ represents the magnitude of the Fourier transform of the t-th frame at the frequency group n.

3.1.2. Spectral Flux. The spectrum flux is generally a measure of the rate of change of the signal spectrum. It is calculated by comparing the spectrum of the current frame with the spectrum of the previous frame. More precisely, it is usually calculated as the 2-norm between two normalized spectrums. Since the spectrum is normalized, the spectrum flux calculated in this way does not depend on the phase; only the amplitudes can be compared. Spectrum flux is generally used to determine the timbre of an audio signal or to determine whether to pronounce. Its mathematical definition is as follows:

$$F_t = \sum_{n=1}^{N} \left(N_t[n] - N_{t-1}[n]\right). \tag{2}$$

3.1.3. Spectral Contrast. Spectral contrast is a feature used to classify music genres. Spectral contrast is expressed as the difference in decibels between peaks and valleys in the frequency spectrum, which can represent the relative spectral characteristics of music. It can be seen from the experimental results of the literature [26] that the spectral contrast has a good ability to discriminate music genres.

3.1.4. Mel-Scale Frequency Cepstral Coefficients. Since the cochlea has filtering characteristics (as shown in Figure 1), different frequencies can be mapped to different positions of the basilar membrane. So the cochlea is often regarded as a filter bank. Based on this feature, psychologists obtained a set of filter banks similar to the cochlear effect through psychological experiments, that is, the Mel frequency filter bank. Since the sound level perceived by the human ear is not linearly related to its frequency, researchers have proposed a new concept called Mel frequency. The Mel frequency scale is more in line with the auditory characteristics of the human ear. The relationship between Mel frequency and frequency f is as follows:

$$f_{\text{mel}} = 25951g\left(1 + \frac{f}{700}\right), \tag{3}$$

where f_{mel} is the converted Mel frequency, f is the frequency, and the unit is Hz.

Firstly, the audio signal is divided into frames, pre-emphasized, and then windowed, and then short-time Fourier transform (STFT) is performed to obtain its frequency spectrum. Secondly, set the Mel filter bank of L channels on the Mel frequency. The L value is determined by the highest frequency of the signal, generally 12 to 16, and each Mel filter has the same interval on the Mel frequency. Let $o(l)$, $c(l)$, and $h(l)$ be the lower limit frequency, center frequency, and upper limit frequency of the l-th triangular filter, respectively; then, the relationship between the three frequencies of adjacent triangular filters is as follows:

$$c(l) = h(l-1) = o(l+1). \tag{4}$$

Pass the linear amplitude spectrum of the signal through the Mel filter to get the output of the filter:

$$Y(l) = \sum_{k=o(l)}^{h(l)} W_l(k)\left|X_m(k)\right|, \quad l = 1, 2, \ldots, L. \tag{5}$$

The frequency features of the filter are

$$W_l(k) = \begin{cases} \dfrac{k - o(l)}{c(l) - o(l)}, & o(l) \le k \le c(l), \\ \dfrac{h(l) - k}{h(l) - c(l)}, & c(l) \le k \le h(l). \end{cases} \tag{6}$$

Take the natural logarithm of the filter output value, and then transform the discrete cosine to MFCC. The expression is as follows:

$$\text{MFCC}_{\text{MFCC}}(n) = \sum_{l=1}^{L} \lg Y(l) * \cos\left[\pi(l - 0.5)\frac{n}{L}\right], \quad n = 1, 2, \ldots, L. \tag{7}$$

3.2. 1D Residual Gated Convolutional Neural Model

3.2.1. Selection of Convolution Kernel. Convolutional neural networks can well identify potential patterns in the data. By superimposing convolution kernels to perform repeated convolution operations, more abstract features can be obtained in the deep layers of the network. One-dimensional convolution is often used to deal with problems related to time series. Unlike two-dimensional convolution that

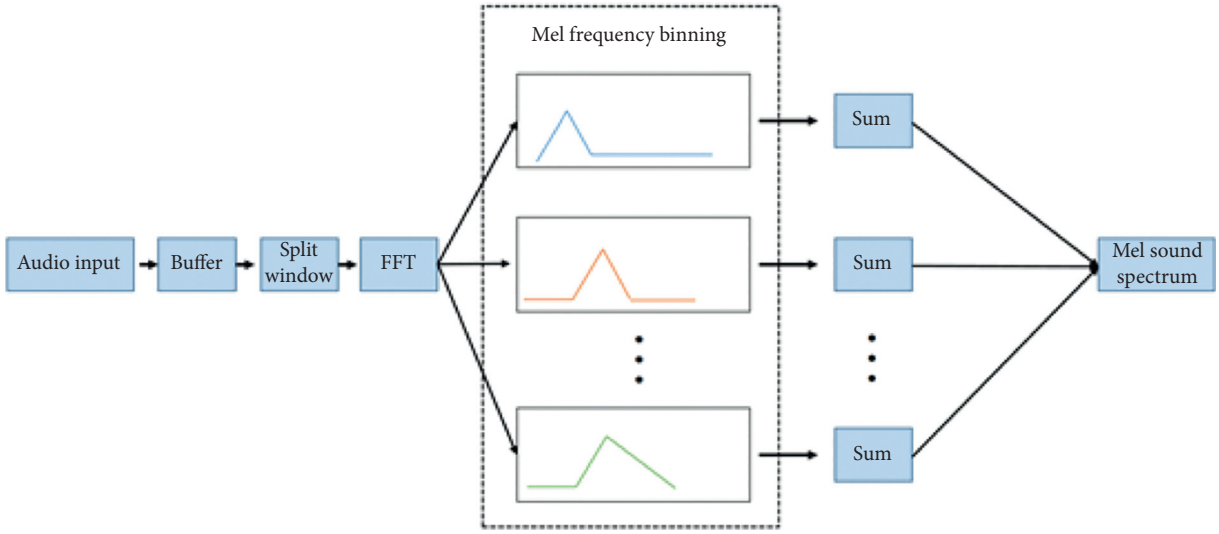

Figure 1: The calculation process of Mel sound spectrum.

attempts to convolve in multiple directions, one-dimensional convolution focuses more on capturing the translation invariance of data features in a specific direction. When dealing with time-related data, this direction is often the direction of time change. One-dimensional convolution is often used to analyze time series or sensor data and is suitable for signal data analysis within a fixed period of time, such as audio signals.

Figure 2 shows the convolution process of one-dimensional convolution and two-dimensional convolution on the sound spectrum. It can be seen that the receptive field of the one-dimensional convolution kernel covers all frequency ranges on the sound spectrum, which is only performed on the time axis. Convolution can capture the percussion components of the musical instruments appearing on the sound spectrum, and their overtones and other musical elements. Unlike the one-dimensional convolution that only convolves in the time direction, the two-dimensional convolution performs convolution in the two dimensions of time and frequency and can extract specific patterns of frequency within a certain time range, such as the rise and fall of pitch. In the field of music classification, many models use two-dimensional convolution as the basic convolution structure of convolutional networks.

The time perception of two-dimensional convolution is not as good as one-dimensional convolution, and the range of perception in the frequency range is not as broad as one-dimensional convolution, and the computational complexity of one-dimensional convolutional neural networks is smaller. In addition, two-dimensional convolution also performs convolution in the frequency dimension of the sound spectrum, which is inexplicable for sound signals. Therefore, the model in this article will use one-dimensional convolution as the basic convolution structure, which is more in line with the fact that the audio signal is expanded in time and has less correlation in the frequency range.

The essential difference between one-dimensional convolution and two-dimensional convolution lies in the translation direction, and its calculation method is not essentially different from that of two-dimensional convolution. Although the original audio signal is a time series, after it is converted into a sound spectrum, its expression is similar to a single-channel grayscale picture, so the calculation of convolution can be expressed by the following equation:

$$a_{ij} = h\left(\sum_{m=0}^{f_w-1}\sum_{n=0}^{f_h-1} w_{mn}x_{i+m,j+n} + b\right), \tag{8}$$

where a_{ij} is the width and height of the feature map, h is the activation function used by the convolution layer, f_w is the width of the convolution kernel, f_h is the height of the convolution kernel, b is the offset of the convolution, and w and x represent the weight matrix and data input of the product core, respectively. In the one-dimensional convolution operation based on the sound spectrum, f_h and the frequency range l of the sound spectrum have the following relationship:

$$l = f_h. \tag{9}$$

That is, the height of the convolution kernel in one-dimensional convolution is equal to the frequency range in the sound spectrum, and the receptive field of the convolution kernel covers the entire frequency axis, so as to capture a specific frequency pattern. Then the convolution operation can be expressed as

$$\text{conv}(X, W) = \sum_{m=0}^{F_w-1}\sum_{n=0}^{L-1} w_{mn}x_{i+m,j+n}. \tag{10}$$

Assuming that the output of the convolution kernel is R and the bias matrix is B, then the convolution operation can be simply expressed as

$$R = \text{conv}(X, W) + B. \tag{11}$$

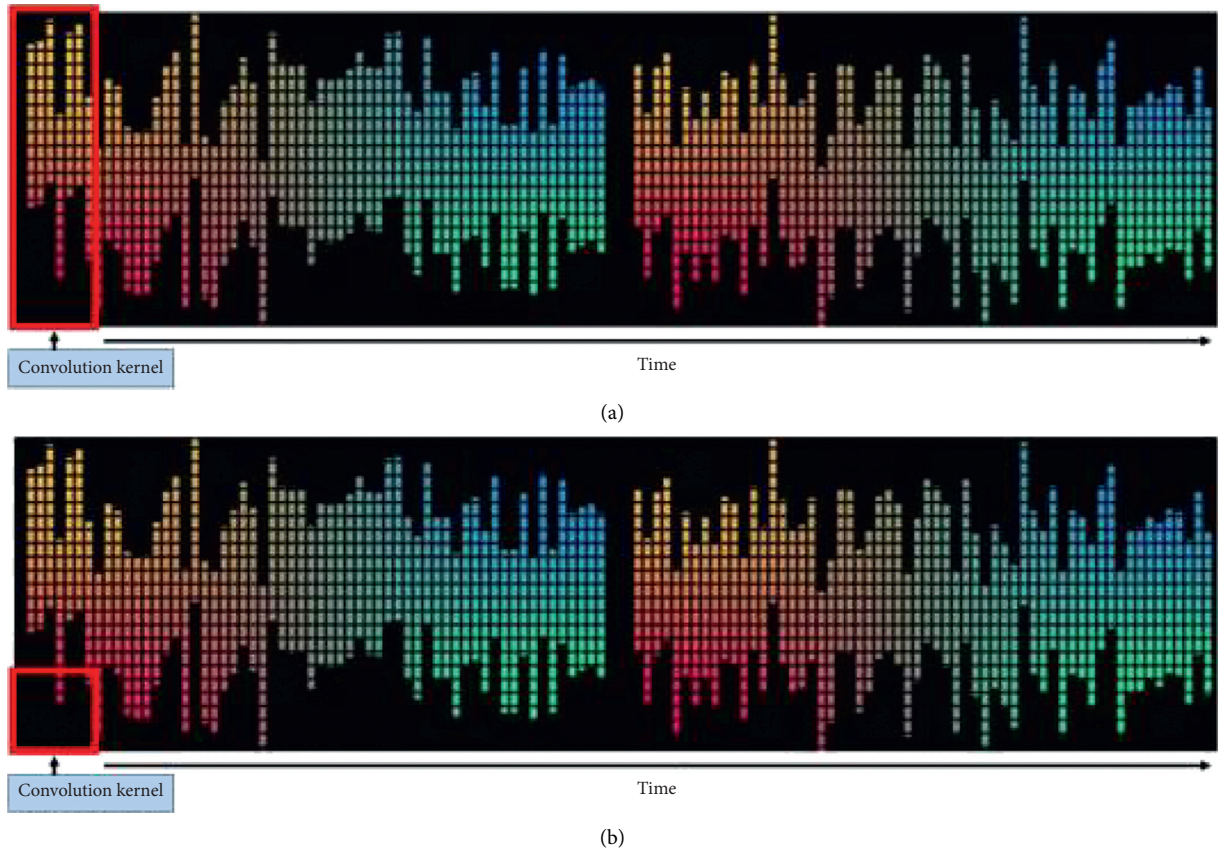

Figure 2: Convolution process comparison of 1D convolution and 2D convolution on the sound spectrum. (a) 1D conv. (b) 2D conv.

The width R_w of R can be obtained by the following formula:

$$R_w = \frac{t - f_w + 2p}{s} + 1, \tag{12}$$

where t represents the length of the sound spectrum on the time axis, that is, the width of the sound spectrum. p represents the size of the padding, and f_w represents the width of the convolution kernel. Since the one-dimensional convolution only performs translation in the time dimension of the sound spectrum, the height R_h of the output feature map R is as follows:

$$R_h = 1. \tag{13}$$

In other words, R_h has nothing to do with the frequency range l of the acoustic spectrum and the high f_h of the convolution kernel. After one-dimensional convolution, the dimension of the acoustic spectrum changes to that of the two-dimensional convolution. After one-dimensional convolution, the specification of the feature graph is also reduced.

3.2.2. Gated Linear Units. Assuming that the sound spectrum sequence to be processed is $X = [x_1, x_2, \ldots, x_n]$ and the output of the convolution kernel is Y, then the gated linear unit can be expressed as

$$Y = \text{Conv1}D_1(X) \otimes \sigma\left(\text{Conv1}D_2(X)\right). \tag{14}$$

The two Conv1D$_1$ and Conv1D in the above formula represent two identical one-dimensional convolutions, but the weights are not shared. $\otimes$ represents the (element-wise) operation, and σ represents the Sigmoid activation function. One of the results after the two convolutions is activated by the Sigmoid function, and the other is not added with the activation function, and then the creation gate is multiplied bit by bit. Formally, it is equivalent to adding a "valve" to each output of one-dimensional convolution to control the flow. The convolution-based gating mechanism is different from the complex threshold mechanism in the LSTM network. It does not need to forget the gate, only an input gate, which also makes the network model based on the gated convolution unit perform better than LSTM in training speed.

Figure 3 shows the basic structure of the one-dimensional gated convolution unit. You can see the data flow inside the one-dimensional gated convolution unit. After the input of the convolution unit undergoes two identical convolutions, one of the 1D convolution kernels is extra. The activation operation of the Sigmoid function is performed, and then the output of another convolution kernel is multiplied bit by bit to produce the output of this layer.

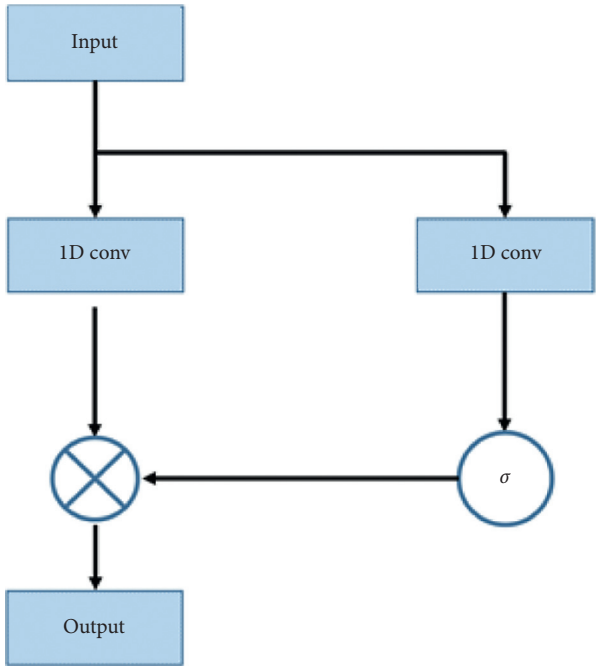

Figure 3: Schematic diagram of 1D convolutional gated unit.

3.2.3. Residual Connection. The entire convolutional neural network can be regarded as a process of information extraction. The more the layers of the network, the stronger the ability of the network to gradually extract from the underlying features to the highly abstract features. When the network layers are deepened, the model is more likely to discover high-level abstract features related to music categories. Increasing the depth of the network too much will cause gradient disappearance and explosion problems to the model. The solution to gradient disappearance and explosion is generally to add regular initialization and an intermediate regularization layer, but the network degradation problem also arises. When the network begins to degenerate, the accuracy on the training set will decrease as the number of network layers increases. This problem is essentially different from overfitting, which will show excellent results on the training set.

The basic residual module is shown in Figure 4. It can be seen that the residual structure has an additional identity mapping channel, so that when the depth of the network increases and it is not conducive to the enhancement of network performance, the network can directly skip these useless layers. Directly accept the output of the upper layer. The calculation equation of the residual structure is as follows:

$$x_{l+1} = x_l + F(x_l, W_l). \tag{15}$$

3.3. Music Feature Extraction and Classification Model. The model in this paper can be divided into GLU stacking layer, global pooling feature aggregation layer, and fully

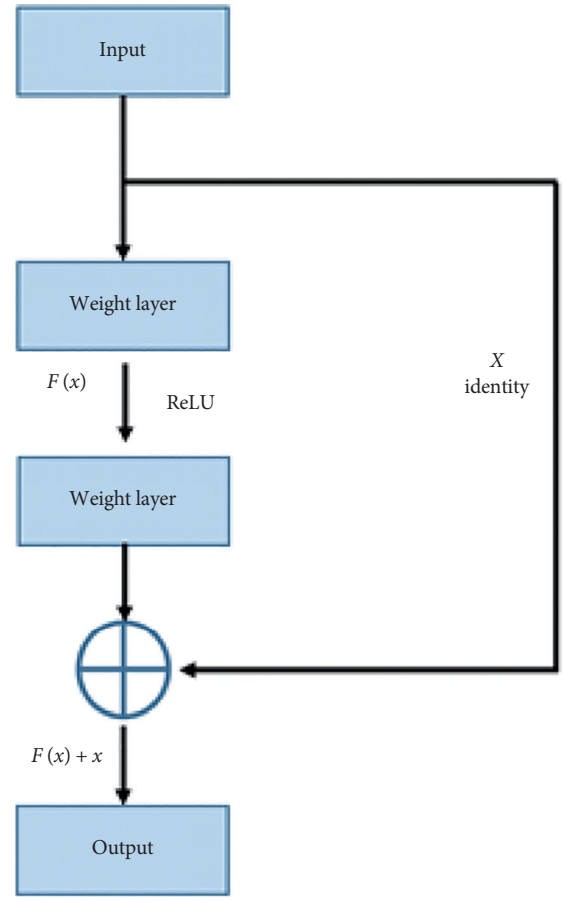

Figure 4: Schematic diagram of residual module.

connected layer from input to output. The overall structure of the network is shown in Figure 5.

To make full use of the statistical information of the pooling layer, the model in this chapter combines the global maximum pooling and the global average pooling to form a global pooling feature aggregation layer. The feature maps obtained from the GLU block stacking layer undergo global average pooling and global maximum pooling to obtain average pooling statistics and maximum pooling statistics, respectively. The results of the pooling operation here are all one-dimensional. In Figure 5, two rectangular blocks of different colors are used to represent these two one-dimensional features, and the two pooled statistical features are spliced into the next layer of fully connected network.

4. Experiments and Results

4.1. Experimental Setup. Due to the repetitive information in the multichannel of the original audio, all audio is converted to mono, and downsampling is performed at a sampling rate of 16 kHz. The Fourier transform window length used when converting the Mel sound spectrum is 512, the window jump size is 256, and the number of frequency bins is 128. The original audio sample is segmented by the segmentation

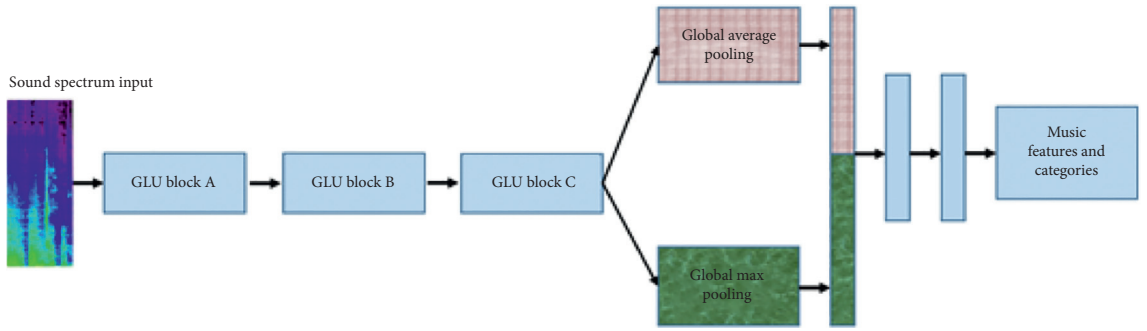

Figure 5: Schematic diagram of the overall model.

method. The slice duration is 5 seconds, and the overlap rate is 50%. The Mel sound spectrum specification of a single slice generated after processing according to the above settings is (313, 128), and each audio sample produces 11 slices of the same size.

4.2. Data Set. The experiment in this chapter uses the GTZAN data set, which is widely used to verify the performance of music classification methods and is the most popular music classification data set. The GTZAN data set has 10 music genre categories (as shown in Table 1). The number of audio samples in each genre category is 100, the sample duration is 30 seconds, and the sampling rate is 22050 Hz.

4.3. Evaluation Index. The classification accuracy rate (Acc) is selected as the evaluation index of the music classification method proposed in this chapter. The calculation method of classification accuracy is as follows:

$$\text{Acc} = \frac{N_C}{N} \times 100\%. \tag{16}$$

4.4. Experimental Results. Different models produce different recognition results by learning different deep features. In order to make a fair comparison, all experiments were implemented in the same environment, and all parameters were retained, comparing the proposed model with SVM, CNN, GLU, RCNN, and RGLU.

The above five types of networks with different structures are tested with the same experimental settings, and the results are shown in Table 2 and Figure 6. The GLU network using the gated structure has higher accuracy than the ordinary convolutional network CNN, which indicates that the stacking of multiple gated convolutions used in the model in this chapter is more conducive to the sound spectrum characteristics than the ordinary convolution learning. The gating structure makes the features passed to the next layer of the network pay more attention to the sound spectrum features that are more important for the music classification task, and the information that is not related to the music classification task is ignored by the gating mechanism. The results of the comparison experiment verify that the gating structure is based on the effectiveness of the sound spectrum

Table 1: Introduction to the GTZAN data set.

ID	Category
1	Rock
2	Reggae
3	Pop
4	Metal
5	Jazz
6	Hip hop
7	Disco
8	Country
9	Classical
10	Bruce

Table 2: Comparative experiment results on the GTZAN data set.

Method	Acc	Std
SVM	0.52	0.03
CNN	0.70	0.01
GLU	0.75	0.03
RCNN	0.80	0.01
RGLU	0.82	0.02
Ours	**0.87**	**0.01**

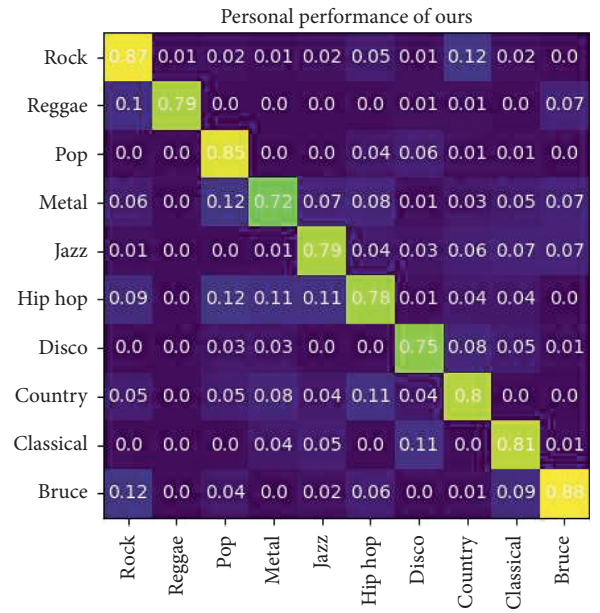

Figure 6: Confusion matrix.

Table 3: Results of ablation studies.

Method	Acc
Global average pooling	0.85
Global max pooling	0.84
Ours	**0.87**

in the task of music classification. From the perspective of information filtering, GLU can be used as another implementation of the attention mechanism. Unlike the RGLU structure that determines an attention weight for each feature map channel, GLU can adaptively determine the time during the network learning process. The attention weight in the one-dimensional convolution is expanded in time; this kind of gated structure that increases attention in the time dimension, combined with the one-dimensional convolution in the time dimension, can get better performance. Compared with CNN and GLU without residual structure, the accuracy of RCNN and RGLU with added residual structure has been improved, which shows that the use of residual connection can improve the accuracy of classification to a certain extent. It is worth noting that the accuracy of RGLU using the residual structure is improved compared to GLU, and the accuracy of RCNN is greater than that of CNN. This indicates that the combination of residual structure and gated convolution is more beneficial for the transmission of information in the network. Therefore, this experiment fully proves the effectiveness and superiority of our algorithm.

4.5. Ablation Study of Global Pooling. This section will compare the classification performance of different pooling features and their combinations in the global pooling feature aggregation layer. We used three pooling methods to conduct experiments, and the experimental results are shown in Table 3.

The aggregation of the two global pooling features can make the model obtain a higher accuracy rate. Using the global average pooling feature alone is more accurate than using the global maximum pooling feature alone, which means the overall statistical information in the spectroscopic feature map is more conducive to classification. The model in this chapter combines two types of global pooling features, which enables the fully connected layer to grasp more statistical information of the features abstracted by the convolutional layer and makes the classification performance of the model stronger.

5. Conclusion

Digital music has grown and exploded with the growing developments of information technology and communication. Music feature extraction and classification are considered as a significant portion of music information retrieval. The design of features requires knowledge and in-depth understanding in the domain of music. The features of different classification tasks are often not universal and comprehensive. Traditional music classification approaches use a large number of artificially designed acoustic features. It is difficult to effectively extract music features due to manual and traditional machine learning methods. Therefore, the contribution of this paper is to convert the audio signal of music into a sound spectrum as a unified representation, avoiding the problem of manual feature selection. According to the characteristics of the sound spectrum, combined with one-dimensional convolution, gating mechanism, residual connection, and attention mechanism, a music feature extraction and classification model based on convolutional neural network is proposed, which can extract and correlate more closely related sound spectrum features of music category. Finally, this paper designs a comparison and ablation experiment. Experimental results show that this method is superior to traditional manual models and machine learning-based methods.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] A. El Saddik, "Digital twins: the convergence of multimedia technologies," *IEEE Multimedia*, vol. 25, no. 2, pp. 87–92, 2018.

[2] I. Gorbunova and H. Hiner, "Music computer technologies and interactive systems of education in digital age school," in *Proceedings of the International Conference Communicative Strategies of Information Society (CSIS 2018)*, pp. 124–128, Atlantis Press, Almaty, Kazakhstan, February 2019.

[3] C. C. S. Liem, E. Gómez, and G. Tzanetakis, "Multimedia technologies for enriched music performance, production, and consumption," *IEEE MultiMedia*, vol. 24, no. 1, pp. 20–23, 2017.

[4] Y. An, S. Sun, and S. Wang, "Naive Bayes classifiers for music emotion classification based on lyrics," in *Proceedings of the 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pp. 635–638, IEEE, Wuhan, China, 2017 May.

[5] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2887–2905, 2021.

[6] M. Russo, L. Kraljević, M. Stella, and M. Sikora, "Cochleogram-based approach for detecting perceived emotions in music," *Information Processing & Management*, vol. 57, no. 5, Article ID 102270, 2020.

[7] M. Chełkowska-Zacharewicz and M. Paliga, "Music emotions and associations in film music listening: the example of leitmotifs from the Lord of the Rings movies," *Roczniki Psychologiczne*, vol. 22, no. 2, pp. 151–175, 2019.

[8] 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.

[9] 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.

[10] A. Rosner and B. Kostek, "Automatic music genre classification based on musical instrument track separation," *Journal of Intelligent Information Systems*, vol. 50, no. 2, pp. 363–384, 2018.

[11] M. D. S. Anisetty, G. Shetty, S. Hiriyannaiah, S. Gaddadevara Matt, K. G. Srinivasa, and A. Kanavalli, "Content-based music classification using ensemble of classifiers," in *Proceedings of the International Conference on Intelligent Human Computer Interaction*, pp. 285–292, Springer, Cham, Germany, 2018 December.

[12] J. Nam, K. Choi, J. Lee, S. Y. Chou, and Y. H. Yang, "Deep learning for audio-based music classification and tagging: teaching computers to distinguish rock from bach," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 41–51, 2018.

[13] B. Matityaho and M. Furst, "Neural network based model for classification of music type," in *Proceedings of the Eighteenth Convention of Electrical and Electronics Engineers in Israel*, pp. 4–3, IEEE, Tel Aviv, Israel, 1995 March.

[14] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, and S. Tian, "Feature refinement and filter network for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, p. 1, 2020.

[15] W. Cai, Z. Wei, R. Liu, Y. Zhuang, Y. Wang, and X. Ning, "Remote sensing image recognition based on multi-attention residual fusion networks," *ASP Transactions on Pattern Recognition and Intelligent Systems*, vol. 1, no. 1, pp. 1–7, 2021.

[16] S. Li, X. Ning, L. Yu et al., "Multi-angle head pose classification when wearing the mask for face recognition under the COVID-19 coronavirus epidemic," in *Proceedings of the 2020 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)*, pp. 1–5, IEEE, Shenzhen, China, May 2020.

[17] 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.

[18] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[19] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, 2006.

[20] M. McKinney and J. Breebaart, "Features for audio and music classification," 2003.

[21] 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*, Wiley, Hoboken, NJ, USA, 2020.

[22] X. Zhang, Y. Yang, Z. Li, X. Ning, Y. Qin, and W. Cai, "An improved encoder-decoder network based on strip pool method applied to segmentation of farmland vacancy field," *Entropy*, vol. 23, no. 4, p. 435, 2021.

[23] Z. L. Yang, S. Y. Zhang, Y. T. Hu, Z. W. Hu, and Y. F. Huang, "VAE-stega: linguistic steganography based on variational auto-encoder," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 880–895, 2020.

[24] X. Ning, X. Wang, S Xu et al., *A Review of Research on co-training. Concurrency and Computation: Practice and Experience*, Beijing Union University, Beijing, China, 2021.

[25] K. I. Molla and K. Hirose, "On the effectiveness of MFCCs and their statistical distribution properties in speaker identification," in *Proceedings of the 2004 IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2004 (VCIMS)*, pp. 136–141, IEEE, Boston, MA, USA, July 2004.

[26] D. N. Jiang, L. Lu, H. J. Zhang, J. H. Tao, and L. H. Cai, "Music type classification by spectral contrast feature," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 113–116, IEEE, Lausanne, Switzerland, August 2002.