Multimedia Tools and Applications (2024) 83:13527–13542
https://doi.org/10.1007/s11042-023-15277-1

Music genre classification based on res-gated CNN and attention mechanism

Changjiang Xie[1] · Huazhu Song[1,2] · Hao Zhu[1] · Kaituo Mi[2] · Zhouhan Li[1] · Yi Zhang[1] · Jiawen Cheng[1] · Honglin Zhou[1] · Renjie Li[1] · Haofeng Cai[1]

Received: 21 May 2022 / Revised: 20 February 2023 / Accepted: 6 April 2023 /
Published online: 6 July 2023

Abstract
The amount of digital music available on the internet has grown significantly with the rapid development of digital multimedia technology. Managing these massive music resources is a thorny problem that powerful music media platforms need to face where music genre classification plays an important role, and a good music genre classifier is indispensable for the research and application of music resources in the related aspects, such as efficient organization, retrieval, recommendation, etc. Due to convolutional networks' powerful feature extraction capability, more and more researchers are devoting their efforts to music genre classification models based on convolutional neural networks (CNNs). However, many models do not combine the musical signal features for effective design of the convolutional structure, which cause a simpler convolutional network part of the model and weaker local feature extraction ability. To solve the above problem, our group proposes a model using a 1D res-gated CNN to extract local information of audio sequences rather than the traditional CNN architecture. Meanwhile, to aggregate the global information of audio feature sequences, our group applies the Transformer to the music genre classification model and modifies the decoder structure of the Transformer according to the task. The experiments utilize the benchmark datasets, including GTZAN and Extended Ballroom. Our group conducted contrastive experiments to verify our model, and experimental results demonstrated that our model outperforms most of the previous approaches and can improve the performance of music genre classification.

Keywords Music genre classification · CNN · Transformer

1 Introduction

With the rapid development and popularity of multimedia and digital technologies, more and more digital music resources are available on the internet. However, the massive amount of music resources and huge online music libraries inspire users to generate a variety of

✉ Huazhu Song
756862517@qq.com

Extended author information available on the last page of the article.

complex music retrieval needs. High-level tags for music genres provide an efficient way to organize and retrieve music, which plays an essential role in music classification [17].

At present, the annotation of music genres is mainly done manually, which requires a high level of music knowledge and music literacy of the annotator. Music media platforms usually hire music experts to perform music tagging, which results in good accuracy but high cost. Sometimes the platforms also allow non-professional users to tag by opening up the tags, and the tagging data will be counted to form the music tags. This saves costs but causes many cases of category labelling errors. Therefore, an accurate and effective music genre classification method is essential for more effective retrieval of music content.

Traditional methods for music genre classification usually use manually extracted features, which often require specialized software to extract the features. Machine learning algorithms also have difficulty coping with today's massive amounts of music data [23]. Deep learning has been widely used in image processing [6, 13, 22], speech recognition [7, 38, 50], and other applications. Deep learning outperforms traditional machine learning algorithms on many tasks, and many scholars have applied deep learning to the field of Music Information Retrieval (MIR) [17]. This makes music genre classification enter a new stage of development.

The Transformer has attracted much attention since it was proposed by Ashish Vaswani et al. [42]. Various existing Transformer-based models are only related to Natural Language Processing (NLP) tasks [35, 45, 53]. However, some recent papers [5, 16, 29] have pioneered the cross-domain application of Transformer models to Computer Vision (CV) tasks with good results. Inspired by this, our group also applies the Transformer to a music genre classification approach.

Transformer excels at capturing global content-based interactions, while CNN effectively exploits local features. Therefore, our model combines CNN and Transformer to model audio sequences' local and global dependencies, which achieves superior results in both aspects. Our contributions are summarized as follows:

(1) To aggregate the global information of audio feature sequences, our group applies the Transformer to the music genre classification method and modifies the decoder structure of the Transformer according to the task.
(2) Our group uses a 1D res-gated CNN structure to extract local information of audio sequences instead of the traditional CNN architecture and achieve good results.
(3) Our model combines CNN and Transformer, which facilitates the extraction of deep features of audio sequences and facilitates the matching and aggregation of patterns of temporal information.

The remainder of this paper is organized as follows. In Section 2, our group analyses the previous related music genre classification works. In Section 3, our group describes the method of constructing our model, the RGLUformer, for music genre classification. In Section 4, our group conducts various experiments based on two datasets and verifies the validity of our proposed architecture. Finally, the conclusion is shown in Section 5.

2 Relative work

Music genre classification is an extensively researched area in MIR and has been studied using machine learning methods by many scholars. Tzanetakis et al. [41] used underlying audio features such as rhythm, timbre, and pitch as feature sets and used algorithms such

as Gaussian mixture models, Gaussian classifiers, and K-nearest neighbour (KNN) [11] for classification selection experiments, the widely used GTZAN dataset was presented in this paper. Based on them, Changsheng Xu et al. [46] constructed a three-layer Support Vector Machine (SVM) for obtaining optimal class boundaries between different music genres by learning from training data. Carlos et al. [40] proposed an integrated classifier for classifying genres of Latin music where they also showed that using segments in the middle of a piece of music for classification gives the highest accuracy. Pradeep et al. [26] used both Fast Fourier Transform (FFT) and the Mel Frequency Cepstrum Coefficient (MFCC) to characterize the data, whose study compared the accuracy of various classifiers, such as logistic regression, KNN, SVM, and decision trees, with SVM achieving the highest accuracy of 83%.

The above studies focus on the selection of feature sets and classifiers. However, the association between audio features and music categories is elusive. The classification results using these underlying features are not stable and depend heavily on the selected feature sets. The rise of deep learning has also attracted extensive attention in music genre classification, and CNNs have had a significant impact on many audio and music processing tasks [1, 24, 25] in recent years. By building deep networks, CNNs possess a powerful ability to learn more representative features of music samples. In addition, CNNs require less engineering effort and have little prior knowledge of a specific domain. Just as CNNs excel in image classification [10], Li et al. [27] demonstrated that musical patterns obtained by certain transforms, such as the FFT and MFCC, also work well in CNNs. They also demonstrated that CNNs are feasible for automatically extracting musical pattern features. Dieleman et al. [14] designed a convolutional network with 1D convolution and 1D max-pooling as the main structure. The Mel-spectrogram as the network input got better classification performance than the original audio waveform signal as the network input.

Simply using CNN to classify music genres ignores the temporal information inherent in music. Many scholars combined it with recurrent neural networks (RNNs). Choi et al. [9] designed a hybrid model called Convolutional Recurrent Neural Network (CRNN), which used a two-layer RNN with gated recurrent units (GRU) [8] combined with CNN for music genre classification, which achieved good results. Yang et al. [49] designed a hybrid model consisting of parallel CNN and Bi-RNN blocks to preserve the original music samples' spatial features and temporal order as much as possible. Wang et al. [44] firstly used a CNN to extract deep abstract features of the spectrogram. Wang scanned the resulting feature maps in multiple directions to generate feature sequences fed into LSTM [31] for music annotation. Dai et al. [12] used LSTM RNN to extract features from scattered spectrograms [2] of audio clips and fused them. Dong et al. [15] proposed a bidirectional convolutional recursive sparse network for music sentiment classification that adaptively learned sentiment salience features containing temporal information from the spectrogram.

Attention mechanism was gradually applied to problems related to sequence data before [42], such as machine translation [32], machine comprehension [21], sentence summarization [37], and word representation [28]. In CV, attention also works well in tasks such as image classification [33], object detection [3], and image caption [47]. In practice, these attention-based approaches follow two rules summarized by [43]: (1) Decide which parts of the input to focus on; (2) Allocate the limited processing resources reasonably. Although attention has been widely used in NLP and CV, few works have been done in music genre classification. Based on the assumption that all segments of a music track are equally crucial for music genre classification, we added attention to the model to aggregate audio sequence features. Experimental results show that the model's performance incorporating attention is higher than that of the model using the RNN structure.

3 Methods

As in many studies on music genre classification [9, 48, 51], our group uses Mel-spectrogram as input to the model, which can be considered a time series containing spectral features. In the paper, our group proposed a novel architecture named RGLUformer, which consists of RGLUBlock, encoder, and decoder, as shown in Fig. 1.

The model uses 1D res-gated CNN to identify potential spectrogram patterns and extracts deep abstract features of the spectrogram by stacking kernels. The feature sequences extracted are fed into the encoder-decoder structure of the model. The encoder is used to code the feature sequences into vector representation to be fed into the decoder and subsequently decoded into music genre labels. The encoder consists of positional encoding, multi-headed attention, and a feed-forward neural network, the same as Transformer's consistency. The decoder feeds feature vectors into a global-pooling feature aggregation layer, passing through several fully-connected layers, and finally outputs the music labels. Further explanations in detail will be given in the following several sections.

3.1 1D res-gated CNN

Our model uses 1D convolution as the basic convolutional structure of a convolutional network. 1D convolution is often used to deal with problems related to time series. Unlike 2D convolution, which convolves in multiple directions, 1D convolution focuses more on capturing the data features in a particular direction. As shown in Fig. 2, the receptive field of 1D convolution covers all frequency ranges of the spectrogram and convolves only in the time dimension, which can capture various musical elements that appeared in the spectrogram. In contrast, the convolution kernel of 2D convolution convolves in both the time dimension and the frequency dimension, which is not interpretable for the sound signal.

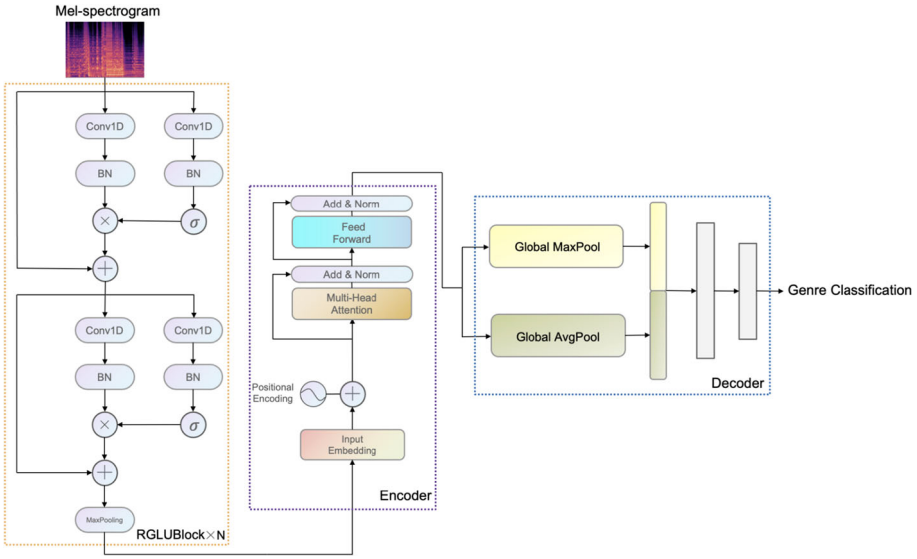

Fig. 1 Architecture of our model

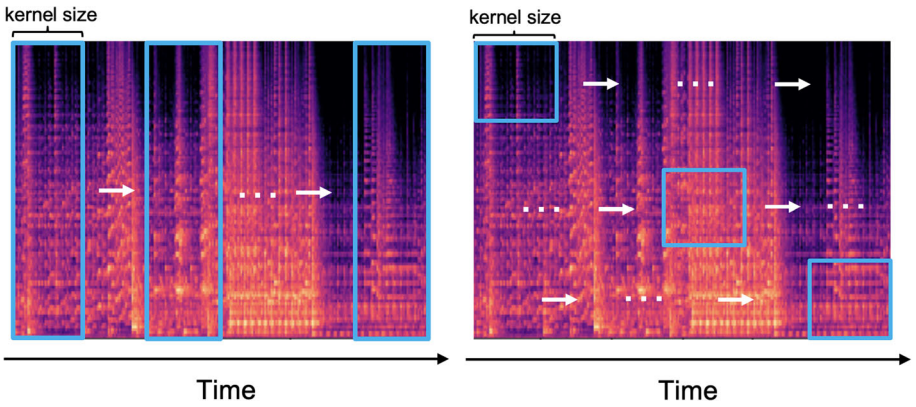

Fig. 2 The process of 1D convolution and 2D convolution

The res-gated convolutional structure is inspired by the Gated Linear Units (GLU) proposed by Dauphin et al. [18]. Figure 3 illustrates the specific structure of a 1D gated convolution unit. The model's input passes through two separate 1D convolutions, which use the same kernels without sharing weights. The output of one of the 1D convolutions passes through Sigmoid (which is more suitable for gated structures) and is element-wise multiplied with the output of the other 1D convolution. Denote a spectrogram sequence as

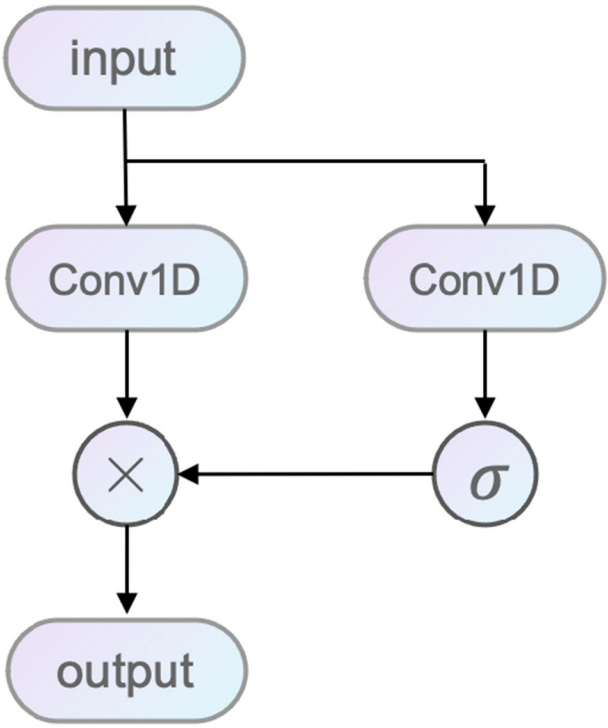

Fig. 3 1D gated convolution structure

$X = \{x_1, x_2, \ldots, x_\tau\}$, the output after 1D gated convolution can be computed by (1):

$$Y = \text{Conv1D}_1(X) \otimes \sigma(\text{Conv1D}_2(X)) \tag{1}$$

Where $\otimes$ denotes element-wise and σ denotes Sigmoid. The output of a 1D convolution activated by the Sigmoid can be regarded as a "valve" that can be multiplied element-wise with another 1D convolution that is not activated to achieve the effect of flow control, which not only effectively reduces the gradient dispersion but also retains the nonlinear capability.

The gating structure can control the amount of information flowing between layers. The information related to the music genre category in the feature representation gets greater attention in the network's learning process. The information not related to the category will be filtered. Meanwhile, Dauphin et al. also proposed that a complex gating mechanism like LSTM is not needed in GLU, and one input gate is sufficient, so the model convergence and training of GLU is simpler and faster.

In practice, the network slowly degrades after adding too many layers, and the training error of the model tends to increase rather than decrease. Kaiming He et al. [20] proposed the Residual Network (ResNet) to deal with this problem. Our group incorporates residual structure into the 1D gated convolutional structure. The network can be skipped when its depth increases but does not benefit the network's performance, preventing its performance from being impaired. In this case, (1) can be reformulated as:

$$Y = X + \text{Conv1D}_1(X) \otimes \sigma(\text{Conv1D}_2(X)) \tag{2}$$

A 1D res-gated convolutional structure consists of a residual network and a gated convolutional structure. Each RGLUBlock contains two 1D res-gated convolutions and a max-pooling layer. The max-pooling layer reduces the feature map's size, which helps reduce the computational effort and prevent overfitting. Using two 1D res-gated convolutional structures per RGLUBlock makes controlling the number of input and output data channels easier. More kernels used also provides greater possibilities for the network to discover more valuable patterns in the spectrogram.

3.2 Encoder

The audio signal is time series; using only 1D convolution to extract abstract features of the spectrogram would lose the temporal information within the music. Wang et al. [31] chose LSTM to summarize the time-domain information after the convolutional structure, allowing the network to model audio sequence relationships effectively. Nevertheless, the training of the LSTM is iterative and serial, which must wait for the current message to be processed before the following message can be processed. In contrast, the training of Transformer is parallel, which can significantly increase the computational efficiency. Moreover, the feature extraction capability of Transformer is better than that of RNN. Therefore, after extracting the music features, our model uses the same encoder-decoder structure as Transformer. The encoder is responsible for mapping the feature vectors extracted from the music feature learning layer to the hidden layer. The decoder is used to decode the hidden layer into music genre labels.

For music, the presence of a particular audio characteristic at different moments in the music has different effects on the music category. For example, it does not bring the same feeling when the same melody appears at the beginning or end of the music. When some musical elements appear in specific positions, it also impacts music classification. Therefore, we use the attention mechanism to aggregate the audio sequence features based on the

assumptions above. The attention mechanism allows the model to capture the abstract features that significantly impact the audio category. The abstract features of all moments will be aggregated into an overall feature vector, which can then be passed to the subsequent network to complete the music genre classification task. Denote an original representation as I, the output O after self-attention can be computed by (3):

$$O = W_v I \times \text{softmax}\left((W_k I)^T W_q I\right) \tag{3}$$

Where W_q, W_k, and W_v can be learned by neural networks, which are all matrices that can be multiplied directly with I.

Our model avoids iterative operations like RNN; each sequence has no order in the model itself. Therefore, our group provides the model with the positional encoding of each sequence before self-attention, enabling the model to learn the dependencies between positions and the temporal properties of the audio sequences.

3.3 Decoder

Our group uses a decoder to map music genre classification labels for the abstract features aggregated by the encoder. Since we study the task of music genre classification, we do not use the decoder structure in [9]. Instead, we use a global-pooling feature aggregation layer and several fully-connected layers stacked together.

The global-pooling feature aggregation layer performs 1D-GlobalAvgPooling and 1D-GlobalMaxPooling on the feature vectors of the previous layer, stacks the pooling results, and sends them to the fully-connected layer, which outputs the labels for audio classification. The process of 1D global-pooling is shown in Fig. 4. The vector of the feature map after 1D-GlobalAvgPooling and the vector after 1D-GlobalMaxPooling are stacked together to form a 1D feature vector, which contains both the overall features and the most significant features of the feature map. Using 1D global-pooling is therefore analogous to matching a pattern over the entire duration of the audio, which can be seen as integrating information from the time dimension of the spectrogram while reducing the size of the features passed into the fully-connected layer.

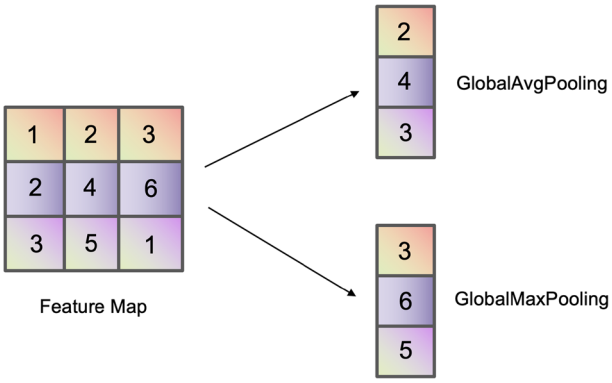

Fig. 4 The process of 1D global-pooling

4 Experiments

To validate the effectiveness of our model, our group conducts experiments on two benchmark datasets: GTZAN and Extended Ballroom. Section 4.1 demonstrates the details of the two above-shown datasets. Section 4.2 summarizes the parameter settings and the detailed configuration of the network during the experiments. Section 4.3 shows the experimental results on these two datasets and the validity verification of the partial structure of the model.

4.1 Dataset

GTZAN [41] dataset collected by Tzanetakis and Cook is widely used for music genre classification. It consists of 1000 audio clips containing ten different music genres: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, and Rock. Each genre consists of 100 excerpts that last about 30 s and are stored as 22,050 Hz.

Extended Ballroom [30] dataset is an improved version of the well-known dataset Ballroom [4], which contains a more significant number of audio tracks and better sound quality than the original. It has 4180 tracks in 13 different genres, with relatively low numbers of tracks in four categories, PasoDoble, Salsa, SlowWaltz, and WcSwing, as shown in Table 1. Each audio clip lasts 30 s and is stored as 16,000 Hz.

4.2 Experimental settings

GTZAN Due to duplicate information in the polyphonic channels of the original audio, all audio was converted to monophonic processing and down-sampled as 16 kHz. When converting the original audio to Mel-spectrogram, we set a Fourier transform window length of 512 and a hop size of 256. The processed Mel-spectrogram will be used as the input of our model with a dimension size of 217×334.

People do not need to listen to the whole song to distinguish the type of music when listening to music. Instead of using the complete audio as the input to the model, our model slices the audio and uses shorter audio slices as the basic unit of prediction. This increases the computational speed of the model and scales up the data. Besides, the final label of the audio is aggregated by the prediction results of each slice and voted jointly, which improves the classification performance. Figure 5 shows the slicing process of audio samples. Each GTZAN and Extended Ballroom music track is cut into 11 smaller clips (lasting 5 s) with 50% overlaps. Then, there are 11 times more music tracks for each genre label than in the original datasets.

Like many similar studies [36, 52], the experiments were performed by dividing the dataset 8:1:1 into training, validation, and test sets. The division was repeated ten times

Table 1 Genre distribution of extended ballroom

Genre	Track number	Genre	Track number	Genre	Track number
Chacha	455	Tango	464	Pasodoble	53
Jive	350	Viennesewaltz	252	Foxtrot	507
Quickstep	497	Wcswing	23	Waltz	529
Rumba	470	Slowwaltz	65		
Samba	468	Salsa	47		

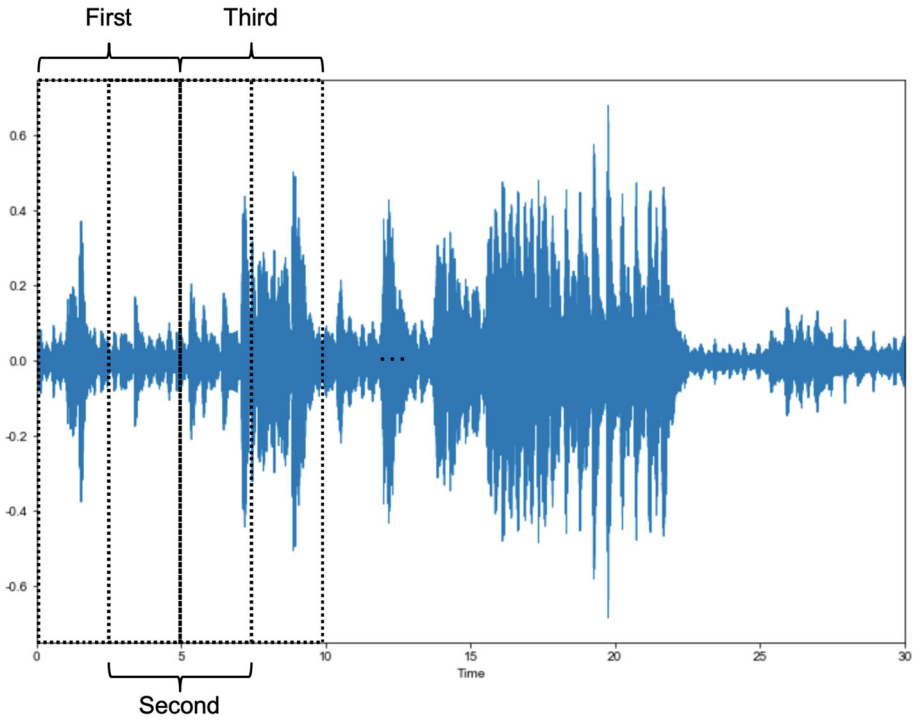

Fig. 5 The slicing process of audio samples

for ten-fold cross-validation, and the final experimental results were averaged over the ten experiments. Figure 6 shows the detailed configuration of each layer of the network, where F denotes the number of kernels, K denotes the size of kernels, and S denotes the stride. N_Head is the number of heads in multi-head attention, and N_Classes is the number of music genres in the dataset. Combined with the characteristics of the spectrogram, we use smaller kernels.

4.3 Experimental results

Define the confusion matrix, as shown in Fig. 7. We used accuracy, precision, recall and F1-score as evaluation indexes of our proposed music genre classification method.

Accuracy represents the proportion of correctly predicted samples to all samples, it can be computed by (4):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision is defined as the ratio of the number of correctly predicted samples to the total number of predicted samples, it can be computed by (5):

$$P = \frac{TP}{TP + FP} \tag{5}$$

Structure	Layer		Parameter	Output size
RGLU Blocks	Input		/	Mel-Spectrogram（217 × 334）
	Conv1D		F: 64, K: 3, S: 1	64 × 334
	RGLU Block 1	RGLU	Conv1D×2: F:64, K: 3, S: 1	64 × 334
		RGLU	Conv1D×2: F: 64, K: 3, S: 1	64 × 334
		Max-Pooling	K: 2	64 × 167
	RGLU Block 2	RGLU	Conv1D×2: F: 64, K:3, S:1	64 × 167
		RGLU	Conv1D×2: F: 128, K: 3, S: 1	128 × 167
		Max-Pooling	K: 2	128 × 83
	RGLU Block 3	RGLU	Conv1D×2: F: 128, K: 3, S: 1	128 × 83
		RGLU	Conv1D×2: F: 128, K: 3, S: 1	128 × 83
		Max-Pooling	K:2	128 × 41
	RGLU Block 4	RGLU	Conv1D×2: F:128, K: 3, S: 1	128 × 41
		RGLU	Conv1D×2: F: 256, K: 3, S: 1	256 × 41
		Max-Pooling	K:2	256 × 20
	RGLU Block 5	RGLU	Conv1D×2: F: 256, K: 3, S: 1	256 × 20
		RGLU	Conv1D×2: F: 256, K: 3, S: 1	256 × 20
		Max-Pooling	K: 2	256 × 10
Encoder	Positional Encoding		N_Head:8	256 × 10
	Multi-Head Attention			
	Feed Forward			
Decoder	1D-GlobalAvgPooling		/	256
	1D-GlobalMaxPooling			256
	Concatenate			512
	Fully-connected		U: 200	200
	Fully-connected		U: 100	100
	Fully-connected		U: N_Classes	N_Classes

Fig. 6 Network structure and parameters

Recall is defined as the ratio of the number of correctly predicted positive samples to the number of actual positive samples, it can be computed by (6):

$$R = \frac{TP}{TP + FN} \tag{6}$$

F1-score is a comprehensive evaluation metric that balances precision and recall to a certain extent. It takes into account both precision and recall to give a comprehensive evaluation of the performance of a model. F1-score can be computed by (7):

$$F1 = \frac{2 \times P \times R}{P + R} \tag{7}$$

4.3.1 Validation of res-gated convolutional structure

To verify the effectiveness of our proposed 1D residual-gated convolutional structure in the music genre classification task, we compare this structure with other forms of convolutional modules, and the experimental results are shown in Table 2. Since we are only verifying the

Confusion matrix		True value	
		Positive	Negative
Predicted value	Positive	TP	FP
	Negative	FN	TN

Fig. 7 Confusion matrix

Table 2 Validation of res-gated convolutional structure

	GTZAN				Extended Ballroom			
Models	ACC	P	R	F1	ACC	P	R	F1
CNN	0.903	0.882	0.890	0.886	0.816	0.795	0.826	0.810
RCNN	0.926	**0.882**	**0.893**	**0.887**	0.849	0.848	0.794	0.820
GLU	0.923	0.874	0.880	0.877	0.850	0.792	**0.854**	0.822
RGLU	**0.931**	0.872	0.884	0.878	**0.852**	**0.902**	0.762	**0.826**

The bold entries indicate the highest value

effectiveness of the convolutional structure, none of the networks in the table is combined with the encoder and decoder.

The four networks with different structures above experimented with the same experimental settings. It is clear from the results that the CNN with the gated convolutional structure works better than the original CNN. This is because the gated structure focuses more on the information related to the music genre categories, which is beneficial to learning spectrogram features. The network becomes more stable with the residual structure added, and the performance can be further improved.

4.3.2 Validation of attention mechanism

To show that the attention mechanism plays a positive role in the network's learning of musical features, we performed experiments using different RNN structures replacing the encoder of our model to obtain the effect of the attention mechanism on the model's classification. The experimental results are shown in Table 3. RGLUformer is illustrated in Fig. 1. RGLUNet is the network after removing the encoder from RGLUformer. RGLU-GRU is the structure of GRU by replacing encoder in RGLUformer. RGLU-BGRU uses a bidirectional structure based on RGLU-GRU. RGLU-LSTM is similar to RGLU-GRU, and RGLU-BLSTM is similar to RGLU-BGRU.

From the results, it can be demonstrated that the RNNs using a bidirectional structure outperform the single structure in terms of classification. This is because perceiving the overall sequence information in both directions helps the model intelligently understand the significance of the features abstracted by the convolutional layer at a particular moment for

Table 3 Validation of attention mechanism

	GTZAN				Extended Ballroom			
Models	ACC	P	R	F1	ACC	P	R	F1
RGLUNet	0.939	0.832	0.869	0.850	0.919	0.862	0.853	0.857
RGLU-GRU	0.944	0.866	0.852	0.859	0.921	0.875	0.853	0.864
RGLU-BGRU	0.948	0.872	0.884	0.878	0.926	0.850	0.882	0.866
RGLU-LSTM	0.951	0.894	0.876	0.880	0.923	0.845	0.891	0.867
RGLU-BLSTM	0.958	0.875	0.906	0.890	0.932	**0.879**	0.892	0.885
RGLUformer	**0.968**	**0.894**	**0.913**	**0.903**	**0.947**	0.873	**0.912**	**0.892**

The bold entries indicate the highest value

the whole music, thus enabling the network to better model the music sequence. The performance of the RGLUformer using the attention mechanism is slightly higher than several other RNNs, which indicates that assigning the corresponding attention weights to the feature sequences at each moment can integrate the sequence information more effectively and obtain a better overall feature representation of the music.

4.3.3 Validation of global-pooling

In this section, our group compares the classification performance of different pooling in the decoder, and the experimental results are shown in Table 4.

From the results, it can be observed that GlobalMaxPooling, which only focuses on the most representative features, performs less well. For example, the most representative feature of electronic music is the various types of electronic instruments. When the upper layer network has captured a particular pattern of electronic instruments, GlobalMaxPooling will highlight this pattern to make it easier to be recognized by the classifier. However, suppose only a particular region in the feature map is related to the music category. The feature map will significantly impact the results of music genre classification, resulting in poor classification performance. GlobalAvgPooling is more oriented to down-sampling the general feature information. It takes each region into account, and the presence of one or two special regions does not interfere with selecting the final model. For example, Blues is represented by a triplet rhythm, while Jazz is a quadruplet, showing a remarkable similarity between the two genres. This difference in the time domain is often reflected in the entire time dimension of the feature map, and GlobalAvgPooling enables the model to make better judgments about this situation. Therefore, global-pooling has higher classification performance when combining the advantages of GlobalAvgPooling and GlobalMaxPooling.

4.3.4 Validation of overall network

In this subsection, we compare our model with some other methods, and the results are shown in Table 5.

From Table 5, we can see that RDNN and BLSTM without convolutional structure exhibit relatively poor classification results, indicating that the convolutional structure can improve the model's ability to extract spectral features. Although KCNN uses a convolutional structure, it plainly stacks convolutional layers, which is not as effective as ResNet, which uses a residual structure. Audeep is based on the fusion of different sets of features, both visual and acoustic. Audeep uses a machine learning approach that struggles to cope with data sets with large amounts of data and has a weak ability to extract features, ulti-

Table 4 Validation of global-pooling

	GTZAN				Extended Ballroom			
Models	ACC	P	R	F1	ACC	P	R	F1
No pooling used	0.939	0.874	0.866	0.870	0.918	0.855	0.867	0.861
GlobalAvgPooling	0.956	**0.924**	0.872	0.897	0.931	0.846	0.878	0.862
GlobalMaxPooling	0.940	0.882	0.911	0.896	0.925	0.841	0.902	0.870
Global-Pooling	**0.968**	0.894	**0.913**	**0.903**	**0.947**	**0.873**	**0.912**	**0.892**

The bold entries indicate the highest value

Table 5 Validation of global-pooling

	GTZAN				Extended Ballroom			
Models	ACC	P	R	F1	ACC	P	R	F1
ResNet [20]	0.926	0.864	0.875	0.870	0.909	0.832	0.856	0.844
KCNN [51]	0.909	0.842	0.866	0.854	0.886	0.828	0.846	0.837
RDNN [39]	0.830	0.792	0.818	0.805	0.830	0.784	0.820	0.802
BLSTM [31]	0.878	0.855	0.847	0.851	0.867	0.802	0.848	0.824
AuDeep [19]	0.874	0.834	0.849	0.841	0.864	0.796	0.832	0.813
CVAF [34]	0.929	0.848	0.906	0.880	0.916	0.862	0.876	0.869
MFMCNN [34]	0.932	0.872	0.903	0.887	0.915	0.871	0.894	0.882
RGLUformer	**0.968**	**0.894**	**0.913**	**0.903**	**0.947**	**0.873**	**0.912**	**0.892**

The bold entries indicate the highest value

mately showing poor performance. CVAF and MFM-CNN rely on different feature fusion strategies to improve classification accuracy, but they do not summarize the temporal information in the audio. The RGLUformer proposed in the paper adopts a structure combining 1D res-gated convolution with attention mechanism and aggregates the global-pooling features of the convolutional layer. The network is more capable of extracting features related to the music category and achieves optimal performance.

5 Conclusion

The current music genres classification model is ineffective due to losing certain music signal features or low ability to extract the local feature in music. Our group proposes a novel model combining CNN and Transformer to recognize the music genre accurately. First, the proposed model uses a structure of 1D res-gated convolution based on the observation of audio sequences, which can extract deeper abstract features in the spectrogram. Second, considering the impact of specific musical elements appearing at different positions, our model utilizes the attention mechanism to assign corresponding weights to the feature sequences at each moment. Last, our model integrates the feature sequences by the global-pooling before the fully-connected layer classifies the music genres. Through the experimentation of the datasets GTZAN and Extended Ballroom, our group evaluated our model and our results show an increased accuracy of 3% to 4% in our model. Furthermore, the results demonstrate that our model is feasible and outperforms most of the previously proposed models in accuracy.

In the future, we will carry out further research from the following aspects. (1) An attempt to use multimodal data as input. In this paper, only audio is used as the data input for classification. In addition to audio, data such as lyrics and music videos also contain a wealth of information that can be used for classification. (2) Pre-training using self-supervised learning. Self-supervised learning can extract knowledge from unlabeled data, capturing a representation of the underlying structure that can help transfer to downstream tasks. (3) An attempt is made to learn the original audio signal directly. It is possible to consider using the original audio signal as the input to the network for music feature extraction, which can help improve the classification performance of the model.

Declarations

No potential conflict of interest was reported by the authors.

References

1. Abdoli S, Cardinal P, Koerich AL (2019) End-to-end environmental sound classification using a 1d convolutional neural network. Expert Syst Appl 136:252–263
2. Andén J, Mallat S (2014) Deep scattering spectrum. IEEE Trans Signal Process 62(16):4114–4128
3. Ba J, Mnih V, Kavukcuoglu K (2015) Multiple object recognition with visual attention. In: ICLR (Poster)
4. Cano P, Gómez E, Gouyon F, Herrera P, Koppenberger M, Ong B, Serra X, Streich S, Wack N (2006) Ismir 2004 audio description contest. Music Technology Group of the Universitat Pompeu Fabra, Tech. Rep
5. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision. Springer, pp 213–229
6. Chen C-FR, Fan Q, Panda R (2021) Crossvit: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 357–366
7. Chen X, Wu Y, Wang Z, Liu S, Li J (2021) Developing real-time streaming transformer transducer for speech recognition on large-scale dataset. In: ICASSP 2021-2021 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5904–5908
8. Cho K, van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: encoder–decoder approaches. In: Proceedings of SSST-8, eighth workshop on syntax, semantics and structure in statistical translation, pp 103–111
9. Choi K, Fazekas G, Sandler M, Cho K (2017) Convolutional recurrent neural networks for music classification. In: 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2392–2396
10. Ciresan DC, Meier U, Masci J, Gambardella LM, Schmidhuber J (2011) Flexible, high performance convolutional neural networks for image classification. In: Twenty-second international joint conference on artificial intelligence
11. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inform Theory 13(1):21–27
12. Dai J, Liang S, Xue W, Ni C, Liu W (2016) Long short-term memory recurrent neural network based segment features for music genre classification. In: 2016 10th International symposium on chinese spoken language processing (ISCSLP). IEEE, pp 1–5
13. Dai Y, Wu Y, Zhou F, Barnard K (2021) Attentional local contrast networks for infrared small target detection. IEEE Trans Geosci Remote Sens 59(11):9813–9824
14. Dieleman S, Schrauwen B (2014) End-to-end learning for music audio. In: 2014 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6964–6968
15. Dong Y, Yang X, Zhao X, Li J (2019) Bidirectional convolutional recurrent sparse network (bcrsn): an efficient model for music emotion recognition. IEEE Trans Multimed 21(12):3150–3163
16. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929
17. Downie JS (2003) Music information retrieval. Ann Rev Inform Sci Technol 37(1):295–340
18. Dauphin YN, Fan A, Auli M, Grangier D (2017) Language modeling with gated convolutional networks. In: International conference on machine learning. PMLR, pp 933–941
19. Freitag M, Amiriparian S, Pugachevskiy S, Cummins N, Schuller B (2017) audeep: unsupervised learning of representations from audio with deep recurrent neural networks. J Mach Learn Res 18(1):6340–6344
20. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
21. Hermann KM, Kocisky T, Grefenstette E, Espeholt L, Kay W, Suleyman M, Blunsom P (2015) Teaching machines to read and comprehend. Advances in Neural Information Processing Systems, 28
22. Hong D, Gao L, Yao J, Zhang B, Plaza A, Chanussot J (2020) Graph convolutional networks for hyperspectral image classification. IEEE Trans Geosci Remote Sens 59(7):5966–5978
23. Kereliuk C, Sturm BL, Larsen J (2015) Deep learning, audio adversaries, and music content analysis. In: 2015 IEEE Workshop on applications of signal processing to audio and acoustics (WASPAA). IEEE, pp 1–5

24. Kim T, Lee J, Nam J (2019) Comparison and analysis of samplecnn architectures for audio classification. IEEE J Selected Topics Signal Process 13(2):285–297
25. Koerich KM, Esmailpour M, Abdoli S, Britto AS, Koerich AL (2020) Cross-representation transferability of adversarial attacks: from spectrograms to audio waveforms. In: 2020 International joint conference on neural networks (IJCNN). IEEE, pp 1–7
26. Kumar DP, Sowmya BJ, Srinivasa KG et al (2016) A comparative study of classifiers for music genre classification based on feature extractors. In: 2016 IEEE Distributed computing, VLSI, electrical circuits and robotics (DISCOVER). IEEE, pp 190–194
27. Li TL, Chan AB, Chun AH (2010) Automatic musical pattern feature extraction using convolutional neural network. Genre 10(2010):1–1
28. Ling W, Dyer C, Black AW, Trancoso I (2015) Two/too simple adaptations of word2vec for syntax problems. In: Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1299–1304
29. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
30. Marchand U, Peeters G (2016) The extended ballroom dataset
31. Malhotra P, Vig L, Shroff G, Agarwal P et al (2015) Long short term memory networks for anomaly detection in time series. In: Proceedings, vol 89, pp 89–94
32. Meng F, Lu Z, Wang M, Li H, Jiang W, Liu Q (2015) Encoding source language with convolutional neural network for machine translation. arXiv:1503.01838
33. Mnih V, Heess N, Graves A et al (2014) Recurrent models of visual attention. Advances in Neural Information Processing Systems, 27
34. Nanni L, Costa YM, Lucio DR, Silla CN Jr, Brahnam S (2017) Combining visual and acoustic features for audio classification tasks. Pattern Recogn Lett 88:49–56
35. Ngai H, Park Y, Chen J, Parsapoor M (2021) Transformer-based models for question answering on covid19. arXiv:2101.11432
36. Pons J, Slizovskaia O, Gong R, Gómez E, Serra X (2017) Timbre analysis of music audio signals with convolutional neural networks. In: 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, pp 2744–2748
37. Rush AM, Chopra S, Weston J (2015) A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 379–389
38. Shi Y, Wang Y, Wu C, Yeh C-F, Chan J, Zhang F, Le D, Seltzer M (2021) Emformer: efficient memory transformer based acoustic model for low latency streaming speech recognition. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6783–6787
39. Sigtia S, Dixon S (2014) Improved music feature learning with deep neural networks. In: 2014 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6959–6963
40. Silla Jr CN, Kaestner CelsoAA, Koerich AL (2007) Automatic music genre classification using ensemble of classifiers. In: 2007 IEEE International conference on systems, man and cybernetics. IEEE, pp 1687–1692
41. Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Trans Speech Audio Process 10(5):293–302
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. Advances in Neural Information Processing Systems, 30
43. Wang F, Tax DMJ (2016) Survey on the attention based rnn model and its applications in computer vision. arXiv:1601.06823
44. Wang Z, Muknahallipatna S, Fan M, Okray A, Lan C (2019) Music classification using an improved crnn with multi-directional spatial dependencies in both time and frequency dimensions. In: 2019 International joint conference on neural networks (IJCNN). IEEE, pp 1–8
45. Xu W, Carpuat M (2021) Editor: an edit-based transformer with repositioning for neural machine translation with soft lexical constraints. Trans Assoc Comput Linguist 9:311–328
46. Xu C, Maddage NC, Shao X, Cao F, Tian Q (2003) Musical genre classification using support vector machines. In: 2003 IEEE International conference on acoustics, speech, and signal processing, 2003. Proceedings.(ICASSP'03)., vol 5. IEEE, pp V–429
47. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning. PMLR, pp 2048–2057

48. Yang H, Zhang W-Q (2019) Music genre classification using duplicated convolutional layers in neural networks. In: INTERSPEECH, pp 3382–3386
49. Yang R, Feng L, Wang H, Yao J, Luo S (2020) Parallel recurrent convolutional neural networks-based music genre classification method for mobile devices. IEEE Access 8:19629–19637
50. Yang C-HH, Qi J, Chen SY-C, Chen P-Y, Siniscalchi SM, Ma X, Lee C-H (2021) Decentralizing feature extraction with quantum convolutional neural network for automatic speech recognition. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6523–6527
51. Zhang P, Zheng X, Zhang W, Li S, Qian S, He W, Zhang S, Wang Z (2015) A deep neural network for modeling music. In: Proceedings of the 5th ACM on international conference on multimedia retrieval, pp 379–386
52. Zhang W, Lei W, Xu X, Xing X (2016) Improved music genre classification with convolutional neural networks. In: Interspeech, pp 3304–3308
53. Zhang T, Gong X, Chen CLP (2021) Bmt-net: broad multitask transformer network for sentiment analysis. IEEE Transactions on Cybernetics

Affiliations

Changjiang Xie[1] · Huazhu Song[1,2] · Hao Zhu[1] · Kaituo Mi[2] · Zhouhan Li[1] · Yi Zhang[1] · Jiawen Cheng[1] · Honglin Zhou[1] · Renjie Li[1] · Haofeng Cai[1]

Changjiang Xie
715480238@qq.com

Hao Zhu
1276004625@qq.com

Kaituo Mi
kaituo.mi@analyses.cn

Zhouhan Li
765975094@qq.com

Yi Zhang
1581554849@qq.com

Jiawen Cheng
1264825375@qq.com

Honglin Zhou
3485997412@qq.com

Renjie Li
1143936357@qq.com

Haofeng Cai
862378176@qq.com

[1] Wuhan University of Technology, WuHan, China

[2] Anngeen Technology Co., ltd., WuHan, China