Hindawi
Mobile Information Systems
Volume 2021, Article ID 3561829, 8 pages
https://doi.org/10.1155/2021/3561829

Research Article

A Generative Adversarial Network Model Based on Intelligent Data Analytics for Music Emotion Recognition under IoT

I.-Sheng Huang,[1] Yu-Hsuan Lu,[2] Muhammad Shafiq ,[3] Asif Ali Laghari,[4] and Rahul Yadav[5]

[1]*College of Music, Huaiyin Normal University, Huai'an 223300, Jiangsu, China*
[2]*College of Music, University of North Texas, Denton 76201, TX, USA*
[3]*Cyberspace Institute of Advance Technology, Guangzhou University, Guangzhou, China*
[4]*Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan*
[5]*Peng Cheng Laboratory, Shenzhen, China*

Correspondence should be addressed to Muhammad Shafiq; srsshafiq@gmail.com

Received 1 September 2021; Accepted 18 October 2021; Published 2 November 2021

Academic Editor: Anand Nayyar

The popularity of the Internet has brought the rapid development of artificial intelligence, affective computing, Internet of things (IoT), and other technologies. Particularly, the development of IoT provides more references for the realization of smart home. However, when people have achieved a certain amount of material satisfaction, they are more likely to want to communicate emotionally. Music contains a lot of emotion information. Music data is an important communication way between people and a better way to convey emotions. Therefore, it has become one of the most convenient and natural interactive ways expected by people in intelligent human-computer interaction. Traditional music emotion recognition methods have some demerits such as low recognition rate and time-consuming. So, we propose a generative adversarial network (GAN) model based on intelligent data analytics for music emotion recognition under IoT. Driven by the double-channel fusion strategy, the GAN can effectively extract the local and global features of the image or voice. Meanwhile, in order to increase the feature difference between the emotional voices, the feature data matrix of the Meyer frequency cepstrum coefficient of the music signals is transformed to improve the expression ability of the GAN. The experiment results show that the proposed model can effectively recognize the music emotion. Compared with other state-of-the-art approaches, the error recognition rate of proposed music music data recognition is greatly reduced. In terms of the accuracy, it exceeds 87% which is higher than that of other methods.

1. Introduction

The 21st century is a new era of information technology, including smart city, 4G communication technology, low-carbon technology, Internet of Things (IoT) [1], 3D display, enhanced display technology (AR), cloud computing, human vaccine technology, motor system energy saving, and combustible ice mining technology. Here, the IoT technology has naturally become a hot topic in the scientific community. Industry experts believe that on the one hand, the IoT can improve the quality of people's life and work and change the way people live. On the other hand, the IoT is an information technology revolution, which will drive the huge development of related industries, promote the progress of ST, and provide technological power for the recovery of the global economy.

Artificial emotion is a research field of simulating, identifying, and understanding human emotional processes by means of information science so that machines can generate human-like emotions and interact with humans naturally and harmoniously [2]. At present, research studies on artificial emotion mainly include two related fields: affective computing and Kansei Engineering. Artificial psychology is to use the means of information science to

simulate human emotion activities; its purpose is to study emotion, cognition, and motivation from the general psychological level of artificial machine realization. In this paper, we focus on the research studies of music emotion.

The music expresses the feelings of the composer and lyricist when it is created. Music is closely related to emotion and conveys a message that is difficult to quantify. With the development of the Internet, music plays an important role in people's life. People begin to pay more and more attention to the music emotion characteristics, and music emotion has also begun to be applied to music retrieval and music recommendation [3–5].

Music emotion recognition refers to the high-level effective emotional state recognition from low-level features of music, which can be regarded as a classification problem based on music sequence. The main processes include the emotion database establishment, the phonetic emotion features extraction, dimensionality reduction and features selection, and emotion classification and recognition. There are many methods for music emotion recognition, which have achieved better effects, such as hidden Markov model (HMM) [6], artificial neural network (ANN) [7], Gaussian mixture model (GMM) [8], support vector machine (SVM) [9], K-nearest neighbor (KNN) [10], and maximum likelihood Bayesian classification [11, 12]. However, the research objects (languages) are different, and there is no unified standard for the corpus database, so the recognition results are greatly different with each other.

SVM and KNN methods are often used in some models with high certainty, while human emotions are complex and uncertain. Therefore, the effect of music emotion recognition is poor. Neural network is a typical nondeterministic model, which has the characteristics of I/O nonlinear mapping, strong generalization ability [13], self-learning, self-organization, and self-adaptation ability. It has unique advantage in dealing with uncertain and nonlinear mapping problems. In the neural network models, convolutional neural network (CNN) [14, 15] is a kind of multilayer feedforward network that is widely used and most successful in pattern recognition. For example, Wal et al. [16] proposed a new emotion recognition method based on deep spatiotemporal analysis of facial geometric features. Nantasri et al. [17] investigated the possibility of using the mean values of MFCCs and their derivatives to create a new set of informative features. Maheshwari et al. [18] proposed the rhythm-specific multichannel convolutional neural network (CNN)-based approach for automated emotion recognition using multichannel EEG signals. Rajapakshe et al. [19] introduced a novel policy called "Zeta policy" which was tailored for speech emotion recognition and applied pretraining in deep reinforcement learning to achieve faster learning rate. Pretraining with cross dataset was also studied to discover the feasibility of pretraining the reinforcement learning agent with a similar dataset in a scenario where no real environmental data was available. Since human emotions have complex and uncertain information, the recognition rate of voice emotion based on convolutional neural network is still not high. In order to increase the feature difference between emotional music information, this paper proposes a generative adversarial network (GAN) model via double-channel fusion strategy based on intelligent data analytics. The main contributions are as follows:

(1) A generative adversarial network (GAN) model based on intelligent data analytics for music emotion recognition under IoT is proposed

(2) Driven by the double-channel fusion strategy, the GAN can effectively extract the local and global features of the image or voice

(3) Meanwhile, in order to increase the feature difference between the emotional voices, the feature data matrix of the Meyer frequency cepstrum coefficient of the music signals is transformed to improve the expression ability of the GAN

(4) The experiment results show that the proposed model can effectively recognize the music emotion

The remaining of this paper is organized as follows. Music emotion feature extraction is introduced in Section 2. Section 3 presents the experiments. Section 4 concludes this paper.

2. Music Emotion Feature Extraction

Traditional music emotion feature extraction methods through global analysis extract music signal pitch frequency, amplitude, energy, speed, and formant parameters. The timing and distribution characteristics of these parameters are analyzed to find the rhythm rules in different emotional sounds, which can be used as the basis for emotion recognition. In literature [20], a 40-dimensional emotional feature vector A is extracted by extracting relevant emotional features in music signals. Its form is as follows:

$$A = \begin{bmatrix} a_{1,1} \cdots a_{1,40} \\ \vdots \\ a_{40,1} \cdots a_{40,40} \end{bmatrix}. \quad (1)$$

From signal analysis, music signal is composed of many different overlapped frequency signals. The spectrum features analysis of signals is also conducive to emotion recognition research. MFCC is an algorithm based on the auditory characteristics of human ears, which uses a nonlinear frequency unit (Mel frequency) to simulate the human auditory system. In recent years, relevant research has applied MFCC to music recognition [21]. This paper also adopts the MFCC feature extraction method to extract music emotion features.

At present, the traditional MFCC feature extraction method takes 256 sampling points as 1 frame length and 160 sampling points as frame shift. The coefficient order is 12. The energy level, first-order difference, and second-order difference of each frame are calculated, respectively, and the average value of each frame coefficient is further calculated. Therefore, a 40-dimensional filter band coefficient is obtained for each frame [22]. The sampling points of each voice sample are not uniform, resulting in inconsistent frame number. In order to obtain a uniform frame number, the following two extracting feature methods are studied in this paper:

Method 1. The common feature extraction scheme is to directly extract feature data from each music sample. Although the frame number of extracted feature is not uniform, most of them are between 140 and 170. In order to retain the features in each music, the feature data of each sample is cut to 160 frames. And the feature data is further converted into a matrix form with size 80×80 as the input of the convolutional neural network.

Method 2. The extracted data is huge in method 1, which leads to long training time. To reduce the dimension of the feature, the music sampling points of each sample are first uniformly adjusted as 7136. Then, the MFCC coefficient is extracted to obtain the unified 40-frame feature data. The feature data is converted into the matrix form A1 with size 40×40 as the input of the convolutional neural network.

Through abundant comparison experiments for the above two methods, it is found that the training time of method 2 is less and the recognition rate is higher. This paper also considers the normalization effect on $A1$. The comparison results show that the recognition rate of $A1$ after normalization is not improved and unstable. Therefore, the original $A1$ is used as the input of the neural network in this paper. The feature vectors A and $A1$ are used as the input of the convolutional neural network model, respectively. The experiment shows that the error recognition rate is lower when extracting MFCC features.

3. GAN

GAN was proposed by GoodFellow in 2014 [23]. The framework consists of two subnetworks: generator G and discriminator D. The corresponding functions are mapping random noise to sample distribution and discriminating real samples and generated samples, respectively. Different from other generative models, GAN adopts an adversarial approach. First, it learns the difference between the real sample and the generated sample through D and then guides G to generate false samples closer to the distribution of the real sample. It adopts alternate training to continuously reduce the difference. At present, GAN mainly optimizes the following maximum and minimum loss function to achieve Nash equilibrium:

$$\min_G \max_D E_{x \sim q_{\text{data}}(x)}[\lg D(x)] + E_{z \sim p(z)}[\lg(1 - D(G(z)))], \tag{2}$$

where $z \in R^{d_z}$ is a potential variable obtained from a distribution $p(z)$ such as Gaussian noise or uniform distribution. The generator G and discriminator D have their own loss functions. During the training, G and D will update their respective parameters and minimize the loss. G and D cannot update each other's parameters, but they need to rely on adverse parameters to update their own. Arjovsky et al. [24] proposed the DCGAN (deep convolutional generative adversarial network) framework in 2016 and applied convolutional neural networks (CNNs) to GAN for the first time. Since then, generator G and discriminator D usually adopt the CNN model. Deep learning-based generative adversarial network has achieved great success in the field of image generation. As a unique image synthesis technology, it is widely used in image generation. GAN has the following advantages:

(1) It can train the unconditional generator only by inputting random noise

(2) It is a new technique for data transfer between different domains and an effective method for unsupervised image conversion between domains

(3) It is a new optimization method and provides an effective image perception loss function

Although GAN has made great progress and effectively generated convincing images, there are still some problems to be solved:

(1) The training process of GAN is extremely unstable, and the network is very sensitive to super parameters, so it is difficult to reach Nash equilibrium

(2) GAN often shows the model collapse phenomenon, which results in the model simulating only a part of the real distribution, rather than all the object distribution

(3) GAN cannot capture the structure and geometric shapes in some categories of images

Most of the existing works are devoted to optimizing the training process of GAN, and some works focus on changing the objective function of GAN. For example, the cross entropy loss replaced the least squares loss in the LSGAN (least squares generative adversarial network) method, which not only improved the stability of training but also shortened the training time. Some works focus on gradient punishment or the gradient of constraint D to ensure that D can provide effective gradient for G. The WGAN (Wasserstein generative adversarial network) model [25] is restricted D to satisfy Lipschitz constraint, which greatly improved the stability of the network. Although WGAN satisfied Lipschitz constraint, it directly restricted the parameter matrix, which destroyed the structure of the parameter matrix. To solve this problem, a new regularization technique was introduced in reference [25], which not only satisfied Lipschitz constraint but also did not destroy the structure of the parameter matrix.

Additionally, some references modify the GAN framework. EBGAN (energy-based generative adversarial network) was the first framework by introducing the energy model into GAN [26]. It regarded D as an energy model and adopted an auto-encoder structure. Real samples were given low energy, and fake generated samples were given high energy. By reducing the reconstruction error of the generated samples, the samples were gradually closer to the real sample distribution. ProGANs (progressive generative adversarial networks) [27] trained a high resolution GAN by gradually enhancing G and D. Training starts with low-

resolution images, and it gradually improved resolution by adding layers to the network. This training method first detected the distribution of large structural images and then shifted attention to finer and finer scale details. It could not learn all the ratios at once. However, it only produced good results on a single feature image. By reinforcing the connection between local and global locations of feature graphs, SAGANs (self-attention generative adversarial networks) [28] attempted to generate high quality images on multicategory images, but it ignored the connection between channels of feature graphs.

In view of the fact that GAN cannot capture features in certain categories of music, this paper proposes a GAN model based on double-channel attention mechanism, which can effectively capture the feature distribution of music through adaptive learning the dependence between local and global features to improve the music emotion recognition accuracy.

4. Double-Channel Attention Mechanism

GAN has made great progress in this field. It is difficult to train the model on larger data sets, and it cannot capture the geometric features that occur many times in some classes. The reason for this problem is that the current model relies much on the dependence between different regions of the convolution simulating image. Due to the local receptive field of convolution, the dependence between a large range of regions can only be obtained through multiple convolution operations. As shown in Figure 1, three 3×3 convolutional layers are required to obtain the feature relations between 7×7 receptive fields. However, in the process of convolution operation, the optimization algorithm may be difficult to coordinate so many convolution layers. And more convolutional layers capture little dependencies. If the size of the convolution kernel is enlarged, for example, a convolution kernel with the size of 7×7 is adopted, the feature dependence between 7×7 receptive fields can be obtained through only one convolutional layer. However, this method is not only less effective than the convolution layer combination of several small filters but also greatly increases the computational burden. Therefore, it is difficult to obtain the dependencies between images only through convolutional layer.

To solve the problem of ineffectively capturing music features by CNN, some scholars introduced an attention model in GAN to make up for the deficiency of CNN framework. The essence of the attention model is to emphasize or select the important information of the object through many attention distribution coefficients (weight coefficient) and suppress some irrelevant details. The attention mechanism can capture the relations between local and global flexibly, improve the representation ability of the model, and reduce the model complexity. Therefore, in order to improve the recognition rate of music emotion, this paper proposes a GAN framework based on double-channel attention mechanism (DCGAN) and introduces two different attention models: feature attention model and channel attention model, which can capture the feature dependency in feature space and channel, respectively.

4.1. Feature Attention Model. In order to add the dependent information between the local feature and the global feature in the feature graph, a feature attention model is introduced. This model enhances its representation capability by encoding extensive global spatial information and adding it to local feature information. The specific framework is shown in Figure 2 where C represents the channel number in the feature graph and H and W represent the height and width of the feature graph, respectively.

Firstly, the feature map $X \in R^{C \times H \times W}$ of the previous layer with 1×1 convolution forms three feature spaces: R, S, and T. The number of channels in each feature space is $C/8$, $C/8$, and C, respectively, in which matrix multiplication is performed on the transpose of feature space R and S. Softmax is applied to obtain the parameters of feature attention layer. The specific parameter values are calculated by the following equation:

$$p_{j,i} = \frac{\exp\left(R_i \cdot S_j\right)}{\sum_{i=1}^{H \times W} \exp\left(R_i \cdot S_j\right)}, \tag{3}$$

where $p_{j,i}$ represents the influence of the feature in the i-th position on the feature of the j-th position. If the features of the two positions are more similar, the correlation between them is bigger. Then, the feature graph of feature attention $P = (P_1, P_2, \ldots, P_j, \ldots, P_{(H \times W)}) \in R^{C \times (H \times W)}$ is obtained by matrix multiplication of the feature space T and the transpose of the feature attention layer.

$$P_j = \sum_{i=1}^{H \times W} p_{ji} T_i. \tag{4}$$

4.2. Channel Attention Mechanism. For feature graphs, each channel can be considered to be a specific class. The different channels are related to each other. Therefore, a channel attention model is proposed to extract the dependencies between different channels. The channel attention model framework is shown in Figure 3.

Feature attention needs to convolve feature graph $R \in R^{C \times H \times W}$. However, channel attention directly uses feature graph X to calculate channel attention feature layer parameters, but the calculation process is similar. The calculation formula is shown in the following equation:

$$p_{j,i} = \frac{\exp\left(R_i \cdot S_j\right)}{\sum_{i=1}^{H \times W} \exp\left(R_i \cdot S_j\right)}, \tag{5}$$

where $q_{m,n}$ is the influence of the n-th channel on the m-th channel. If the features of the two channels are very close, the dependency between them is greater [29]. In addition, matrix multiplication is performed for channel attention feature layer and transpose of input feature space X, and finally channel attention feature graph $Q = (Q_1, Q_2, \ldots, Q_m, \ldots, Q_{(H \times W)}) \in R^{C \times (H \times W)}$ is output.

$$Q_m = \sum_{n=1}^{H \times W} q_{mn} X_n. \tag{6}$$

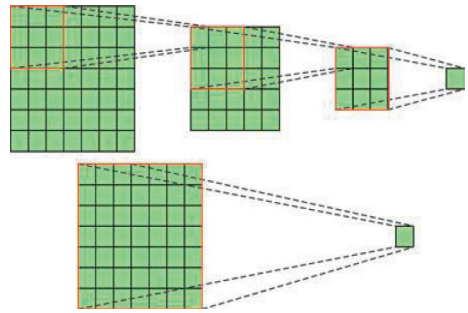

Figure 1: Schematic diagram of obtaining 7×7 receptive fields by different convolution kernels.

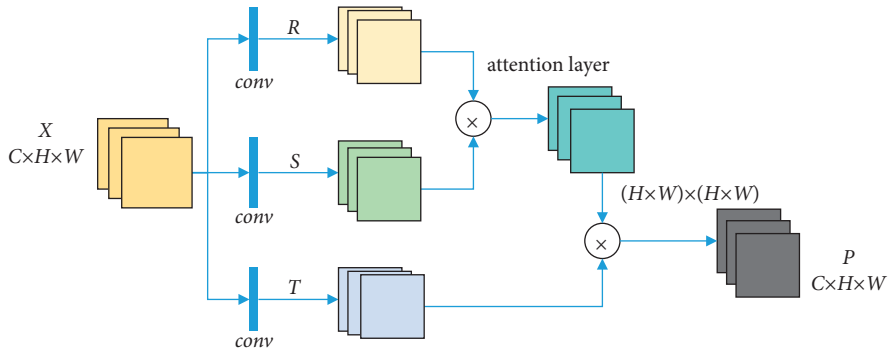

Figure 2: Feature attention model.

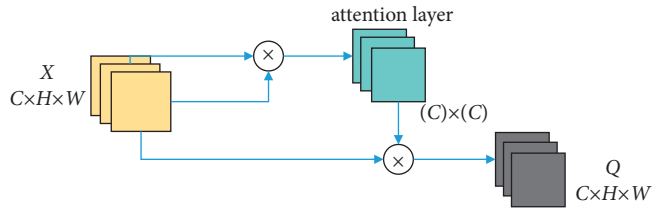

Figure 3: Channel attention model.

4.3. Double-Channel Attention Model. Figure 4 shows the double-channel attention model. The input feature graph is fused with the output P (obtained by feature attention model) and Q (obtained by channel attention model) to obtain the feature space $E \in R^{C \times H \times W}$ with local and global feature dependence information as well as class dependence information. Its calculation formula is shown in the following equation:

$$E = \alpha P + \beta Q + X, \tag{7}$$

where α and β are the hyperparameters of P and Q, respectively, and they are initialized as 0 and updated through back propagation. In the process of network training, while the two attention models start from the simple feature dependence, they gradually learn to complex dependencies; α and β of P and Q are gradually increased; and the weight feature map learned by the attention module is added to the original feature map. Thus, the feature graph that needs to be applied attention is emphasized. In the high-level networks of G and D, the double-channel attention mechanism acts as an auxiliary structure of GAN, cascaded after CNN [29]. Figure 5 shows the training flow chart of proposed network, where CNN represents the convolution operation and DC represents the introduced double-channel attention mechanism. Through continuous cyclic alternate training of G and D, G generates more and more realistic images.

5. Experiments and Analysis

The experimental songs are mainly from the mood song lists recommended by various music websites on the Internet, such as Kuwo Music Box, BaiDu Heartlisten, and other music websites. A total of 637 songs are used in this experiment, among which 445 pop musics (215 happy musics and 230 sad musics) are used to train the music emotion classifier and 192 pop musics (93 happy musics and 99 sad musics) are used as test songs [30]. The music emotion is divided into sad, happy, quiet, lonely, and miss in this paper [31]. The extracted music features are used to construct training and testing samples. In these experiments, we also

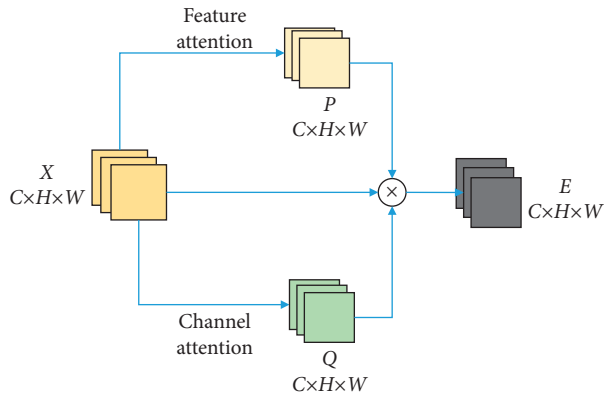

Figure 4: Double-channel attention model.

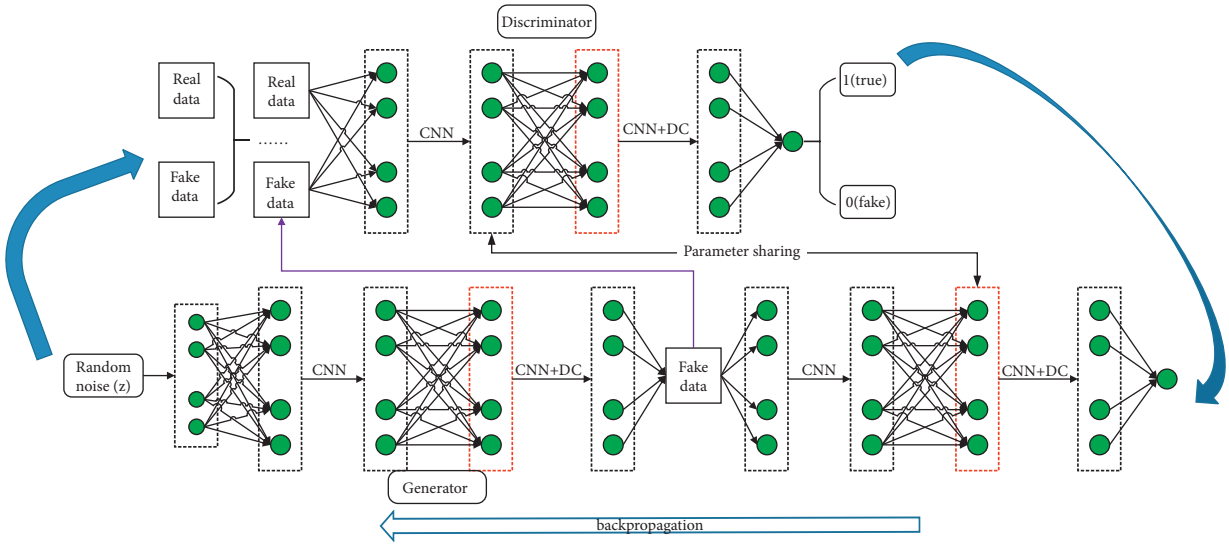

Figure 5: Double-channel GAN.

Table 1: The recognition rate comparison with different models.

Model	Average accuracy (%)	Time (s)
Feature attention	86.3	0.8
Channel attention	87.8	0.7
Double-channel attention	93.4	1.2

make comparison with CLSTM [32], RNN [33], and HTG [34].

First, the feature attention, channel attention, and double-channel attention models are compared, and the results are shown in Table 1. Figure 6 is the visualization result.

From Table 1 and Figure 6, we can see that the double-channel attention model achieves the 93.4% accuracy, which improves by 7.1% and 5.6% than that of feature attention and channel attention, respectively. However, the running time of the double-channel attention model is 1.2 s, which is a little longer than that of feature attention and channel attention due to the two channels' combination. Table 2 shows the comparison with different methods.

Table 2 shows that the proposed method has the best results than other methods. The bold values in Table 2 are the best values. Especially, the sad and happy recognition rates exceed 90%, because the double channels are utilized to extract the local and global features. The error rate is shown in Table 3.

As can be seen from the comparison of error recognition rates in Table 3, although A and B begin to converge in the 700-th iteration, their convergences are not stable. The proposed model is relatively stable in the 500-th iteration. This is because the two channel attention mechanisms are adopted, so the weight of the convolution kernel gradually becomes stable with the increase in iteration times, and the recognition rate of the model converges steadily, and the

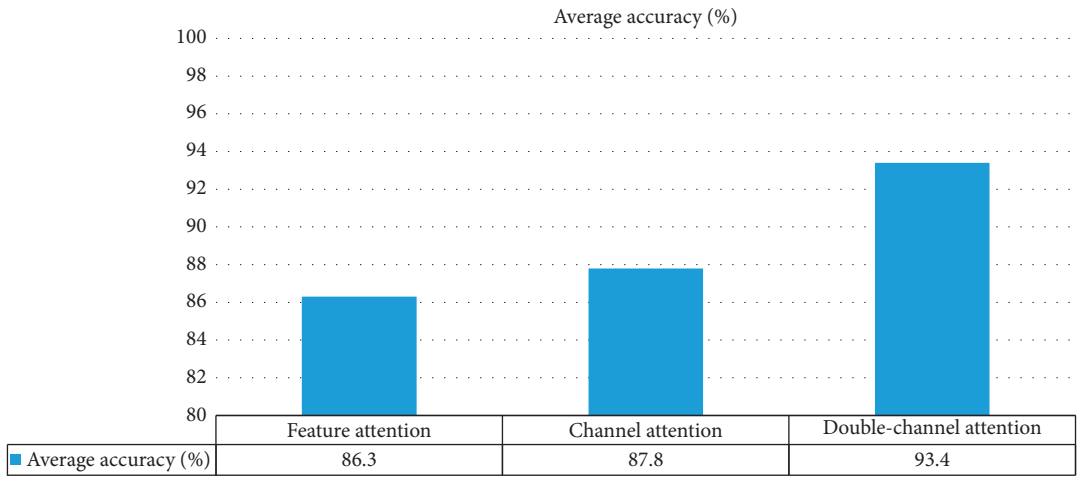

Figure 6: Visualization result for accuracy.

Table 2: The recognition rate comparison with different methods (%).

Model	Sad	Happy	Quiet	Lonely	Miss
CLSTM	79.3	76.4	71.8	72.2	69.5
RNN	82.5	80.9	77.2	79.8	75.6
GAN	83.6	82.1	79.4	80.5	79.4
HTG	88.1	86.4	81.6	82.5	80.7
Proposed	**91.2**	**90.5**	**88.7**	**89.2**	**87.6**

Table 3: The error rate with different methods.

Iteration number	CLSTM	RNN	HTG	Proposed
100	59.2	58.5	57.6	56.1
200	52.3	50.2	47.6	44.8
300	51.8	49.4	46.2	43.7
400	45.8	37.6	35.2	29.7
500	35.4	28.7	26.5	9.2
600	35.4	28.6	26.1	9.2
700	31.6	27.2	26.1	9.2
800	31.6	27.2	26.1	9.2

error recognition rate decreases significantly. Increasing the input feature data by one time can better reflect the difference of music emotion features. The error recognition rate is lower, so the convergence is faster.

6. Conclusions

To solve the problem that traditional CNN cannot effectively extract the dependence between music features and different categories, a generative adversarial network model based on double-channel attention mechanism is proposed, which includes feature attention and channel attention. Driven by attention mechanism, the two submodels are used to model the dependencies between local features and global features and the dependencies between classes, respectively. Then, it realizes the task of generating music feature library. The results show that the proposed framework can obtain the feature information of music more comprehensively than other frameworks and improve the ability of music emotion recognition.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] M. Shafiq, Z. Tian, A. K. Bashir, X. Du, and M. Guizani, "CorrAUC: a malicious bot-IoT traffic detection method in IoT network using machine-learning techniques," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3242–3254, 2021.

[2] R. Savery, R. Rose, and G. Weinberg, "Establishing human-robot trust through music-driven robotic emotion prosody and gesture," in *Proceedings of the 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1–7, New Delhi, India, October 2019.

[3] Y. Jing, H. Li, and S. Yin, "Dynamic gesture recognition based on deep learning in human-to-computer interfaces," *Journal of Applied Science and Engineering*, vol. 23, no. 1, pp. 31–38, 2020.

[4] F. Pan, L. Zhang, Y. Ou, and X. Zhang, "The audio-visual integration effect on music emotion: behavioral and physiological evidence," *PLoS One*, vol. 14, no. 5, Article ID e0217040, 2019.

[5] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 142–156, 2012.

[6] P. Chen, L. Zhao, Z. Xin, Y. Qiang, M. Zhang, and T. Li, "A scheme of MIDI music emotion classification based on fuzzy theme extraction and neural network," in *Proceedings of the 2016 12th International Conference on Computational*

Intelligence and Security (CIS), pp. 323–326, Wuxi, China, December 2016.

[7] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "Modeling the affective content of music with a Gaussian mixture model," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 56–68, 2015.

[8] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio Speech and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.

[9] R. Xu, L. Ye, and J. Xu, "Reader's emotion prediction based on weighted latent Dirichlet allocation and multi-label k-nearest neighbor model," *Journal of Computational Information Systems*, vol. 9, no. 6, pp. 2209–2216, 2013.

[10] Y. Sun, S. Yin, H. Li, L. Teng, and S. Karim, "GPOGC: Gaussian pigeon-oriented graph clustering algorithm for social networks cluster," *IEEE Access*, vol. 7, pp. 99254–99262, 2019.

[11] C. Chen and Q. Li, "A multimodal music emotion classification method based on multifeature combined network classifier," *Mathematical Problems in Engineering*, vol. 2020, Article ID 4606027, 11 pages, 2020.

[12] S. Yin, H. Li, and L. Teng, "Airport detection based on improved faster RCNN in large scale remote sensing images," *Sensing and Imaging*, vol. 21, 2020.

[13] M. Shafiq, Z. Tian, A. K. Bashir, X. Du, and M. Guizani, "IoT malicious traffic identification using wrapper-based feature selection mechanisms," *Computers & Security*, vol. 94, Article ID 101863, 2020.

[14] S. Yin, Y. Zhang, and S. Karim, "Region search based on hybrid convolutional neural network in optical remote sensing images," *International Journal of Distributed Sensor Networks*, vol. 15, no. 5, Article ID 155014771985203, 2019.

[15] Y. Huang, J. Xiao, K. Tian, A. Wu, and G. Zhang, "Research on robustness of emotion recognition under environmental noise conditions," *IEEE Access*, vol. 7, pp. 142009–142021, 2019.

[16] J. V. Wal, A. A. Kauffman, and Z. A. Soulliard, "Differences in alexithymia, emotional awareness, and facial emotion recognition under conditions of self-focused attention among women with high and low eating disorder symptoms: a 2 X 2 experimental study," *Journal of Eating Disorders*, vol. 8, 2020.

[17] P. Nantasri, E. Phaisangittisagul, J. Karnjana et al., "A lightweight artificial neural network for speech emotion recognition using average values of MFCCs and their derivatives," in *Proceedings of the 2020 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 41–44, Phuket, Thailand, June 2020.

[18] D. Maheshwari and S. K. Ghosh, R. K. Tripathy, M. Sharma and U. R. Acharya, "Automated accurate emotion recognition system using rhythm-specific deep convolutional neural network technique with multi-channel EEG signals," *Computers in Biology and Medicine*, vol. 134, 2021.

[19] T. Rajapakshe, R. Rana, S. Khalifa, B. W. Schuller, and J. Liu, "A novel policy for pre-trained deep reinforcement learning for speech emotion recognition," 2021, https://arxiv.org/abs/2101.00738.

[20] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[21] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, pp. 2672–2680, 2014.

[22] M. Shafiq, Z. Tian, A. A. Bashir, A. Jolfaei, and X. Yu, "Data mining and machine learning methods for sustainable smart cities traffic classification: a survey," *Sustainable Cities and Society*, vol. 60, 2020.

[23] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, https://arxiv.org/abs/1511.06434.

[24] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 214–223, Sydney, Australia, August 2017.

[25] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, https://arxiv.org/abs/1802.05957.

[26] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial networks," 2016, https://arxiv.org/abs/1609.03126.

[27] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, https://arxiv.org/abs/1710.10196.

[28] S. Wu, J. Yang, Y. Shan, and X. Bingbing, "Research on generative adversarial networks using twins attention mechanism," *Journal of Frontiers of Computer Science and Technology*, vol. 14, no. 5, pp. 833–840, 2020.

[29] M. Shafiq, Z. Tian, Y. Sun, X. Du, and M. Guizani, "Selection of effective machine learning algorithm and Bot-IoT attacks traffic identification for internet of things in smart city," *Future Generation Computer Systems*, vol. 107, pp. 433–442, 2020.

[30] Y. Deng, Y. Lv, and M. Liu, "Music emotion recognition based on middle and high level features," *Computer Engineering and Design*, vol. 38, no. 4, pp. 1029–1034, 2017.

[31] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Systems*, vol. 24, no. 4, pp. 365–389, 2018.

[32] S. Hizlisoy, S. Yildirim, and Z. Tufekci, "Music emotion recognition using convolutional long short term memory deep neural networks," *Engineering Science and Technology an International Journal*, vol. 24, no. 3, 2020.

[33] J. Grekow, "Static music emotion recognition using recurrent neural networks," in *Lecture Notes in Computer Science*, D. Helic, G. Leitner, M. Stettinger, A. Felfernig, and Z. W. Raś, Eds., vol. 12117, pp. 150–160, Springer, Cham, Germany, 2020.

[34] D. Chaudhary, N. P. Singh, and S. Singh, "Automatic music emotion classification using hashtag graph," *International Journal of Speech Technology*, vol. 22, no. 3, pp. 551–561, 2019.