Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Music Genre Classification Using Deep Learning

ZHIJIE CAO[1], (Fellow, IEEE)

[1]Nanjing University of Posts and Telecommunications School of Computer Science, Nanjing, China (e-mail: 1023040914@njupt.edu.cn)

Corresponding author: First A. Author (e-mail: 1023040914@njupt.edu.cn).

This paragraph of the first footnote will contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

ABSTRACT Music genres are classification labels created by humans to describe musical fragments. The characteristic of music genres is that their members share the same characteristics. These characteristics are usually related to the instruments, rhythm, and style of the music being played. This article aims to explore big data analysis methods and applications for music classification. We will propose an effective music classification method based on large-scale music datasets, utilizing machine learning and data mining techniques, and apply it to practical scenarios.

INDEX TERMS music genre classification, convolutional neural networks, data miningsignal analysis

I. INTRODUCTION

MUSIC is an important component of human culture, which integrates emotions, expression, and creativity, providing people with rich and diverse artistic experiences. However, with the advent of the digital age, music data is exploding, and how to effectively classify and organize music has become an urgent need. Traditional music classification methods often rely on manual annotation and professional knowledge, but this method faces significant challenges when dealing with large-scale music data.

With the rapid development of big data and machine learning technology, using big data analysis methods for music classification has become increasingly attractive. Big data analysis can help us extract features from massive music datasets and build models that can automatically classify music. Through this approach, we can better understand and discover different types of music, providing users with personalized music recommendations and related services.

This article aims to explore big data analysis methods and applications for music classification. We will propose an effective music classification method based on large-scale music datasets, utilizing machine learning and data mining techniques, and apply it to practical scenarios. Specifically, we will introduce methods for music data collection and preprocessing, discuss commonly used music feature extraction techniques, and the process of constructing and optimizing music classification models. We will also explore how to apply music classification technology to practical scenarios, such as music recommendation systems, music style analysis, and music copyright protection.

Through this study, we hope to provide new ideas and methods for the research and practice of music classification. I believe that big data analysis has great potential in music classification, providing users with a better music experience and promoting the development of the music industry. At the same time, we also recognize the limitations and challenges of current methods, such as cross domain music classification and multimodal data fusion. Therefore, future research work will need to further explore these directions in order to improve the accuracy, efficiency, and scalability of music classification.

CNN is one of the most popular deep learning methods in the past six years, because it can provide very high accuracy in the process of extracting information from images. CNN can observe image features, and with the deepening of the hierarchy, the features that need to be learned become more and more complex. By observing these features, CNN can classify images. In audio classification, the CNN based model with time frequency as input is very popular. CNN has also demonstrated the most advanced performance in speech recognition and music segmentation.

The advantage of CNN is that it can automatically learn the key features in image and audio data without manually extracting features. This makes CNN an ideal choice for processing large datasets and achieving high accuracy. In addition, CNN can effectively reduce the number of parameters through convolution and pooling, thus reducing the complexity of the model and improving the computational efficiency.

In the audio field, CNN is widely used in music genre classification, emotion analysis, voice recognition and other tasks. Its advantage is that it can capture the time-domain and frequency-domain characteristics of audio signals and identify different audio categories from them. In addition, CNN can also handle long-term dependence, and has a good perception of rhythm, melody, harmony and other features in music.

In a word, CNN, as a powerful deep learning method, has achieved remarkable results in both image and audio fields. With the further development of technology, we can expect the application and performance improvement of CNN in more fields.

II. RELATED WORKS

Music genre classification is a popular task in the field of deep learning. In the field of music classification, researchers have conducted extensive work and explored various methods and techniques to solve this problem. Here is a brief introduction to some related work.

In [1], the author used the residual neural network (RNN) to train the model on the 3-second audio clip extracted from the GTZAN dataset. The overlapping characteristics of different schools are also considered, and the accuracy rate reaches 94%. Similarly, the author compared various algorithms in [2], and proposed a new near real-time classification using RNN, but the accuracy rate is low, 64%. The authors used the mean and covariance of MFCC to train their model.

After comparing various pre-existing models [3], the author tries to find the best machine learning algorithm for music genre classification. The model is trained with MFCC (Mel frequency cepstrum coefficient) and other features of songs. Among them, the convolutional neural network (CNN) has the highest accuracy, 88.5%.

Some authors use convolutional neural networks to classify genres. There are three different ways to visualize video - spectrogram, chromatogram and MFCC; Different authors use different visualization methods.

The author [4] compares two types of models, in which the CNN architecture (VGG-16) is used for the spectrum of audio signals and the time and frequency domains of audio. The data set used by the author is 10 seconds, and audio clips are extracted from 2.1 million YouTube videos. The author uses a new classifier, and obtains 65% accuracy, AUC is 0.894.

In [5], the author extracted mel spectrogram and used it as input. The author uses repeated convolution layer, in which the output passes through different pooling layers, and makes statistical analysis.

Another author [6] used MFCC and compared CNN and Long Short-Term Memory (LSTM) models, in which CNN model produced better accuracy. CNN has five convolution layers, each of which has 32 nodes, a fully connected layer has 128 nodes, and then an output layer of 10 nodes. The LSTM model has five layers, of which the first layer has 128 nodes, and the remaining layers have 32 nodes.

However, it is worth mentioning that they used convolutional neural networks with ReLU activation on song segments preprocessed into MFCC spectrograms. Gwardys et al. [7] demonstrated an interesting method involving transfer learning. They initially trained models for image recognition on ILSVRC-2012 [8], and then reused Model 1 for type recognition on the MFCC spectrogram. The architecture used in this article consists of five convolutional layers, with the first two and the last also having a maximum pool. Finally, three fully connected layers. The author used midi, pitch, and duration as features of music for classification in [9] to achieve good results. Considering the following literature, we propose a system that automatically classifies music into different genres.

III. PROBLEM STATEMENT

When it comes to music classification, an important issue is how to automatically categorize music works into the correct genres. Music genres are defined based on the style, characteristics, and expression of music. Accurately categorizing music works into appropriate genres is of great significance for areas such as music recommendation, music library management, and music research.

The current music classification methods face some challenges. Firstly, music is a multimodal data that contains various forms of information such as audio, text, and images, while traditional music classification methods often only utilize a single data modality. How to effectively integrate multimodal data and fully utilize the information in each data modality is a problem worth exploring.

Secondly, there is subjectivity and ambiguity in the definition and division of music genres. Different people may have different classification criteria and interpretations for the same musical work. Therefore, how to establish an objective, consistent, and universally applicable music genre classification system is a difficult problem that needs to be solved.

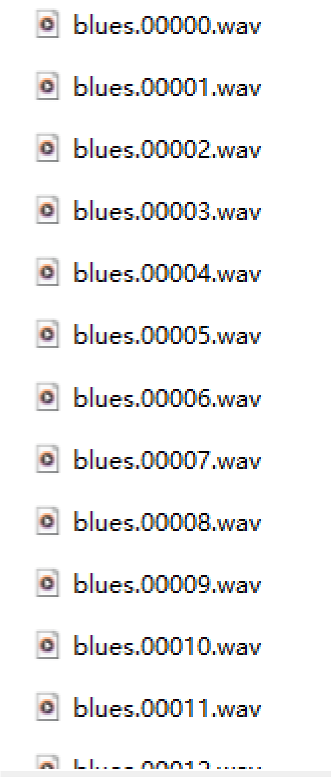
In addition, feature extraction of music is also a key step in music classification. The current feature extraction methods are usually based on manually designed rules, such as pitch, rhythm, lineage, etc. However, these features are often subjectively chosen by domain experts and may not fully capture the complexity and diversity of music. Therefore, how to use machine learning and deep learning methods to automatically learn more representative and discriminative features from raw music data is a problem that needs to be solved.

In summary, the core of music classification is how to accurately classify music works into the correct genres. When solving this problem, it is necessary to overcome challenges such as multimodal data fusion, subjectivity in genre definition and partitioning, and feature extraction. Solving these issues will help improve the accuracy of music recommendation systems, optimize the management of music libraries, and promote the development of music research.

IV. SOLUTIONS

A. DATASET

In order to complete the task of music type classification, we used the GTZAN dataset [5], which has been used as a benchmark for various music type classification studies. The GTZAN dataset provides us with 100 music clips, each lasting 30 seconds with 10 types, as shown in Figure 1.



blues.00000.wav

blues.00001.wav

blues.00002.wav

blues.00003.wav

blues.00004.wav

blues.00005.wav

blues.00006.wav

blues.00007.wav

blues.00008.wav

blues.00009.wav

blues.00010.wav

blues.00011.wav

blues.00012.wav

FIGURE 1. The GTZAN dataset

B. CONVOLUTIONAL NEURAL NETWORK(CNN)

For many years, deep learning has played an indispensable role in many studies, and CNN has become one of the main algorithms in deep learning.

CNN is a feedforward neural network in which each neuron is only connected to the neurons in the previous layer, and the input data from the previous layer is transmitted to this layer for calculation, resulting in output data that is transmitted to the next layer. It has a deep structure. At the same time, CNN has convolutional computing power and representation learning ability.

A fully connected neural network is actually a single switch that connects all inputs and outputs. When using fully connected networks to process large-sized images, problems such as loss of spatial information, low training efficiency, and easy overfitting may occur. However, CNN has optimized the shortcomings of fully connected neural networks and effectively solved these potential problems.

In the field of image recognition, CNN has been widely applied and its performance is better than other general deep neural networks in Figure 2.

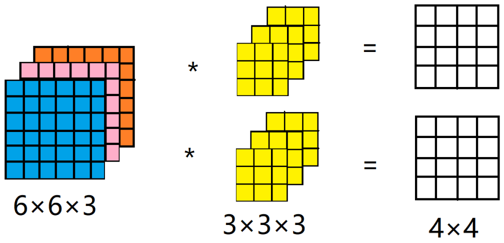


FIGURE 2. Convolution operation

CNN has a feature extractor composed of convolutional and pooling layers, which is also the main difference between CNN and other neural networks. The layers of CNN are mainly divided into three dimensions: width, height, and depth. In CNN, the convolutional kernel mainly calculates the weighted average of a region of input data and a weight function during image processing. This weight function is called the convolutional kernel, also known as a filter, to obtain the output data. Convolutional kernels can obtain a suitable weight during the training process of CNN, which can reduce the interconnection of layers in the CNN and greatly avoid overfitting. Performing subsampling operations in the pooling layer, also known as pooling operations. Pooling operation is actually a relatively special convolution operation. The entire neural network model can be greatly simplified through the operation of convolutional and pooling layers, while also reducing the number of parameters to a certain extent.

CNN mainly includes convolutional layers, linear rectification layers, pooling layers, and fully connected layers. By stacking these network layers layer by layer, a CNN structure model that can work properly can be constructed. In general, the convolutional layer and the linear rectification layer are collectively referred to as the convolutional layer, because the convolutional layer needs to go through an activation function when performing convolution operations.

The convolutional layer is the core layer in the CNN structure, which is composed of several convolutional units. The convolutional layer has the following functions: its parameters are composed of a portion of learnable convolution kernels, and the width and height of each convolution kernel are not significant compared to the input data, but its depth is equal to the depth of the input data. Simply put, CNN activates the convolutional kernel when it captures certain specific features. Convolutional layers can be considered as the output of a neuron, which only calculates a small area of data. Due to the use of the same convolutional kernel, in three-dimensional space, a neuron has the same weight on both sides, which can reduce the number of parameters and thus reduce computational consumption, avoiding overfitting caused by too many parameters. However, most of the computation in the entire CNN network is generated from convolutional layers.

The linear rectification layer, also known as the excitation layer, is used to excite the features extracted by the convolutional layer. As the operation of linearly transforming the input data and convolution kernel is a convolutional operation carried out by the convolutional layer, it is necessary to add a linear rectification layer in which a nonlinear function is used to map the results nonlinearly.

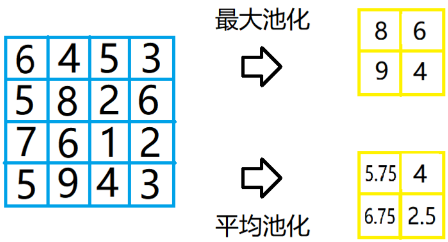


FIGURE 3. Pooling operation

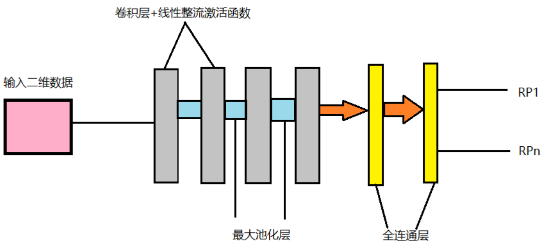


FIGURE 4. Convolutional neural network structure

The pooling operation of the pooling layer, also known as downsampling, is mainly used to reduce the size of feature data. By reducing the parameters of CNN in the pooling layer to reduce the computational complexity of the entire network, overfitting can be avoided to a certain extent.However, the calculation of filters in the pooling layer is different from that in the convolution layer. In the convolution layer, the inner product is calculated, while in the pooling layer, the simpler maximum value calculation or average value calculation is used. The most commonly used pooling layer operation is the maximum pooling operation. This pooling layer is called the maximum pooling layer, while the pooling layer using the average pooling operation is generally called the average pooling layer. The fully connected layer plays a role similar to a classifier in CNN, which is a linear feature mapping process.

In the convolution layer, linear rectification layer and pooling layer, the initial data is mapped into a feature matrix, while in the full connection layer, the obtained features are mapped to the sample tag space, mainly to fit the obtained features again, so as to reduce the loss of feature information in the CNN structure. In fact, the main process of the full connection layer is to multiply matrix vectors and linearly map from one feature space to another. In the CNN structure, the full connection layer is generally set in front of the output layer to feature weight the data obtained in the previous process to get the final result.

In CNN, full connection layer can be converted to convolution layer. For the convolution layer, turn the weight of the convolution core into a large matrix, and most of the data in the matrix is 0. Except for the receptive field area, many areas have the same weight due to the weight sharing, so that a convolution layer can be transformed into a full connected layer. Conversely, a full connection layer can also be transformed into a roll up layer.

C. EQUATIONS

Number equations consecutively with equation numbers in parentheses flush with the right margin, as in (1). To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Use parentheses to avoid ambiguities in denominators. Punctuate equations when they are part of a sentence, as in

$$E = mc^2. \tag{1}$$

The following 2 equations are used to test your LaTeX compiler's math output. Equation (2) is your LaTeX compiler' output. Equation (3) is an image of what (2) should look like. Please make sure that your equation (2) matches (3) in terms of symbols and characters' font style (Ex: italic/roman).

$$\frac{47i + 89jk \times 10rym \pm 2npz}{(6XYZ\pi Ku)Aoq\sum_{i=1}^{r} Q(t)} \int_{0}^{\infty} f(g)\mathrm{d}x \sqrt[3]{\frac{abcdelqh^2}{(svw)\cos^3\theta}}. \tag{2}$$

$$\frac{47i + 89jk \times 10rym \pm 2npz}{(6XYZ\pi Ku)Aoq\sum_{i=1}^{r} Q(t)} \int_{0}^{\infty} f(g)\mathrm{d}x \sqrt[3]{\frac{abcdelqh^2}{(svw)\cos^3\theta}}. \tag{3}$$

Be sure that the symbols in your equation have been defined before the equation appears or immediately following. Italicize symbols (T might refer to temperature, but T is the unit tesla).

D. ALGORITHMS

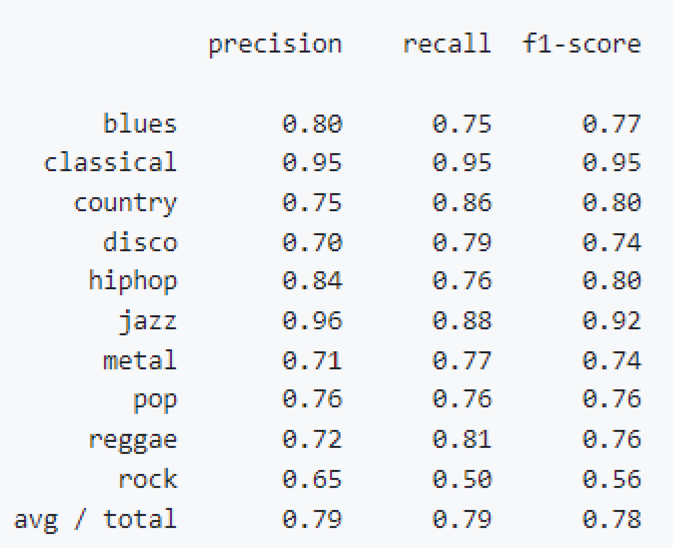
We use convolutional neural networks for implementation. The network receives 599 intermediate frequency bean vectors, each containing 128 frequencies describing its window. The network consists of three hidden layers, and I have done a maximum pooling between them. Finally, there is a fully connected layer, followed by softmax, and finally a 10 dimensional vector representing our 10 type classes.

We use a reshaping layer to change the corresponding size. In two CNN models, we use a Dense layer that uses the softmax activation function to track the CNN layer. We use librosa to process audio data and export MFCC features, and we use keras to implement our models and train them.

V. EVALUATION

Our experiment was completed using the Anaconda package in Python. The TensorFlow package is used for deep learning. The system configuration for implementing this algorithm is Intel Xeon CPU E5-2630 v4, which has 2.20GHz, 10 cores, 20 logic processors, and 32GB RAM.We use the GTZAN dataset. We first divide the dataset into a training set and a testing set, with a testing set size of 20%. Then we divide the training set into a training set and a validation set. The model was built using Keras.

Before training the classification model, we must convert the original data in the audio sample to a more meaningful



```
               precision    recall  f1-score

       blues       0.80      0.75      0.77
   classical       0.95      0.95      0.95
     country       0.75      0.86      0.80
       disco       0.70      0.79      0.74
      hiphop       0.84      0.76      0.80
        jazz       0.96      0.88      0.92
       metal       0.71      0.77      0.74
         pop       0.76      0.76      0.76
      reggae       0.72      0.81      0.76
        rock       0.65      0.50      0.56
 avg / total       0.79      0.79      0.78
```

FIGURE 5. Experimental Results

representation. The audio clip needs to be converted from the. au format to the. wav format to make it compatible with the python waveform module used to read audio files. We can open source SoX modules for conversion.

Then, we need to extract meaningful functions from audio files. In order to classify our audio clips, we will select five features, namely zero crossing rate, spectral centroid, spectral attenuation, Mel frequency cepstrum coefficient and chromaticity frequency. Then attach all the functions to the. csv file so that the classification algorithm can be used.

Once the features are extracted, we can use existing classification algorithms to classify songs into different types. We can use spectral images directly for classification, or we can extract features and use classification models on them.

We use cross validation or partition of training set and test set to evaluate the selected music classification model. The evaluation index can include precision, recall rate, F1-score, etc. We can see the experimental results from Figure 5.

The jazz and rural types provide the best results, with jazz accuracy reaching around 96% and rural accuracy reaching around 95%. The accuracy of other schools ranges from 60% to 90%, and these results have achieved considerable success, especially in the case of metals and rural areas. As for why these genres allow for easier classification, it can be inferred that these genres contain lexical features that differ from other genres. For example, there are countless jokes about how country music commonly mentions beer, trucks, girls, or their combinations. Similarly, jazz genres are renowned for their noble themes, hence their quite unique lyrics. As for other genres, there is no clear expectation of different lyrical trends. Especially for popular music, it contains quite elusive lyrics, which is meaningful because its goal is to attract as many people as possible.

We investigated the impact of feature extraction methods on music classification performance. The current feature extraction methods are mainly based on manually designed rules, such as pitch, rhythm, lineage, etc. By comparing the experimental results using different feature extraction methods, we found that automatic feature extraction methods based on convolutional neural networks have better classification performance compared to traditional manual design methods. This indicates that learning more representative and discriminative features from raw music data can improve the accuracy of music classification.

VI. CONCLUSION

This work demonstrates the potential of a music genre automatic classification system based on convolutional neural networks. According to current research results, this method has the potential to accurately classify a large database of songs into their respective genres.

Future work will focus on developing an emotion based song classification system. This system will help determine which type of music can effectively alleviate personal stress while listening to music. This is of great significance for music therapy as it can play specific music based on an individual's stress level. This work needs further expansion to achieve the improvement and application of such a system. By combining sentiment analysis techniques and music feature extraction methods, we can more accurately identify and classify music related to emotions, and provide individuals with more personalized and effective music therapy experiences. This will be an exciting and promising research direction with positive implications for promoting people's mental health and happiness.

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in American English is without an "e" after the "g." Use the singular heading even if you have many acknowledgments. Avoid expressions such as "One of us (S.B.A.) would like to thank . . . ." Instead, write "F. A. Author thanks . . . ." In most cases, sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page, not here.

REFERENCES

[1] [1] Bisharad, Dipjyoti, and Rabul Hussain Laskar. "Music Genre Recognition Using Residual Neural Networks." In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pp. 2063-2068. IEEE, 2019.

[2] [1] Zhang, Scott, Huaping Gu, and Rongbin Li. "MUSIC GENRE CLASSIFICATION: NEAR-REALTIME VS SEQUENTIAL APPROACH."(2019).

[3] Chillara, Snigdha, A. S. Kavitha, Shwetha A. Neginhal, Shreya Haldia, and K. S. Vidyullatha. "Music Genre Classification using Machine Learning Algorithms: A comparison." (2019).

[4] Bahuleyan, Hareesh. "Music genre classification using machine learning techniques." *arXiv preprint arXiv:1804.01149* (2018).

[5] Yang, Hansi, and Wei-Qiang Zhang. "Music Genre Classification Using Duplicated Convolutional Layers in Neural Networks." In *INTERSPEECH*, pp. 3382-3386. 2019.

[6] Gessle, Gabriel, and Simon Åkesson. "A comparative analysis of CNN and LSTM for music genre classification." (2019).

[7] D. G. Grzegorz Gwardys. Deep image features in music information retrieval. 2014 10.O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A.

[8] Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211– 252, 2015.

[9] Eve Zheng, Melody Moh, Teng-Sheng Moh, Music Genre Classification: A N-gram based Musicological Approach. 7th International Advance Computing Conference, 672-677, 2017.

• • •

FIRST A. AUTHOR received the B.S. and M.S. degrees in aerospace engineering from the University of Virginia, Charlottesville, in 2001 and the Ph.D. degree in mechanical engineering from Drexel University, Philadelphia, PA, in 2008.

From 2001 to 2004, he was a Research Assistant with the Princeton Plasma Physics Laboratory. Since 2009, he has been an Assistant Professor with the Mechanical Engineering Department, Texas A&M University, College Station. He is the author of three books, more than 150 articles, and more than 70 inventions. His research interests include high-pressure and high-density nonthermal plasma discharge processes and applications, microscale plasma discharges, discharges in liquids, spectroscopic diagnostics, plasma propulsion, and innovation plasma applications. He is an Associate Editor of the journal *Earth, Moon, Planets*, and holds two patents.

Dr. Author was a recipient of the International Association of Geomagnetism and Aeronomy Young Scientist Award for Excellence in 2008, and the IEEE Electromagnetic Compatibility Society Best Symposium Paper Award in 2011.