Music Genre Classification Based on Convolutional Neural Network

Xiang Haojie
1023040917
Nanjing University of Posts and Telecommunications
School of Computer Science
Nanjing, China

***Abstract*—The text describes a music genre classification method based on Convolutional Neural Networks (CNNs) to address issues such as misclassification, omissions, and incorrect categorizations in a model built on a single feature. The proposed algorithm utilizes Mel-Frequency Cepstral Coefficients (MFCC) extracted from audio signals to form the feature matrix input for the CNN. The CNN is trained on these features, obtaining an optimal classifier to be used as a training model. The algorithm is then tested through simulations on audio information from four music genres: classical, country, heavy metal, and rock.The experimental results indicate that the average classification accuracy of the CNN-based method reaches 88%, demonstrating a significant improvement in processing speed and a reduction in misclassification and error rates.**

***Index Terms*—music classification, audio features, feature extraction,longword convolutional neural network**

I. Introduction

The rise of mobile internet has led to an explosive growth in information, posing the challenge of efficiently extracting valuable information from this vast pool. Minimizing the cost of searching for relevant information has become a pressing demand for internet users. Providing effective classification services has become a necessary responsibility for major information service providers in various sectors such as e-commerce, film and television, news, etc. Music, being an indispensable part of people's daily lives, requires rational and effective classification. This allows users to quickly and accurately find their desired music styles and provides a guarantee for service providers to effectively push content to users. After undergoing traditional manual classification, music classification has witnessed a significant improvement in efficiency and accuracy with the introduction of new technologies such as machine learning and deep learning.

Li [1] and others conducted a detailed comparative analysis of various factors influencing the performance of automatic music classification. They utilized machine vision technology to enhance the selection of classification features. Zhong Wei [2] addressed the issue of incomplete matching by proposing a hybrid recommendation algorithm combining collaborative filtering and music gene-based recommendation algorithms. Gao Linjie [3] proposed a music classification method based on entropy and support vector machines. This method decomposes music segments into different frequency channels using filters, and calculates information entropy after discrete Fourier transform to obtain a spectrogram. Matiyaho [4] and others introduced a music genre recognition method based on a multi-layer neural network, demonstrating higher recognition rates compared to simple models based on neural networks.

This paper focuses on extracting signal features from audio signals in music, utilizing deep learning tools for efficient automatic selection. The goal is to achieve effective classification through algorithmic processing of music audio signals.

II. Related Works

Music genre classification is a comprehensive system involving numerous factors and holds significant importance as the foundation for searching and recommendations. Traditional music classification often employs labeling methods. For instance, the foreign music platform Pandora utilizes a team of musicians or music engineers to annotate music based on various dimensions such as melody, rhythm, and arrangement, establishing different tags for classification. However, the massive workload involved in this process, coupled with the rapidly evolving nature of the internet, renders its efficiency and results less than satisfactory.

In contrast, the UK-based music platform Last.fm does not rely on expert annotations. Instead, it achieves music annotation through user feedback, reducing the investment and time required by service providers. However, this approach results in a subjective and personalized classification method with arbitrary and diverse genres, leading to inaccuracies in classification.

In recent years, with the introduction of deep learning, the value of automated algorithms for music genre classification has gradually become apparent. As a crucial component in the field of music information retrieval, it is gaining increased attention.

Building a classification model based on music genres involves, firstly, extracting music features to describe essential music information. These features can be categorized into aspects such as energy, time domain, and frequency domain features. Music classification is achieved by combining different features, with Mel-frequency cepstral coefficients (MFCC) gradually gaining popularity. Secondly, besides feature extraction, another crucial aspect is establishing a music classifier. Traditional classifiers include rule-based music classification and pattern matching methods, each having its limitations.

Fig. 1. The general classification of flow

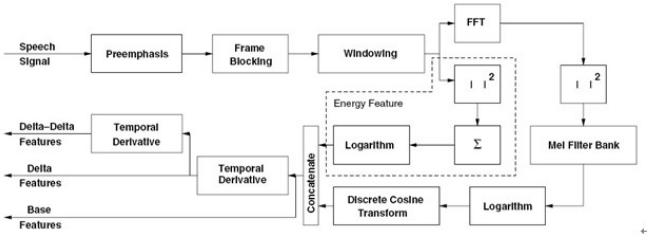

Fig. 2. The basic process of MFCC parameters extraction

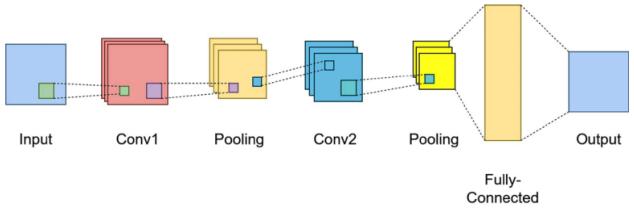

Fig. 3. The basic framework of convolution neural network

Rule-based methods are suitable only for simple audio files, while pattern matching requires establishing standard patterns, resulting in low accuracy and high computational complexity.

III. Problem Statement

New classification methods based on machine learning, such as Hidden Markov Models [5], neural networks [6], and support vector machines [7], have significantly improved efficiency and accuracy. In the realm of neural network-based music classification, Backpropagation (BP) neural networks are widely used. [8] However, compared to BP neural networks, Convolutional Neural Networks (CNNs) [9] reduce the number of required parameters for neural network training through receptive fields and weight sharing, providing advantages in classification algorithms.Therefore, the classification process begins with extracting MFCC as feature vectors from audio files. Subsequently, a CNN serves as the classifier, employing t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm [10] for data dimensionality reduction. Finally, genre determination is accomplished through a voting decision. The classification steps can be implemented using Python libraries such as Librosa and TensorFlow, ensuring a simple and efficient process.

No matter the method, music classification fundamentally involves a pattern recognition process, and thus, it can be applied to the general pattern recognition workflow as depicted in Figure 1 [11]. Initially, data is collected from audio files. Then, based on the characteristics of the collected data, features and models are selected for extraction. Following this, a portion of the data is used to train the classifier. Finally, the classifier's parameters are adjusted and ultimately determined based on test results [12].

The classifier plays a crucial role in determining the quality of classification. In the field of image recognition and speech analysis, Convolutional Neural Networks (CNNs) have shown superiority over traditional deep neural network structures. The unique features of CNN, including partially connected networks, convolution, and pooling, contribute to its excellent performance during the training process.

IV. Algorithms

A. The Extraction of Music Features

MFCC is primarily applied in the field of audio processing, built upon Fourier and cepstral analysis. By sampling the audio and subsequently applying Fourier transformation to the sampled points, it obtains the energy distribution of audio frames in the frequency domain. It represents a logarithmic transformation of the non-linear Mel scale, reflecting the auditory characteristics of the human ear. Using MFC as a feature for music classification enhances the accuracy of the classification [13]. The general extraction process of MFCC from audio to energy is illustrated in Figure 2 and has become well-established [14]. Substantial research and validation have been conducted by Benyamin Matiyaho [15] at Tel Aviv University. MFCC's feature parameters primarily depict the static characteristics of music signals, while the dynamic features can be described through static differences, combining first-order and second-order differences as dynamic features. The synergy between static and dynamic features complements each other, enhancing the system's performance.

B. CNN algorithm

CNN were proposed by Japanese scholar Fukushima [16] in 1984 and have since been widely applied in various fields, achieving breakthrough results in image and speech processing. CNNs are characterized by two main features: local perception and weight sharing. During each convolution operation, local perception is performed, and the obtained response forms a feature map with shared parameters across multiple feature maps. With multiple convolution operations, the receptive field expands, gradually forming global characteristics and achieving higher-level representation.

A typical CNN consists of convolutional layers, pooling layers, and fully connected layers. Convolution and pooling layers are responsible for input and feature extraction, while fully connected layers map the features to dimensional space. The basic structure is illustrated in Figure 3. Based on the two-dimensional numerical convolution calculation, the CNN algorithm can be broken down into the following steps:

1) Convolutional Layer:

The convolutional kernel, acting as a weight matrix, moves according to the designed stride, performing weighted sum-

mation of the data at the corresponding position to generate the output value in the feature map.

$$f_{i,j} = h(\sum_{m=0}^{F_\omega - 1} \sum_{m=0}^{F_H - 1} \omega_{m,n} x_{i+m,j+n} + b) \quad (1)$$

Type: f characterized the ith line in the figure, the first j column value;W is weight matrix;X is the input matrix.B is the convolution of bias;H is the activation function. Can be simplified as:

$$Y = conv(X, W) + B \quad (2)$$

2) Pooling Layer: Type: X, Y, W, respectively for convolution matrix, the input data, and after the convolution kernel weight matrix. The pooling layer is responsible for downsampling the feature map, achieving dimensionality reduction (reducing computational load, lowering complexity, decreasing size, etc.), and implementing invariance, enlarging receptive fields, and so on. Similar to convolution, the pooling process requires setting the size and stride of the pooling region and aggregating numerical values. Common pooling methods include max pooling and average pooling. In this case, max pooling is employed, and the process involves selecting the maximum value within the pooling region.

3) Fully Connected Layer:

After the final layer of convolution or pooling, all feature maps are concatenated into a one-dimensional vector, serving as the global feature. This vector is then mapped, along with other fully connected layers, to a probability vector.

By extracting MFCC features from the data source and utilizing Principal Component Analysis (PCA) for dimensionality reduction, the distribution of music genres in the graph can be observed, as depicted in Figure 5.

The feature extraction of audio information from various music genres is achieved through a feature extraction algorithm based on MFCC. Subsequently, PCA is employed for dimensionality reduction and fusion processing of the feature set. Experimental results, as shown in Figure 5, demonstrate that classical and metal genres can be effectively distinguished.

4) Convolutional Neural Network Backpropagation Training:

The backpropagation training of CNN involves computing the gradients of the loss function with respect to each parameter while simultaneously updating these parameters using the backpropagation algorithm.

The backpropagation process primarily involves calculating the error at each layer during the backward pass, facilitating the correction of each weight.

V. Evaluation

A. The Simulation Results

1) Music Data Source:

This study utilized the authoritative GTZAN dataset as the experimental sample library for music genre classification. The music audio samples chosen for this study encompassed four music genres: classical, country, heavy metal, and rock, with each genre consisting of 100 samples, totaling 400 samples.

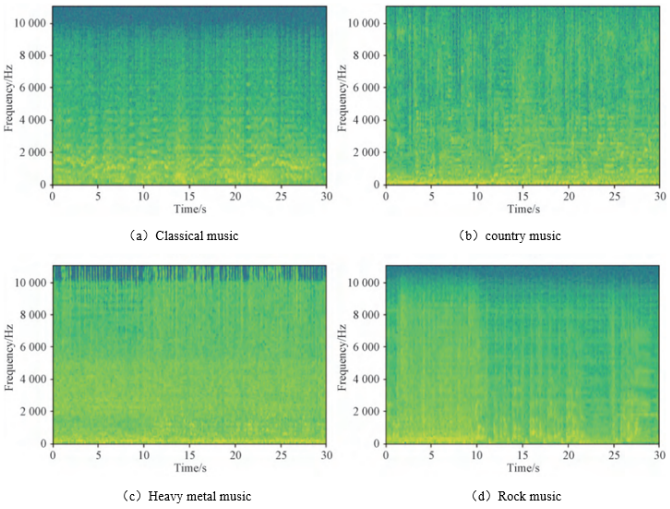

Fig. 4. Four different genres of music audio features

Each sample had a sampling frequency of 22.05 kHz, a 16-bit monaural digital signal, and a duration of 30 seconds.

The dataset was divided into two portions in a 3:1 ratio, with 75% allocated for the training set and 25% for the testing set.

2) Music Characteristic Signal:

At random in the data source to extract various schools of a song, can get different audio features.As shown in figure 4.

MFCC features were extracted from the data source, and by employing the principal component analysis (PCA) method, the dataset was dimensionally reduced. This process revealed the distribution of music genres in the graph depicted in Figure 5.

The feature extraction of audio information from various music genres was accomplished by selecting an MFCC-based feature extraction algorithm. Additionally, the PCA method was applied to perform dimensionality reduction and fusion processing on the feature set. Experimental results, as shown in Figure 5, indicate a clear separation between classical and metal genres.

B. CNN Classification Accuracy

After training the classifier with the four music genres, testing was conducted using the test set. Specifically, 100 songs from each different genre were used to assess the accuracy of the classifier training. The results are presented in Table 1.

TABLE I
THE ACCURACY OF THE MUSIC GENRE

Music Genre	***Accurate Classification***	***Accuracy***
classic	93	93%
country	86	86%
heavy metal	94	94%
rock	82	82%
Total	355	88%

It is evident that the classification performance after feature fusion surpasses that of classification based on individual

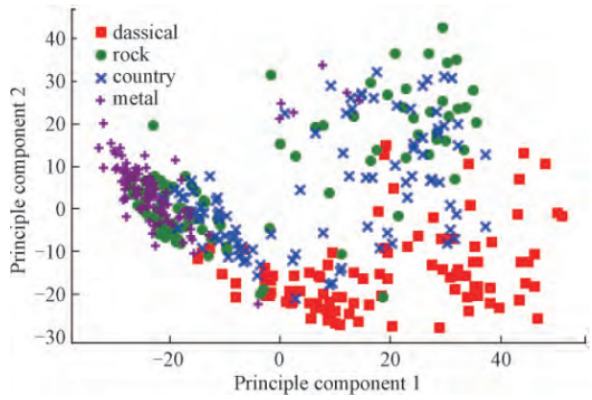

Fig. 5. The distribution of 4 kinds of different music genres MFCC

features. The accuracy of the classification results is relatively high, indicating robust stability. Therefore, in the field of music genre classification, CNNs show great potential in achieving high accuracy.

VI. Summary

The classification of music genres holds significant engineering applications and research significance in the field of multimedia applications. In this study, the cepstral coefficients method was employed for feature extraction, and CNN was utilized for classification. The average classification accuracy reached 88%, with particularly strong performance in the classification of heavy metal and classical music, achieving 97% and 90%, respectively. This indicates that the proposed method is rational, effective, and yields notable results in music genre classification.

References

[1] Tzanev, T., & Lit, A. . “Factors in automatic musical genre classification of audio signals.” *In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* IEEE, 2003

[2] W. Zhong, ”Design and implementation of a hybrid music recommendation system based on music genes,” Ph.D. dissertation, Anhui University, 2014.

[3] Gao, Liangjie, et al. ”Research on music classification method based on entropy and support vector machine,” in Computer Systems and Applications, 2014, vol. 23, no. 5, pp. 83-88.

[4] Matiushkin, Furst M. ”Neural network based model for classification of music type,” *in Eighteenth Convention of Electrical and Electronics Engineers in Israel,* IEEE, 1995.

[5] X. Liang, *Practical Recommender Systems,* Beijing: People’s Posts and Telecommunications Press, 2012.

[6] Y. Zhang, Z. Tang, Y. Li, et al., ”Research on Music Classification Method Based on MFCC and HMM,” *Journal of Nanjing Normal University (Engineering and Technology Edition),* vol. 8, no. 4, pp. 112-114, 2008.

[7] L. Tian, X. Lu, and S. Bai, ”Fast Neural Network Algorithm-Based Non-Specific Speaker Recognition,” *Control and Decision,* vol. 2002, no. 1, pp. 65-68.

[8] H. Han, Y. Wang, B. Wang, et al., ”Research on Pop Music Classification Based on Probabilistic Neural Network,” *Digital Technology and Applications,* vol. 2013, no. 8, pp. 64-65.

[9] X. Wang, G. H. Geng, P. Li, et al., ”An Algorithm for Constructing Neural Network Music Emotion Classifier Based on Correlated Feedback,” *Journal of Northwest University: Natural Science Edition,* vol. 2012, no. 1, pp. 30-35.

[10] P. E. Rauber and A. C. Telea, ”Visualizing Time-Dependent Data Using Dynamic t-SNE,” *Eurographics,* 2016.

[11] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), *Secaucus: Springer-Verlag* New York, Inc., 2006.

[12] Yandong Li, Zongbo Hao, and Hang Lei, ”Research overview of convolutional neural networks,” *Computer Applications,* vol. 36, no. 9, pp. 2508-2515, 2016.

[13] Chitu, A.G., Rothkrantz, L.J.M., Wiggers, P., et al., ”Comparison between different feature extraction techniques for audio-visual speech recognition,” *Journal on Multimodal User Interfaces,* vol. 1, no. 1, pp. 7-20, 2007.

[14] Davis, S.B., ”Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 28, no. 4, pp. 65-74, 1980.

[15] Fukushima, K., ”Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological Cybernetics,* vol. 36, no. 4, pp. 193-202, 1980.

[16] Ambikapathy, S., Singh, A. V., et al., ”Fundamental Concepts of Neural Networks and Deep Learning of Different Techniques to Classify the Handwritten Digits,” *Advances in Systems, Control and Automation,* vol. 442, pp. 287-297, 2018.