Music Classification

Peihan Cai
NJUPT

***Abstract*—This study focuses on audio data feature extraction and music genre classification. In this research, I used audio data sets and processed them with segmentation. By increasing the data size, I aimed to mine the feature information of audio data more comprehensively. I used PyTorch to reconstruct previous work for comparison with my PyTorch-based CNN and LSTM models. The results show that compared with the CNN model and the previous neural network model, the LSTM model shows a more significant effect in the music classification task. This finding provides new insights into the field of music classification and also provides a valuable reference for future research.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

In the contemporary digital society, music, as a cross-cultural language, is not only a carrier of artistic expression, but also an important part of social culture. However, with the explosive growth of music production, traditional manual classification and management methods have been unable to meet the processing needs of large-scale music data. Therefore, the development of automated music classification techniques has become crucial. The challenge of music classification task lies in the complexity and diversity of music itself. The tracks contain rich information, including multiple musical features such as melody, chord and rhythm, and the combination and variation of these features lead to the diversity of music, which makes effective classification more complex. Therefore, finding a method that can comprehensively capture these features has become one of the core issues in music classification research. In this study, we start with the segmentation processing of audio files, and aim to mine the feature information of music data more comprehensively by increasing the data scale and deep learning model. In particular, we focus on comparing the performance of convolutional neural Network (CNN) and Long Short-Term memory network (LSTM) with previous neural networks on music classification tasks, and explore the ability of different models to extract and classify music features. This not only helps to solve the technical challenges in music classification, but also provides new ideas and methods to advance the field of music information processing. By deeply exploring the relationship between music data and deep learning models, we expect to bring new enlightenment to the field of music classification, and provide useful reference for the development and application of music information processing technology in the future. The results of this research will help to build a more intelligent and efficient music management system, and provide more accurate and powerful support for applications such as music recommendation and personalized music experience. Through my research, we aim to elucidate the significance of advanced feature extraction methods and LSTM networks in music genre classification systems, particularly concerning the comparative analysis with CNN approaches. This study provides valuable insights for the development of future models in audio data processing and classification.

II. Related Work

Prior research in the field of music genre classification has witnessed the application of various machine learning methodologies aimed at effectively categorizing audio data into distinct genres. Among these, Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) have emerged as pivotal methodologies, each offering unique strengths and encountering specific challenges in the context of audio content analysis.

A. SVM

The Support Vector Machine (SVM) is a supervised learning algorithm utilized for solving classification and regression problems. In the domain of music genre classification, SVM has been extensively explored and applied. Its core concept involves finding an optimal decision boundary within a feature space to effectively separate different music genres. The functionality of SVM involves constructing a hyperplane in the feature space, allowing the data to be divided into distinct categories while maximizing the margin, known as the separation between classes. The selection of this hyperplane is achieved by maximizing this margin while ensuring good generalization to unknown data. In the context of music genre classification, SVM typically involves converting audio data into feature vectors derived from attributes such as spectral characteristics, tonal aspects, and rhythmic patterns. SVM operates effectively in high-dimensional spaces and performs well with small sample datasets, effectively handling data with complex boundaries. However, in scenarios involving large-scale datasets and high-dimensional feature spaces, SVM's computational complexity might increase, resulting in longer training times. Additionally, in some cases, SVM's performance might be sensitive to parameter selection and tuning, requiring careful optimization to achieve optimal performance. Overall, SVM has been extensively researched in music genre classification and often demonstrates good classification performance. Its strengths lie in its ability to handle high-dimensional feature spaces and small sample datasets. However, challenges may arise in dealing with large-scale and high-dimensional data, necessitating careful considerations and optimizations.

B. CNN

Convolutional Neural Networks (CNNs) are deep learning models primarily recognized for their excellence in processing and analyzing visual data, such as images. However, they have also found substantial applications in the domain of audio analysis, including music genre classification. CNNs are characterized by their ability to automatically learn hierarchical representations from data. They consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers. In the context of audio, CNNs are adapted to capture temporal and frequency features, making them suitable for processing audio spectrograms. In music genre classification, CNNs are often applied to spectrogram representations of audio signals. Spectrograms depict the frequency content of audio signals over time and are transformed into two-dimensional images suitable for CNNs. By convolving and pooling these spectrogram representations, CNNs can effectively capture patterns and features that differentiate various music genres. The advantage of CNNs lies in their ability to automatically extract relevant features from raw input data, reducing the need for manual feature engineering. They excel in capturing local patterns and are robust to translation and distortion, making them suitable for audio data with varying characteristics. However, CNNs may face challenges in capturing long-term dependencies in sequential audio data due to their fixed-size convolutional filters. Additionally, they might require a substantial amount of labeled data for effective training, and the architecture design and hyperparameter tuning can significantly impact their performance. In summary, CNNs have shown promise in music genre classification by automatically learning discriminative features from spectrogram representations. Their capacity to capture local patterns and robustness to variations in audio data make them valuable tools in audio content analysis, although addressing long-term dependencies remains an ongoing area of research.

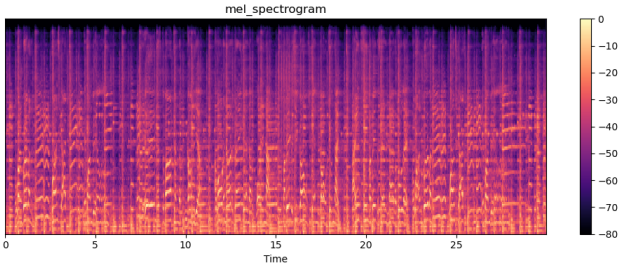
C. RNN

Recurrent Neural Networks (RNNs) are a category of neural networks designed to effectively model sequential data by maintaining an internal memory or state. They have been widely explored in various sequential data applications, including natural language processing, time series analysis, and also in the domain of audio processing, such as music genre classification. RNNs are distinguished by their ability to capture temporal dependencies in sequential data. Unlike feedforward neural networks, RNNs have connections that form directed cycles, allowing them to maintain memory of previous inputs while processing the current input. This inherent memory enables them to model time-based patterns and dependencies in the data. In the context of music genre classification, RNNs are adapted to handle sequential audio data, where the temporal ordering of features plays a crucial role. They are well-suited for analyzing sequences of audio features extracted from music tracks, enabling them to capture long-term dependencies that may exist in music patterns and styles across time. One of the key advantages of RNNs is their ability to handle variable-length sequential data, making them effective in modeling audio sequences of different lengths. They can learn from past inputs to inform predictions about future inputs, which is beneficial in capturing the sequential nature of music data. However, traditional RNN architectures like vanilla RNNs may suffer from the vanishing or exploding gradient problem, limiting their ability to capture long-range dependencies effectively. To address this, more advanced RNN variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) have been introduced, which have better memory retention capabilities and can mitigate these issues to some extent. In summary, RNNs, especially LSTM and GRU variants, have shown promise in modeling sequential audio data for tasks like music genre classification. Their ability to capture temporal dependencies makes them suitable for understanding the sequential nature of music, though addressing issues related to long-range dependencies remains a focus of ongoing research.

III. SOULTION

A. Feature Extract

During the initial phase of feature extraction, Mel spectrograms were considered as the primary feature representation for the audio data. Mel spectrograms offer a detailed insight into the frequency content of audio signals over time, providing a powerful foundation for capturing essential acoustic characteristics. By employing Mel-frequency cepstral coefficients (MFCCs), derived from the Mel spectrogram, distinct spectral features were extracted, enabling the representation of audio signals in a condensed and informative manner. The Mel spectrogram transformation involves partitioning the audio signal into short overlapping frames, calculating the Fast Fmyier Transform (FFT) for each frame, and subsequently mapping the resulting spectrum onto the Mel scale. This process generates a Mel spectrogram, which encapsulates the varying intensity of different frequencies over time, thus encoding critical information about the audio's frequency distribution. The subsequent extraction of MFCCs from the Mel spectrogram involves capturing the cepstral coefficients representing the magnitude of distinct frequency bands. These coefficients convey essential spectral features, such as timbral characteristics, pitch, and harmonic content, contributing significantly to the discriminative power necessary for music genre classification. The details are listed in Figure 1.



However, due to the input being a 30-second file, the resulting feature matrix is of dimensions (128,1292). This

high-dimensional matrix prompted me to reconsider my approach. Considering this situation, I made the following processing for the feature extraction project of the data set. Each audio file in the GTZAN data set is about 30s long. After loading it using Libsora library, the list length is 66000+. The detailed process is as follows. The final size of the matrix is (128,129). The details are shown in Figure 2. The amalgamation of these acoustic features forms the foundation of my feature extraction process, aiming to provide a comprehensive representation of audio content for music genre classification tasks. Through the comprehensive utilization of different features, my objective is to empower music genre classification models with a more accurate and targeted feature set, thereby enhancing the performance and reliability of classification systems.