Exploring CNN and LSTM for Music Genre Classification: A Study Using the GTZAN Dataset

Dongze Li
1023040813
Nanjing University of Posts and Telecommunications
Jiangsu, Nanjing China

***Abstract*—Music genre classification is a crucial task in music information retrieval. With the exponential growth of digital music, there is an increasing demand for efficient and accurate music classification, and the development of neural network technology has made it capable of addressing this task. This study explores the performance of Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) in music genre classification. The research utilizes the well-known GTZAN dataset, which comprises 1000 audio segments spanning 10 different music genres. After data preprocessing, feature extraction and evaluation are conducted. Feature analysis includes both individual feature performance analysis and experiments with feature combinations. In the experiments, the CNN model achieves an average accuracy of 46.60%. On the other hand, the LSTM model, tailored for time series data, excels in capturing dynamic music changes with an average accuracy of 54.40%.**

***Index Terms*—Music Genre Classification, Convolutional Neural Networks, CNN, Long Short-Term Memory, LSTM.**

I. Introduction

TMusic classification, a core task in the field of Music Information Retrieval, has gained increasing attention with the advent of the digital music era. The surge in audio data and the growing commercial and analytical needs for music have emphasized its importance. Music classification not only assists users in discovering their preferred music genres but is also crucial for deep data analysis. However, efficiently and accurately classifying music remains a challenging and valuable problem. Music, as an art form rich in emotion and expression, includes complex elements like melody, harmony, and rhythm, which together form the unique characteristics of a music style. Traditional music classification methods mainly rely on manually extracted features, but these are often limited by subjective judgment and technical constraints.

In recent years, neural network technology has become an important tool for solving complex classification problems in multiple fields, including computer vision, natural language processing, bioinformatics, and music genre classification. Neural networks are advantageous because they can automatically learn and recognize complex patterns and features from data, especially in problems where traditional methods fall short. In music genre classification, neural networks, through end-to-end learning, effectively extract more abstract and advanced features from music data, significantly enhancing the accuracy and robustness of music classification, and providing new solutions to this problem. Applying neural networks in music classification has become a vibrant and attention-garnering research field.

This study aims to compare the performance of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks in music classification. Focusing on a 10-genre classification problem for different music styles, this study first extracts features from music data, then designs and conducts a series of experiments to analyze and evaluate the performance of LSTM and CNN models in music classification tasks. In 10 different sampling experiments, the LSTM model showed an average accuracy rate of xx%, while the CNN model achieved an average accuracy rate of xx%. Through these experiments, this study aims to delve into and compare the advantages and limitations of these two neural networks in music classification.

II. Related Works

Music genre classification, as an important topic in the fields of machine learning and music information retrieval, has attracted the attention of many researchers. Early studies focused on using various machine learning algorithms for the automatic classification of music genres.

A milestone study [1] not only first utilized machine learning algorithms for music genre classification but also established the renowned GTZAN music dataset. This dataset, containing various music genres, has become a standard in music genre classification research. The introduction of the GTZAN dataset greatly promoted the development of music classification research, providing a common benchmark for subsequent studies. Another significant study [2] introduced an automatic music genre classification system based on Convolutional Neural Networks (CNN). This system uses Mel Frequency Cepstral Coefficients (MFCC) to calculate the feature vectors of music. MFCC is a feature widely used in audio signal processing, particularly suitable for extracting the spectral properties of audio signals. This study demonstrated the powerful potential and application value of deep learning technology in music genre classification by introducing CNN.Combined with CNN, [3] compared the performance of traditional machine learning models, such as logistic regression, random forests, gradient boosting, and support vector machines, in music genre classification. Additionally, the research investigated the relative importance of different features.

In recent years, with the rapid development of deep learning technology, more studies have begun to explore the use of deep neural networks, especially CNN and RNN (Recurrent Neural Network), and even more complex structures like Long Short-Term Memory networks (LSTM), to address music classification problems. [4] applied Long Short-Term Memory Networks (LSTM), a deep learning technique, to music genre classification. Experiments indicated that this method showed improvement in accuracy compared to methods such as Boltzmann machines and convolutional neural networks. [5] employed unsupervised learning techniques with an automatic predictive analysis approach. It enhanced the ability to sequentially learn advanced features from data. [6] utilized transfer learning techniques, employing a Multiframe approach for a detailed analysis of music.

These studies have shown that deep learning can process and classify large-scale and complex music datasets more effectively compared to traditional machine learning methods.

This summary outlines some key research and development trends in the field of music genre classification, providing a solid foundation for the background and theoretical basis of this study. These related works not only showcase the technical progress in the field of music classification but also offer valuable references and inspiration for this research.

III. Problem Statement

Music genre classification, as a specialized domain of data mining, presents unique challenges and opportunities. The task revolves around processing and interpreting complex audio data to categorize it into distinct genres. In this study, the dataset comprising 100 audio samples from 10 different music genres, each 30 seconds long, offers a comprehensive overview of diverse musical styles.

These audio samples represent high-dimensional time-series data, encapsulating a rich tapestry of sonic elements from simple beats to complex harmonies. This complexity arises from the multitude of sample points in each audio file, capturing the subtleties and nuances that define each genre. Extracting meaningful features from these data points is crucial for effective classification. The process involves not just identifying the most informative features but also understanding how they interact to create the distinctive sound of each genre.

Moreover, the challenge extends to dealing with the inherent variability within and across music genres. This includes understanding the influence of cultural, historical, and technological factors on music production and perception. The goal is to develop a system that can accurately and reliably categorize these diverse audio samples into their respective genres, considering the multidimensional nature of music.

In summary, this study aims to address these challenges by applying advanced data mining techniques to extract, analyze, and classify complex audio data, thereby contributing to the broader field of music information retrieval and analysis.

A. Audio Feature Extraction

We selected a series of commonly used features in audio analysis, such as MFCC (Mel-frequency cepstral coefficients), Spectral Centroid, Chroma Features, Spectral Contrast, Zero Crossing Rate, Root Mean Square (RMS), and statistical features of the amplitude spectrum. These features capture different aspects of the audio signal, providing rich information for subsequent classification.

MFCC is a widely used audio spectral feature that extracts characteristics related to human auditory perception through dimensionality reduction and redundancy removal. We calculate MFCC using the librosa.feature.mfcc function.

The Spectral Centroid describes the “center of mass” of the audio spectrum, used to determine whether the audio is high-pitched or low-pitched. We compute this using the librosa.feature.spectral_centroid function.

Chroma Features help analyze audio’s tonal and scale relationships, suitable for music emotion analysis and chord recognition. We extract these features using the librosa.feature.chroma_stft function.

Spectral Contrast describes the intensity differences between different frequency bands, helpful for distinguishing different audio features. We calculate this using the librosa.feature.spectral_contrast function.

Zero Crossing Rate and Root Mean Square value describe the signal’s high-frequency content and overall volume, respectively, essential for audio quality assessment.

B. Feature Analysis

After feature extraction, we transform the extracted features into vector form. Dimensionality reduction not only reduces the computational complexity of the model but also helps mitigate the risk of overfitting.

In subsequent feature analysis, we trained neural network models based on specific feature selection modes (positive or negative) and showcased the contribution of different features to the accuracy of classification results under each corresponding mode. In each run, during the positive mode, we retained the selected feature data and removed the rest. In the negative mode, we exclusively excluded the selected feature data. We recorded the contribution assessment results for each feature set under CNN and LSTM networks and computed the average of these results.

C. Neural Network Models

In our study, we employed two different neural network models - Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) - for music genre classification.

The CNN model structure includes multiple convolutional and pooling layers, along with fully connected and Dropout layers, using ReLU as the activation function. Specifically, we applied the softmax activation function in the final layer of the model, combined with L1 regularization and the Adam optimizer for training. This structure allows the CNN model to excel in processing audio spectral features, aptly capturing the subtle changes in music.

Conversely, our constructed LSTM model consists of two LSTM layers and a fully connected layer, paired with dropout and recurrent dropout to prevent overfitting. The LSTM model uses the softmax activation function, combined with L2 regularization and the Adam optimizer for training. The LSTM model is particularly suited for processing the time series characteristics of music data, effectively capturing the dynamic changes in music.

During the experiments, we focused on the classification capabilities and accuracy of both models across various music genres. Through the exploration and application of different feature combinations, both models showed significant effectiveness in the task of music genre classification.

Overall, our research not only demonstrates the effectiveness of CNN and LSTM in music genre classification but also opens up possibilities for the future use of other features and machine learning algorithms in this field, aiming to further improve classification performance and accuracy.

IV. Solution

A. GTZAN Dataset Overview

The GTZAN dataset contains 1000 audio clips, covering 10 different music genres, with 100 clips for each genre, each clip lasting 30 seconds. These genres include rock, pop, jazz, blues, country, hip-hop, metal, classical, electronic, and world/international music. The audio is stored in 16-bit, 44.1kHz .wav format, and the dataset files are organized by music genre, facilitating classification and retrieval.

B. Data Preprocessing Process

In the data preprocessing stage, we first defined key parameters, including the dataset split ratio and seed for the random number generator, to ensure balanced distribution and randomness of the dataset. Then, we processed the audio files in the GTZAN dataset using the librosa library and recorded and cleaned up unreadable files to maintain the integrity and quality of the dataset. We employed a series of data augmentation techniques, such as random sampling and data transformation, to enhance the model's generalization capability. We also removed any unavailable audio files. Additionally, the audio files were divided into training, validation, and test sets, ensuring the effectiveness and reliability of model evaluation.

C. Feature Extraction and Selection

Feature extraction is a key step in the music classification task. We used the Librosa library to extract a range of audio features, including MFCC, Spectral Centroid, Chroma Features, Spectral Contrast, Zero Crossing Rate, Root Mean Square value, and amplitude spectrum. These features describe the audio characteristics from different dimensions, providing rich input information for the classification model. We normalized the extracted features to reduce the impact of dimensional differences between different features on classification results. Furthermore, we explored feature selection methods to identify the subset of features most impactful on classification results, further enhancing model performance.

D. Data Visualization and Analysis

To better understand the dataset characteristics and feature distribution, we used Matplotlib to plot distribution graphs and heatmaps of the features. This not only helped us understand the data's intrinsic structure but also provided intuitive references for feature selection and model optimization.

In summary, our data preprocessing and feature extraction process provided a solid foundation for subsequent music classification models. Through carefully designed data processing strategies and feature engineering, we ensured the quality of the dataset and the effectiveness of the features, laying the groundwork for achieving high accuracy in music genre classification.

E. Analysis of Experimental Results

In conducting experiments on music genre classification, we adopted a multi-step approach to ensure accuracy and reliability. Below is a detailed analysis of the experimental results:

1) **Feature Analysis Experiment**
   a) Single Feature Performance Analysis: We first evaluated each individual feature to determine their contribution to classification accuracy. By training models on each feature separately and recording the average accuracy over three runs, we found certain features like MFCC and Spectral Centroid to be particularly important for differentiating music genres. In contrast, other features like Zero Crossing Rate and Spectral Contrast contributed less to the final classification accuracy.
   b) Feature Combination Experiment: Next, we experimented with combinations of different feature sets to assess their cumulative effect. By sequentially excluding each feature and analyzing its impact on overall classification accuracy, we identified an optimal set of feature combinations that demonstrated higher performance in classification.
   c) Final Selected Features: After a series of experiments, we selected MFCC, Spectral Centroid, Chroma Features, etc., as the primary input features, as they consistently showed importance in classification tasks across multiple experiments.
2) **Neural Network Classification Results**
   a) CNN Model Construction and Evaluation: In our research, we constructed a Convolutional Neural Network (CNN) model using the Keras framework. This model comprises multiple convolutional layers, pooling layers, fully connected layers, and Dropout layers to enhance the model's generalization ability and prevent overfitting. The activation function chosen is ReLU, while the softmax activation function is employed in the output layer to address multi-class classification problems. We selected Adam as the optimizer and incorporated L1 regularization. During

the training process, the model exhibited good learning capability, with its validation accuracy steadily increasing with the training epochs, demonstrating the effectiveness of the CNN model in music genre classification tasks.

b) LSTM Model Construction and Evaluation: For the LSTM model, we also constructed it using the Keras framework. The model consists of two LSTM layers and a fully connected output layer. The design of the LSTM layers aims to capture the time-series features within the music data, which is particularly important for understanding the dynamic changes in music. Similarly, the model utilizes the softmax function as the activation function and employs the Adam optimizer. Furthermore, to mitigate the risk of overfitting, L2 regularization has been introduced into the model. The LSTM model demonstrates outstanding performance in handling complex time-series data, especially excelling in the dynamic recognition of music styles.

c) Model Performance Comparison: After a thorough comparison of the CNN and LSTM models, we found each model had unique advantages in classifying different music types. Using confusion matrices and classification reports, we conducted a comprehensive assessment of each model's performance. The results showed that while the two models had similar performances in some music types, they differed significantly in others.

d) Results Visualization and Saving: For a more intuitive display of the accuracy changes during the model training, we used the Matplotlib library to plot accuracy curves. We also saved the confusion matrix plots for subsequent analysis and reporting.

In conclusion, our experiments involved meticulous analysis of feature selection, as well as the construction and evaluation of two different neural network models. Through these experiments, we gained deeper insights into the impact of different features and models on music genre classification, providing valuable insights and data for future research.

V. Evaluation

A. Data Characteristics

The GTZAN dataset, utilized in our music genre classification experiment, presents a well-rounded and diverse structure, crucial for robust and unbiased analysis. It encompasses a broad spectrum of musical genres, featuring 10 distinct styles including rock, pop, jazz, and others, thus capturing a comprehensive range of musical diversity. This variety is essential in ensuring that the classification model is not skewed towards a particular genre or style.

The dataset is meticulously balanced, with each genre represented by 100 samples, each lasting 30 seconds. This uniformity in sample count and duration across genres is a critical aspect, as it prevents any bias towards a genre with more data points. Such balance is vital for the accuracy and fairness of the classification algorithm.

Moreover, the dataset maintains a high level of consistency in its audio file format, with all files stored as 16-bit, 44.1kHz .wav files. This uniformity in audio quality and format is crucial for maintaining the integrity of the data analysis, ensuring that variations in classification results are due to the inherent differences in musical genres, rather than discrepancies in audio file quality or format.

In summary, the GTZAN dataset's diversity, balance, and consistency make it a robust and reliable resource for experimenting with and evaluating music genre classification models. These characteristics ensure that the dataset provides a comprehensive and unbiased foundation for the development and testing of advanced machine learning algorithms in the field of music information retrieval.

B. Feature Analysis Results

After extracting features, the visualization results for some features are shown in Figure 1. From bottom to top, the visualizations include MFCC, spectral centroid, chroma, spectral contrast, zero-crossing rate, and standardized RMS.

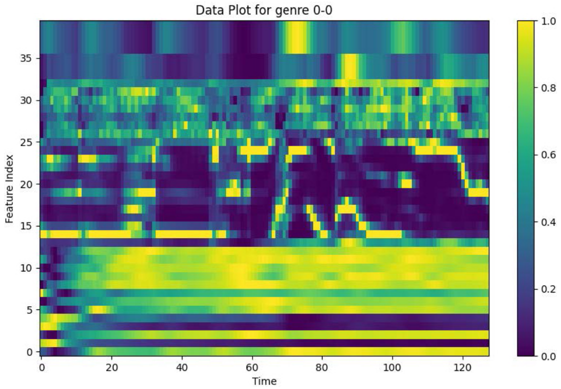

Fig. 1. Visualization of music features.

In the single-feature mode, that is, when extracting individual features separately, the results of 5 experimental runs are shown in Fig 2. As shown in the figure, MFCC, spectral centroid, spectral contrast, and zero-crossing rate demonstrated notable contribution values.

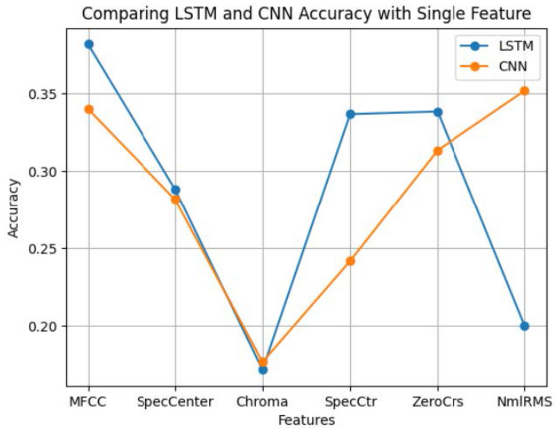

Fig. 2. Comparing LSTM and CNN accuracy with single feature.

Subsequently, we removed one feature during each training session and extracted the average results from 5 runs. The experimental results are shown in Figure 3. From the figure, it can be observed that in this mode, MFCC, chroma, and spectral contrast appear to be indispensable.

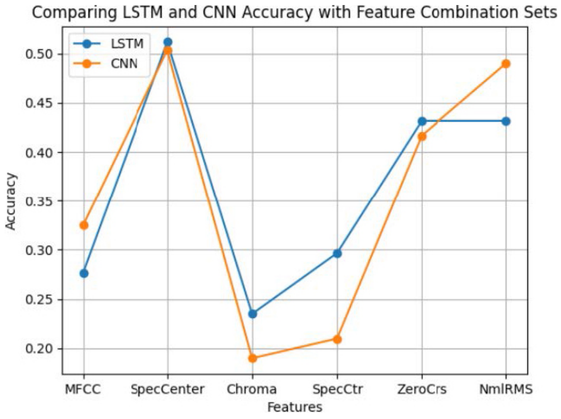

Fig. 3. Comparing LSTM and CNN accuracy with feature combination sets.

From the above experiments, it can be observed that the relatively important features may include MFCC, chroma, spectral contrast, spectral centroid, and zero-crossing rate. Through the analysis of feature contribution values, we can systematically optimize feature selection and preprocessing steps, thereby enhancing the overall model performance.

C. Experimental Results

1) **Accuracy**
In the CNN model, we observed an average accuracy of approximately 46.60%, indicating that CNN performs well in processing audio spectral features.The LSTM model showed higher accuracy in processing the time series features of music, with an average accuracy of 54.40%, particularly suitable for capturing the dynamic changes in music. The accuracy results from 5 training sessions are illustrated in Fig 4.

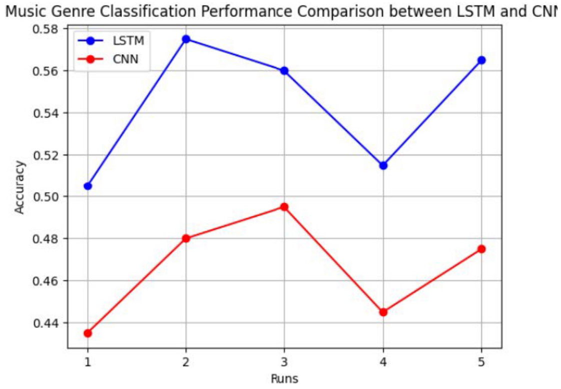

Fig. 4. Music Genre Classification Performance Comparison between LSTM and CNN.

2) **Time Complexity**
The CNN model has a relatively shorter training time, with an average training duration of about 0.242 seconds per epoch.The LSTM model, due to its complex time series processing capability, takes longer, averaging about 4.047 minutes per epoch.

3) **Space Complexity**
Both models have a moderate number of parameters, with the CNN model having about 52.29 thousand parameters, and the LSTM model about 107.46 thousand.The storage requirements of the models depend on the number of parameters and data types, and our models run efficiently on standard hardware.

VI. Conclusion

In this comprehensive study, we have explored the application of two prominent neural network models, CNN and LSTM, in the realm of music genre classification using the GTZAN dataset. The research reveals distinct advantages of each model, highlighting the versatility and complexity inherent in the task of music classification.

The CNN model, with its proficiency in handling spectral features, demonstrates significant efficacy in identifying and classifying genres based on sound texture and quality. This aspect of the CNN model is particularly beneficial in scenarios where the spectral components of music, such as timbre and tone, play a pivotal role in genre determination. On the other hand, the LSTM model's strength lies in its ability to process time-series data, making it exceptionally adept at capturing the temporal dynamics and rhythmic patterns in music.

Furthermore, the study delves into the time and space complexity of these models. The CNN model's training efficiency makes it a practical choice for scenarios with limited computational resources or where rapid model deployment is necessary. The LSTM model, while more resource-intensive, offers unparalleled depth in processing complex sequences, making it ideal for detailed analysis of intricate musical pieces.

Looking ahead, the potential for future research in this area is vast. By combining the strengths of CNN and LSTM models, it is possible to create a more accurate and powerful classification system. Such a hybrid model would leverage the spectral analysis capabilities of CNN and the sequential data processing advantages of LSTM, offering a comprehensive analysis of musical pieces. The exploration of more advanced deep learning techniques, such as attention mechanisms or transformer models, could further enhance classification accuracy. Additionally, integrating other forms of data, like lyrics or user-generated metadata, could provide a more holistic approach to music genre classification. The incorporation of unsupervised learning methods could also uncover deeper insights into music genres, potentially leading to the discovery of sub-genres or new genre classifications.

References

[1] G. Tzanetakis and P. Cook, ”Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293-302, July 2002, doi: 10.1109/TSA.2002.800560.

[2] Vishnupriya and K. Meenakshi, ”Automatic Music Genre Classification using Convolution Neural Network,” *2018 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2018, pp. 1-4, doi: 10.1109/ICCCI.2018.8441340.

[3] I. Pathania and N. Kaur, ”Classification of Music Genre Using Machine Learning,” *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, 2022, pp. 1-5, doi: 10.1109/GCAT55367.2022.9972105.

[4] L. He and B. Yuan, “Classification of music genres using long short term memory networks,” *Computer Technology and Development*, vol. 29, no. 11, p. 5, 2019. 10.3969/j.issn.1673-629X.2019.11.038

[5] K. Manikandan and G. Mathivanan, ”An Intelligent Music Genre Classification Method with Feature Extraction based on Deep Learning Techniques,” *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, Bengaluru, India, 2023, pp. 902-907, doi: 10.1109/IDCIoT56793.2023.10053460.

[6] K. S. Mounika, S. Deyaradevi, K. Swetha and V. Vanitha, ”Music Genre Classification Using Deep Learning,” *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, Coimbatore, India, 2021, pp. 1-7, doi: 10.1109/ICAECA52838.2021.9675685.