Music Classification Using Convolutional Recurrent Neural Network Method

Liang Pengyuan
1024041118
Nanjing University of Posts and Telecommunications
School of Computer Science
Nanjing, China

***Abstract*—This paper introduces a music classification method based on Convolutional Recurrent Neural Networks (CRNN), leveraging the strengths of both Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to enhance the accuracy and robustness of music genre classification. Traditional approaches using Mel Frequency Cepstral Coefficients (MFCC) as features and employing CNN for classifying music data have shown proficiency in capturing spectral characteristics but fall short in handling temporal dependencies within audio sequences, leading to limited classification performance. While Deep Neural Networks (DNN) can effectively classify music data, they struggle with processing long-term dependencies and complex musical structures. The CRNN model integrates Long Short-Term Memory (LSTM) layers, inheriting CNN's advantage in extracting local features while significantly enhancing the ability to capture temporal dependencies in audio sequences. Experimental results demonstrate that under identical settings for training epochs, dataset splits, and other parameters, the CRNN model achieved an accuracy of approximately 72% on the test set, notably higher than the 50% accuracy of CNN-only models and nearly 70% accuracy of DNN models. Furthermore, confusion matrix analysis reveals that the CRNN excels in distinguishing music genres with complex temporal characteristics, such as Classical vs. Jazz and Pop vs. Rock.**

***Keywords*—Classification, CRNN, MFCC, Accuracy, Confusion Matrix.**

I. Introduction

The rise of mobile internet has led to an explosion of information, increasing the demand for efficient tools to extract useful data and reduce the cost of finding relevant content. Robust classification services are now essential for e-commerce, media, and news providers.Music, as a vital part of daily life, benefits greatly from effective classification, allowing users to quickly find their preferred genres and enabling service providers to offer more accurate recommendations. With the advent of machine learning and deep learning, music classification has become more efficient and accurate than traditional manual methods.This article will employ the CRNN (Convolutional Recurrent Neural Network) method for music classification, while also comparing it with conventional CNN (Convolutional Neural Network) and DNN (Deep Neural Network) approaches.

II. Related Works

In the field of music classification, researchers have proposed various methods for feature extraction and classification. The following provides a comprehensive overview of these methods:

A. Audio Feature Extraction

Mel-frequency cepstral coefficients (MFCC) features emulate the auditory characteristics of the human ear, transforming audio signals into a set of coefficients that carry perceptual information relevant to music. These coefficients typically encompass spectral features of the audio, such as pitch, timbre, and volume. Since musical genres often possess distinct spectral characteristics, MFCC features can capture these differences, thereby providing useful information for classification tasks.And fig 1 illustrates the temporal representation of Mel-frequency cepstral coefficients of a blue genre music.

One advantage of MFCC is its ability to reduce the dimensionality of feature vectors while retaining key information, leading to faster processing speeds and a lower risk of overfitting. However, MFCC also has limitations, such as the potential loss of certain detailed information, which might impact model performance. Therefore, it is important to consider these factors in specific applications. While MFCC can provide valuable classification insights in music genre classification tasks, its effectiveness may be influenced by the subjectivity of the task and other musical features.

In summary, MFCC offers a robust method for extracting meaningful features from audio data, but its application should be carefully evaluated considering both its strengths and limitations.

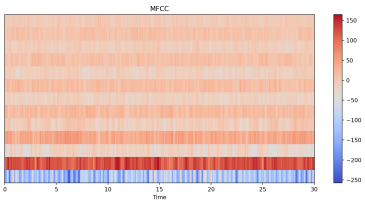

Fig. 1. MFCC of a blue genre music

Zero Crossing Rate (ZCR) measures the rate at which an audio signal changes sign—from positive to negative or vice versa. This metric serves as a fundamental indicator of transient behavior in audio signals and is especially valuable for analyzing rhythmic and percussive characteristics in music. Higher ZCR values typically indicate more pronounced rhythmic elements, suggesting a greater presence of fast transients and percussive sounds.

In music classification, ZCR can help distinguish between genres with varying levels of rhythmic intensity. For instance, genres like rock or hip-hop tend to have higher ZCRs due to their strong percussion components, whereas classical music might exhibit lower ZCRs because of its smoother, more sustained tones. Therefore, ZCR provides a simple yet effective way to capture and quantify rhythmic patterns, aiding classifiers in differentiating between various musical styles.

Chroma features represent the distribution of pitch classes within a segment of music, providing valuable insights into harmonic content and tonal characteristics. These features are derived from the audio signal's spectrogram and map each frequency bin to one of the twelve pitch classes, creating a chromagram that reflects the relative prominence of each pitch class over time.

The computation of chroma features involves several steps:

- Short-Time Fourier Transform (STFT): The audio signal is transformed into the frequency domain using STFT.
- Pitch Class Mapping: Frequencies are mapped to their corresponding pitch classes based on the chromatic scale.
- Aggregation: The energy in each pitch class is aggregated over time to form a chromagram.

Chroma features are particularly useful for identifying chords, estimating the key, and performing content-based music retrieval. By capturing the harmonic structure of music, they enable classifiers to recognize common chord progressions and tonal patterns, thereby enhancing the accuracy of music genre classification. Additionally, chroma features are invariant to changes in octave, allowing for a more generalized representation of the music's harmonic content.

B. Traditional Classification Methods

Early studies relied on manually annotated datasets and domain expertise for music classification. However, this approach was inefficient and highly subjective. With the advancement of machine learning, automated methods gradually replaced manual classification techniques.

Support Vector Machines (SVM) and other shallow models such as K-Nearest Neighbors (KNN) and Random Forests have been used in music classification tasks. While these methods are effective in certain cases, they have limitations in handling complex data, especially with the increasing dimensionality of features and the complexity of music data.

C. Deep Learning Approaches

In recent years, deep learning techniques have significantly enhanced the performance of music classification. Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs) have demonstrated superior performance in music classification tasks.

Convolutional Neural Networks (CNN): Initial attempts to use CNN with only MFCC as input features did not yield the expected results. However, when combined with additional feature types such as chroma features, rhythm features, and more, CNN showed promising potential. The integration of these diverse features allowed CNN to better capture the complexities of musical content.

Deep Neural Networks (DNN): Compared to CNNs, DNNs exhibited better performance in music classification, successfully extracting relevant features from audio data. Experiments showed that DNN models achieved accuracy rates of up to 60 percent on test sets, outperforming earlier attempts with CNNs. This highlights the potential of DNNs in achieving higher accuracy in music classification tasks.

D. Toolkits and Libraries

Librosa: As a powerful Python library, Librosa offers a wide range of functionalities for audio and music analysis, including loading audio files in various formats, extracting MFCCs, chroma features, and rhythm patterns, among other tasks. Its comprehensive feature set greatly simplifies the feature engineering process in music classification tasks.

III. CONVOLUTIONAL RECURRENT NEURAL NETWORK

Convolutional Recurrent Neural Network (CRNN) is a hybrid architecture that combines the strengths of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory networks (LSTMs),shown as fig 2. CRNNs are particularly effective for tasks that require both spatial feature extraction from input data and temporal sequence modeling, such as video analysis, handwriting recognition, and time-series prediction.

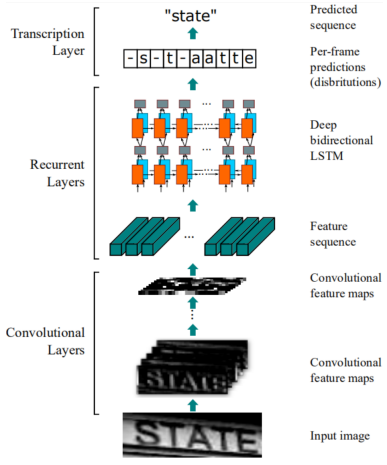

Fig. 2. Structure of CRNN

A. Convolutional Layers

Convolutional Layers are a fundamental component of CRNNs and play a crucial role in automatically learning hierarchical features from the input data. These layers apply convolution operations where filters (also known as kernels) are systematically slid over the input data. At each position, the filter performs element-wise multiplication with the input data, and the results are summed to produce feature maps. In the context of CRNNs, these layers are instrumental in capturing local patterns and relationships within the data, especially when dealing with sequential data like audio signals or time-series data.

For example, in music classification, convolutional layers can detect notes, melodic fragments, rhythmic patterns, and harmonic progressions, which form the basis for subsequent temporal analysis. The use of convolutional layers allows the network to effectively leverage shared weights within the filters, enabling it to learn and generalize features across different regions of the input data. This contributes to the network's robustness and efficiency in feature extraction, making it highly adaptable to various types of input.

B. Temporal Modeling with LSTM Layers

After extracting spatial features using convolutional layers, LSTM Layers come into play to model the temporal dependencies within the extracted features. Unlike traditional CNNs that treat each input independently, LSTMs allow CRNNs to maintain information across time steps through their memory cells and gating mechanisms, which can selectively remember or forget information. This capability is essential for understanding the context and dynamics over time, which is critical for tasks like speech recognition, music classification, and video analysis.

In music classification, LSTMs help capture the evolution of musical elements over time, such as melody development, harmonic changes, and emotional expression. By processing the sequence of feature maps generated by the convolutional layers, LSTMs enable CRNNs to understand the context and dynamics of the music, leading to more accurate and nuanced classifications.

C. Flattening

Following the convolutional and LSTM layers, the multi-dimensional output must be flattened into a one-dimensional vector to prepare it for input into the fully connected layers. Flattening discards the spatial structure but retains all learned features in a sequential format, ready for further processing. For music classification, this step ensures that the temporal and spatial features extracted by the previous layers are integrated into a single vector, facilitating the next stages of classification.

D. Fully Connected Layers

Fully Connected Layers, also known as dense layers, follow the flattening process. In these layers, every neuron is connected to all neurons in the previous layer, allowing the network to integrate and refine the features extracted by the convolutional and recurrent components. These layers are responsible for learning complex patterns and making predictions based on the processed input. For classification tasks, the final fully connected layer typically uses an activation function like softmax to produce probability scores for each class.

In the case of music classification, fully connected layers help the network make sense of the extracted information and produce meaningful predictions regarding the genre, emotion, or other characteristics of the music. By leveraging the interconnected nature of the neurons in these layers, CRNNs can effectively capture complex patterns and relationships, contributing to accurate classification and recognition.

E. Dropout

The Dropout layer is a regularization technique frequently used to prevent overfitting. It achieves this by randomly dropping (setting to zero) a fraction of input units during each training iteration. During training, the Dropout layer randomly sets a proportion of input units to zero at a given probability (commonly between 0.2 and 0.5). This helps reduce co-adaptation between neurons and encourages each neuron to learn more robust and independent features.

By randomly setting outputs from neurons to zero, the Dropout layer simulates multiple different sub-networks during training. Each of these sub-networks can learn different combinations of features, leading to a model that learns more robust and generalized feature representations.

Dropout also has the effect of implicitly performing model averaging across a large number of neural network architectures, which aids in preventing overfitting. In summary, the Dropout layer enhances the generalization ability of the model by ensuring that the network does not rely too heavily on any single neuron, thus promoting a more distributed representation of learned features. This technique contributes significantly to building models that perform well on unseen data.

This approach forces the model to learn more robust and generalized feature representations during training, as it cannot depend on any specific neuron being present. The result is a model that performs better when encountering new, previously unseen data.

IV. MUSIC CLASSIFICATION

A. Dataset

The dataset comprises multiple music categories, with each category containing up to 100 audio files. Specific categories include blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock—representing diverse musical styles. This diversified dataset helps ensure that the model can learn the unique characteristics of different music genres.

B. Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCC): To capture the spectral characteristics of audio, MFCC is used as the primary feature. MFCC mimics the sensitivity of the human

auditory system to different frequency components, transforming audio signals into a set of coefficients that carry perceptual information relevant to music. These coefficients typically include pitch, timbre, and volume. Feature Standardization: All extracted MFCC features are adjusted to a fixed length of 1000 to ensure consistency in input dimensions. Additionally, the StandardScaler is applied for standardization, ensuring that feature values vary within similar ranges, thereby enhancing the effectiveness of model training. Data Splitting The dataset is divided into training and testing sets at a ratio of 0.8:0.2. This split allows the model to be trained on a substantial portion of the data while reserving a separate set for evaluation. By validating the model on unseen data, this approach improves its generalization capability.

C. Model Architecture

In the music classification task, the model utilizes a Convolutional Recurrent Neural Network (CRNN) structure, combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. This architecture is effective in extracting both spatial and temporal features from music data, allowing for efficient classification. Below is the detailed construction process of the model:

The initial layers of the model consist of multiple convolutional layers, whose purpose is to extract spatial information from the music features using filters. The convolutional layers are capable of recognizing local patterns in the audio signal (such as spectral features), which are critical for music data classification.

- Conv2D Layer 1: The input data first passes through a 64-filter, 3x3 convolutional layer, with the ReLU (Rectified Linear Unit) activation function, allowing the model to learn non-linear features. The padding is set to 'same', which means the input and output sizes are the same, preventing information loss.
- BatchNormalization: This step normalizes the data in each batch, accelerating training and improving stability.
- MaxPooling2D: The pooling layer reduces the size of each feature map by half, lowering the data dimensions and reducing the computational load.

This structure is repeated twice, using 128 and 256 convolutional filters, respectively. Each layer undergoes BatchNormalization and MaxPooling to help extract more complex features.

After the convolutional layers process the spatial features of the input data, the subsequent layers are Long Short-Term Memory (LSTM) networks. LSTM layers are primarily used to capture temporal dependencies in sequential data. Since audio signals typically exhibit temporal characteristics (i.e., spectral features change over time), LSTM layers are well-suited to handle these temporal relationships.

- LSTM Layer 1: The first LSTM layer contains 256 units and is set to `return_sequences=True`, meaning it returns outputs for all time steps, not just the final output. This is necessary for the subsequent LSTM layer, which relies on outputs from each time step.
- LSTM Layer 2: The second LSTM layer contains 128 units and is set to `return_sequences=False`, meaning it only outputs the result from the final time step, which typically represents the final state of the sequence.

After the LSTM layers, the model includes a fully connected layer that maps the output from the LSTM layers into a 64-unit space, using the ReLU activation function for non-linear transformation. Additionally, a Dropout layer with a rate of 0.5 is added to reduce overfitting and improve the generalization ability of the model.

The final layer is a fully connected output layer with a softmax activation function. The size of this layer is 10, corresponding to the 10 music classes. The softmax function converts the predictions for each class into a probability distribution, and the final predicted class is the one with the highest probability.

V. EXPERIMENTAL RESULTS

1) Accuracy: We trained models using both CNN and DNN methods on the same dataset, with important parameters such as the number of epochs, the division of the training and test sets, and other settings being consistent with those used in the CRNN model. The CRNN model demonstrated significant performance improvements in classifying different music genres. After training for 70 epochs, the model achieved an accuracy of approximately 72% on the test set (as shown in Fig 3). Although there was some instability, this performance was notably higher than the roughly 50% accuracy achieved using a CNN model in previous studies (as shown in Fig 4), and the approximately 65% accuracy achieved using a DNN model (as shown in Fig 5). This improvement highlights the effectiveness of adding the LSTM layer to capture temporal dependencies in the audio sequence. Training Accuracy: Throughout the training process, the accuracy on the training set steadily increased, reaching nearly 98% by the end of the 70th epoch. Validation Accuracy: The accuracy on the validation set followed a similar trend, but stabilized after approximately the 50th epoch, indicating that the model was able to learn effectively without overfitting the training data.

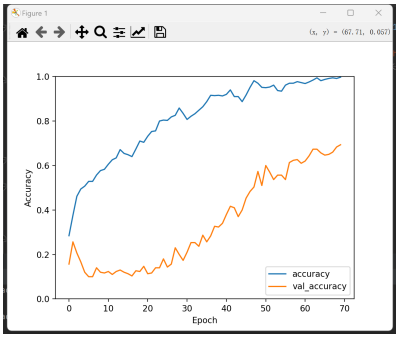

Fig. 3. Accuracy of CRNN

2) Confusion Matrix: The final confusion matrix for the CRNN model (Fig 6) was plotted to evaluate the model's performance across different categories. The confusion matrix

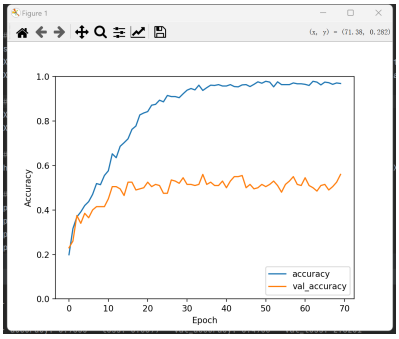

Fig. 4. Accuracy of CNN

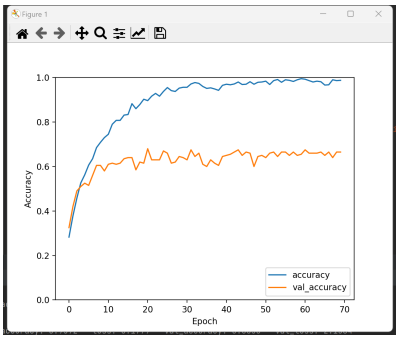

Fig. 5. Accuracy of DNN

shows that the model performed well in most categories, but some misclassifications occurred between genres with similar acoustic characteristics, such as ”rock” and ”metal.” However, the model was able to accurately distinguish between genres like ”classical” and ”hip-hop,” which have significantly different audio features. The analysis of the confusion matrix indicates that the model overall performed well and was effective in handling the classification task across various music genres. Additionally, we also plotted the confusion

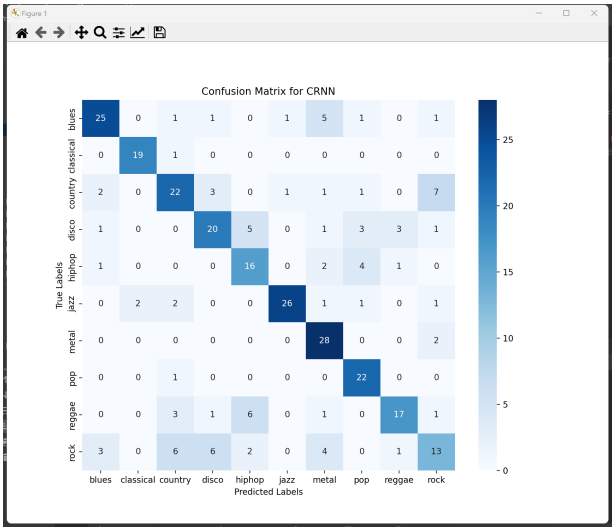

Fig. 6. Confusion Matrix for CRNN

matrices for the CNN (Fig 7) and DNN (Fig 8) models. Both models exhibited similar performance to the CRNN, with misclassifications occurring between genres with similar acoustic features. However, the overall classification performance of the CNN model was better than that of the DNN, although it still fell short compared to the CRNN model.

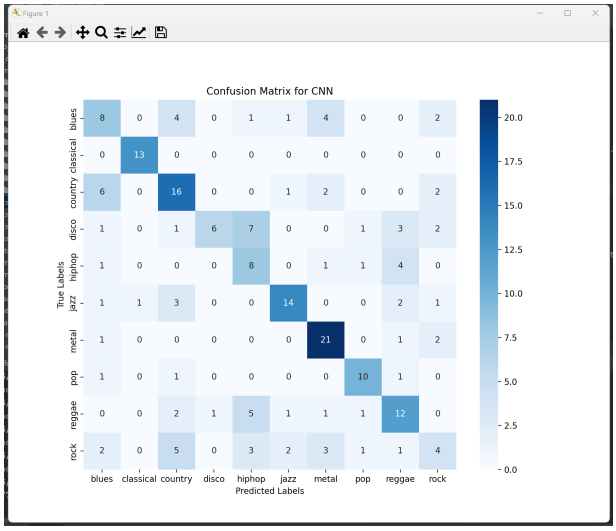

Fig. 7. Confusion Matrix for CNN

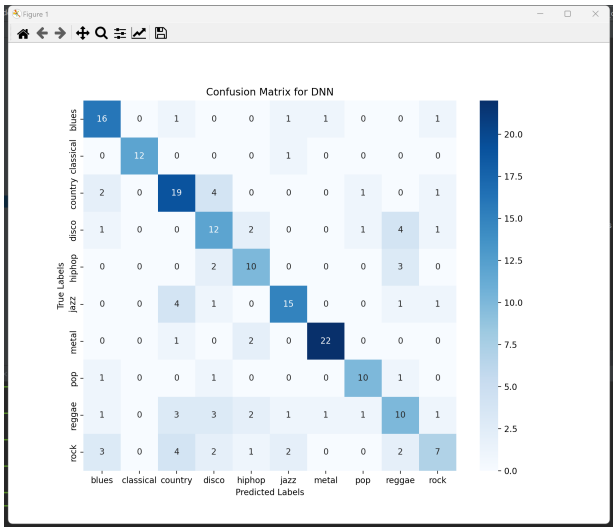

Fig. 8. Confusion Matrix for DNN

VI. SUMMARY

This study underscores the significant potential and practical value of machine learning in music classification. While current research faces limitations—such as the need to enhance classification accuracy across diverse music styles and bolster model generalization—the CRNN model demonstrates superior performance.

By integrating the strengths of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), particularly through Long Short-Term Memory (LSTM) layers,

the CRNN adeptly captures both spatial and temporal audio features. This dual capability allows the CRNN to excel in handling complex musical structures and time-dependent patterns, significantly outperforming traditional CNN and DNN models. The results highlight the CRNN's proficiency in distinguishing nuanced genres such as Classical vs. Jazz and Pop vs. Rock.

Future work will focus on refining the CRNN architecture and incorporating advanced audio features to further enhance classification accuracy and model robustness.

REFERENCES

[1] Cheng, Y. H., Chang, P. C., Nguyen, D. M., & Kuo, C. N. (2020). Automatic Music Genre Classification Based on CRNN. *Engineering Letters*, *29*(1).

[2] J.Choi, K., Fazekas, G., Sandler, M., Cho, K. (2017, March). Convolutional recurrent neural networks for music classification. In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 2392-2396). IEEE.

[3] Li, J., Han, L., Li, X., Zhu, J., Yuan, B., Gou, Z. (2022). An evaluation of deep neural network models for music classification using spectrograms. Multimedia Tools and Applications, 1-27.