Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/96/20241394

Research on the Construction of Music Recognition Model Based on Improved CNN Convolutional Neural Network

Jing He

The University of Queensland, Queensland, Australia

jing.he3@uqconnect.edu.au

Abstract. In the context of today's informationized society, music recognition technology within the field of intelligent audio processing has received extensive attention and become a research focus. However, traditional music recognition techniques are difficult to adapt to the complex changes of music signals due to the inherent defects in feature extraction and model construction. To address these difficulties, this study constructs an innovative music recognition algorithm, which is based on an optimized convolutional neural network (CNN). The efficacy of feature extraction is enhanced by fusing the hash convolutional neural network (Hash-CNN), and the temporal data is processed using the long short-term memory network (LSTM) to improve the accuracy of recognition. In the preprocessing stage of music signals, we applied various noise reduction and normalization means to enhance the data quality to a high standard. According to the experimental data, the model performs well in recognizing complex music segments, especially when analyzing multi-level and multi-dimensional music features, it shows strong robustness. Meanwhile, compared with traditional techniques, the model proposed in this study shows significant improvement in both recognition efficiency and accuracy, confirming its great potential and broad prospect in practical applications.

Keywords: Improved CNN, convolutional neural network, music recognition model.

1. Introduction

Music recognition technology has received a significant boost with the growing demand for intelligent audio processing. Continuous technological innovation has led to the maturity of this field. However, conventional techniques are often limited in processing audio features and time series data, especially when dealing with complex music signals. Convolutional neural networks have demonstrated excellent performance in the image domain, which has also attracted the interest of audio processing researchers. Its hierarchical structure and ability to automatically extract features have opened up new paths for music recognition, but traditional CNN architectures still have limitations in capturing the time-series features of music. Therefore, how to combine CNNs with other deep learning techniques in order to build more robust and efficient music recognition models has become a major topic. This paper proposes an improved version of a new model based on CNN and LSTM, and incorporates hashed convolutional neural networks to improve accuracy and efficiency. Experimental analysis shows that the innovative model shows excellent performance in coping with complex musical phrases, and points to a new direction for future development in this field[1].

2. Recognition methods for basic neural networks

2.1. Neural Networks

Neural network, as an information processing model of bionic nervous system (see Fig. 1), has been widely used in various pattern recognition tasks. In music recognition, it is able to extract effective features from complex music signals by virtue of its powerful nonlinear mapping capability. However, traditional neural network structures often face the problem of insufficient feature expressiveness when dealing with these signals. Due to the highly time-dependent and multilevel structure of music signals, it is difficult for traditional planar-structured neural networks to capture the important information in them. Therefore, in this context, convolutional neural networks (CNNs) have gradually become the focus of research[2]. Through the unique convolutional operation, CNN can extract local features from the input signal while maintaining spatial invariance, which significantly enhances the feature extraction effect on music signals.

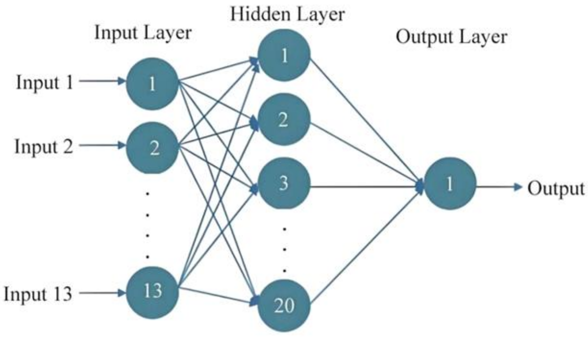

Figure 1. Neural network

Conventional convolutional neural networks have limitations in grasping time-series features, especially in processing long-delayed music clips. The researchers propose to combine convolutional neural networks with other structures, such as long short-term memory (LSTM) units, to better process time-series information of music signals. This combination enhances model robustness and also effectively addresses recognition challenges due to signal complexity. Therefore, incorporating improved versions of neural networks in music recognition is a new path to optimize traditional methods and explore high-precision and high-efficiency recognition approaches[3].

2.2. Hashish convolutional neural network

Among the basic neural network recognition methods, Hash Convolutional Neural Network (Hash-CNN) provides an efficient and innovative solution for feature extraction and classification in music signals by fusing classical convolutional neural networks with advanced hashing techniques. The framework of this unique approach is composed of a multi-layered, multi-stage complex neural network, which mainly consists of a pre-trained Fully Convolutional Network (FCN-5), a Long Short-Term Memory (LSTM) model, a specially-designed hashing layer, and ultimately a softmax classifier used in the decision-making process of the classification, as shown in Figure 2.

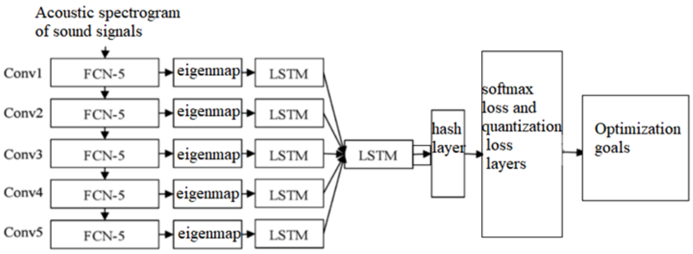
Acoustic spectrogram of sound signals
Conv1
Conv2
Conv3
Conv4
Conv5
FCN-5
eigenmap
LSTM
LSTM
hash layer
softmax loss and quantization loss layers
Optimization goals

Figure 2. Music recognition idea

Specifically, in this approach, the input music signal is first passed through a fully convolutional network (FCN) that has been pre-trained to perform complex and in-depth convolutional operations to extract significant feature maps that can represent the characteristics of this signal. These extracted feature maps are then further transformed into feature sequences through a series of operations of bilinear interpolation processing and similarity selection strategies, and corresponding hash codes are generated. This hash code generation not only reduces the computational complexity of the model to a large extent, but also preserves the key information in a lower dimensional space, thus significantly improving the processing efficiency[4]. Next, these generated feature sequences are fed into the Long Short-Term Memory (LSTM) layer to fully utilize the natural advantages of LSTM for time-series data processing, capturing the complex temporal relationships and diverse characteristics of music signals. By accurately modeling the temporal dependence of signals, LSTM can identify key patterns in music signals, enabling the model to maintain a high degree of accuracy even in the face of complex and varied music data. Afterwards, these feature sequences processed by the LSTM layer will enter the hash layer, further compressing the data through efficient and precise hashing techniques, and finally completing the final classification process of music signal categories in the softmax classifier. Compared with the traditional sense of the convolutional neural network, the use of this improved version of the convolutional neural network including hash technology, in the disposal of large-scale, huge music data information shows more efficient and accurate[5]. With the introduction of advanced and powerful hash technology, the model not only optimizes the utilization of computational resources, but also shows stronger robustness in the actual signal classification, effectively dealing with various types of noise interference problems. The application of this emerging technology in the field of music signal recognition provides an innovative idea, which not only improves the recognition accuracy and efficiency, but also promotes the development and wide application of the whole music recognition technology.

3. Construction of music recognition based on hash convolutional neural network

The music recognition model based on hash convolutional neural networks demonstrates significant advantages, particularly in handling complex and large-scale music data. The introduction of hash technology greatly optimizes convolution operations, enhancing the efficiency of computational resource usage while ensuring the accuracy of feature extraction. This combined model excels not only in traditional music classification tasks but also shows exceptional adaptability when dealing with diverse and intricate music data. By incorporating LSTM networks, the model effectively captures long-term dependencies in music signals, making rhythm, melody, and harmony recognition more precise. Notably, the model exhibits robust generalization capabilities across various music types, such as classical, pop, and electronic music, adapting to different audio features and style variations. Compared to traditional convolutional and recurrent neural network methods, this improved model not only enhances recognition accuracy but also significantly reduces computation time and memory consumption. Its practicality and cost-effectiveness are particularly evident in real-world applications, such as online music streaming services and music recommendation systems, driving further development and innovation in music recognition technology. For the music industry, this technological advancement signifies a more intelligent and personalized user experience and provides strong technical support for music creation and analysis.

3.1. Music signal preprocessing

In the construction process of music recognition based on hashed convolutional neural networks, the preprocessing stage appears to be particularly important, which aims at transforming the raw audio data into a format and features suitable for model processing. In this crucial stage, we need to select a suitable audio input, such as an MP3 audio file with a length of 29.14 seconds. Next, the audio signal needs to be transformed in the frequency domain using the fast Fourier transform (FFT) .This is an efficient and useful algorithm[6]. This is an efficient and useful algorithm to extract the various frequency components of the audio by converting the time-domain signal into a frequency-domain signal, thus laying the foundation for feature extraction and analysis in subsequent steps. Table 1 details the specific parameters, which include sampling rate, window size, overlap rate, etc., which directly affect the quality of the preprocessing results and the overall recognition system performance.

Table 1. Pre-processing parameters

Duration	29.12s
FFT frame size	512
Hop Size	256
Number of Mcl filters	96
Size of Md-Time matrix	1×96×1366
Extended boundary number	74
Expanded Mel-Time	1×96×1400

In practice, through the Fast Fourier Transform (FFT), a spectral analysis tool, the audio signal will be decomposed into a number of different frequency components, each independent component corresponds to a specific frequency and its amplitude in the original signal. This complex but effective process is able to accurately capture a wide range of spectral features hidden within the audio signal, making a large amount of audio data that would otherwise be quite complex and difficult to understand and analyze more structured, clear and easy to analyze and process[7]. These detailed and colorful frequency domain data after conversion can be used as important input features for the subsequent convolutional neural network for further deep learning processing, thus helping the powerful neural network to more accurately identify and classify different types of music signals. When pre-processing audio signals, other advanced and efficient techniques that are widely used in the professional field, such as Mel-Frequency Cepstral Coefficients (MFCC), can also be considered. These additional methods can significantly enhance the feature extraction effect and make the model more stable in the face of a variety of changing and complex environments. Through these meticulous, multi-layered pre-processing steps, redundant information and clutter in the original waveform will be effectively filtered out, while the key information retained will greatly improve the performance of the hash convolutional neural network. This is not only a simple data preparation work, but also a crucial part in the whole music recognition task based on hash convolutional neural network, which plays an indispensable role in improving the overall model performance and prediction accuracy[8].

3.2. Convolutional feature based feature extraction

In the construction process of music recognition based on hash convolutional neural network, feature extraction is one of the key steps, which is mainly through the extraction of convolutional features to realize the accurate recognition of music signals. In order to obtain the best performance, the layer activation method is used for feature extraction of music signals. Specifically, this process extracts the activation maps of multiple convolutional layers through a pre-trained fully convolutional network (FCN-5) to form a sequence of feature maps.

The FCN-5 network structure, shown in Figure 3, consists of multiple convolutional layers, each of which is capable of capturing different features of the input signal. After the music signal is input into the network, low-level features such as spectral features are first extracted through the initial convolutional layers. As the network layers deepen, the features extracted by the convolutional layers gradually become more abstract and complex, and are able to capture high-level features in the music signal, such as rhythm, pitch and melody[9].

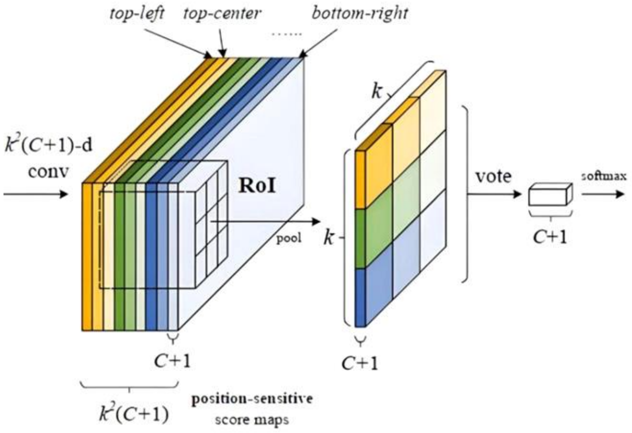

Figure 3. FCN-5 network structure

These activation images demonstrate localized properties of the music signal and capture more complex patterns and relationships through convolutional operations and nonlinear activation functions. Convolutional feature extraction constructs a sequence of multi-level, multi-scale feature maps that provide rich information for subsequent hash coding and classification. Optimization strategies such as bilinear interpolation and similarity selection can be used in the feature extraction process, which help to create tight connections between feature maps and improve resolution as well as expressiveness. After generating this sequence, it will be fed into the Long Short-Term Memory (LSTM) network. the LSTM handles time-series data, which better captures the temporal information in the music signals and ensures that the model understands their dynamic changes and complex structure. Combined with the hash layer functionality, this featured image sequence is finally converted into a concise hash code, while ensuring the integrity of the information and significantly reducing the computational difficulty.

3.3. LSTM-based music recognition

3.3.1. Multi-layer LSTM for hashing

The application of long short-term memory networks (i.e., LSTMs) appears to be extremely critical and important in the construction of hash-based convolutional neural networks for music recognition, especially in generating multilevel, multilayered LSTM models for use in hash codes. This new and innovative approach to music recognition proposed in this paper incorporates a pre-trained Convolutional Neural Network (CNN), which is designed to extract convolutional feature maps at multiple levels and different depths to form a complete and complex sequence of feature maps. Next, a multilayer recurrent neural network (RNN), in particular a multilayer LSTM model consisting of two long short-term memory units stacked and shaped, is used to process these rich and informative feature maps and ultimately generate hash codes. Specifically, a sequence of these detailed and refined processed convolutional feature maps is first fed into the first LSTM network. This specially designed and optimized LSTM network processes each specific and independently existing feature maps one by one, step by step, through time steps to obtain a systematic and comprehensive sequence of feature vectors. Equation (1) exhaustively describes this process, where a set of a series and diverse feature vectors over a set of time steps is obtained. In this way, this well-established large-scale LSTM neural network is able to fully capture the important temporal dependencies and internal dynamic patterns implicitly present in these complex and dynamically changing data.

$$H_{abstract} = \{(h_1^1, h_2^1, \cdots h_p^1), (h_1^2, h_2^2, \cdots h_p^2), \cdots, (h_1^s, h_2^s, \cdots h_p^s)\} \tag{1}$$

After successfully capturing the feature vector sequences from the first LSTM layer, these feature-rich and information-dense vector sequences are further fed into the second LSTM network for deeper processing. This second LSTM network, named LSTM_abstrgt, further refines and captures higher-level, more abstract and complex information features by combining the feature vector sequences generated by the aforementioned first LSTM. In Equation (2), we can clearly see the details of the operation performed by the second LSTM, where hend represents the last implicit state finally obtained by this second LSTM network, while W2 denotes the set of weight matrices corresponding to this. Meanwhile, v2 is the associated set of bias vectors. Through this step-by-step, iterative refinement process, the second fine and accurate LSTM is able to effectively summarize the critical and valuable information in the initially large and diverse sequence of feature vectors, and capture the more complex, variable, and highly abstract patterns hidden behind the data.

$$h_{end} = LSTM_{encode}^{abstract}(H_{abstract}, W_2, v_2) \tag{2}$$

Eventually, the hidden layer state of the second LSTM is tightly connected to the hash layer used for hash processing, thus generating the hash code in its final concrete form. This specially designed hashing layer transforms this complex data into a very compact and useful hash code by efficiently compressing and accurately encoding the high-dimensional features generated by the LSTM. Equation (3) explicitly describes this hash code generation process, in which these unique and highly representative hash codes defined and generated by mathematical formulas not only retain the most critical and important information of the original features, but also significantly reduce the large number of dimensions of the data and the complexity of the computational resources consumed by the data with the help of an extremely efficient method.

$$q = \emptyset(W_H^T h_{end} + v_H) \tag{3}$$

This advanced hash coding method based on multilayer LSTM has shown significant advantages in the field of music recognition. On the one hand, the LSTM network is able to capture a long time-dependent time series of features in the music signal, which makes the model particularly good at dealing with continuous changes; on the other hand, the introduction of a specially designed and powerful independent practical application of an important part of the process - the so-called "advanced" hash layer - not only greatly improves the overall feature representation, but also greatly improves the overall feature representation, which is the so-called "advanced" hash layer. On the other hand, the introduction of the so-called "advanced version" of the hash layer, which not only greatly improves the compactness and computational efficiency of the entire feature representation process, but also effectively improves and enhances the accuracy of the recognition results as well as the overall robustness of the results through the reduction of redundant and useless information as well as various types of noise interference[10].

3.3.2. Loss function design

In the process of constructing a complex music recognition model based on hash convolutional neural networks, the careful design of the loss function is a very critical step, which directly and significantly affects the training effect of the entire model and the final performance of music recognition. Since the traditional hash method will face the constraint problem of discrete values when obtaining the binary encoding results, it is necessary to adopt more relaxed and flexible constraints for optimization, in order to improve the training efficiency and stability of the overall model. Specifically, the binary encoding in the traditional method strictly requires that the encoding result must be absolutely 0 or 1, but this strict restriction is likely to lead to a large computational complexity in the optimization process, and it may also cause the problem of gradient disappearance. In order to effectively solve this series of problems, we can appropriately relax the rigid constraints of the "binary code", no longer extreme requirements can only be coded as 0 or 1, but to provide that its value can fall within a continuous range. In the actual optimization process, the first generation of these relatively relaxed and flexible "binary code", and to ensure that these code values can be optimized within a predetermined range, so as to reduce the adverse impact of discrete values on the model training process. After the optimization is completed, the relaxed "binary code" is then quantized to obtain a real binary number that meets the initial requirements and standards. This improved strategy not only improves the overall training efficiency, but also enhances the performance of the system in all aspects during operation.

Let the binary code $b^{(n)}$ of the music signal be the input to the softmax layer, the probability of predicting the label y(n) can be defined as shown in Equation (4):

$$p\left(y^{(n)} = m \middle| b^{(n)}\right) = \frac{exp(z_m)}{\sum_{i=1}^{M} exp(z_l)}, m = 1,2,\cdots,M \tag{4}$$

In this text, W and v are the weight matrix and bias vector representing the weight of the softmax layer, respectively, while K represents the total number of classes in the classification task. With this particular approach, the loss function can be designed as a cross-entropy loss, thus minimizing the difference between the predicted labels and the true labels, which enables optimizing the model recognition erformance. This relatively relaxed binary coding optimization method not only improves the efficiency during model training, but also effectively mitigates the problem of information loss that may arise during information binarization. By quantizing these slack codes, it can be ensured that the final generated hash code can express the features efficiently as well as have good recognition ability and robustness. When designing the loss function, not only the factor of coding accuracy should be considered, but also the computational complexity and whether the optimization process is feasible.

4. Experiments and Analysis

4.1. Comparison of RNNs with different hierarchies

In the construction of this hash-based convolutional neural network-based music recognition system, RNNs (Recurrent Neural Networks) with different hierarchical structures have significant and obvious effects on the recognition performance. In order to analyze their effects deeply and comprehensively, this paper compares in detail three RNNs with different structures and multi-layers, including Simple RNN, Gated Recurrent Unit (GRU), and Long Short-Term Memory Network (LSTM), and conducts a series of rigorous and accurate performance tests on the same dataset.

The experimental results are shown in Figure 4, from which it can be clearly seen that the RNN, which employs multiple LSTM hierarchies and stacks them up in multiple layers to form a complex model, performs most optimally and prominently in the music recognition task. Specifically, the large deep multilayer RNN architecture containing two LSTM layers performs particularly well in capturing and processing long time dependencies and their complex features in music signals, and its overall recognition accuracy and robustness outperforms that of the other compared structures. This is mainly due to the fact that LSTM is able to process long time series data more efficiently, effectively avoiding the gradient vanishing problem, which is a common and troubling problem for researchers in Simple RNN, and it is also able to capture more, richer, and more subtle information features with significant meaning than GRU.

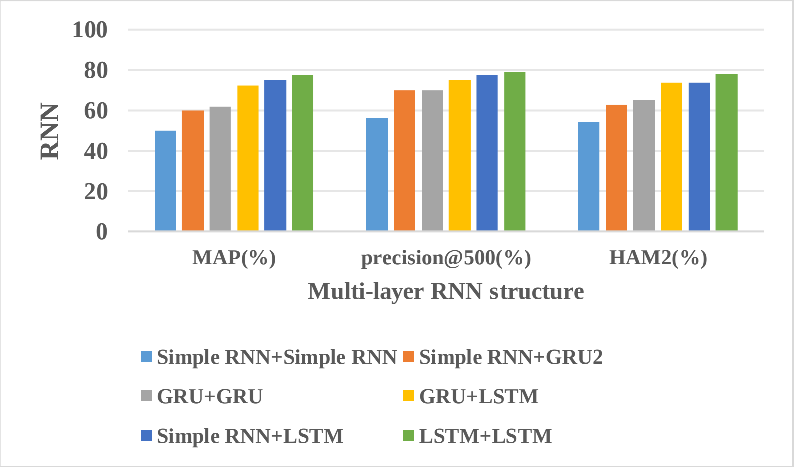

Figure 4. Comparison of recognition results for different RNN structures

In contrast, the worst performance is that of Simple RNN, which is mainly due to its inability to effectively deal with the problem of long time dependency, which leads to its less satisfactory performance in recognizing complex music signals. As for GRU, although it alleviates the problem of gradient vanishing to a certain extent, its ability to capture features is still not as strong as that of LSTM due to its relatively simpler structure.

4.2. Comparison of different combinations of feature maps

In a study of music recognition based on hashed convolutional neural networks, the effects of a number of different combinations of feature maps are analyzed through comparative experiments, and the results show the advantages of feature map sequences in terms of recognition accuracy. Specifically, this paper compares and analyzes three different combinations of feature maps: a feature map that includes all five convolutional layers, separate feature maps for each of the different convolutional layers, and a method that consists of a sequence of feature maps from multiple convolutional layers and is applied to a recurrent hash recognition model. Based on the data and results presented as inserted in the fifth image of this paper, the results show that the recognition accuracy using the serialized feature maps is significantly higher than the other methods.

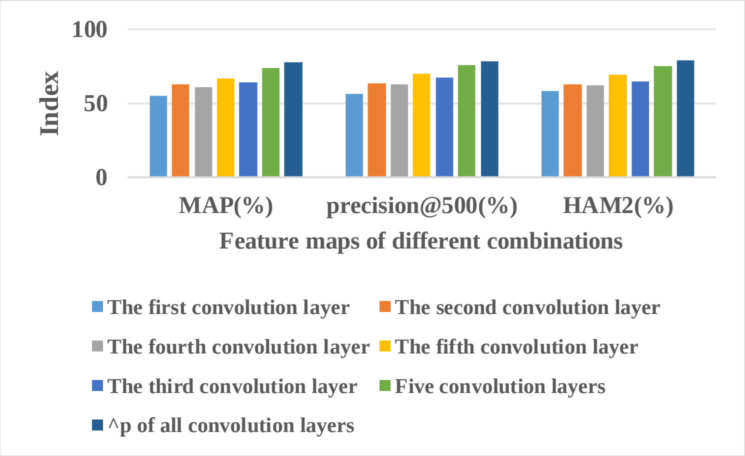

Figure 5. Comparison of recognition results for different convolutional feature maps

This result can be attributed to the rich and detailed spatial details and deep semantic information contained in the feature map sequence approach. Compared with the feature maps generated by a single convolutional layer, a series of feature maps formed by multiple convolutional layers can capture more detailed and complex image features and higher-level and more abstract semantic information, which is especially important and critical for complex pattern recognition in music signals. Although the feature maps generated by convolutional layers retain some of the spatial information to a certain extent, they lack sufficiently broad and in-depth contextual semantics, so their recognition accuracy is relatively low and not as effective as the multiple convolutional layers approach.

4.3. Comparison with other methods

In the study of music recognition based on hashing convolutional neural networks, this experiment compares the feature approach of convolutional recurrent hashing with the traditional kernel-based hashing (KSH) method. The experimental results are shown in Fig. 6, showing the significant advantages of convolutional recurrent hashing in music recognition.

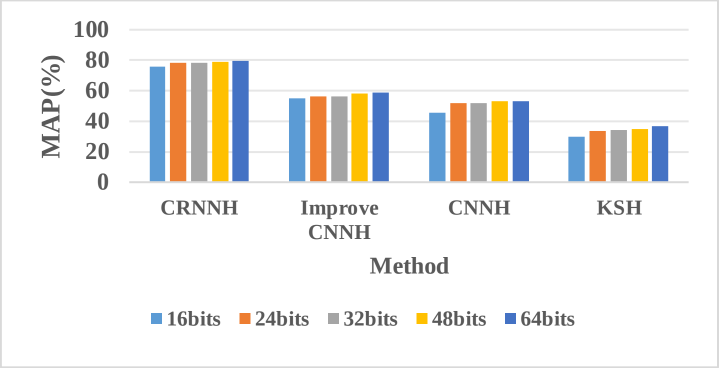

Figure 6. Comparison of average mean progress of different methods

Specifically, the method of convolutional cyclic hashing is designed and optimized by carefully designing and optimizing a specific loss function, thus making it perform particularly well in terms of recognition accuracy. In contrast, the KSH method, although it also performs quite well in some tasks, performs significantly worse than the convolutional hashing method in the extremely complex task of music recognition. This is mainly due to the fact that the convolutional cyclic hashing method is able to extract and fully utilize the two key features, spatial and temporal, in music signals more efficiently. The method extracts features through a multilevel convolutional network and combines it with a recurrent neural network to capture temporal information, which enables the convolutional recurrent hashing method to generate richer and more representative feature vectors, thus dramatically improving the ability to accurately recognize music signals.

5. Conclusion

With the ever-changing development of music recognition technology, it has always been a challenging problem to achieve efficient and accurate recognition in those complex, changing and unpredictable music signals. In this paper, by introducing advanced techniques - hash convolutional neural network and LSTM model - and proposing an improved CNN model, the accuracy and efficiency of the music recognition process are greatly enhanced. The experimental results show that this improved model shows excellent performance when dealing with different levels and various music features, especially when dealing with complex music segments with complicated and redundant timing information, it shows strong robustness ability, which not only significantly improves the accuracy of the recognition, but also makes a big breakthrough in processing speed compared with the traditional methods, providing innovative Ideas. In the future research direction, the model structure can be further optimized and adjusted to explore more new combinations with unlimited potentials in order to cope with more complicated and tricky experiences. In addition, this successful application not only brings useful references to other audio processing tasks, but also lays a solid foundation for technological progress in related fields.

References

[1] Jia X. A music emotion classification model based on the improved convolutional neural network[J]. Computational Intelligence and Neuroscience, 2022, 2022(1): 6749622.
[2] Hizlisoy S, Yildirim S, Tufekci Z. Music emotion recognition using convolutional long short term memory deep neural networks[J]. Engineering Science and Technology, an International Journal, 2021, 24(3): 760-767.
[3] Chen J. Construction of Music Intelligent Creation Model Based on Convolutional Neural Network[J]. Computational Intelligence and Neuroscience, 2022, 2022(1): 2854066.
[4] Zhang Y. Music recommendation system and recommendation model based on convolutional neural network[J]. Mobile Information Systems, 2022, 2022(1): 3387598.

[5] Lei L. Multiple Musical Instrument Signal Recognition Based on Convolutional Neural Network[J]. Scientific Programming, 2022, 2022(1): 5117546.

[6] Miao Z, Cheng C. Construction of multimodal music automatic annotation model based on neural network algorithm[C]//Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023). SPIE, 2023, 12779: 482-488.

[7] Li T L, Chan A B, Chun A H. Automatic musical pattern feature extraction using convolutional neural network[J]. Genre, 2010, 10(2010): 1x1.

[8] Villarreal M, Sánchez J A. Handwritten music recognition improvement through language model re-interpretation for mensural notation[C]//2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE, 2020: 199-204.

[9] Schlüter J, Böck S. Improved musical onset detection with convolutional neural networks[C]//2014 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2014: 6979-6983.

[10] Li J, Soradi-Zeid S, Yousefpour A, et al. Improved differential evolution algorithm based convolutional neural network for emotional analysis of music data[J]. Applied Soft Computing, 2024, 153: 111262.