Music Genres Classification

Di Feng
fengdi2022@shanghaitech.edu.cn

Yaoyu He
heyy2022@shanghaitech.edu.cn

Peijun Xu
xupj1@shanghaitech.edu.cn

Abstract

Genre classification plays a crucial role in music recommendation systems which enables people to search the vast amount of music data more quickly and efficiently according to their preferred music style. We aim to create a strong music genre classifier that will elevate classification accuracy. By leveraging Convolutional Neural Networks (CNNs) and Convolutional Recurrent Neural Networks (CRNNs), we constructed a classifier that are first decentralized trained on the three features of Mel spectrogram, Chroma feature, and Mel Frequency Cepstral Coefficients (MFCC), and then classified based on them jointly. Our approach notably augments the precision of classifying comparing to basic CNN. The classifier's architecture is uniquely designed to utilize the existing categorization data in music thoroughly, implementing an innovative strategy for joint feature classification.

1 Related Work

1.1 Traditional Method

Music genre classification has been studied since the early stage of machine learning. Pioneering research in the early 21st century had already revolved around feature extraction techniques as a pre-processing step for music analysis. An excellent approach was to extract of three distinct feature sets - timbral texture, rhythmic content, and pitch content first. These features were then utilized in genre classification with Expectation-Maximization (EM) algorithm within the framework of Gaussian Mixture Models (GMM) or K-Nearest Neighbors (KNN) algorithm [10]. Additionally, there was a also method focusing on audios' Daubechies wavelet coefficients (DWCHs) , which included both local and global musical information. These features then were used by Support Vector Machines (SVM) or Linear Discriminant Analysis (LDA) for genre categorization [7]. A novel approach proposed in 2005 involved synthesizing the outcomes from these effective classifiers to construct a confusion matrix. This matrix then served as a basis for inferring genre relationships, thereby augmenting the precision and efficiency of the classification process [6].

1.2 Neural Network

In recent years, the advent of ReLU activation functions propelling deep learning frameworks to the forefront of preferred models for a variety of machine learning tasks. Particularly within the task of music genre classification, CNNs have become a popular tool. A seminal architecture in this domain is the Bottom-up broadcast neural network [8] for music genre classification. This architecture is distinguished because of it contains bridge of information transfer from shallow to deep layers, simultaneously reducing parameter count and dataset size requirements for training. In another research, the extraction and utilization of Mel Frequency Cepstral Coefficients (MFCCs) as CNN inputs for automatic classification have been explored [2], drawing on early experiences in music genre classification.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Beyond these developments, RNNs – known for their capacity to capture temporal dynamics in data – has also been a useful tool for music genre classfication. A well-known research has combined LSTM RNNs with support vector machines (SVMs), using extended signal MFCC features as inputs [3]. This joint approach has shown promise in music genre classification tasks. Additionally, the extraction and training of Mel spectrograms using CRNN-based models have also yielded positive outcomes [1], which showed the effectiveness of these combined methodologies in the field of music genre classification.

2 Method

2.1 Feature Extraction

Previous classification researches has shown the important role of music feature extraction in the task of music genre classification, so we decided to extract them for our training dataset. We chose to use the GTZAN database [10] as our training dataset, which contains 10 different music styles with about 100 songs in each style. These styles include Blues, Classical, Country, Disco, Hip hop, Jazz, Metal, Pop, Reggae and Rock. Their music files are usually available in WAV or MP3 format, and each song is approximately 30 seconds in length.

The three main audio features we chose for our training were the Mel spectrogram, Chroma feature, and Mel Frequency Cepstral Coefficients (MFCC). The Mel spectrogram represents the power spectrum of a sound, mapped onto the Mel scale which is a perceptual scale of pitches judged by listeners to be equal in distance from one another. MFCCs are highly used in voice recognition and other audio-related fields because they efficiently represent the shape of the sound's spectral envelope. Chroma features represent the energy content of a signal in each of the 12 different pitch classes. They capture the harmonic and melodic characteristics of music, making them particularly useful for tasks like chord detection and music similarity.

We use the librosa [9] package to extract the music features, and in order to facilitate the subsequent combination of the features for classification, we set the number of samples in each audio file to 1293. Therefore, the three feature vectors mentioned above corresponding to each audio file are $(1293, 128)$, $(1293, 12)$, $(1293, 20)$ respectively. Our output is a 10-dimensional vector, with each value on the vector corresponding to its probability of belonging to that genre

2.2 Convolutional Neural Networks

We created CNN models and CRNN models for each of the three features. The whole structure of our CNN model is shown in Fig 1, where convolution operation represents the part of convolutional layer we built, and N represents the size of the vector dimensions corresponding to the three features, i.e., 128, 12, and 20. The output shape of each layer are shown beneath the picture. Detailed configurations of specific convolutional layers are presented in Fig 2.Our architecture encompasses four convolutional layers i.e., three connected structures shown in Fig 2 (the CNN model for MFCC features only have two MaxPool layers). The final output of convolutional layers has 256 channels. The convolutional layers are followed by the integration of a normalization layer and a dropout layer, enhancing the model's efficacy. For this classification task, we employed the cross-entropy loss function and Adam optimizer for optimization.

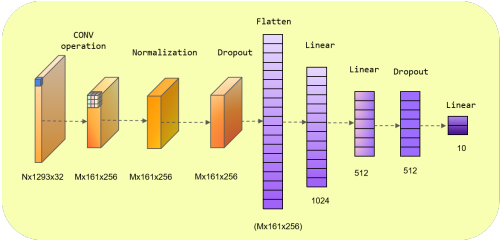


Figure 1: Basic CNN.

In our study, we also introduce a Convolutional Recurrent Neural Network (CRNN) architecture, as shown in Fig 3. This structure retains the convolutional layer of the conventional CNN; how-

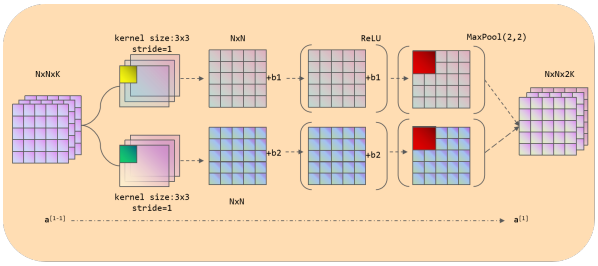


Figure 2: convolutional layers.

ever, it incorporates a novel adaptation after the convolutional stage. Specifically, we integrate a TimeDistributed layer, which serves to reduce the dimensionality of the convolutional layer's output while specifying the time step dimension simultaneously. Subsequent to this layer, we employ a Long Short-Term Memory (LSTM) network. Since our samples corresponded to features at different times in the audio, we utilize the dimension in the convolutional layer's output vector, corresponding to the sample count (161 in Fig 3), as the time step dimension. For optimization, we also use the cross-entropy loss function coupled with the Adam optimizer, aligning with the CNN model training.

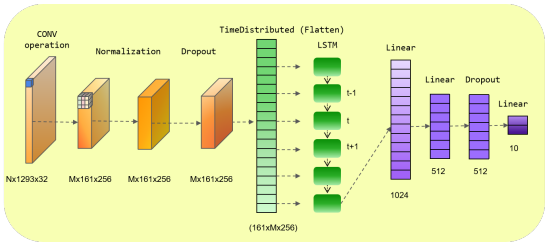


Figure 3: Basic CRNN.

2.3 Classification Network

After gathering the classification outputs from three distinct features, we hope to integrate these outputs to accurately identify the genre of a given audio file. Initial attempts employing basic combination techniques, such as vector summation followed by selecting the genre corresponding to the maximum value, or implementing a weighted average with a similar selection criterion. However, they all lead to unsatisfactory results. We thought that this inadequacy came from the fact that each feature's output vector primarily captures the genre relationships in same feature, but neglecting the features connections with same genre. In our approach, we integrate the outputs from three distinct features into a unified 30-dimensional vector. This vector is subsequently processed through a series of nine one-dimensional convolutional layers, each characterized by a kernel size of 3 and a stride of 1. The utilization of convolutional layers serves a dual purpose. Firstly, it further enhance the combination of categorization information between different genres in the same feature. Secondly, it establishes a linkage between the categorical information of identical genres across different features and interconnects the categorical information of diverse genres and features. Additionally, in deeper layers, there is a progressive enhancement in the integration of categorical information from different features, leading to a more comprehensive representation of the mixed information.

Based on the analysis, our classification network is shown in Fig 4. The 10-dimensional output values finally turn into genre probabilities through a Softmax function, which yields the corresponding genre of the audio file. We also utilize the cross-entropy loss function and Adam optimizer for the network.

3 Experiments

3.1 Training

In our experiment, we first employed CRNN as well as CNN across three distinct features, comparing their results to select the appropriate models for features. After selecting the most approparitae

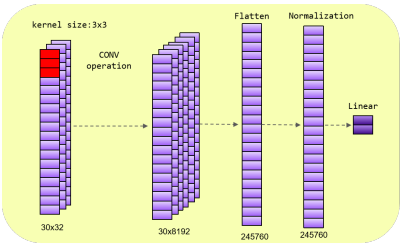
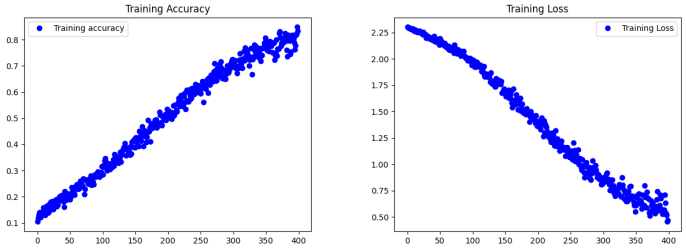


Figure 4: Classification Network.

models (Mel:CRNN, Chroma:CNN, MCFF:CRNN), we trained for 400 epochs using the GTZAN database dataset, with 100 segments of each genre selected as training data. The training accuracy and loss are shown in Fig 5

3.2 Testing

We tested on a novel corpus of 6400 music clips from FMA dataset [4], distinct from the training set. Since the test set is characterized by human-annotated music genres and the overall musical style of data in each of its genres was slightly different from the training set. Thus, we decide to consolidate them into the four categories of pop, classic, jazz, and rock for classification. Since there is no benchmark that overlaps exactly with the types we used in test set, we compare it with the benchmark in this paper [5], which covers almost the same categories with 1800 music clips. The comparison of the test results of our method with other classical methods is shown in Table 1



(a) Training Accuracy (b) Training Loss

Figure 5: Training Results.

Method	Test Accuracy
Deep Belief Network	51.88%
Naive Bayes	43.69%
KNN	53.23%
C4.5	45.44%
CNN	46.87%
Our method	58.54%

Table 1: Comparsion

4 Conclusion

In conclusion, while our approach marks a advancement beyond the conventional CNN methodology, it is evident that there remains a scope for further refinement. Optimizing the integration of the three distinct features presents a promising avenue for enhancement. We may anticipate delving into a broader array of questions to augment our understanding and application of this method.

contribution: FengDi:pre and code(35%).HeYaoyu:pre and code(45%).XuPeijun:paper and pre(20%)

References

[1] Bisharad, D. and Laskar, R. H. (2019). Music genre recognition using convolutional recurrent neural network architecture. *Expert Systems*, 36(4):e12429.

[2] Ceylan, H. C., Hardalaç, N., Kara, A. C., and Firat, H. (2021). Automatic music genre classification and its relation with music education. *World Journal of Education*, 11(2):36–45.

[3] Dai, J., Liang, S., Xue, W., Ni, C., and Liu, W. (2016). Long short-term memory recurrent neural network based segment features for music genre classification. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE.

[4] Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. (2016). Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*.

[5] Homburg, H., Mierswa, I., Möller, B., Morik, K., and Wurst, M. (2005). A benchmark dataset for audio classification and clustering. In *ISMIR*, volume 2005, pages 528–31.

[6] Li, T. and Ogihara, M. (2005). Music genre classification with taxonomy. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v–197. IEEE.

[7] Li, T., Ogihara, M., and Li, Q. (2003). A comparative study on content-based music genre classification. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 282–289.

[8] Liu, C., Feng, L., Liu, G., Wang, H., and Liu, S. (2021). Bottom-up broadcast neural network for music genre classification. *Multimedia Tools and Applications*, 80:7313–7331.

[9] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25.

[10] Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302.