Harmonizing Genres: An Ensemble Approach Using CNNs and RNNs for Music Genre Prediction

Authors: Jatin Kulkarni, Anthony Ceja, Saiprathik Chalamkuri

Abstract

This project explores the challenging task of predicting music genres using a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to analyze audio data. Starting with the GTZAN audio dataset augmented by a comprehensive dataset from Hugging Face, our study initially utilized various graphical representations of audio---waveforms, spectrograms, log mel spectrograms, and chromagrams---to train separate CNN models. Recognizing the limitations in melody recognition which resulted in suboptimal genre classification accuracy, the project pivoted to integrating RNNs with LSTM to better capture the sequential nature of melodies. Through a meticulous ensemble of these models, the research aimed to improve prediction accuracy by leveraging the strengths of each model type. Our findings contribute to the understanding of genre prediction complexities and highlight the potential of mixed-model approaches in the field of music analysis.

1) Introduction

1.1 Background

The background for the project stems from the increasingly important role of machine learning in various domains, including music analysis. Accurately predicting music genres holds significant implications for enhancing music discovery, refining recommendation systems, and facilitating efficient audio indexing. However, despite notable advancements in machine learning techniques, genre prediction remains a challenging endeavor due to the multifaceted nature of musical attributes. Musical genres often exhibit inherent diversity and subjectivity, making it difficult to develop models that can effectively capture the nuanced characteristics defining each genre.

Our project aims to address this challenge by leveraging state-of-the-art machine learning methods, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to analyze audio data comprehensively. We draw inspiration from existing scientific papers and research in the field of music genre prediction, particularly those exploring ensemble learning techniques and the integration of CNNs and RNNs for improved performance. Some of the key papers informing our project include “Music Recognition and Classification Algorithm considering Audio Emotion” and “Music Score Recognition Method Based on Deep Learning”. By building upon the insights and methodologies outlined in these papers, we strive to advance the current understanding of genre prediction complexities and contribute to the development of more robust and accurate music analysis models.

1.2 Objective

Our project aimed to harness the power of ensemble learning by integrating different types of neural network models—specifically CNNs for image-based audio analysis and RNNs for sequence-based melody understanding—to enhance the accuracy of music genre prediction. This approach seeks to capitalize on the distinct advantages of each model type, potentially offering superior performance over using any single model approach.

2) Methodology

2.1 Data Preprocessing

When constructing our dataset, we aimed to ensure diversity, representativeness, and standardization to facilitate robust model training and evaluation. Our dataset comprises multiple sources, each chosen with specific considerations in mind.

Firstly, we incorporated the GTZAN Dataset, a well-established benchmark in the field of music genre classification. This dataset consists of 30-second audio clips spanning various genres, providing a solid foundation for training our models.

Additionally, we integrated the Hugging Face 'ccmusic-database/music_genre' dataset, which offers full-length songs categorized by genre. To maintain consistency with the GTZAN Dataset, we processed these songs into 30-second snippets, ensuring uniformity across the dataset.

For our experiment datasets, we manually curated testing datasets for acapella and instrumental versions of songs from diverse genres. These datasets were sourced from YouTube, allowing us to include a broader range of musical styles and characteristics not fully represented in existing datasets.

To introduce variability and assess model robustness, we manipulated select songs from the original dataset by applying effects such as speeding up songs. This created a new test set with altered audio characteristics, challenging the models to generalize beyond standard audio features.

All audio snippets were stored in the WAV format to preserve high-quality audio data and minimize compression artifacts. This choice ensures that our models are trained and evaluated on pristine audio representations, crucial for accurate genre classification.

Overall, our dataset construction process prioritized diversity, standardization, and quality, enabling comprehensive model training and thorough evaluation across various musical genres and audio characteristics.

2.2 Convolution Neural Network (CNN) Models

We chose to employ convolutional neural networks (CNNs) in our model architecture due to their effectiveness in capturing spatial and temporal patterns in data. Our model comprises separate CNN architectures optimized for processing various types of audio representations: waveform, spectrogram, log mel spectrogram, and chromagram. Each architecture is meticulously designed to exploit the inherent structure and characteristics of the input data, ensuring that the model can effectively capture the unique features relevant for accurate genre prediction.

1. **CNN for Waveforms:** This component processes raw audio waveforms, focusing on temporal and amplitude patterns. It utilizes 1D convolutional layers optimized for time-series data, allowing the model to capture nuances in the raw audio signal.

2. **CNN for Spectrograms:** Spectrograms provide a visual representation of the spectrum of frequencies in an audio signal as they vary with time. The CNN architecture for spectrograms incorporates 2D convolutional layers, enabling the model to analyze spectral patterns across both time and frequency dimensions.
3. **CNN for Log Mel Spectrograms:** Log mel spectrograms are derived from spectrograms but emphasize frequencies perceptible to humans. This component of our model utilizes 2D convolutional layers optimized to capture subtle variations in scaled frequency content, enhancing the model's sensitivity to perceptual aspects of sound.
4. **CNN for Chromagrams:** Chromagrams focus on capturing the intensity of different pitch classes in an audio signal, providing insights into harmonic and melodic content. The CNN architecture for chromagrams features 2D convolutional layers tailored to analyze pitch class intensity patterns, crucial for identifying genre-specific harmonic structures.

To train these models, we employed rigorous techniques such as stochastic gradient descent (SGD) or Adam optimization, along with categorical cross-entropy loss functions. Dropout regularization was also applied to prevent overfitting and ensure generalization across different datasets. Our design choices aim to extract genre-relevant features efficiently while maintaining computational effectiveness, ultimately leading to superior performance in genre prediction tasks.

2.3 Ensemble Learning

After training individual CNN models for each graphical representation, we employed ensemble learning techniques to combine their predictions. Ensemble methods, such as weighted averaging or stacking, were used to aggregate the outputs of the CNN models and produce a final prediction for each audio snippet.

2.4 Transition to Recurrent Neural Networks

Despite the promising results obtained with CNN models, we observed limitations in accurately capturing the sequential nature of melodies, particularly in distinguishing between genres with similar rhythmic patterns. To address this issue, we transitioned to recurrent neural networks (RNNs), specifically long short-term memory (LSTM) networks.

Details on the implementation and training of the RNN models are provided in the subsequent sections.

2.4.1 RNN Models

We explored the dynamics of audio data through Recurrent Neural Networks (RNNs), leveraging different types of feature representations: Chroma, Log Mel Spectrogram, Spectrogram, and Waveform. Each feature captures unique aspects of the audio signal, making them suitable for our RNN-based genre classification approach.

Chroma RNN: The Chroma feature captures the intensity of each of the 12 different pitch classes and is particularly useful for analyzing music where harmony is more defining than timbre. We implemented an RNN to model these features as they provide insight into the harmonic progression of a piece, which is critical for genre identification.

```
# Example Python code snippet for Chroma RNN model
```

```
model = Sequential()
model.add(LSTM(256, return_sequences=True,
input_shape=(input_length, 12)))  # '12' for 12 pitch classes
model.add(LSTM(256))
model.add(Dense(10, activation='softmax'))  # '10' for the
number of genres
model.compile(loss='categorical_crossentropy', optimizer='adam',
metrics=['accuracy'])
```

Log Mel Spectrogram RNN: Log Mel Spectrogram reduces the frequency dimension of standard spectrograms while retaining the most perceptually relevant information. The RNN model designed for this representation focuses on capturing textural details that are essential for distinguishing between genres like electronic and rock, which can have similar pacing but differing texture.

```
# Example Python code snippet for Spectrogram RNN model
model = Sequential()
model.add(LSTM(64, return_sequences=True, input_shape=(frames,
frequency_bins)))
model.add(LSTM(64))
model.add(Dense(num_genres, activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer='adam',
metrics=['accuracy'])
```

Waveform RNN: Direct modeling of waveforms with RNNs allows the capture of raw audio signals' temporal dynamics. This method is beneficial for genres where beat, rhythm, and tempo are more defining characteristics than harmonic or melodic content.

```
# Example Python code snippet for Waveform RNN model
model = Sequential()
model.add(GRU(256, input_shape=(audio_length, 1)))
model.add(Dense(10, activation='softmax'))
model.compile(optimizer='adam', loss='categorical_crossentropy',
metrics=['accuracy'])
```

These implementations show how different audio features are processed and modeled using RNN architectures, each tailored to capture the most genre-relevant characteristics of the data. By employing these diverse approaches, our project aims to robustly classify music genres across a broad spectrum.

3) Experiments

3.1 Genre Prediction across Song Variants

The experiment aimed to assess the neural network's ability to classify different versions or renditions of songs, particularly focusing on variants such as sped-up versions. We hypothesized that the neural network, trained on a diverse dataset containing various song variants, would demonstrate the capability to accurately classify these variants into their respective genres. By conducting this experiment, we aimed to evaluate the robustness and generalization capability of the model across different song variations, contributing to our understanding of its performance in real-world scenarios where song versions may vary.

3.2 Genre Classification Using Instrumental and Acapella Music

This experiment aimed to investigate how the inclusion of instrumental and acapella music affects the performance of our genre classification models. We hypothesized that the absence of vocals in instrumental music and the exclusive presence of vocals in acapella music could potentially alter the genre predicted by the model.

3.3 Results

The experiments yielded similar, discouraging results. The CNN achieved a classification accuracy of approximately 26-28% across the tested genres (pop, rock, opera/classical, and R&B) in both experiments. Figure 1 illustrates the classification accuracy achieved for each genre in the combined results.

The consistent performance of the model across both experiments suggests that the model struggles to classify sound based solely on the input provided. The inability to effectively differentiate between genres based on song variants or instrumental/acapella characteristics indicates a limitation in the model's ability to capture genre-specific features from audio signals.

3.4 Discussion of Results

The poor performance of the CNN in genre prediction across the tested genres underscores the challenges inherent in music classification tasks. The low accuracy rates suggest that the model struggled to discern genre-specific features effectively, resulting in inaccurate predictions across diverse musical styles.

From these experiments, we have learned that the task of genre prediction remains a complex challenge, particularly when dealing with diverse genres and variations in audio characteristics. The disappointing results highlight the need for further research and exploration of more sophisticated models or feature representations to improve genre classification accuracy.

In conclusion, while the experiments did not yield favorable outcomes, they provide valuable insights into the limitations of current CNN-based approaches in music genre prediction. The results emphasize the importance of continued research and innovation to address the complexities of genre classification in music analysis tasks.

4) Insights and Next Steps

4.1 Insights:

1. **Different Predictive Outcomes:** Our experiments uncovered distinctive predictive outcomes between CNNs and RNNs, underscoring the importance of selecting the appropriate model architecture for music genre prediction tasks. This highlights the necessity of understanding the nuances of different architectures to optimize performance.
2. **Understanding the Song Components:** Through our experiments, we gained insights into the critical role of comprehending various song components, including notes, frequencies, and their sequencing, in accurately predicting music genres. This underscores the importance of feature engineering and data preprocessing techniques to extract meaningful information from audio data.
3. **Commonality in Notes and Frequencies:** Despite variations in song variants and characteristics, we observed a commonality in notes and frequencies across different genres. This discovery suggests the existence of underlying patterns that models can potentially leverage for more effective genre prediction. It emphasizes the need for models to capture genre-specific features while remaining robust across diverse musical styles.
4. **Importance of Notes Sequencing:** Our experiments highlighted the significance of capturing the sequential nature of notes in melodies for accurate genre classification. This emphasizes the limitations of CNNs in understanding long-form melodies and underscores the necessity of exploring alternative architectures, such as RNNs, to better capture temporal dynamics.
5. **Issues with Unbalanced Dataset Size:** An additional insight from our experiments was the disparity in dataset sizes for each genre, particularly evident with opera having only 300 songs for training compared to other genres with 1000-2000 songs. This imbalance can affect model performance and underscores the importance of dataset balancing techniques to ensure equitable representation across all genres.
6. **Selective Activation Implementation:** We have implemented selective activation techniques to enhance our model's performance. By selectively activating specific neurons based on learned patterns, our model can better capture genre-specific features and improve classification accuracy. This approach leverages recent advancements in neural network architectures and has shown promise in optimizing genre prediction models.

In conclusion, our experiments have provided valuable insights into the complexities of music genre prediction tasks and the challenges associated with employing different model architectures. By exploring the predictive outcomes of CNNs and RNNs, understanding the significance of song components, recognizing commonalities in notes and frequencies, and addressing issues with dataset size imbalances, we have deepened our understanding of genre prediction in music analysis. Furthermore, the implementation of selective activation techniques represents a step forward in enhancing model performance and adapting to the intricacies of music data. Moving forward, these insights will guide our continued efforts to refine our models, optimize prediction accuracy, and contribute to advancements in the field of music analysis and classification.

Works Cited

Liang, Dawen et al. “Music Genre Classification with the Million Song Dataset 15-826 Final Report.” (2011).

Lin Q. Music Score Recognition Method Based on Deep Learning. Comput Intell Neurosci. 2022 Jul 7;2022:3022767. doi: 10.1155/2022/3022767. PMID: 35845890; PMCID: PMC9282982.

Luo, Xining. (2023). Automatic Music Genre Classification based on CNN and LSTM. Highlights in Science, Engineering and Technology. 39. 61-66. 10.54097/hset.v39i.6494.

Wang Na, Fang Yong, "Music Recognition and Classification Algorithm considering Audio Emotion", Scientific Programming, vol. 2022, Article ID 3138851, 10 pages, 2022. https://doi.org/10.1155/2022/3138851