**Deep Learning for Automated Music Genre Classification Using Audio and Lyrical Features**

|  |  |  |
| --- | --- | --- |
| *Manav Mandal* | *Hrushikesh Attarde* | *Jayanand Hiremath* |

**Abstract**

This study investigates the potential of a multi-modal deep learning approach that combines audio and lyrical features to enhance the accuracy of automated music genre classification compared to single-modality methods. While prior research has achieved significant success using convolutional neural networks (CNNs) with spectrograms for audio-based classification, lyrical content has been largely overlooked. To bridge this gap, we propose a novel multi-modal architecture that leverages both audio and lyrical features. Our approach utilizes the GTZAN Genre Collection dataset for audio tracks and the musiXmatch dataset for lyrics, ensuring comprehensive coverage through manual collection where necessary. The model architecture consists of a CNN for extracting audio features, a BERT-based model for processing lyrical content, a fusion layer for feature integration, and a final classification layer. Audio features are extracted using mel-spectrograms and Mel-frequency cepstral coefficients (MFCCs), while lyrics are tokenized and encoded using a pre-trained BERT model. We train and evaluate the model using 5-fold cross-validation and compare its performance against single-modality baselines. Through attention visualization, we analyze the contributions of audio and lyrical features to the classification process. Our approach aims to achieve higher classification accuracy and provide deeper insights into the complementary nature of audio and lyrical modalities in music genre classification.

**1 Introduction**

Music genre classification is a critical task in the field of Music Information Retrieval (MIR), with widespread applications in recommendation systems, playlist generation, and music organization. The ability to accurately classify the genre of music has significant implications for enhancing user experience in streaming platforms and improving the personalization of music discovery tools. Traditional approaches to music genre classification have predominantly relied on handcrafted audio features such as timbre, rhythm, and pitch. However, with the rise of deep learning, there has been a shift towards using raw audio data, allowing models to automatically learn and extract meaningful features. Convolutional Neural Networks (CNNs) have played a pivotal role in this shift, as demonstrated by Choi et al., who achieved state-of-the-art results using CNNs on mel-spectrogram representations of audio data from the GTZAN dataset.

Despite these advancements, existing methods have primarily focused on audio-based features while overlooking the rich contextual information embedded in song lyrics. Lyrics often contain semantic and cultural cues that are indicative of genre, such as the storytelling nature of country music or the repetitive hooks in pop and hip-hop. Previous studies, such as those by Oramas et al., have highlighted the value of textual data for genre classification. By analyzing album reviews, Oramas et al. demonstrated that textual features can complement audio-based models, providing a more holistic view of the music. This insight serves as a motivation for the current study, which aims to combine the strengths of both audio and lyrical features within a unified, multi-modal deep learning framework.

The primary objective of this study is to develop a multi-modal deep learning approach that integrates both audio and lyrical features to improve the accuracy of music genre classification. Our approach addresses the limitations of single-modality models, which rely exclusively on either audio or lyrical data, and aims to demonstrate the effectiveness of a combined approach. The study will utilize two key datasets: the GTZAN Genre Collection for audio data and the musiXmatch dataset for lyrical content. The GTZAN dataset contains 1000 audio tracks across 10 genres, each 30 seconds long, while the musiXmatch dataset provides lyrics corresponding to the Million Song Dataset. Where direct matches between GTZAN and musiXmatch are unavailable, lyrics will be manually collected to ensure full coverage of the audio dataset.

The proposed method follows a systematic approach, beginning with audio and lyric preprocessing. Audio tracks will be converted into mel-spectrograms and Mel-frequency cepstral coefficients (MFCCs) using the librosa library, allowing for effective extraction of temporal and spectral features. Lyrical data will be tokenized and encoded using a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model, capturing the contextual and semantic features of the text. Our multi-modal model architecture will consist of a CNN-based audio feature extractor, a BERT-based lyric analysis module, a fusion layer to combine features from both modalities, and a final classification layer. The model will be trained using a 5-fold cross-validation strategy, and its performance will be benchmarked against single-modality models that rely solely on audio or lyrical features.

In addition to measuring classification accuracy, we will conduct an in-depth analysis of the model's performance across different music genres. Using attention visualization techniques, we aim to identify the relative importance of audio and lyrical features in genre classification. This analysis will offer valuable insights into how different features contribute to the prediction process and reveal the complementary nature of these modalities.

By incorporating both audio and lyrical features, this study aims to advance the state of the art in music genre classification. The findings have the potential to improve the accuracy of recommendation systems and playlist generation tools used by streaming platforms. Moreover, this work highlights the importance of leveraging multi-modal data for complex classification tasks, serving as a foundation for future research in multi-modal deep learning for music information retrieval.

**Related Works**

1. **Choi et al. (CNNs for Music Genre Classification using Spectrograms)**

Choi et al. introduced a convolutional neural network (CNN) approach for music genre classification, utilizing spectrograms as input features. Their method shifted the paradigm from handcrafted audio features to automatically learned feature representations from raw audio data. The CNN architecture effectively captured local and hierarchical patterns within the spectrograms, enabling the model to recognize genre-specific characteristics. Their approach achieved state-of-the-art results on the widely used GTZAN dataset, demonstrating the power of deep learning in music genre classification. This study serves as a foundational reference for the audio-based component of our multi-modal approach.

1. **Oramas et al. (Use of Album Reviews for Genre Classification)**

Oramas et al. explored the potential of textual information for music genre classification by incorporating album reviews into the classification process. Their work highlighted the value of natural language content, demonstrating that reviews and textual descriptions provide semantic insights that complement audio features. By processing and embedding these textual inputs, Oramas et al. achieved notable improvements in classification accuracy, especially in genres where lyrical and contextual cues are prominent. This study underscores the significance of textual features in music classification and motivates the inclusion of lyrical analysis in our multi-modal framework.

1. **GTZAN Genre Collection (Audio Dataset)**

The GTZAN Genre Collection is a widely used benchmark dataset for music genre classification, containing 1000 audio tracks spanning 10 genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each track is 30 seconds long and stored in .wav format with a 22050 Hz sampling rate. This dataset has been the basis for several landmark studies in genre classification, including work by Choi et al., making it a key resource for developing and benchmarking the audio component of our multi-modal approach. Its balanced distribution across genres ensures fair evaluation of model performance.

1. **musiXmatch Dataset (Lyrics Dataset)**

The musiXmatch dataset, a component of the Million Song Dataset, contains lyrical data for a vast collection of songs. It provides access to lyrics that can be used as a textual representation of songs, enabling genre classification from a linguistic perspective. While the dataset does not have a direct one-to-one mapping with GTZAN audio tracks, it provides a rich source of lyrical data. For unmatched tracks, we will manually collect corresponding lyrics to ensure full coverage of the GTZAN dataset. The musiXmatch dataset supports the development of the lyrical analysis component of our multi-modal approach, offering contextual and semantic information that complements audio features

1. **Methodology**

Our study proposes a multi-modal deep learning approach for music genre classification by integrating both audio and lyrical features. This methodology is designed to overcome the limitations of single-modality models and to leverage the complementary nature of these two information sources. The key aspects of our methodology are categorized into three main components: Components, Datasets, and Approach.

* 1. **Components**

The architecture of our multi-modal model is structured into four key components:

* **Audio Feature Extraction**

**Input**: Raw audio tracks from the GTZAN dataset (.wav files at 22050 Hz).

**Preprocessing**: Audio tracks are transformed into mel-spectrograms and Mel-frequency cepstral coefficients (MFCCs) using the librosa library. Mel-spectrograms provide a time-frequency representation of audio, while MFCCs capture perceptually relevant features, often used in speech and music processing.

**Feature Extraction Model**: We implement a convolutional neural network (CNN), inspired by the architecture proposed by Choi et al. The CNN extracts high-level temporal and spectral features from the mel-spectrograms, enabling the model to capture genre-specific audio patterns.

* **Lyrical Feature Extraction**

**Input**: Lyrics corresponding to the GTZAN audio tracks, obtained from the musiXmatch dataset.

**Preprocessing**: Lyrics are tokenized and encoded using a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model. BERT allows for context-aware embedding of words, ensuring that lyrical meanings are captured effectively.

**Feature Extraction Model**: The BERT model processes each set of song lyrics, generating a fixed-length embedding that encodes semantic and contextual information. This feature vector serves as the input for the next stage of the model.

* **Fusion Layer**

**Input**: Feature vectors from the CNN (audio) and BERT (lyrical) modules.

**Fusion Strategy**: A concatenation-based fusion layer merges the feature vectors from the audio and lyrical components. This fused representation captures the combined knowledge from both modalities. Alternatively, advanced fusion techniques like attention-based fusion may be explored to weight the relative importance of audio and lyrical features.

* **Classification Layer**

**Input**: Fused multi-modal feature vector.

**Architecture**: The concatenated features are passed through fully connected dense layers with ReLU activation, followed by a softmax output layer to predict the music genre.

**Output**: A probability distribution over the 10 possible genres (blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock).

* 1. **Datasets**

Our approach utilizes two key datasets to provide comprehensive coverage of both audio and lyrical information.

* **GTZAN Genre Collection**
* **Audio Data:** The GTZAN dataset contains 1000 30-second audio tracks distributed across 10 music genres (blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock).
* **Format:** The files are stored in .wav format with a sampling rate of 22050 Hz.
* **Usage:** This dataset serves as the primary source for the audio component of our model, with each track being converted into a mel-spectrogram and MFCCs for CNN-based processing.
* **musiXmatch Dataset**
* **Lyrics Data:** The musiXmatch dataset, which is part of the Million Song Dataset, contains lyrical content for a large collection of songs.
* **Matching with GTZAN:** Since musiXmatch does not have a one-to-one correspondence with the GTZAN tracks, efforts are made to identify and map lyrics for each GTZAN track. For tracks that cannot be matched, manual collection of the lyrics is undertaken to ensure full coverage of the GTZAN dataset.
* **Usage:** Lyrics from this dataset are tokenized and embedded using BERT to generate semantic feature vectors for the lyrical component of the multi-modal model.
  1. **Approach**

The approach for building the multi-modal model follows a series of well-defined steps, starting from data preprocessing to model evaluation.

* **Data Preprocessing:**
* **Audio Preprocessing:**
  + Load audio tracks from the GTZAN dataset.
  + Convert each track into **mel-spectrograms** and **MFCCs** using the librosa library.
  + Normalize the spectrograms and coefficients to ensure consistent input for the CNN.
* **Lyrics Preprocessing:**
  + Extract corresponding song lyrics from the musiXmatch dataset or collect lyrics manually where no match is available.
  + Tokenize and encode the lyrics using a **pre-trained BERT model** to obtain fixed-length embeddings.
  + Clean text to remove unnecessary symbols, stopwords, and non-informative content.

* **Model Design**
* **Audio Stream (CNN):**
  + Use a **3-layer convolutional neural network (CNN)** to process the mel-spectrograms and extract hierarchical audio features.
  + Employ **batch normalization** and **dropout** to improve generalization and prevent overfitting.
  + Extract the final feature vector from the flattened CNN layer to be used in the fusion step.
* **Lyrics Stream (BERT):**
  + Use the **pre-trained BERT model** to convert tokenized lyrics into contextualized embeddings.
  + Use the final BERT embedding as the lyrical feature vector for the fusion step.
* **Fusion Strategy**
* **Concatenation-based Fusion:** The extracted feature vectors from the CNN (audio) and BERT (lyrical) components are concatenated into a single, combined feature vector.
* **Attention-based Fusion (optional):** In addition to concatenation, an attention mechanism may be applied to assign different weights to audio and lyrical features, allowing the model to focus more on the most relevant modality for each song.
* **Classification Layer**
* The fused feature vector is passed through **fully connected dense layers** with ReLU activation to create a robust feature space for classification.
* **Softmax activation** is used in the final layer to predict the probability distribution over 10 possible genres.
* **Training and Evaluation**
* **Training:** The model is trained using **5-fold cross-validation** to ensure a reliable evaluation of performance. This method splits the dataset into five subsets, using four for training and one for testing, rotating the roles of the subsets.
* **Loss Function:** We use **categorical cross-entropy** as the loss function, as it is well-suited for multi-class classification tasks.
* **Optimizer:** The Adam optimizer is used for efficient convergence, with learning rate decay to prevent overfitting.
* **Evaluation Metrics:** Classification accuracy is the primary metric, but we also track precision, recall, and F1-score to better understand performance across genres.
* **Model Analysis**
* **Baseline Models:** Two baseline models (audio-only and lyrics-only) are created to compare with the multi-modal model. The audio-only model consists of the CNN-based architecture, while the lyrics-only model uses the BERT-based architecture.
* **Attention Visualization:** Attention maps are generated to visualize which audio and lyrical features are most influential in genre classification. This analysis provides insights into the importance of audio and lyrical information in predicting different genres.
* **Genre-Specific Analysis:** We evaluate the model's performance for each of the 10 genres separately, investigating how well the model can differentiate between genres like classical and hip-hop, which have distinct characteristics.