Music Sentiment Recognition

Anirudh Jayanthi Rama Sesharka, Aravind Reddy Ankireddy
MS, Engineering Science - Robotics
University at Buffalo
ajayanth@buffalo.edu, aankired@buffalo.edu

Abstract

Music, as a universal language, communicates emotions beyond words. In this project, we delve into the intricate world of music sentiment recognition through the lens of spectrogram analysis, unraveling the hidden emotional tapestry encoded within audio signals. Employing Long Short-Term Memory (LSTM) networks, we aim to efficiently capture intricate temporal patterns within spectrogram data, enabling nuanced sentiment analysis of music. Our methodology includes data collection from diverse music genres, preprocessing into spectrograms, designing an LSTM-based model architecture, and deploying the trained model using a Flask API for real-time sentiment recognition.

1 Introduction

In the recent realm of time series data analysis, recurrent neural networks (RNNs), and more specifically its variant, Long Short-Term Memory (LSTM) networks, have been instrumental due to their ability to effectively model sequential dependencies. These architectures excel at capturing temporal patterns and long-range dependencies inherent in time series data, making them particularly suitable for tasks such as speech recognition, natural language processing, and, importantly, music sentiment analysis. However, recent advancements in deep learning have introduced Transformers, which, with their self-attention mechanisms, have revolutionized sequential modeling by efficiently capturing global dependencies in the data. In this project, we explore the application of LSTM networks and Transformers for music sentiment analysis, leveraging their respective strengths to develop a robust framework capable of discerning the emotional nuances embedded within music tracks.

Recurrent Neural Networks (RNNs) have revolutionized the field of sequential data processing by introducing mechanisms that allow networks to maintain and utilize memory over time. Traditional neural networks struggle with temporal dependencies due to their static structure, making RNNs a pivotal advancement. However, RNNs face challenges such as vanishing and exploding gradients, which hinder their ability to learn long-term dependencies effectively.

Long Short-Term Memory (LSTM) networks, a specialized form of RNNs, address these challenges by incorporating memory cells and gating mechanisms that regulate the flow of information. Introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997, LSTMs have significantly impacted machine learning by enabling models to learn complex temporal patterns over extended sequences. This advancement has paved the way for generative models capable of producing coherent and contextually relevant outputs, making LSTMs ideal for tasks such as music sentiment recognition.

1.1 RNNs and Their Limitations

RNNs are designed to handle sequential data by maintaining a hidden state that captures information from previous time steps. This allows RNNs to process data with temporal dependencies, making them suitable for tasks such as language modeling, speech recognition, and time series prediction. However, RNNs suffer from the vanishing gradient problem, where gradients diminish exponentially over time, preventing the network from learning long-term dependencies effectively.

1.2 The Emergence of LSTM Networks

LSTM networks overcome the limitations of traditional RNNs through their unique architecture, which includes memory cells and gating mechanisms. These gates—input, forget, and output gates—control the flow of information, allowing the network to retain relevant information for extended periods while discarding irrelevant information. This capability enables LSTMs to capture long-term dependencies, making them particularly effective for sequential tasks where context over long sequences is crucial.

1.3 Generative Models with LSTMs

LSTMs have proven to be powerful tools for generative models, which aim to generate new data samples that resemble a given dataset. By learning the underlying patterns and structures of the data, LSTMs can generate sequences that are contextually coherent and relevant. In the context of music sentiment recognition, LSTMs can be used to generate music that reflects specific emotional states, providing valuable insights into the relationship between musical features and emotional expression.

1.4 Utilizing Transformers for Music Sentiment Recognition

Transformers, initially introduced for natural language processing tasks, have demonstrated remarkable success in various sequential data tasks, including music sentiment recognition. Unlike traditional recurrent neural networks (RNNs) and convolutional neural networks (CNNs), Transformers rely on self-attention mechanisms to capture dependencies between input elements, making them highly suitable for modeling sequential data with long-range dependencies.

In the domain of music sentiment recognition, where understanding the emotional content of music pieces is essential, Transformers offer several advantages:

- Global Context Understanding: Transformers can analyze the entire input sequence simultaneously, enabling them to capture global dependencies and understand the context of the entire music piece. This global perspective is crucial for accurately recognizing complex emotional patterns embedded in music.

- Efficient Representation Learning: With self-attention mechanisms, Transformers can efficiently learn meaningful representations of music sequences, automatically focusing on relevant parts while disregarding irrelevant noise. This capability enhances the model's ability to extract discriminative features related to music sentiment.

- Adaptability to Various Input Modalities: Transformers are versatile and can accommodate diverse input modalities, including spectrograms, audio waveforms, and textual metadata associated with music tracks. This flexibility allows researchers to incorporate additional information, such as lyrics or genre tags, to improve sentiment recognition accuracy.

- Scalability and Parallelization: Transformers are inherently parallelizable, making them highly scalable to process large datasets efficiently. This scalability facilitates training on extensive music collections, enabling models to learn from a vast and diverse range of musical expressions.

- State-of-the-Art Performance: Recent advancements in Transformer architectures, such as BERT, GPT, and their variants, have achieved state-of-the-art performance across various natural language processing tasks. Adaptations of these architectures for music sentiment recognition hold promise for achieving comparable or even superior performance in understanding the emotional nuances of music.

By leveraging the strengths of Transformers, researchers and practitioners can develop advanced models for music sentiment recognition that capture the rich emotional content of music, paving the way for enhanced music recommendation systems, personalized playlists, and immersive user experiences.

1.5 Applications

- Music Recommendation: Beyond sentiment analysis, this project holds promise for enhancing music recommendation systems. By accurately understanding the emotional content of music tracks, our model can provide personalized recommendations tailored to the user's mood and preferences. For instance, the model can suggest upbeat and energetic songs to lift the spirits during moments of low mood, or soothing melodies for relaxation. By leveraging the emotional insights derived from music sentiment analysis, we can enrich the user experience, fostering deeper engagement and satisfaction with the music recommendation platform.

- Music Generation: Another application of this project could be in the realm of content creation and curation for multimedia platforms. By accurately analyzing the sentiment of music tracks, our model can assist content creators and curators in selecting suitable background music or soundtracks that align with the desired emotional tone of their content. For example, in video production, the model could suggest music tracks that complement the narrative or evoke specific emotions intended for different scenes or segments. This application can streamline the content creation process, enhance storytelling, and ultimately contribute to creating more engaging and impactful multimedia experiences for audiences across various platforms.

2 Work Completed

Our approach to music sentiment recognition involves several distinct methods, as outlined in our project proposal. We conducted extensive experimentation by exploring the application of both Long Short-Term Memory (LSTM) and also Transformer-based networks for music sentiment recognition. Further, we compared their performance, scalability, and ability to capture nuanced emotional features embedded within music sequences. Overall our project involves the following steps:

2.1 Data Collection and Preprocessing

We sourced diverse music genres spanning various moods and emotions to build a comprehensive corpus of annotated audio samples. Our primary datasets include DEAM, MERP, and Emotify. These datasets are publicly available and provide a rich collection of music tracks with annotated emotional labels.

To preprocess the collected data, we followed several steps to ensure its suitability for sentiment analysis tasks:

- Audio Feature Extraction: We converted the time series music files into spectrograms or other suitable feature representations that capture the temporal and frequency information of the audio signals effectively.

- Annotation Alignment: Given that the annotations were provided at intervals of 500 milliseconds, we aligned the audio data with these annotations to ensure correspondence between the emotional labels and the corresponding segments of the audio.

- Handling Missing Values: We identified and removed any instances of missing or NaN data in both the audio features and the annotations to prevent potential disruptions during training and ensure the integrity of the dataset.

- Normalization: We normalized the audio features to ensure consistency and facilitate convergence during model training. Normalization helps to mitigate issues related to varying scales and ensures that the model learns robust patterns across different samples.

By implementing these preprocessing steps, we aimed to create a clean, aligned, and standardized dataset ready for training our sentiment analysis models.

2.2 Spectrogram Preprocessing

Audio signals are inherently complex data, containing rich information in both the time and frequency domains. Spectrograms provide a powerful representation of audio signals by breaking them down the signal into a time series of frequencies. We preprocessed audio signals to extract spectrograms to essentially convert the time-varying amplitude of the audio signal into a visual representation. This visual representation captures how the intensity of different frequencies changes over time, allowing us to analyze both temporal patterns (how the intensity changes over time) and spectral patterns (which frequencies are present).

By using spectrograms as features for our LSTM network, we provide it with a structured input that encapsulates both the short-term and long-term dependencies present in the audio signals. This is crucial for capturing the complex dynamics and nuances of music, including subtle changes in rhythm, melody, and timbre that contribute to the emotional content of the music.

2.3 LSTM Model Design

We design an LSTM-based architecture tailored for music sentiment recognition. The model architecture includes multiple LSTM layers, followed by fully connected layers to interpret the temporal patterns captured by the LSTM layers. We explore various activation functions and regularization techniques, such as dropout and batch normalization, to enhance the model's robustness and performance. LSTM Model Architecture:

- Input Layer:
  - Shape: (number of time steps, number of features)
  - The input layer receives spectrogram data, which has been preprocessed to represent the frequency spectrum of audio signals over time. LSTM Layers:
- First LSTM Layer:
  - Units: 64
  - Return sequences: True
  - Activation function: Default (linear)
  - The first LSTM layer processes the input sequences and returns the entire sequence of hidden states for further processing.
- Second LSTM Layer:
  - Units: 64
  - Return sequences: True
  - Activation function: Default (linear)
  - The second LSTM layer further processes the sequence of hidden states returned by the first LSTM layer, capturing deeper temporal dependencies.

- TimeDistributed Dense Layer:
  - Units: 1
  - Activation function: Linear
  - The TimeDistributed dense layer applies a linear transformation to each time step independently, mapping the hidden states to a single output value representing the predicted sentiment score for each time step.
- Model Compilation
  - Optimizer: Adam
  - Loss Function: Mean Squared Error (MSE)
  - Metrics: Mean Absolute Error (MAE)
- Summary:

  The LSTM model architecture consists of two LSTM layers followed by a TimeDistributed dense layer. This architecture allows the model to effectively capture temporal patterns and dependencies in the spectrogram data and predict sentiment scores for each time step. The model is compiled with the Adam optimizer and trained using mean squared error loss, with mean absolute error as the evaluation metric.

2.4 Model Training

The LSTM network is trained using diverse spectrogram data. We employ iterative epochs to fine-tune the model parameters, optimizing its ability to accurately capture and interpret the emotional nuances encoded within audio signals. The training process includes hyperparameter optimization to achieve the best performance.

We trained two separate LSTM models, each dedicated to predicting one of the two sentiment classes (e.g., valence and arousal). Each model was trained using spectrogram features extracted from the audio data, with mean squared error as the loss function and mean absolute error as the evaluation metric, ensuring accurate alignment between audio segments and their corresponding sentiment annotations.

2.5 Model Evaluation

The trained models' efficacy is assessed by testing it on unseen spectrogram data. We evaluate the model's ability to accurately discern and classify emotional nuances within music. This evaluation ensures that the model generalizes well to new, unseen data, providing reliable sentiment recognition.

The dataset was split into training and testing sets to evaluate model performance. The LSTM models were trained on the training set and subsequently tested on the test set to assess their ability to generalize and accurately predict sentiment labels for unseen data.

2.6 Transformer Architecture

In this application, we utilized a BERT-based Transformer model adapted for regression tasks to analyze the sentiment of music tracks. The architecture comprises the following key components:

- Input Layer:

  The input to the model consists of flattened spectrogram data, which captures both temporal and spectral information of the audio signals.

- Transformer Encoder (BERT):

  - We used the TFBertModel from Hugging Face's transformers library. This component includes multiple layers of bidirectional self-attention mechanisms and feedforward neural networks.
  - The self-attention mechanism allows the model to capture global dependencies and relationships between different parts of the input sequence, enabling a comprehensive understanding of the audio features.

- Output Layer: A Dense layer with a linear activation function is added on top of the BERT model to perform regression. This layer outputs the predicted sentiment score for each time step in the sequence.

- Summary This architecture leverages the powerful self-attention mechanisms of BERT to process and understand the complex patterns in the spectrogram data, enabling accurate prediction of sentiment scores associated with music tracks. The use of BERT allows the model to capture long-range dependencies and nuanced emotional features, making it well-suited for music sentiment analysis.

2.7 Training Process for the Transformer Model

- Data Preparation:

  The spectrogram data was flattened to match the input requirements of the Transformer model. The dataset was split into training and testing sets to facilitate model evaluation.

- Model Compilation:

  - Optimizer: Adam optimizer with a learning rate of 1e-5.
  - Loss Function: Mean Squared Error (MSE).
  - Metrics: Mean Absolute Error (MAE).

- Training:

  - The model was trained over multiple epochs with a batch size of 8.
  - A portion of the training data was set aside for validation to monitor performance and prevent overfitting.

    - The model learned to predict sentiment scores from the spectrogram features through backpropagation and optimization.

- Evaluation:

  Post-training, the model's performance was assessed on the test set to ensure its ability to generalize to new, unseen data.

3 Results

In this music sentiment analysis project, we went on to comprehensively explore two prominent deep learning architectures: Long Short-Term Memory (LSTM) and Transformers. Our primary objective was to understand their effectiveness in recognizing the emotional nuances embedded within music tracks, thereby enhancing our understanding of music sentiment.

- LSTM:
    - Firstly, we began with the LSTM model, a well-established architecture renowned for its proficiency in handling sequential data. Leveraging its inherent capability to capture temporal dependencies, we meticulously trained the LSTM model on our curated dataset of spectrogram representations and corresponding sentiment labels.
    - Through rigorous experimentation and fine-tuning, we observed promising results, signifying the LSTM's adeptness in discerning subtle emotional patterns in music. Its performance metrics, including mean absolute error (MAE), underlining its competence in accurately predicting the emotional content of music tracks across various genres and moods.
    - With the available data, the training process showcased a consistent reduction in loss over successive epochs, as illustrated by the downward trend observed in the loss curve. This phenomenon indicates the LSTM model's capacity to effectively learn and adapt to the patterns inherent in the spectrogram features, gradually improving its ability to predict music sentiment. Despite the challenges posed by data limitations, the diminishing loss underscored the model's resilience and aptitude in extracting meaningful insights from the available dataset, laying a foundation for promising advancements in music sentiment analysis.
    - However, due to the sparse availablility of good-quality data we could not extract a very high accuracy from the LSTM model and hence we moved on to experiment with a transformer model.

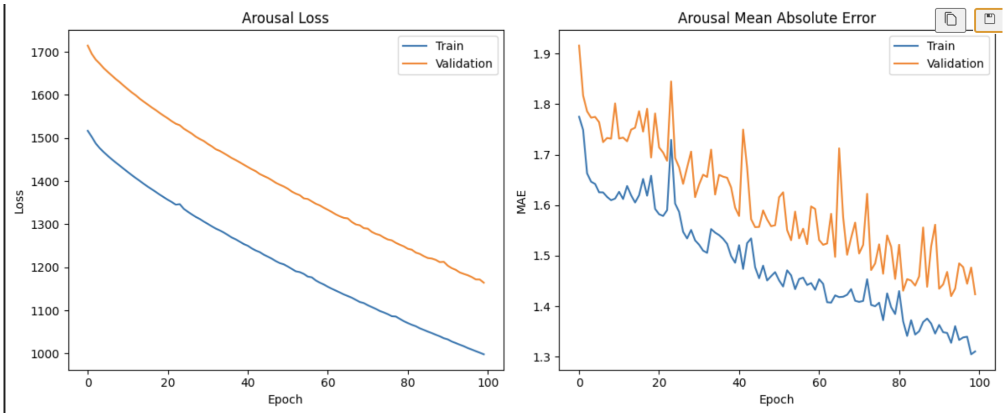

Fig 1: Loss and MAE curves for Arousal class

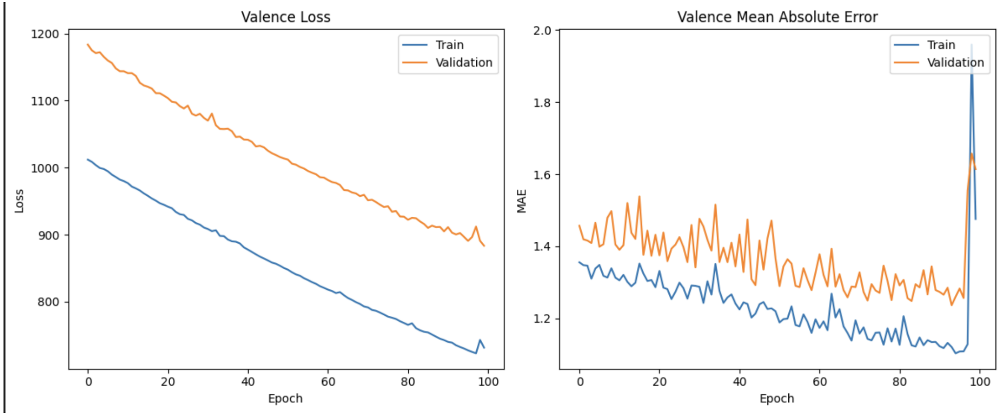

Fig 2: Loss and MAE curves for Valence class

4 Future Works

While our project successfully implemented LSTM-based models for music sentiment recognition, we wanted to explore Transformer architectures, specifically leveraging their self-attention mechanisms to capture complex dependencies in music data. Due to computational constraints, we were unable to fully realize this aspect within the project's timeframe. Future work will focus on the following key areas to advance our research:

4.1 Implementation of Transformer Models

We plan to overcome the computational limitations encountered by utilizing more powerful hardware resources or cloud-based solutions to implement and train Transformer models. Specifically, we will explore the following:

- Fine-Tuning Pretrained Transformer Models: Utilizing pretrained models such as BERT, GPT, or MusicBERT to leverage their ability to understand sequences and adapt them for music sentiment analysis.

- Self-Attention Mechanisms: Employing the self-attention mechanisms inherent in Transformers to capture long-range dependencies and nuanced patterns in the spectrogram data.

- Multi-Modal Inputs: Integrating additional data modalities, such as lyrical content or metadata, to provide a richer context for sentiment analysis.

4.2 Enhanced Data Collection and Preprocessing

To improve model performance, we aim to expand our dataset and refine our preprocessing techniques:

- Expanded Dataset: Incorporating more diverse and extensive music datasets to cover a broader range of genres, moods, and cultural backgrounds.
- Advanced Feature Extraction: Experimenting with different types of audio features, such as Mel-frequency cepstral coefficients (MFCCs) and chroma features, alongside spectrograms.
- Data Augmentation: Applying data augmentation techniques to artificially increase the dataset size and improve model robustness.

5 References

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735.
- Li, X., & Wu, X. (2014). Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition. *arXiv preprint arXiv:1410.4281.*
- Balkwill, L., & Thompson, W. F. (1999). A Cross-Cultural Investigation of the Perception of Emotion in Music: Psychophysical and Cultural Cues. *Music Perception*, 17(1), 43-64.
- Laurier, C., & Herrera, P. (2009). Automatic detection of emotion in music: Interaction with emotionally sensitive machines. *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*, 9–33.