Attention-guided Spectrogram Sequence Modeling with CNNs for Music Genre Classification

AI2100 Deep Learning | Course Project Presentation

Aditya Sridhar
ES21BTECH11004

Abstract

Music, an integral part of human culture, encompasses diverse genres, each carrying unique sonic and cultural characteristics. Understanding music and its genres is fundamental to human expression, communication, and cultural identity.

The proposed architecture for music genre classification processes and ***splits spectrogram images of music pieces into tokens***, encoding them through a ***CNN for subsequent attention-based analysis***. Experimental validation demonstrates ***competitive classification accuracy***, while interpretability uncovers ***intriguing genre-specific encoding patterns.***

In the short term, with adequate resources, I aim to optimize model performance and ***explore diverse applications***. Additionally, we aspire to expand the model's utility towards ***mood and emotion understanding and recommendation systems, as well as better music generating systems***. Ultimately, our long-term vision is to deepen our understanding of music's impact on human experiences, fostering advancements in music-related technologies and cultural discourse.

Goals of the Project

Short-Term Goals

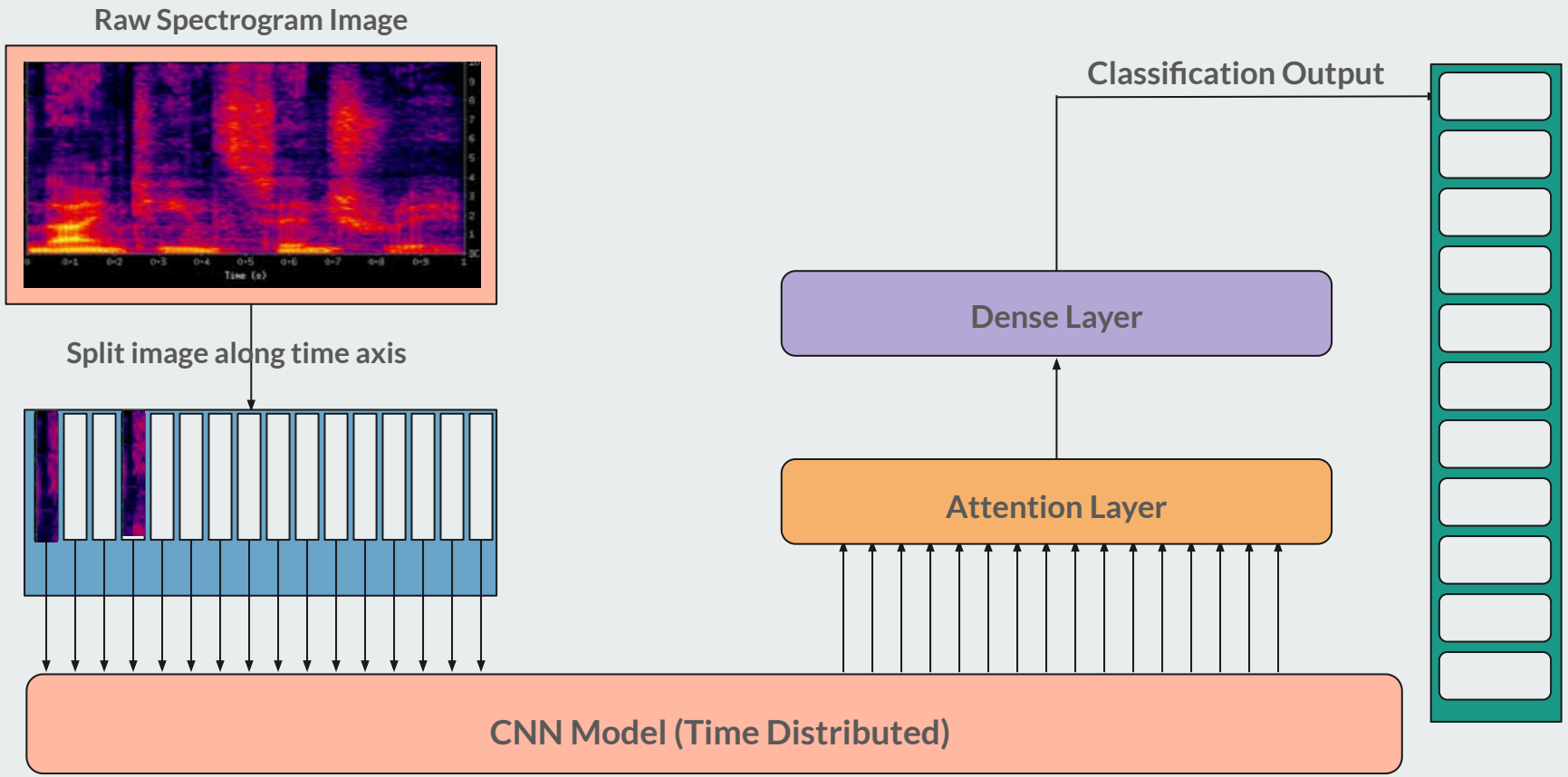
1. Enhance classification accuracy by refining model architectures and optimizing training processes.
2. Analyze and interpret the encoding patterns of music to gain insights into genre-specific characteristics.
3. Investigate the temporal dynamics of music to discern which segments contribute most significantly to genre distinctions.

Long-Term Goals

1. Develop advanced recommendation systems that leverage nuanced understandings of music genres and listener preferences.
2. Enhance music generation techniques by utilizing improved encodings to create more diverse and compelling compositions
3. Foster a deeper understanding of music's emotional and mood-related dimensions to enrich human experiences and cultural discourse.

Part I
The Model Architecture

Proposed Model Architecture



Reference Papers	Respective Proposed Architectures	Shortcomings
• **Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification (2018)**	This model uses a simple Convolutional Neural Network (CNN) to classify short segments of music waveforms.	Temporal Dependencies are not captured.
• **Music Genre Classification with Transformer Classifier (2020)**	This model utilizes transformer-based models to perform music genre classification.	CNNs, which are proved to have much better accuracy, have not been used in this architecture.
• **Deep attention based music genre classification (2019)**	This model integrates a CNN with NetVLAD and self-attention to capture local information across levels and learn their long-term dependencies	This model considers the entire spectrogram image as a unified sequence and aim to identify influential “spatial regions” for each genre without segmenting the images into sequences along the time axis to examine the temporal significance of music segments on genre classification.

Advantages of Proposed Architecture Over Existing Models

- **Capturing Temporal Dependencies**: Unlike the CNN-based model from 2018, our architecture incorporates an attention mechanism that captures temporal dependencies by processing sequences of spectrogram images, enabling the model to understand the context and evolution of music segments over time.
- **Segmentation for Contextual Understanding**: By segmenting spectrogram images into sequences along the time axis, our architecture offers a nuanced understanding of the temporal significance of music segments on genre classification. This approach contrasts with the 2019 and 2024 models, which treat the entire spectrogram as a unified sequence, potentially overlooking crucial temporal cues.
- **Leveraging CNN and Attention**: While the 2020 transformer-based model lacks CNNs, my proposed architecture combines the strengths of CNNs for local feature extraction with the attention mechanism for capturing long-term dependencies.
- **Visualizing Genre Differences:** The trained model facilitates visualization of similarities and differences between genres using dimensionality reduction techniques. This capability allows for a deeper exploration of genre characteristics and aids in understanding the underlying patterns contributing to genre classification accuracy.

Part II
Experiments and Results

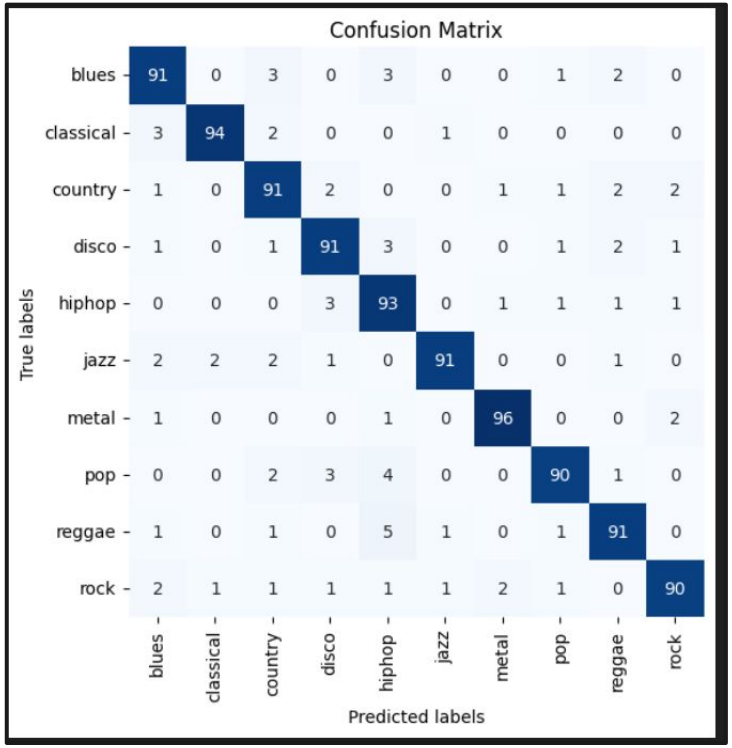
Section I : Classification Accuracy

2.1 | Experiments and Results

Training and Experiments were done on the *GTZAN Dataset* (obtained from Kaggle): The GTZAN dataset is a collection of 1,000 audio tracks, each 30 seconds long, used for musical genre classification. It encompasses 10 distinct genres, each represented by 100 tracks, with all tracks being 22050 Hz Mono 16-bit audio files in WAV format.

During experimentation, the model achieved a notable classification accuracy of **91.90%**. This performance, while near the accuracy of more resource-intensive models with access to superior computational resources, underscores the efficacy of the proposed architecture in handling complex musical data.

The confusion matrix provides a comprehensive overview of the model's performance across different genres, offering insights into its strengths and potential areas for improvement.



Part II
Experiments and Results

Section II : CNN Genre Encoding Analysis

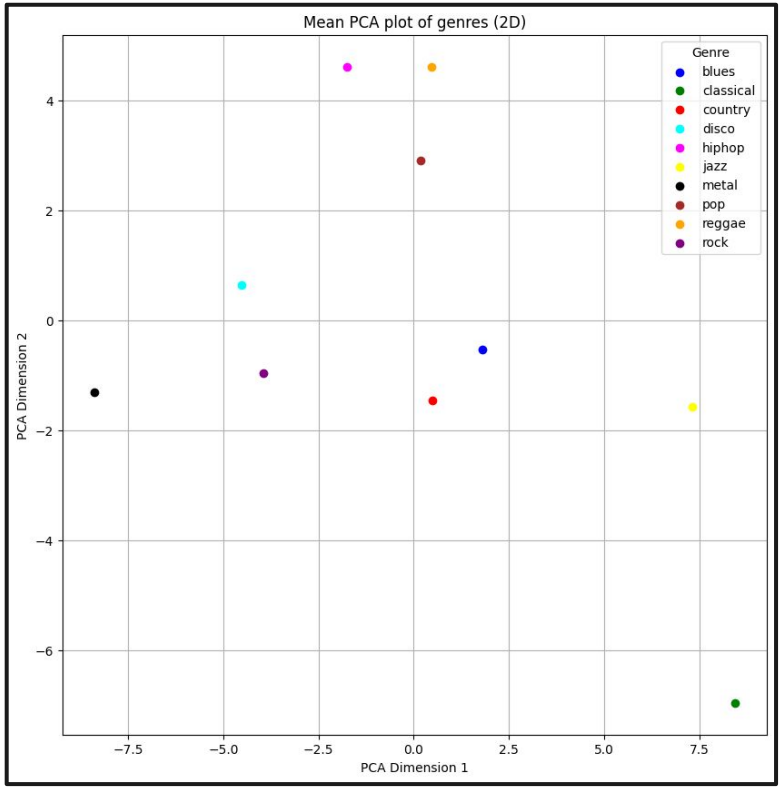
2.2 | Experiments and Results

Dimensionality Reduction using Principal Component Analysis (PCA) was applied to the output embeddings of the ***CNN model, which generates vectors of size 128***. This technique transformed the high-dimensional CNN output vectors into a **two-dimensional space**.

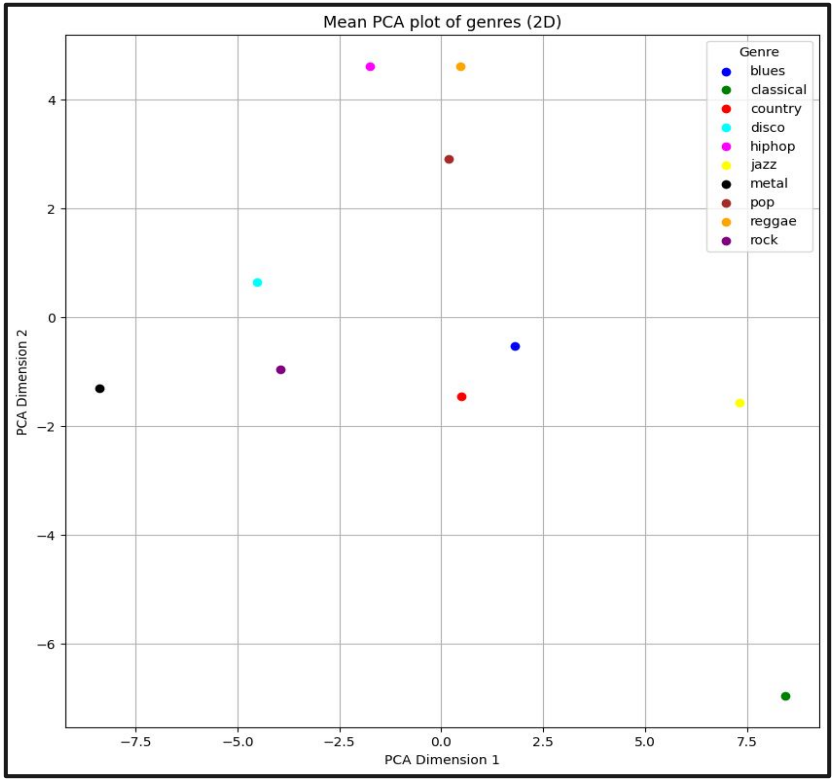
Plotting the **means** of the reduced 2-dimensional CNN output vectors for each genre on a graph provided valuable insights into the similarities and differences between genres:

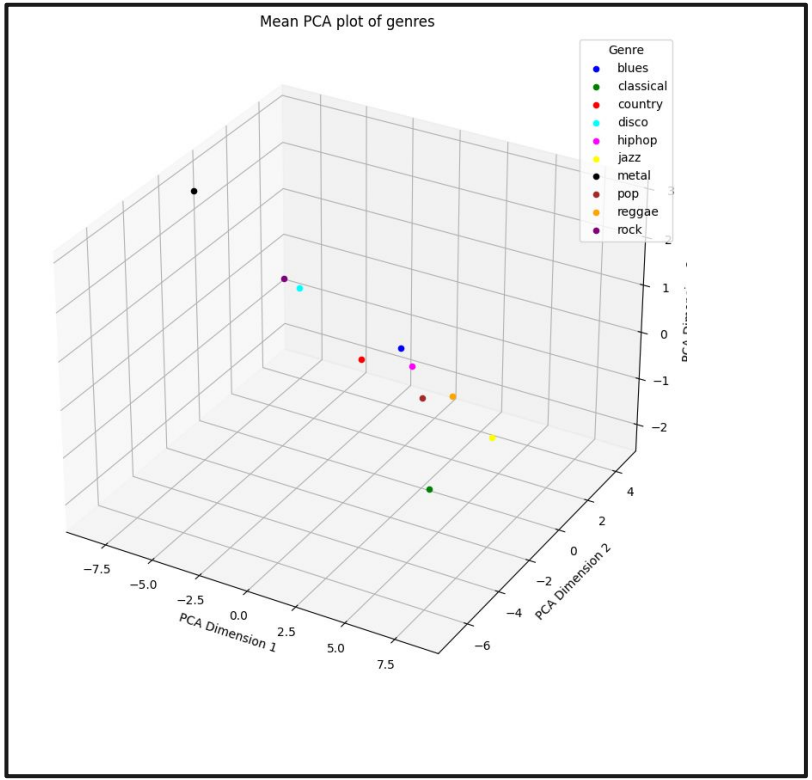
a. ***Hip-hop, pop, and reggae genres appeared closer to each other*** in the reduced space, suggesting similarities in their encoded features.
b. ***Classical music emerged as the furthest genre*** from most others, indicating distinct characteristics that set it apart.
c. ***Rock and Metal genres exhibited proximity to each other,*** implying shared attributes in their embeddings.
d. ***Blues and Country genres demonstrated similarity***, reflecting commonalities in their encoded representations.

All these observations that we have seen mathematically are also ***intuitively true to the human ear***. Such concordance between mathematical analysis and human perception reinforces the validity and reliability of the proposed architecture for music genre classification.



2.2 | Experiments and Results





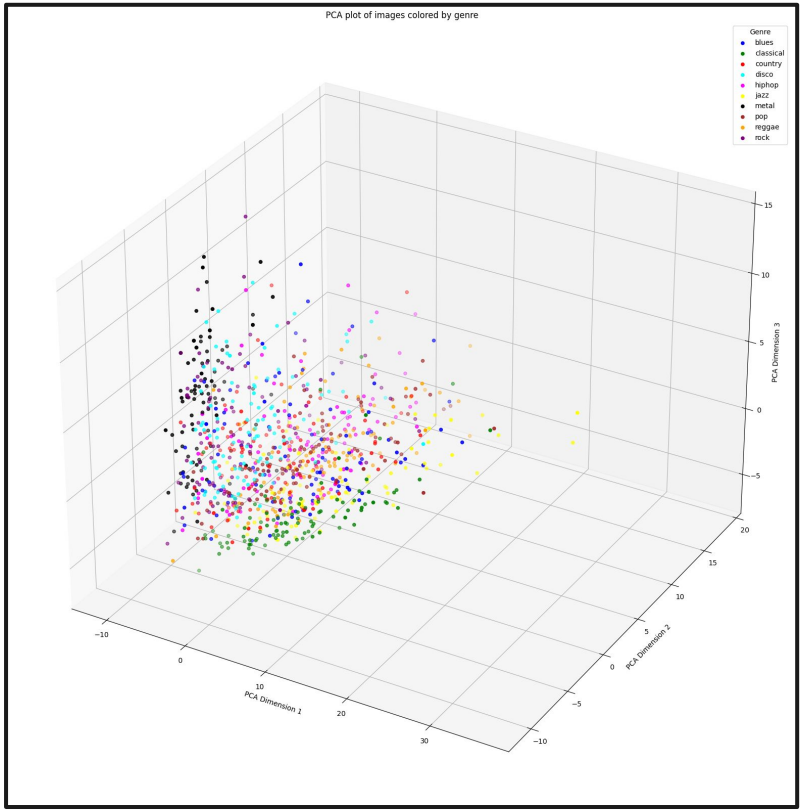
2.2 | Experiments and Results

Dimensionality Reduction using Principal Component Analysis (PCA) was applied to the output embeddings of the ***CNN model, which generates vectors of size 128***. This technique transformed the high-dimensional CNN output vectors into a **three-dimensional space.**

The vectors representing each music piece were plotted on a graph for visual analysis, although the sheer volume of data points limited clarity. Despite this, several discernible patterns emerged (only two have been mentioned):

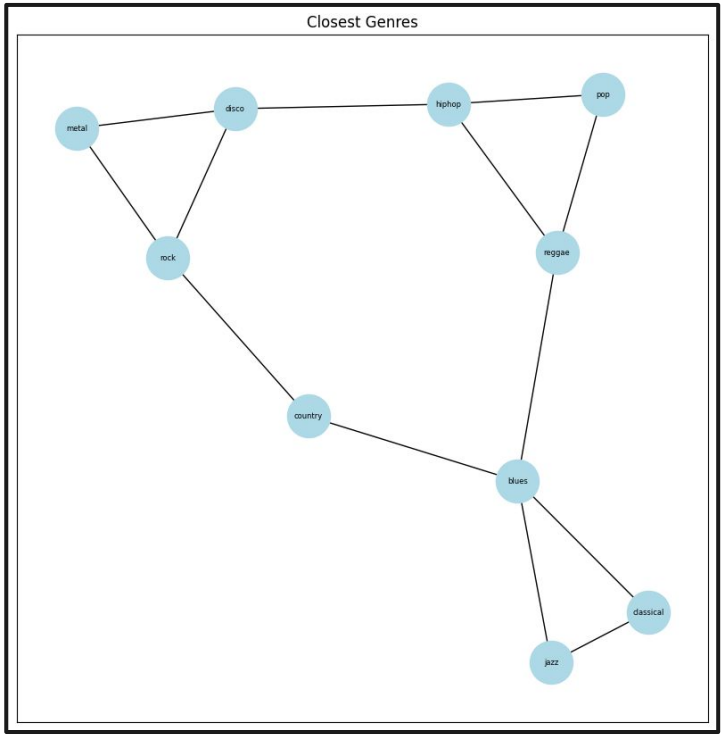
a. Blues and Metal genres exhibited **distinct directions** on the graph. ***Blues music tended to cluster towards one end of PCA Dimension 1***, while ***Metal music leaned more towards PCA Dimension 3***.
b. **Blues music** appeared closely associated with ***Classical, Jazz, and Country*** genres, which are characterized by **slower tempos**. This could suggest that **PCA Dimension 1 may heavily reflect the tempo or overall "pace" of the music.**

Again, the results seem fairly consistent with the known information about music genres.



2.2 | Experiments and Results

```
The closest genres to blues are country and reggae
The closest genres to classical are jazz and blues
The closest genres to country are blues and rock
The closest genres to disco are rock and hiphop
The closest genres to hiphop are pop and reggae
The closest genres to jazz are classical and blues
The closest genres to metal are rock and disco
The closest genres to pop are hiphop and reggae
The closest genres to reggae are hiphop and pop
The closest genres to rock are disco and country
```



2.2 | Experiments and Results

Blues & Country/Reggae
- Shared musical elements like rhythm or melodic motifs contribute to the proximity between blues and country/reggae.

Classical & Jazz/Blues:
- Similar harmonic complexity or structural elements make classical closely related to jazz and blues.

Country & Blues/Rock:
- Blending rhythmic and instrumental characteristics with blues and rock defines the proximity of country music.

Disco & Rock/Hiphop:
- Common rhythmic and percussive elements shape the close relationship between disco and rock/hiphop.

Hip Hop & Pop/Reggae:
- Shared stylistic elements in beats, instrumentation, or lyrical themes contribute to the proximity of hip-hop with pop and reggae.

2.2 | Experiments and Results

Analyzing the equations derived from PCA values (3-D) involving music genres provides interesting insights into the relationships and compositional elements that define these genres:

```
Based on the PCA plot, the
following equations were found:
1. blues-country+pop = reggae
2. blues-classical+jazz = hiphop
```

1. **Blues - Country + Pop = Reggae:**
   - ***Blues:*** Rooted in African-American folk music, emphasizing soulful expression and rhythmic patterns.
   - ***Country:*** Characterized by storytelling, acoustic instrumentation (e.g., guitar, banjo), and folk melodies.
   - ***Pop:*** Often features catchy hooks, electronic elements, and upbeat rhythms.
   - ***Reggae:*** Known for its offbeat rhythms, emphasis on bass and percussion, and socio-political lyrics.
   - ***Subtracting the folk and storytelling aspects of country music*** from blues (**retaining its rhythmic and emotional core**) and adding the ***catchy hooks and upbeat rhythms of pop*** leads to an **outcome resembling reggae.**
   - This equation suggests that ***reggae*** incorporates ***elements of blues' rhythmic foundation***, while the ***influence of pop adds a more accessible and melodic dimension*** to the resulting genre.

2.2 | Experiments and Results

2. Blues - Classical + Jazz = Hip-hop:

- ***Blues:*** Known for its emotive lyrics, soulful melodies, and often a 12-bar structure.
- ***Classical:*** Characterized by complex harmonies, orchestral arrangements, and thematic development.
- ***Jazz:*** Emphasizes improvisation, syncopation, and swing rhythms.
- ***Hiphop:*** Defined by rhythmic beats, spoken word, and sampling.
- In this equation, subtracting the characteristics of classical music (***perhaps complexity and orchestral arrangements***) from blues (***soulful, rhythmic***) and adding the **improvisational and rhythmic elements of jazz** leads to an **outcome resembling hip-hop.**
- This suggests that ***hiphop*** may inherit elements of ***blues' emotive expression*** and ***jazz's improvisational nature*** while ***diverging from classical music's orchestral complexities.***

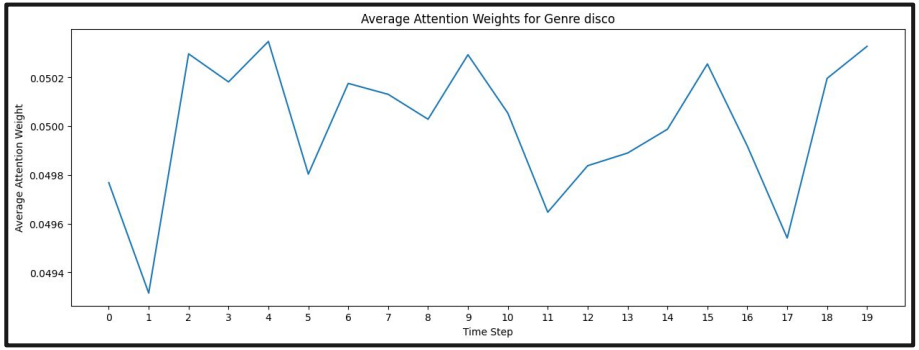
In summary, these equations provide a creative way to conceptualize genre influences based on their PCA values.

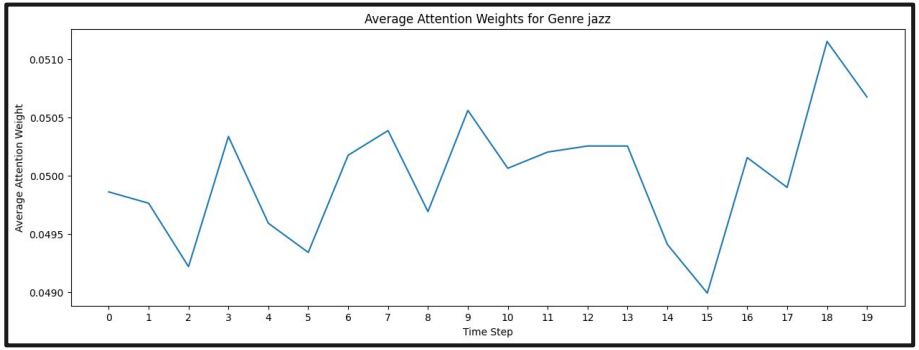
```
Based on the PCA plot, the
following equations were found:
1. blues-classical+jazz = hiphop
2. blues-country+pop = reggae
```

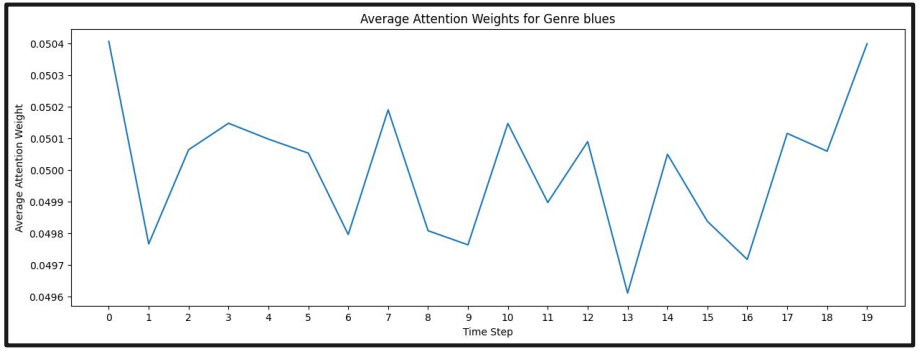
Part II
Experiments and Results

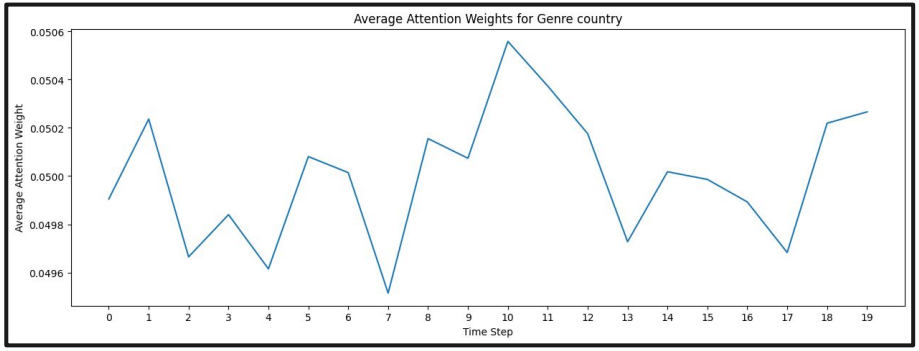
Section III : Temporal Attention Weights Analysis

2.3 | Experiments and Results









2.3 | Experiments and Results

```
Top 3 attention time step indices for disco: 2, 4, 19
Top 3 attention time step indices for hip-hop: 6, 16, 19
Top 3 attention time step indices for pop: 0, 5, 10
Top 3 attention time step indices for rock: 4, 18, 19
Top 3 attention time step indices for country: 10, 11, 19
Top 3 attention time step indices for jazz: 9, 18, 19
Top 3 attention time step indices for reggae: 11, 15, 16
Top 3 attention time step indices for blues: 0, 7, 19
Top 3 attention time step indices for metal: 10, 11, 19
Top 3 attention time step indices for classical: 7, 9, 16
```

2.3 | Experiments and Results

Common High-Attention Time Steps Across Genres:

- ***Time step 19*** appears frequently across ***multiple genres (disco, hip-hop, rock, country, jazz, blues, metal)***. This might indicate a ***climactic or characteristic moment*** in music pieces that resonates with a broad range of genres.
- Time steps **10 and 11** are also notable, appearing in ***country, reggae, metal, and pop***. This suggests that certain rhythmic or melodic elements around these time steps might be foundational or particularly defining in various genres.

Distinctive Attention Patterns for Specific Genres:

- Genres like **disco, hiphop, and rock** show attention around ***earlier time steps (e.g., steps 2, 4, 6)*** which could indicate a focus on ***rhythmic elements or specific beats characteristic of these genres.***
- **Pop music** displays attention at time steps ***0, 5, 10***, suggesting an emphasis on the *beginning and middle of the music piece*, potentially highlighting **catchy hooks or melodic phrases**.
- **Classical music** exhibits attention at ***later time steps (7, 9, 16)***, *potentially indicating a buildup* towards climactic moments or variations in thematic material.

Consistency and Variability:

- Some genres, such as **reggae and metal**, share attention at *similar time steps (e.g., 11, 16)* despite having different stylistic elements, indicating ***potential rhythmic or structural similarities.***
- **Jazz and blues** share some attentional ***focus at time steps 9 and 18***, which might reflect shared improvisational or thematic characteristics inherent to both genres.

Part III

Experiments to be Done

3 | Experiments to be Done

- **Mood Classification with Embeddings**: Utilize pre-trained embedding weights to classify music based on intuitive "moods," assessing the model's ability to generalize across different datasets without retraining.
- **Visualizing Temporal Dependencies**: Create additional visualizations to explore how music segments contribute to genre classification over time, providing insights into the relationship between music structure and genre.
- **Testing on Music Recommendation and Generation Systems (Long Term)**: Employ insights from previous experiments to assess the model's performance in real-world applications such as music recommendation and generation systems. This long-term goal aims to leverage the model's understanding of temporal dependencies and genre characteristics to enhance music-related technologies.

Part IV

Strategies to Increase Classification Accuracy

4 | Strategies to Increase Classification Accuracy

1. **Hyperparameter Tuning:**
   - The model's performance is ***highly sensitive*** to hyperparameters such as *regularization constants*, *learning rate schedules*, *dropout rates*, *sequence lengths*, *size of latent representation*, *and batch sizes.*
   - Conducting a ***thorough grid search*** to optimize these parameters can significantly enhance model performance.
2. **Increasing Dataset Size:**
   - With a **limited dataset size (100 pieces per genre)**, there's a ***risk of overfitting*.**
   - ***Scaling up the dataset by incorporating more samples per genre*** can lead to *better generalization* and **improved model robustness.**
3. **Utilizing Diverse Datasets:**
   - Employing the model architecture across ***multiple datasets*** of similar types can promote *better generalization*.
   - By training on a variety of data sources, the model can learn ***more comprehensive patterns*** and features associated with the target classification task.

4 | Strategies to Increase Classification Accuracy

4. **Regularization of Latent Space:**
    - Implementing a Variational Autoencoder (VAE)-like architecture to ***regularize encodings*** can enhance accuracy.
    - The hypothesis of this approach is that VAEs enforce a ***structured latent space*** compared to traditional autoencoders, potentially leading to **more meaningful representations** for classification tasks.
5. **Forcing Sparseness of Latent Space:**
    - Introducing sparsity constraints in the latent space can encourage the model to learn more robust and discriminative representations.
    - Sparse latent representations can highlight important features while filtering out irrelevant or noisy information, thereby improving classification accuracy.
6. **Variation of VAE Architecture:**
    - Exploring novel VAE architectures, such as predicting the ***next "token" or sequence element*** instead of reconstructing the same input, can be advantageous.
    - This approach is particularly promising for **music generation systems**, where capturing *sequential dependencies in the latent space* is crucial for ***generating coherent musical sequences.***

These strategies collectively aim to optimize model performance and enhance the accuracy of classification tasks by leveraging advanced techniques in hyperparameter optimization, dataset scaling, regularization, and innovative model architectures tailored to specific domain requirements like music generation.

Part V
Project Summary

5 | Project Summary

In this research endeavor, a novel Deep Learning Architecture for Music Genre Classification has been proposed, leveraging the fusion of Convolutional Neural Networks (CNNs) and Attention Mechanisms on sequences of spectrogram images.

Our model addresses shortcomings of existing architectures by ***capturing temporal dependencies***, ***integrating local and global information***, and segmenting spectrogram sequences to discern temporal significance for genre classification.

Through meticulous experiments, including ***genre similarity analysis***, dimensionality reduction techniques, ***attention weight analysis***, and visualization of temporal dependencies, we unveiled the model's robustness, interpretability, and potential for real-world applications.

With a vision for long-term advancements in music-related technologies, including recommendation systems and music generation, this architecture serves as a step towards unlocking the intricate nuances of music understanding and interpretation.