A Multimodal Approach to Music Genre Classification

Anwika Bhandary
Georgia Institute of Technology
Atlanta, USA
abhandary6@gatech.edu

Kritika Gupta
Georgia Institute of Technology
Atlanta, USA
kgupta340@gatech.edu

Adarsh Honawad
Georgia Institute of Technology
Atlanta, USA
avishwanath30@gatech.edu

Abstract

Music classification tasks such as genre or mood classification have been tackled previously using the audio or the lyrics separately. While many of these have been fairly successful, it appears to be more intuitive to approach these tasks by combining both components of music. Our goal is to train a multimodal based model that uses both the audio spectrograms and lyrics of songs to predict their genre. We use the 4MULA dataset to train single modal approaches using audio spectrogram only and lyrics only, along with multimodal approaches like concatenation and CMA (Cross Modal Attention), and compare their performances through several experiments and visualizations.

1. Introduction

Existing methods for music classification have largely been based on using either the audio or lyric data. As an extension to these approaches and by observing the recent success of multimodal neural networks, this paper argues that combining both of these features into a single model will improve the accuracy over models that are trained on a single modality.

To test this hypothesis, we present a multimodal approach to music genre classification. First, we report baseline performances of single modal methods, i.e., methods that utilize only audio data or only lyric data of music tracks to predict their genres. Second, we report the performance of our multimodal approaches. We experiment with two techniques of fusing modalities - concatenation of single modal embeddings and cross modal attention networks.

Sec. 2 discusss a set of related works in the domain of music genre classification. In sec. 3, we introduce the dataset used and explain the rationale behind choosing it. Sec. 4, describes the architecture of our baseline single modal architectures, and the multi modal architectures. In sec. 5, we report the accuracies obtained by training our baseline and multi modal methods. The results are further visualized using t-SNE plots and confusion matrices. We include specific examples from the dataset that illustrate the benefit of multimodal learning over single modal. We discuss our final conclusions in sec. 6 and present areas of future research.

The main contributions of this work are:

1. Present a multi modal approach that fuses audio features and lyrics of songs to predict the song's genre.
2. Provide baseline results using single model approaches on a relatively new dataset, 4MULA [7]. To our knowledge, no music genre classification results have been published on this dataset.
3. Report results of multi modal approaches on the 4MULA dataset.

1.1. Motivation

Classification tasks such as image or text classification are fairly simple as they usually contain a single modality of representation (such as a 2D or 3D matrix for images and a 1D array for text). However, when it comes to representing music, we have to account for two modalities namely, audio and lyrics.

For a long time, approaches to music classification tasks such as genre or mood classification have been tackled using only audio or lyrics. Since music inherently contains both audio and lyrics, it is more intuitive to consider both of these features for classification tasks. With the advent of multimodal based neural networks, it is now possible to combine these two features into a single model in the hope for better results.

2. Related Works

Previous works in music genre classification [3] have popularized Convolutional Neural Networks for classifying music tracks into various genres by utilizing the melspectrogram of the audio signal. A melspectrogram is a visual representation of an audio that depicts the frequencies that make up the sound. We use a CNN-based approach to music genre classification as a single modal baseline for our work.

There has also been a lot of work looking into classifying music based on the lyric information [1] [10], which is especially useful in extracting the emotion or mood of a song - characteristics that make it a popular choice for genre classification tasks as well.

The motivation behind our work is to build a multimodal approach, using the strengths of the audio and lyric based models and compare its performance with the single modal audio-only and lyric-only models. Existing multimodal approaches either use different modes of input, such as audio, album covers and textual metadata (e.g. album reviews) [5] or combine the audio signal and lyrical information for a different task, like emotion classification [8]. Both [5] and [8] show that a multimodal approach performs better than single modal learning methods. Our approach is the first of its kind, to use audio and lyrical information for the music genre classification task.

We identify 2 major ways of performing the multimodal fusion of audio and lyric modalities. It is crucial to identify a suitable way of combining audio and lyric features to create a joint feature that captures the relevant information from both modalities.

The first method is to train audio and lyric models independently on the task and then extract the top level embeddings and combine them into a single embeddding and train a third model on the same task. In [5], a multimodal deep learning network for music genre classification is proposed using this technique. 3 neural networks, one for the audio, album cover and textual metadata (e.g. album reviews) each are trained and then combined into a multimodal task. In [8], the authors combine the audio signal and lyrical information of Western pop music to classify the emotions humans experience on listening to an audio by fusing the output features from a CNN for the audio and a BERT model for the lyrics.

The second technique is to use crossmodal attention[9] in which attention masks from one modality are used to highlight the extracted features in another modality. This fuses the low-level features of the different modalities, which can lead to better results while fusing audio and lyric data as it will incorporate the low-level meaningful elements across modalities, for instance, it directly captures the difference between the same word being accompanied by different style of music. In [11], crossmodal attention modules are used to combine audio, lyrics and text-based context of songs to perform the same task as [8].

Genre	Number of Tracks
Pop	2000
Rock	2000
Indie	2000
Heavy Metal	1493
Gospel / Religious	1129

Table 1: Genre distribution

3. Data

While the FMA dataset [2] is more commonly used for the music genre classification task, we noticed that most songs in this dataset did not have any associated lyrics as they largely contained just musical pieces. Hence, to enable our multi modal approach that requires lyrics of songs, we chose the 4MULA dataset [7] containing both the audio mel spectrograms and the lyrics of tracks in English, Spanish or Portuguese. The dataset consists of 90,000 songs spanning 76 genres. The two features of our interest in the dataset are the lyrics and the audio mel spectrograms. The former is provided in full for every song, while the latter is extracted from a 30 second snippet of the song. Information that represents the entire audio of the songs is not available due to copyrights.

To enable training our models on the compute resources available to we reduce the dataset to a selection of around 8622 songs and 5 genres as seen in Table 1. These are the primary or distinctive genres that are similar to those used in the more common FMA dataset. The distribution of the tracks per genre can be seen in the table below.

The mel spectrograms provided by the dataset that were created for 30 second clips of the tracks, have size (128, 1292). We trimmed these mel spectrograms to cover 10 second clips, of size (128, 431), as seen in Fig. 1. Our assumption is that a 10 second clip still contains enough information about the audio features to not impact the model severely. However, we still continue to use the complete lyrical data to be able to capture the entire sentiment of the song.

4. Methods

4.1. Audio Only Baseline

We set the first baseline for genre classification using the more popular approach for single modal classification that uses CNNs to recognize the characteristic patterns of genres in mel spectrograms (as seen in Fig. 1). Our model involves a simple CNN based classifier as specified in Fig.

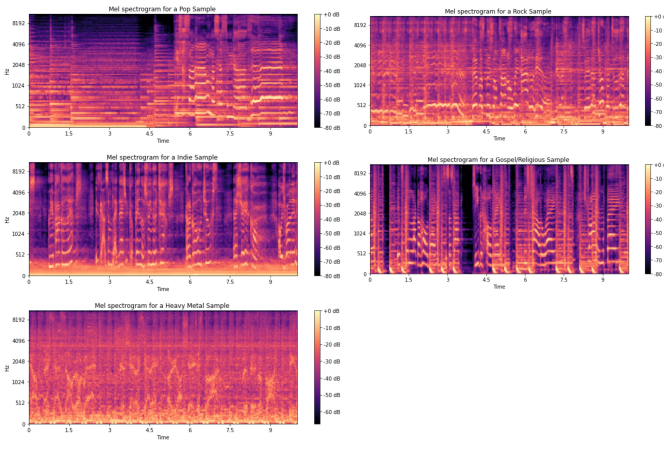


Figure 1: We can see that different genres show their own characteristic features when visualized with their respective spectrograms.

2. The input spectrogram image is passed through 3 Convolutional modules each consisting of a Convolutional Layer (k=5, stride=1), Average Pooling Layer (k=2, stride=2) and a dropout layer (p=0.4). This is followed by two linear layers to output the logit score for each genre class.

4.2. Lyric Only Baseline

The second baseline is a classifier that predicts the genre of a song based on its lyrics only. We use a pre-trained BERT model [4] to tokenize the lyric input and extract BERT embeddings. The pre-trained BERT has been known to act as a good language representation black box that outputs semantically relevant embeddings. The `[CLS]` token for each lyric represents a lyric-level embedding that is passed through 3 fully connected layers to obtain the final classifier output of logit scores for each genre class (Fig. 2).

4.3. Multimodal Fusion using Concatenation

The first multimodal fusion approach we present uses concatenation to combine the embeddings of the spectrogram features and the lyric features generated by the single modal architectures. As shown in Fig. 2, embeddings of size 64 are extracted from the single modal architecture pipelines at the final stage (before being passed into the final classifier layer), concatenated to form an embedding of size 128, which is passes through a fully connected layer and then into the final classifier to output the class logits.

4.4. Multimodal Fusion using Cross Modal Attention

We also followed another approach stated in [9]. The paper discusses combining modalities using cross-modal attention (CMA) that attempts to combine low level features from a 'source' modality into a 'target' modality using an attention layer across the two modalities. In our context, we have two modalities that we need to combine, namely, the spectrogram images and the song lyrics.

We experimented by passing outputs from the CNN model (for spectrogram analysis) and the pre-trained BERT model (for lyrics analysis). Do note as per the architecture diagram presented in Figure 2, that our inputs are not taken from the linear layers of the respective models. Instead, we take the output of the final convolutional layer from the CNN model and splice it along the channel dimension to produce a (N, S, C) dimensional tensor where C = number of channels and S = width * height of the feature map. The output of the final BERT layer already contains feature vectors for each token and we can directly send that to our CMA module.

We also perform ablation studies on this architecture with small modifications such as using pre-trained CNN vs training from scratch, using self-attention layers vs flattening the output of the last layer and discuss the results in future sections.

Model	Accuracy
Single Modal - Audio Spectrogram	42.7 %
Single Modal - Lyric	43.8 %
Multimodal - Concatenation	55.6 %
Multimodal - Cross Modal Attention	60.2 %

Table 2: Overall accuracies for all methods

5. Experiments and Results

For all experiments, we use a 8:1:1 train, test and validation split of the dataset. All models were trained with the Categorical Cross Entropy Loss function. We used the Adam optimizer with a learning rate of 0.001 and batch size of 64. We can see the overall accuracies achieved by each of the four methods, in Table 2

To further evaluate our experiments and compare the different modalities, we will be using t-SNE plots and confusion matrices. t-SNE (t-distributed stochastic neighbour embedding) is a visualization technique that is very helpful in finding patterns in data. We can visualize feature embeddings, even in intermediate layers of trained models, by registering forward hooks in PyTorch's model graphs at desired points. We can then loop through the test dataset, take a forward pass through the network, extract the embeddings and predictions and visualize them. t-SNE plots will be used in the following section to show the progression of our experimentation. Further, we plot confusion matrices for all the models to compare the multi-class classification perfor-

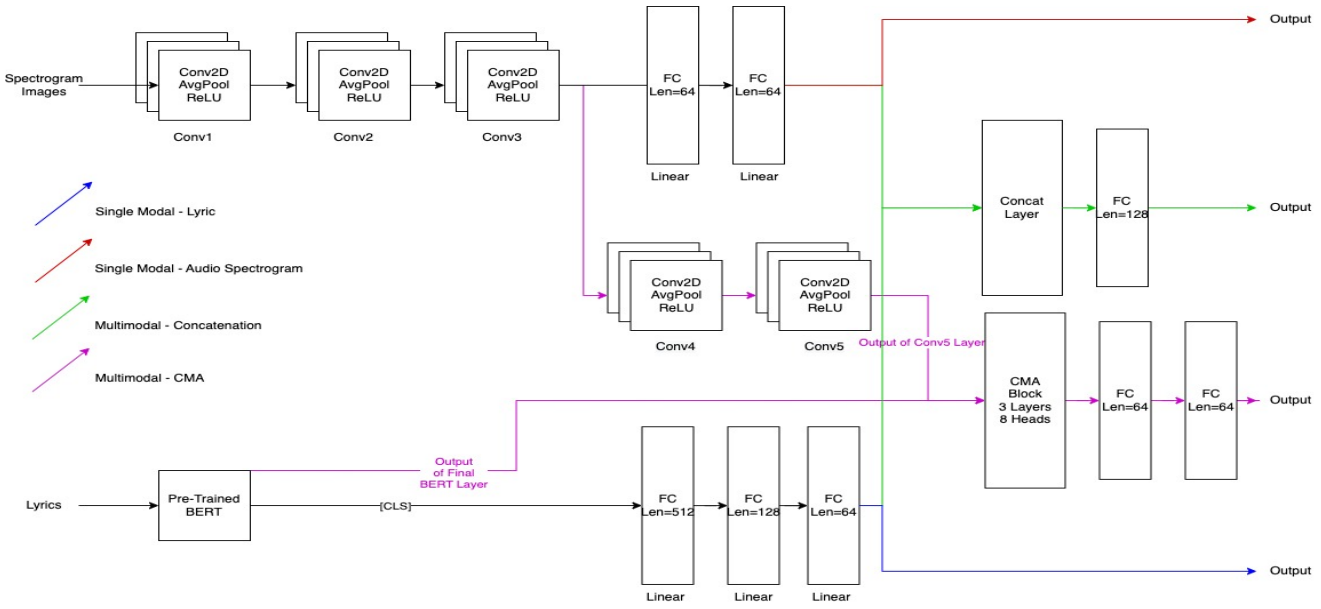


Figure 2: Architecture diagram of all 4 methods showing the components common to multiple methods - audio-only single modal baseline (red), lyric-only single modal baseline (red), multimodal fusion using concatentaion (green), multimodal fusion using cross modal attention (magenta). The final 'Output' represent a final fully connected layer that predicts the logits for each genre class.

mance across classes, detailing per-class accuracy and highlighting the classes that confuse the model the most.

5.1. Single Modal Approaches

5.1.1 Audio Only Baseline

On training the audio only model with the original 8,622 (spectrogram, genre) data pairs, we observed significant overfitting, indicating that there was insufficient data for the model to perform well on unseen data. The accuracy of the model plateaued at around 20%, which is equivalent of a model that predicts randomly for a five class classification model. To deal with this issue, we increased the data through augmentation techniques applied on the spectrogram images. The SpecAugment package [6] works by masking part of the spectrogram, which forces the network to pay attention to other features, thereby expanding its capabilities to generalize to unseen data. There are various different hand-crafted policies that can be used to create the augmentations, which vary on the Frequency Mask Parameter (F), Time Mask Parameter (T), Number of Frequency Masks (n_F) and Number of Time Masks (n_T). We use the following two policies : LibriSpeech basic (LB), LibriSpeech double (LD), and their parameters can be seen in Table 3.

We add 4,000 (spectrogram, genre) data pairs using the LB augmentation policy, and another 4,000 using the LD augmentation policy, resulting in a total dataset size of 16,622. After retraining the model with this new dataset and with the same network parameters, we see an improvement in the performance of the model to 42.7%.

Policy	W	T	n_F	n_T
LB	80	100	1	1
LD	80	100	2	2

Table 3: Augmentation policy parameters

Since our approach is the first to use the 4MULA dataset for genre classification, we don't have an apples to apples comparison of these results with previous works, but by merely glancing at numbers from previous CNN based approaches [10, 5], we can say that this is a reasonable accuracy to set as a baseline.

From the confusion matrix in Fig. 4, we can see that the audio spectrogram model classifies Pop with the highest accuracy (83.9% accuracy), however it also confuses other genres as Pop a lot. 60% of Indie and Gospel/Religious samples and 42.4% of Rock samples are classified as Pop. However only 18% of Heavy Metal samples are classified as Pop. Fig. 3 corroborates these observations with Heavy Metal being distinctly separated from Pop, and all other genres overlapping with Pop. As a consequence, the accuracies of Rock, Indie and Gospel/Religious are low. In fact, the audio-only model cannot classify Gospel/Religious

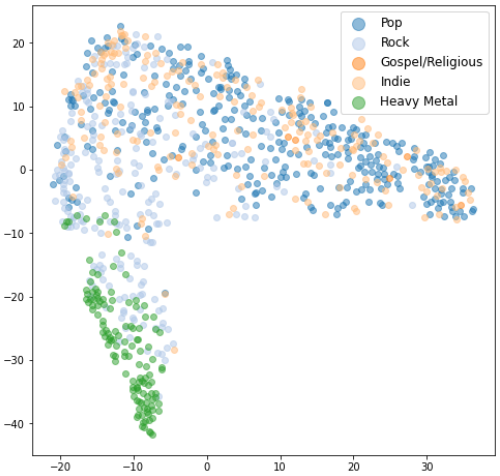


Figure 3: t-SNE plot for embeddings from the Audio Spectrogram model

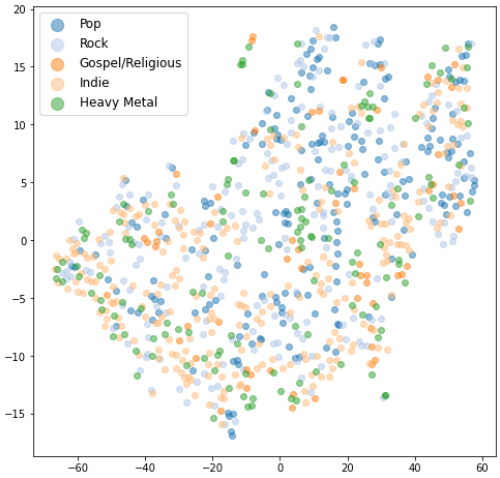


Figure 5: t-SNE plot for the BERT embeddings

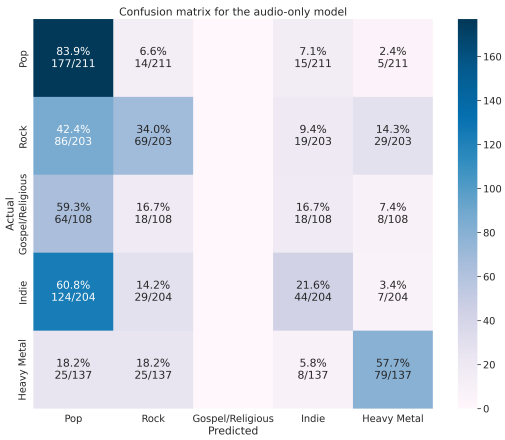


Figure 4: Confusion Matrix for the Audio Spectrogram model

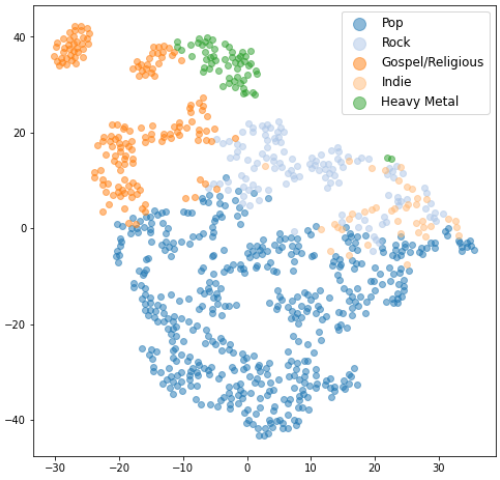


Figure 6: t-SNE plot for embeddings from the Lyric model

tracks at all.

5.1.2 Lyric Only

On training the lyric only baseline, we achieve an overall accuracy of 43.8%. From the confusion matrix in Fig. 7, we can see that the model does an excellent job at identifying the Pop (84.8% accuracy) and Gospel / Religious (75.9% accuracy) genres. If we take a look at the t-SNE plot in Fig. 6, we can see a clear distinction of Pop and Gospel / Religious. This is significantly different from the audio-only model and fewer Gospel/Religious tracks are being misclassified as Pop. This can be attributed to the fact that Gospel/Religious songs talk about a very specific topic, while they can sound very similar to Pop songs.

On the other hand, Indie genre behaves similarly to the audio-only model, achieving an accuracy of 9.8% only. It can be seen that 74% of the Indie samples are misclassified as Pop. The distinction between the lyrics of Indie and Pop samples is not high, this can be seen by the proximity and overlap of their embeddings in Fig. 6 too.

For the Rock genre, the accuracy decreases compared to the audio-only model, and more Rock samples are now misclassified as Pop.

To further investigate the effectiveness of the lyric-only model, we compare the t-SNE plot of the BERT embeddings (Fig. 5 with the t-SNE plot of the final linear embeddings (Fig. 6) of the lyric-only model. The BERT embeddings have no separation based on genre, but training a few linear layers on them leads to a clearer separation across genres.

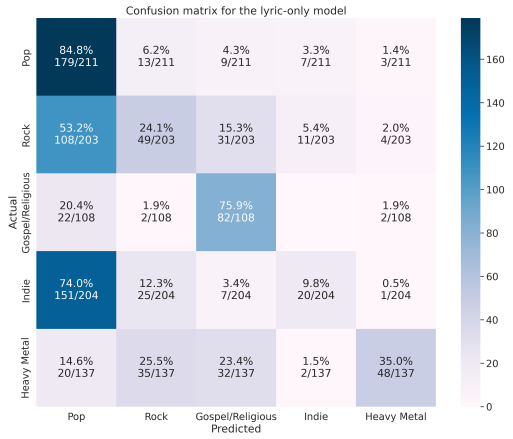


Figure 7: Confusion Matrix for the Lyric model

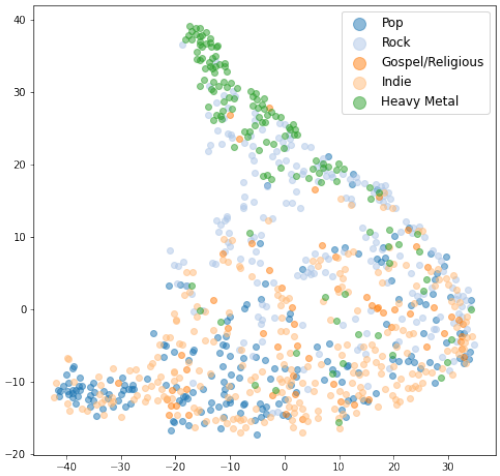


Figure 8: t-SNE plot for embeddings from the concatenated model, with a forward hook at the spectrogram linear layer

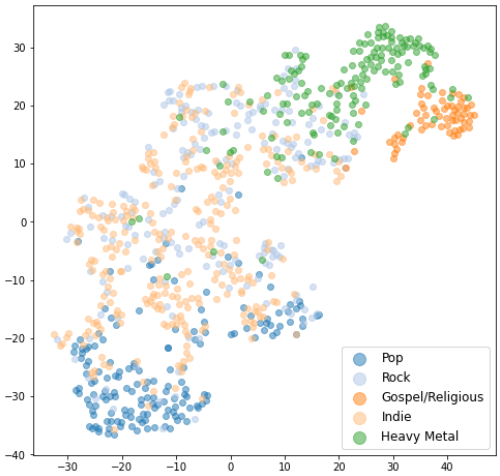


Figure 9: t-SNE plot for embeddings from the concatenated model, with a forward hook at the lyric linear layer

5.2. Multimodal Approaches

5.2.1 Concatenation

Fusing modalities using concatenation significantly improves the identification of the lower performing genres (Rock, Gospel/Religious, Indie) in the single modal approaches, as seen in Table 4.

Fig. 11 shows fewer number of samples of other genres are misclassified as Pop. Indie, in particular, shows a significant improvement compared to the single modal approaches, showing distinct separation in Fig. 10 which was not achieved by either of the single modal methods.

Looking at the intermediate embeddings in this model, i.e., the audio and lyric embeddings before concatenation reveals characteristics that are true to the nature of what information each modality conveys for every genre.

In Fig. 10, Heavy Metal and Gospel / Religious lie close in this embedding space. However in Fig. 8, they are far apart, indicated that aurally they sound different, and in Fig. 9, they are closer in the embedding space, which is attributed to the fact that both share common themes in their lyrics. This shows that the multimodal method is able to capture the strengths of each modality and its contribution in classifying genres.

It's also quite interesting to see that Pop shows a very distinct separation in terms of its intermediate lyric embeddings, and yet when it comes to its intermediate spectrogram embeddings, Pop is very widely spread. This shows the evolution of Pop over recent times, and how it's become more fluid as a genre, incorporating facets of many other genres over the years.

5.2.2 CMA (Cross Modal Attention)

We have adopted the CMA Module using the architecture discussed in [9]. However, the paper discusses fusing 3 modalities and uses video sequence instead of still images. To adopt it to our requirements we have made some modifications to the CMA module.

Firstly, we decided to add two additional convolutional layers as the output from the 3rd convolutional layer was too large to be processed in the CMA module. Since still images do not have a temporal dimension, we decided to splice the output from the convolutional layer along the channel dimension such that each pixel is represented by a vector of C features (C = no. of output activations). As the spectrogram images already contain data along the time dimension, we feel this would in some way act as the temporal dimension for the songs. Doing this took our model accuracy from 55% to around 58% as was expected. In addition, the original CMA module also uses a self-attention layer to fur-

	Audio-only	Lyric-only	Multimodal (concat)	Multimodal (CMA)
Pop	83.9%	84.8%	54.5%	67.8%
Rock	34.0%	24.1%	51.2%	55.7%
Gospel/Religious	0%	75.9%	51.9%	73.1%
Indie	21.6%	9.8%	56.9%	40.7%
Heavy Metal	57.7%	35.0%	65.0%	74.5%

Table 4: Per-genre accuracies for all methods.

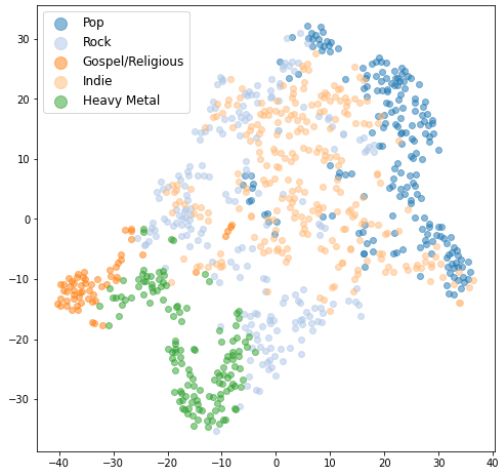


Figure 10: t-SNE plot for embeddings from the concatenated model, using concatenated spectrogram and lyric embeddings

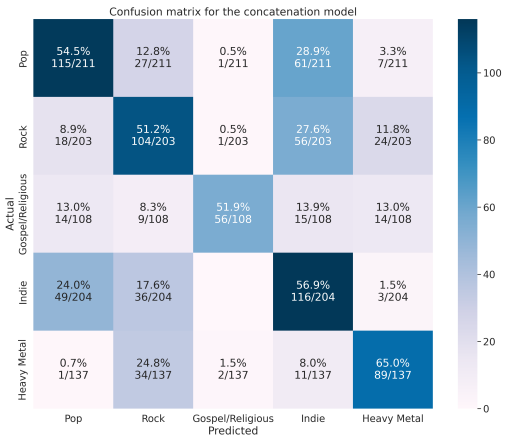


Figure 11: Confusion Matrix for the Concatenated model

ther fuse the two source modalities which we felt was not needed in our scenario as we have a single source modal in both cases (audio - lyrics and lyrics - audio). Removing this self attention layer further took our model accuracy from 58% to 60.2%. We also discuss further potential improvements that can take advantage of this architecture in the Conclusion and Future Work section.

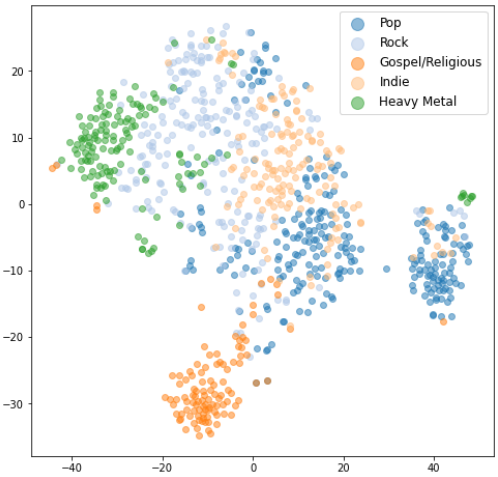


Figure 12: t-SNE plot for embeddings from the CMA model

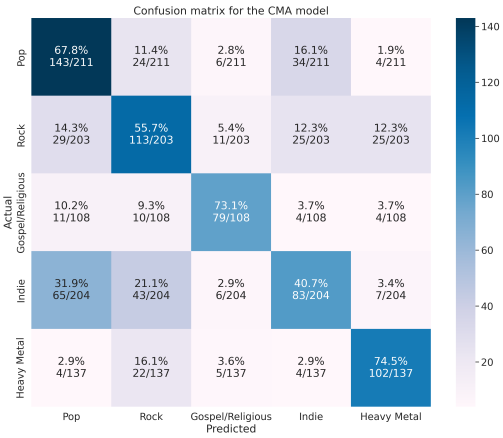


Figure 13: Confusion Matrix for the CMA model

From the confusion matrix in Figure 13, we can see that CMA has even further improved the identification of all the genres. As we can see from the t-SNE plot in Figure 12, the genres are distinct, though we notice one clear change from the t-SNE of the concatenated model - here, Heavy Metal and Gospel / Religious are far apart.

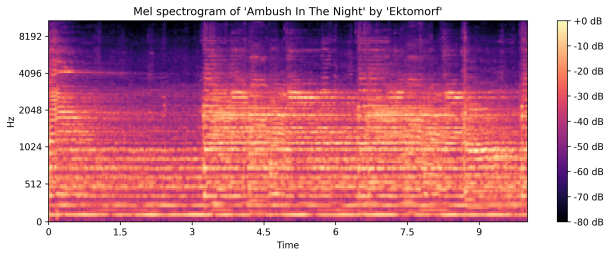
5.3. Visualization of randomly chosen samples

We pick a pair of examples from the test set to better illustrate how our multimodal approach follows our motivation (Fig. 14). The figure shows the spectrograms and parts of lyrics of two Heavy Metal songs. 14a is classified as a Rock song by the audio-only model. This is incorrect and can be attributed to the fact that the spectrogram has percussive characteristics closer to other Rock songs. It is classified a Heavy Metal song by the lyric-only model, which is evident by the nature of themes in the lyrics. 14b is classified as a Pop song by the lyric-only model, which is incorrect. This may be due to the nature of themes in the lyrics, which are more typical to Pop songs than other genres. It is classified as a Heavy Metal song by the audio-only model due to the heavy distortion characteristics seen in most Heavy Metal songs. Both the multimodal approaches classify both songs correctly as Heavy Metal. This shows that though songs can resemble different genres in individual modals, the multimodal method fuses the information of each modality and improves the prediction of genre. It is not unconditionally biased to one modality more than the other, otherwise the either the inaccuracies in lyric-only models would permeate through, or the inaccuracies in audio-only models would permeate through. This example shows that this is not the case.

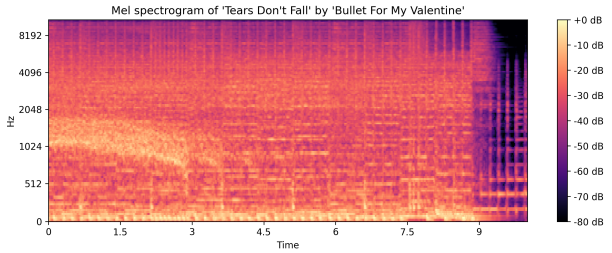
6. Conclusion and Future Work

Through the course of our work, we have implemented and evaluated four different approaches for music genre classification. First, a single modal approach using a convolutional network that inputs audio spectrograms and outputs the genre of the corresponding track. In addition, we performed data pre-processing and data augmentation to improve upon this accuracy and make it a feasible baseline to compare against our multimodal approaches. Next, we evaluated the performance of a single modal lyric based model, using pretrained BERT embeddings. After experimenting with the single modal approaches, we moved onto multimodal approaches, starting with a model that concatenated the spectrogram and lyric embeddings. Finally, we evaluated a cross modal attention based approach for the fusion of the two modalities.

We were able to see a significant improvement in the accuracies of both the multimodal approaches over the single modal approaches. Using visualization techniques like t-SNE to better understand the feature embeddings, and using confusion matrices to better compare the multi-class classification, we were able to observe interesting patterns across the genres. For instance, some genres may be closer together lyrically but not aurally, and vice-versa. In order to have a more comprehensive genre classification model, it's important to carefully consider both aspects, and our results



(a) 'Ambush In The Night': `Two thousand soldiers. Hate and anger. Violent -- brutality. Racist ***. . Segregation. Discrimination. Is not a life. They wanna survive. . Ambush in the night. Ambush in the night. . Woman cries. Her child has died. Revenge and rage. Endless pain. . Rebel. Hunger. One hopeless nation. Lives in the **** of poverty. . Ambush in the night. Ambush in the night. Ambush in the night. Ambush in the night. . They gonna rise. They gonna rise. They gonna rise. . From the pain`



(b) 'Tears Don't Fall' :`Your tears don't fall, they crash around me. Her conscience calls, the guilty to come home. . The moments die, I hear no screaming. The visions left inside me are slowly fading. Would she hear me if I called her name?. Would she hold me if she knew my shame?. . There's always something different going wrong. The path I walk's in the wrong direction. There's always someone *** hanging on. Can anybody help me make things better?. . Your tears don't fall, they crash around me.`

Figure 14: Visualization of two Heavy Metal examples. Song (a) is classified as a Rock song by the audio-only model, a Heavy Metal song by the lyric-only model and multimodal models. Song (b) is classified as a Pop song by the lyric-only model, a Heavy Metal song by the audio-only model and multimodal models.

show that the multimodal approaches have the upper hand over single modal.

We believe there is a lot of potential for future work for this project. In terms of improving the performance of the model alone, we believe that including longer spectrogram

snippets could help improve the overall model accuracy. Additionally, if we are using the CMA fusion process, sequences of 10-second spectrogram snippets can be used as inputs instead of a single long spectrogram image. Another such extension is to compare the performance of the different modalities in a multi-language setting, to analyze how genres transfer between languages.

7. Team Contributions

Team Member	Contribution
Kritika Gupta	Literature Survey, dataset creation and preprocessing, implementing and training lyric-only and multimodal using concatenation model, visualization of results, report writing.
Anwika Bhandary	Literature Survey, dataset augmentation, implementing and training audio-only model, generating t-SNE plots, visualization of results, report writing.
Adarsh Honawad	Literature Survey, implementing and training multimodal using cross modal attention model and its variations, ablation studies, visualization of results, report writing.

Table 5: Team Contributions

References

[1] Hasan Akalp, Enes Furkan Cigdem, Seyma Yilmaz, Necva Bolucu, and Burcu Can. Language representation models for music genre classification using lyrics. In *2021 International Symposium on Electrical, Electronics and Information Engineering*, pages 408–414, 2021.

[2] Kirell Benzi, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. *CoRR*, abs/1612.01840, 2016.

[3] Snigdha Chillara, AS Kavitha, Shwetha A Neginhal, Shreya Haldia, and KS Vidyullatha. Music genre classification using machine learning algorithms: a comparison. *Int Res J Eng Technol*, 6(5):851–858, 2019.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[5] Sergio Oramas, Francesco Barbieri, Oriol Nieto Caballero, and Xavier Serra. Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.*, 2018.

[6] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA, sep 2019.

[7] Angelo Cesar Mendes Da Silva, Diego Furtado Sliva, and Ricardo Marcondes Marcacini. 4mula: A multitask, multimodal, and multilingual dataset of music lyrics and audio features (small version), 2020.

[8] Bo-Hsun Sung and Shih-Chieh Wei. Becmer: A fusion model using bert and cnn for music emotion recognition. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 437–444, 2021.

[9] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[10] Alexandros Tsaptsinos. Lyrics-based music genre classification using a hierarchical attention network. *arXiv preprint arXiv:1707.04678*, 2017.

[11] Jiahao Zhao, Ganghui Ru, Yi Yu, Yulun Wu, Dichucheng Li, and Wei Li. Multimodal music emotion recognition with hierarchical cross-modal attention network. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022.