Deep Learning for Music Genre Classification and Music Emotion Recognition

Queen Mary University of London

Luke Abela

l.abela@se20.qmul.ac.uk

Introduction

Music, an arrangement of sounds having melody, rhythm, and usually harmony, may be subdivided into subcategories. A music genre is a category to which pieces of music belong to depending on characteristics. Classifying music into genres is itself a subjective task, as many songs have elements reminiscent of multiple genres. (Costa, et al., 2011)

The digitisation of music has resulted in mass storage of songs on the Internet. Services such as Spotify achieved commercial success by providing delivery of large amounts and varied types of music. Such a system, which is ultimately a database of songs, requires criteria by which to classify songs. Potential criteria include emotion and genre. (Rajana, et al., 2016)

Traditionally, assigning a genre tag has been a human task, with the results of tagging varying from person to person. Examples of common musical genres include rock, and funk music.

Music Emotion Recognition (MER) is a subfield of Music Information Retrieval. MER systems find use in applications such as automatic playlist generation e.g. ‘Just Happy Tunes’ or ‘Only Sad Songs’.

Datasets

The dataset selected for this genre classification was the GTZAN dataset. The dataset consisted of 1000 tracks. Each track was 30 seconds in length. The dataset contained 10 genres, with each genre containing 100 tracks. The genres were blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae and rock. The GTZAN dataset has appeared in over 100 different published works and is a highly used public dataset in music genre research. (Sturm, 2012)

The dataset used for emotion recognition was the 4-Quadrant dataset. This was a dataset consisting of 900 approximately 30 second long clips gathered from the ALLMusic API. The

collected songs were split into 4 ‘quadrants’ based on the Russell Valence-Arousal quadrants. (Widowati & Nuguroho, 2018)

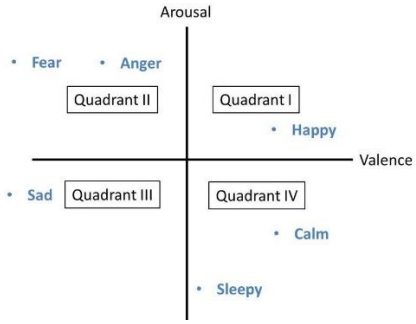


Figure 1: Russell Valence-Arousal Quadrants

Methods

Data Preparation

The respective genre/emotion of each song was used as the label for the data required by the supervised learning method. Deep Learning methods are powerful, but require large datasets. Each audio track was split into multiple tracks to augment the data. To increase the number of data samples available for the training, TensorFlow’s Image Data Generator method was used.

Method 1 – Mel-spectrogram Input based Network

Published literature has inferred various features about each individual track proceeds by feeding these features as a numerical vector into the network. For this method, a Mel-spectrogram was generated for each track.[1]

The data instances and their labels were split into training, validation and test sets. The Mel-spectrograms were normalised and reshaped, whilst the labels were processed with categorical encoding.

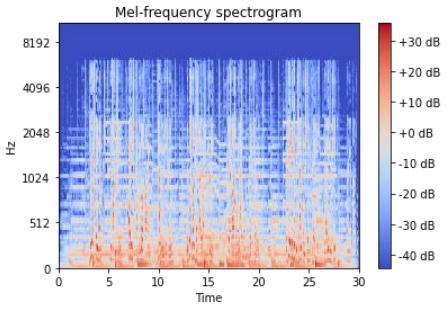


Figure 2: Mel-Spectrogram

[1] A Mel-spectrogram is a spectrogram on the Mel-scale, that is the computed Fast Fourier Transform on overlapping window segments of a signal, whose frequencies are converted to the Mel-scale, a non-linear measurement. Each Mel-spectrogram was effectively a visual representation of the track from which it was inferred.

Method 1 – Networks

Network A – Convolutional

The Convolutional layer was used to extract features in the Mel-spectrogram inputs. Successive convolutional layers were used to extract finer features as the input data moved through the architecture. Max-Pooling layers were used as a form of down-sampling. Dropout layers were used to limit over-fitting.

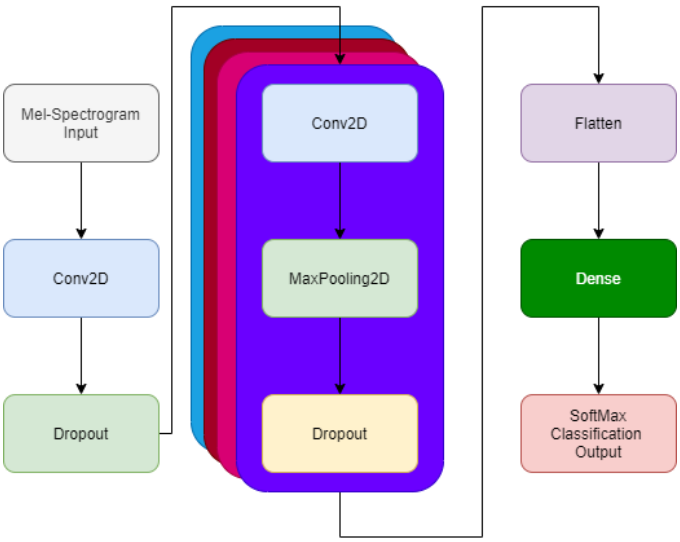


Figure 3: Convolutional Architecture

Network B – Bidirectional Long Term Short Term Memory

Bidirectional LSTMs were used as an extension of LSTMs. The Bidirectional layer may improve model performance for sequence classification. When all timesteps are available, the Bidirectional LSTMs trains two LSTMs on an input sequence. One LSTM would train on the standard input sequence, whilst the second would train on a reversed copy of the input sequence. This architecture fed the Mel-Spectrogram into a Bidirectional LSTM to allow it to extract sequences which would indicate the class to which the input belonged to.

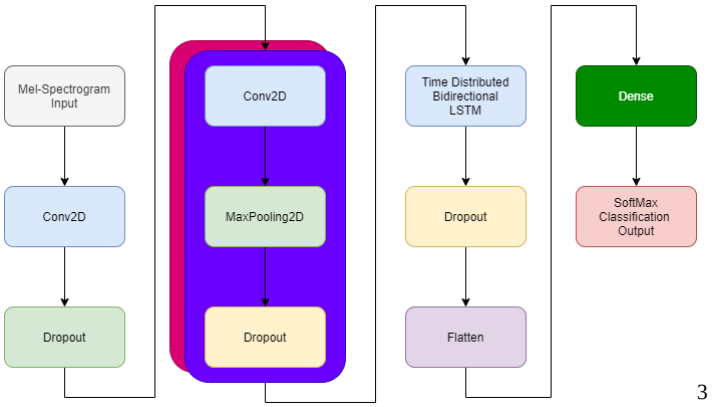
Figure 4: Bidirectional LSTM Layer

Network C – Convolution Bidirectional Long Term Short Term Memory

Research literature has previously shown the promise of architectures featuring both convolutional layers for feature extraction followed by the use of an LSTM to aid in sequence

[2] An equivalent model was created replacing the LSTM with a GRU. A Gated Recurrent Unit is similar to the LSTM with a forget gate, including fewer parameters due to the lack of an output gate. GRUs have been shown to have comparable or even superior performance to LSTMs depending on the task and dataset.

classification. Such a model is typically considered appropriate for sequence tasks the data of which is of a spatial nature.



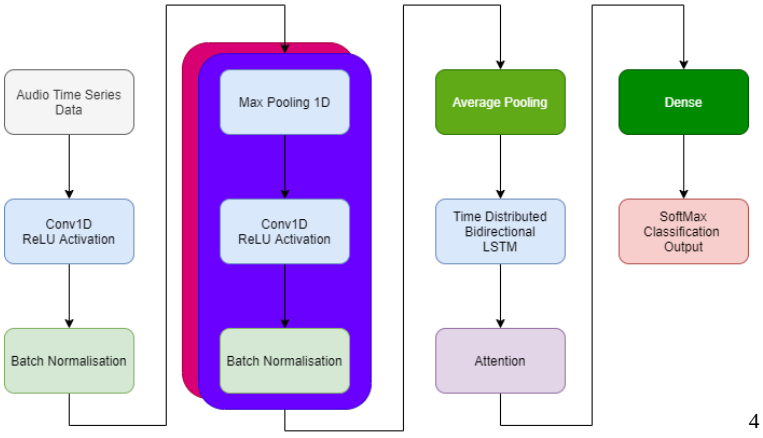
[3]

Figure 5: Convolutional BLSTM Architecture

Method 2 – An End-to-End Network

Literature has shown that raw audio could be fed into an appropriate network and encouraging performance achieved. The model used here was adapted from ConflictNET: End-to-End learning for Speech-based Conflict Estimation. (Rajan, et al., 2019)

Direct audio was used as an, hence required the adaptation of the model from 2D to 1D. An attention layer was also included. Attention was a technique designed with the intention of mimicking cognitive attention. This allowed the system to focus on sections of data deemed more important. The method used here was dot-product attention.

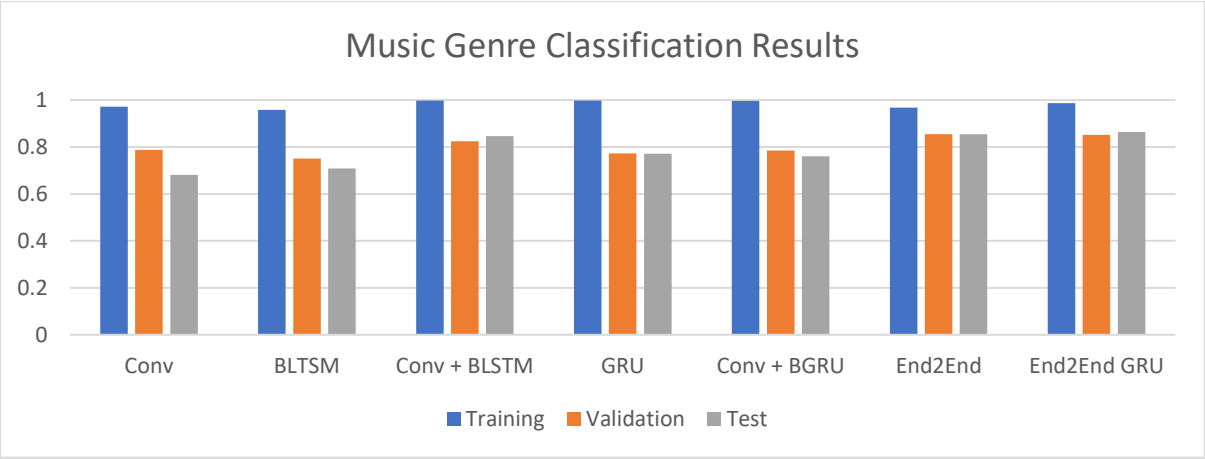


[4]

Figure 6: End-to-End Deep Learning Architecture Featuring Attention

[3] An equivalent model was created replacing the LSTM with a GRU
[4] An equivalent model was created replacing the LSTM with a GRU

Results



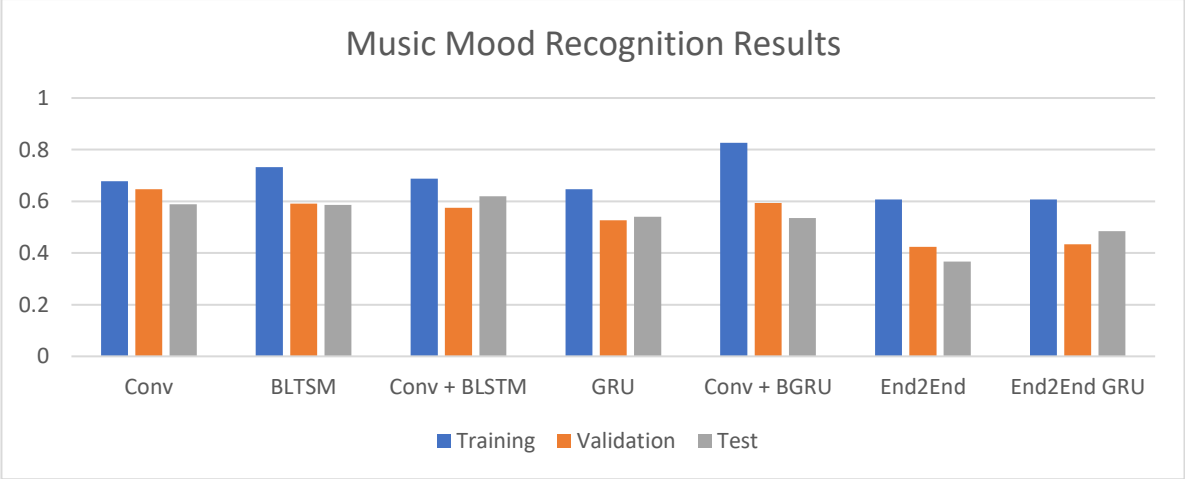
5

Figure 7: Music Genre Classification Results

From Figure 7, it was noted that the End-to-End method performance surpassed that of the Mel-spectrogram based methods. The highest training accuracy was achieved by a Mel-spectrogram input to a convolutional + BLSTM architecture, however, this overfit. The End-to-End methods had comparable performance, however, the GRU-based End-to-End model made use of less parameters and hence less training time than the standard End-to-End model, making the former preferable.

One observation to note was the classification accuracies of the individual genres[6]. Certain genres outperformed others depending on the architecture, with the rock genre being poorly classified specifically by the Convolutional architecture. This would clearly indicate that music has temporal and sequential properties, and that the recognition of a genre such as rock was dependent on past values.

[5] See Annex 1 – Genre for table of numerical Results
[6] See Annex 1- Genre for Classification Matrices



[7]

Figure 8: Music Mood Classification Results

From Figure 8, the optimal training accuracy for Mood recognition was achieved by the convolutional and bidirectional GRU architecture, however this overfit. Comparable results were achieved by the convolution architecture, and the convolution + bidirectional LTSM architectures.

Mood recognition performance was a less successful task than genre classification. This was due to various factors – the lack of inclusion of lyrics could have hampered performance. A two input system with one input handling music and the second input handling lyrics could improve performance. Furthermore, as the best result was obtained by a convolutional architecture, this indicated that sequential and temporal properties were not helpful in this task, and actually reduced performance.

The End-to-End architecture also suffered, with its direct audio input system performing worse than the Mel-spectrogram based input in all cases. The End-to-End classification matrix showed that the architecture overwhelmingly predicted song moods to be one specific category, indicating the attention mechanism seemed to focus on aspects of songs which lent themselves to such a category.

The convolutional architecture also produced the most balanced classification matrix results. [8]

Audio File Case Studies

The selected models were the End-to-End Architecture with a GRU for Genre Classification and the convolutional model for Mood Recognition.

The models were designed to take 3-second songs inputs. Hence, to assess the genre/mood of the song, 100 random 3-second snippets of each song used were used as input to the models. This was beneficial was songs often have influences derivative of many genres, thus, plotting a histogram of the classification of different snippets of songs would allow a distribution of genres. Similarly, many songs are story-like in nature, with each song including various emotions.

[7] See Annex 2 – Mood for table of numerical Results
[8] See Annex 2- Mood for Classification Matrices

Song 1 was ‘Somebody Made for Me’[9] by a group known as ‘The Explorers Club’, who are considered a rock band.[10] The song was primarily classified as rock, with elements of other genres evident. The song featured percussive elements, a rhythmic baseline, and a melody reminiscent of classic rock.

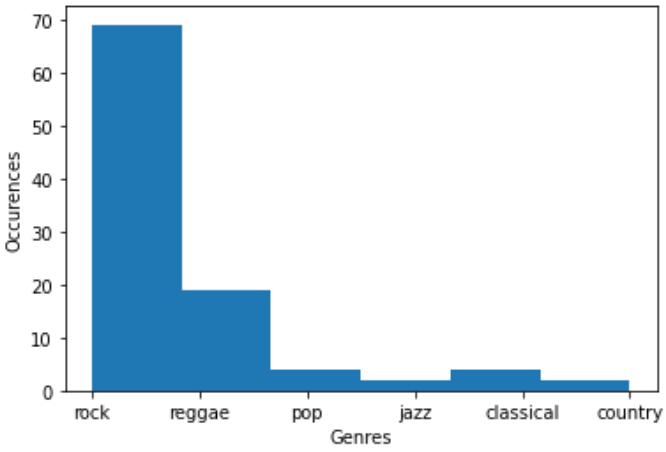


Figure 9: Somebody Made for You by the Explorers Me Genre Classification Histogram

A mood histogram was also generated showing that the song was predominantly happy. This was expected due to the song’s joyful tone. The song had slower moments which were accounted for by the recognition of another moods. This classification fit well with the message promoted by the lyrics, a joyful proclamation of finding a special person.

Figure 10: Somebody Made for Me Mood Histogram

Song 2 was ‘Bohemian Rhapsody’ by Queen, a song streamed more than 1.6 billion times.[11] This song featured differing sections. Evident from the histogram, no genre occurred ina majority of cases. The most prominent genre was classical, this was due to the piano introduction and operatic section. Elements of rock and metal were also detected as song featured passages of percussive elements and heavy electric guitar.

[9] (61) Somebody Made For Me - YouTube
[10] The Explorers Club | Introducing The Explorers Club
[11] (61) Queen – Bohemian Rhapsody (Official Video Remastered) - YouTube

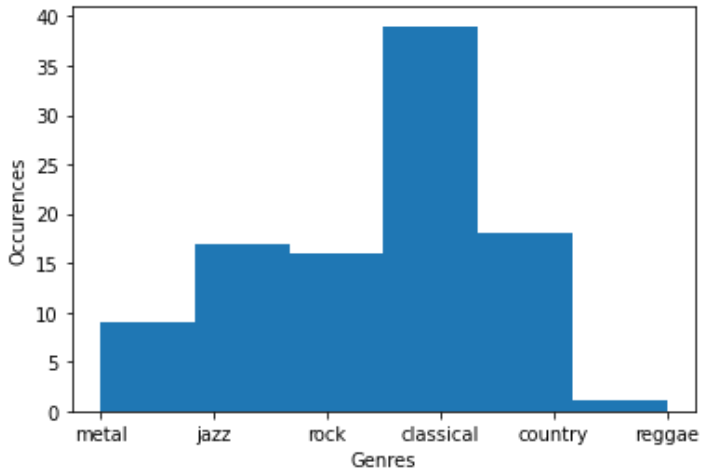


Figure 11: Bohemian Rhapsody Genre Classification Histogram

The mood histogram for Bohemian Rhapsody had a more even distribution. Predominantly ‘Sad/Bored/Sleepy’ and ‘Annoying/Angry/Nervous’, evidenced by the band Queen stating "Bohemian Rhapsody" was about a young man who has accidentally killed someone. Such themes and ideas lend themselves to more negative moods, which would make the classification appropriate as opposed to ‘Happy’.

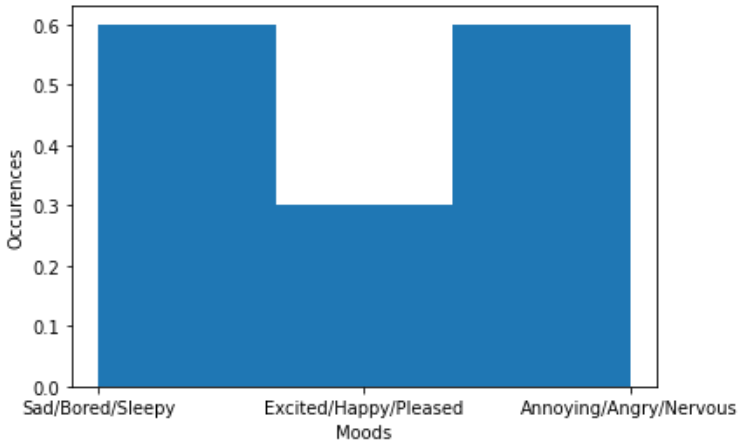


Figure 12: Bohemian Rhapsody Mood Recognition Histogram

Bibliography

Costa, Y. M., Oliviera, L. S. K. A. L. & Gouyon, F., 2011. Music Genre Recognition using Spectrograms. Sarajevo, s.n.

Rajana, A. R., Aryafar, K., Shokoufandeh, A. & Ptucha, R., 2016. Deep Neural Networks: A Case Study for Music Genre Classifcation. s.l., IEEE.

Rajan, V., Brutti, A. & Cavallaro, A., 2019. ConflictNET: End-to-End Learning for Speech Based Conflict Intensity Estimation. IEEE Signal Processing Letters, 26(11), pp. 1668-1672.

Sturm, B. L., 2012. An analysis of the GTZAN music genre dataset. New York, ssociation for Computing Machinery, pp. 7-12.

Widowati, F. U. & Nuguroho, F. S. S. G. F., 2018. Classifcation of Music Moods Based on CNN. s.l., iSemantic.

Panda, P., Malheiro, R. & Paiva, R. P. (2018). “Novel audio features for music emotion recognition”. IEEE Transactions on Affective Computing (accepted for publication).

Note : the annex was included to provide further figures. This section is not strictly part of the report and was included should an interested reader wish to see further detail

Annex 1 - Genre

Music Genre Classification Results							
Accuracy	Conv	BLTSM	Conv + BLSTM	GRU	Conv + BGRU	End2End	End2End GRU
Training	0.9719	0.9583	**0.9974**	0.9983	0.9962	0.9673	0.9873
Validation	0.7878	0.7507	0.8248	0.7728	0.7848	**0.8539**	0.8519
Test	0.6817	0.7087	0.8458	0.7718	0.7608	0.8539	**0.8639**

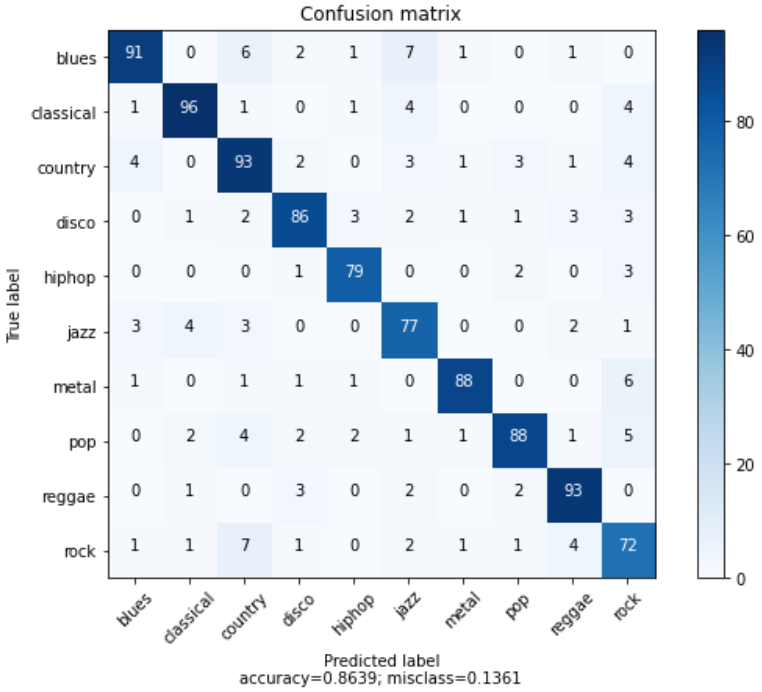


Figure 13: Genre End-to-End using GRU

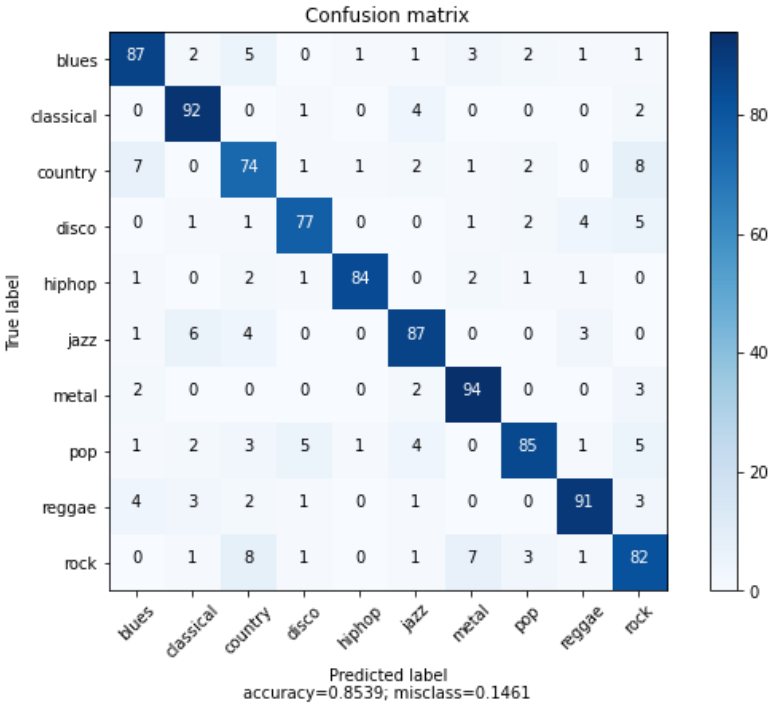


Figure 14: Genre End to End

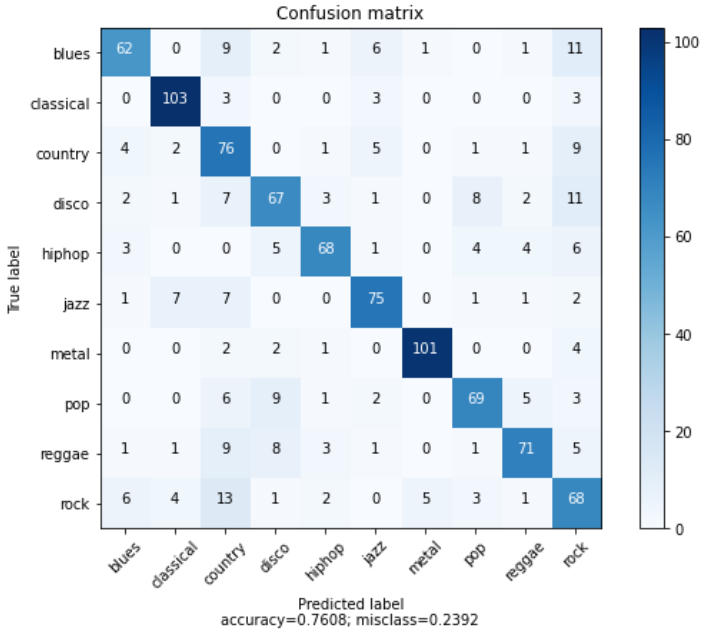


Figure 15: Genre Convolution + BGRU

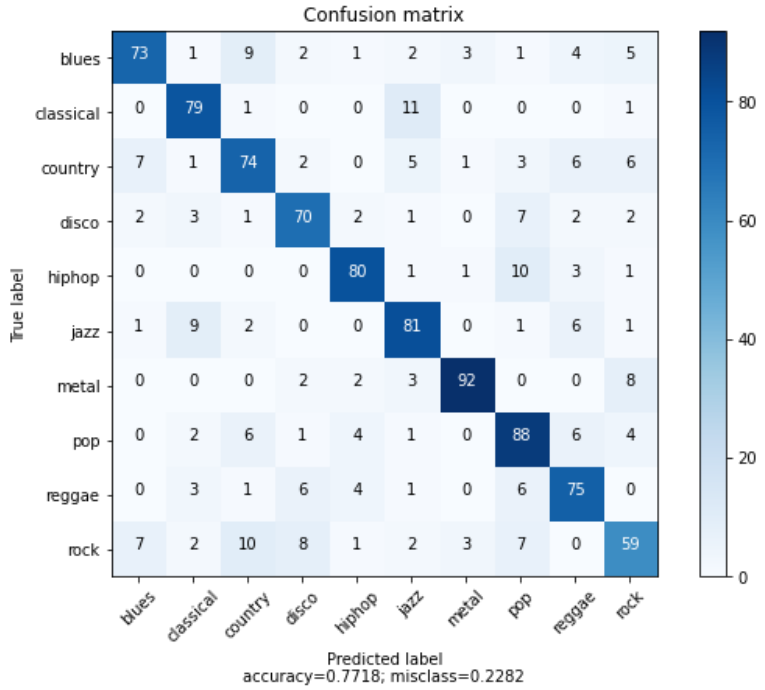


Figure 16: Genre Bidirectional GPU

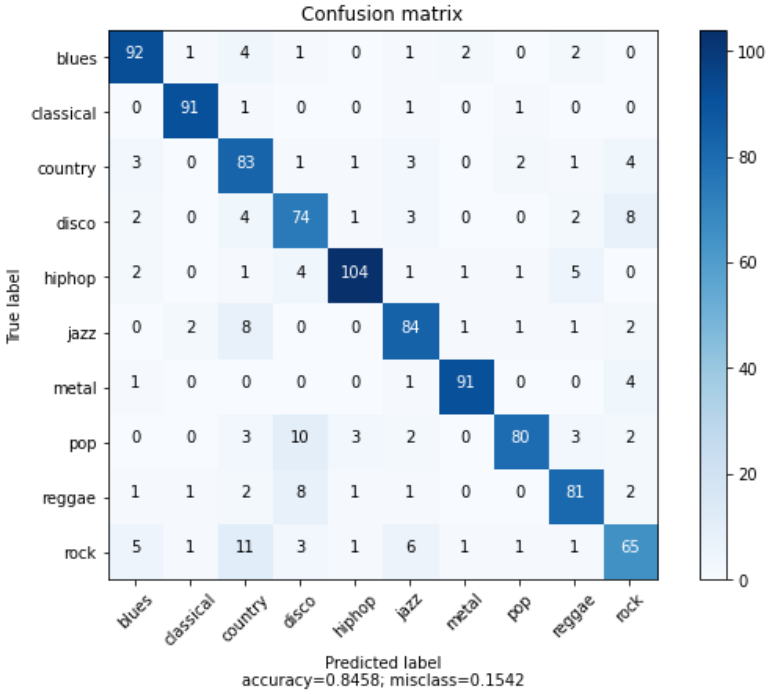


Figure 17: Genre Convolution + Bidirectional LSTM

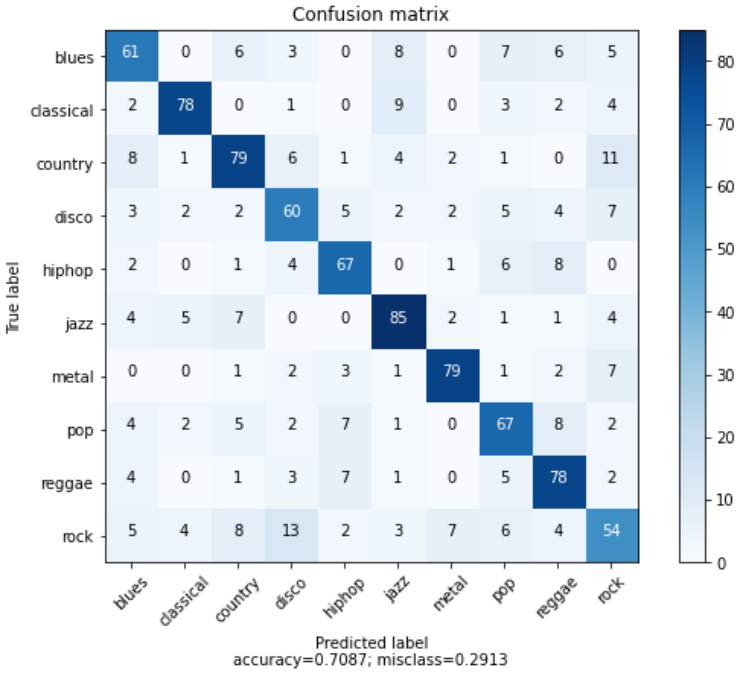


Figure 18: Genre Bidirectional LSTM

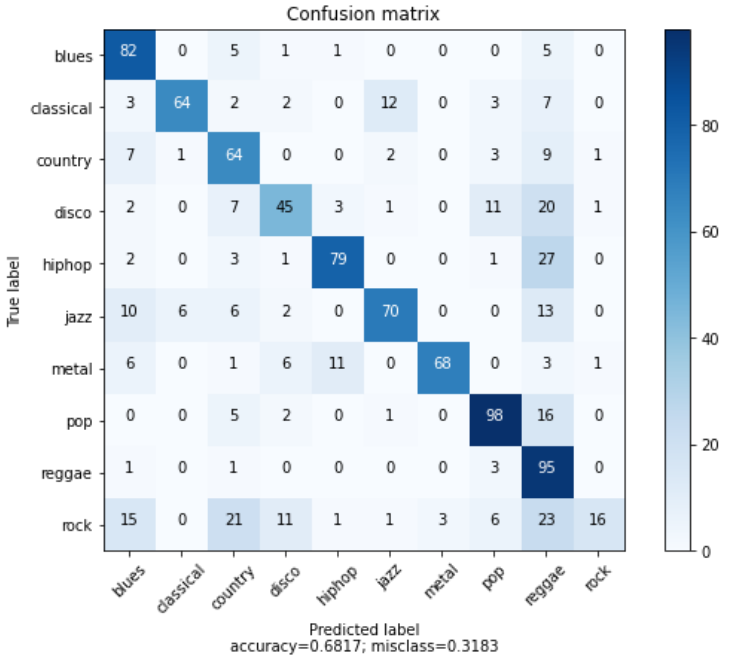


Figure 19: Genre Convolution

Annex 2 - Mood

Music Mood Recognition Results							
Accuracy	Conv	BLTSM	Conv + BLSTM	GRU	Conv + BGRU	End2End	End2End GRU
Training	0.6779	0.7324	0.6878	0.6469	**0.8264**	0.6074	0.6072
Validation	**0.6467**	0.5913	0.5753	0.5267	0.5933	0.4244	0.4333
Test	0.5889	0.5868	**0.62**	0.5400	0.5356	0.3667	0.4844

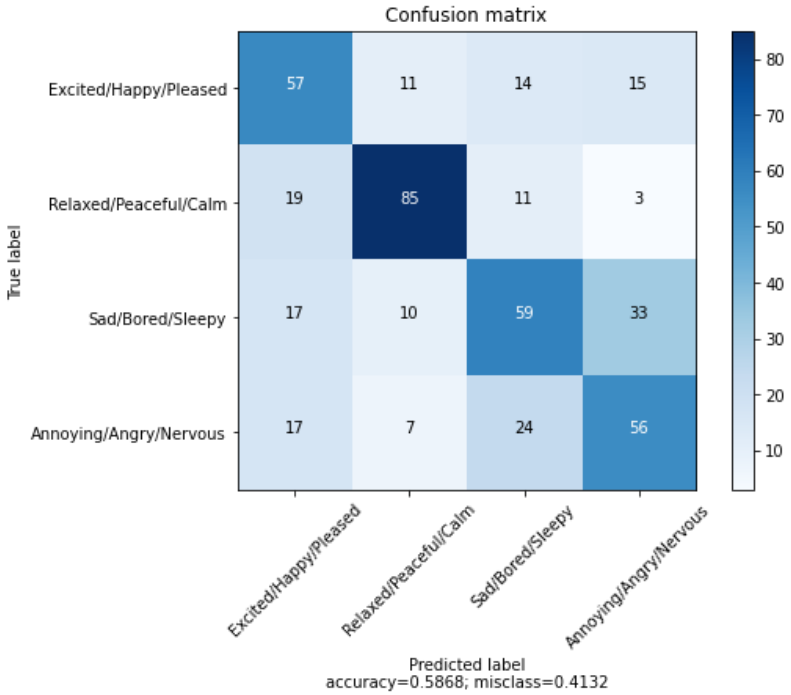


Figure 20: Mood BLSTM

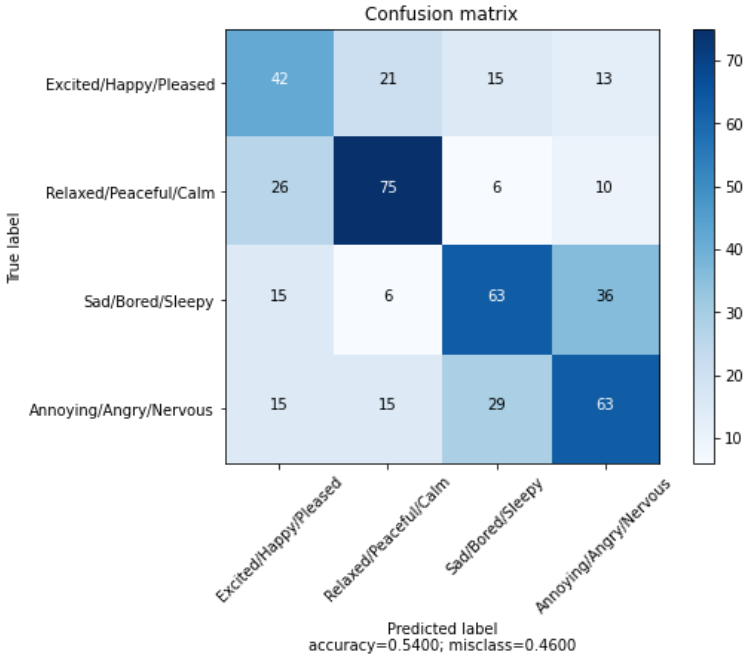


Figure 21: Mood GRU

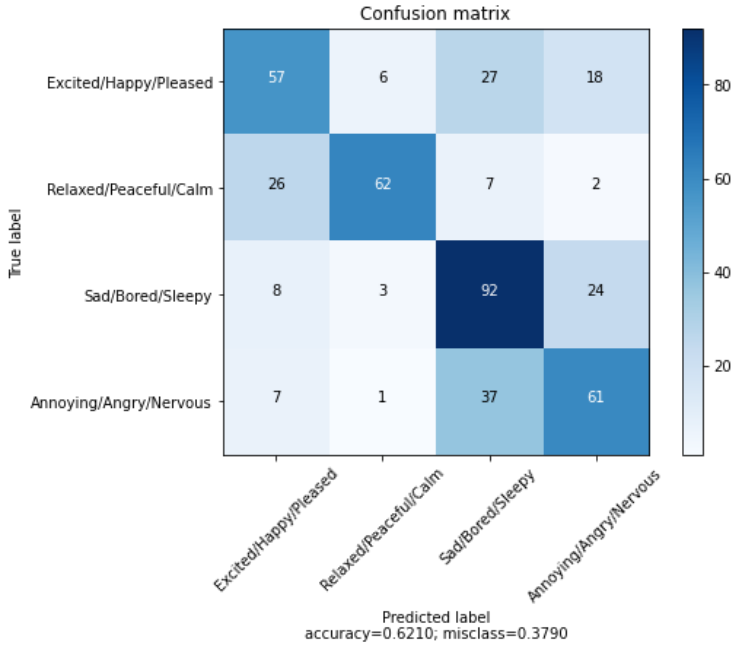


Figure 22: Mood Convolution + Bidirectional LSTM

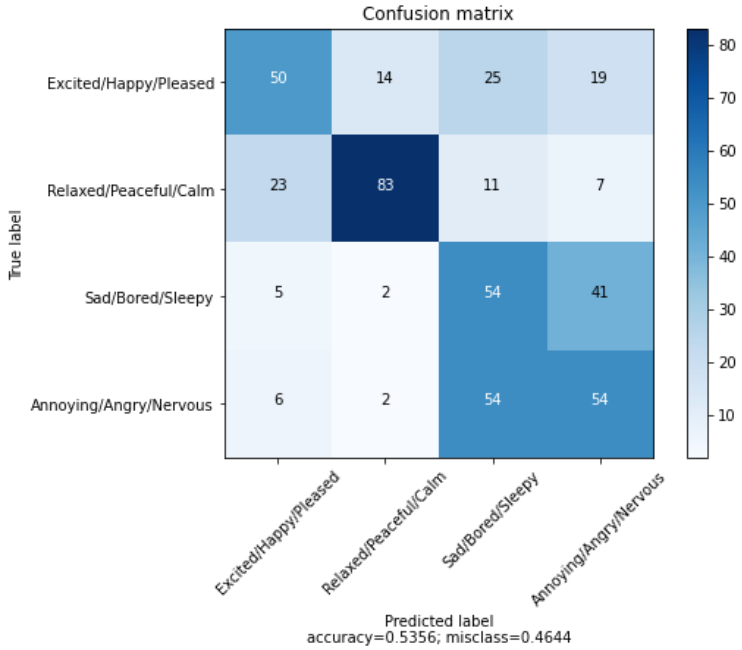


Figure 23: Mood Convolution + Bidirectional GRU

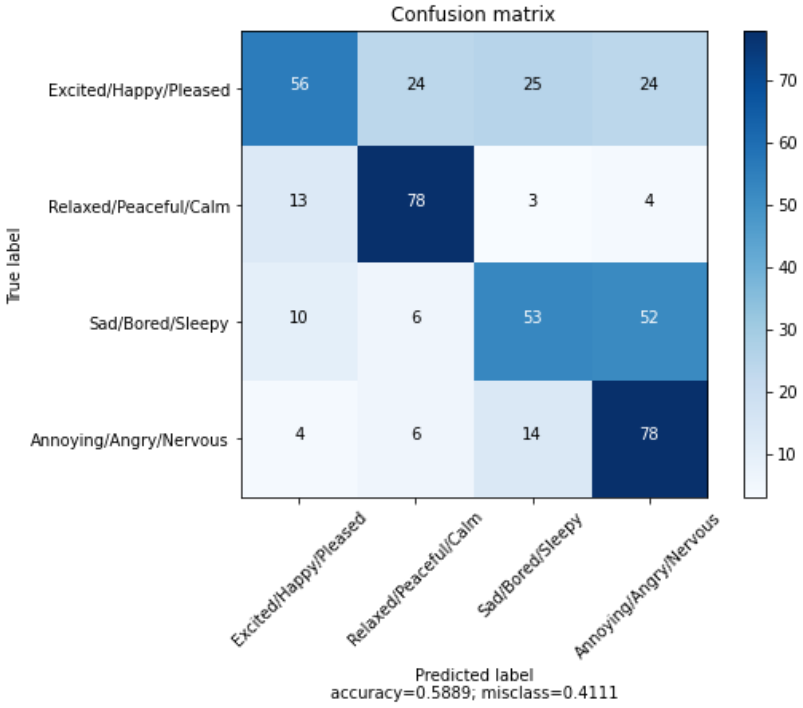


Figure 24: Mood Convolution

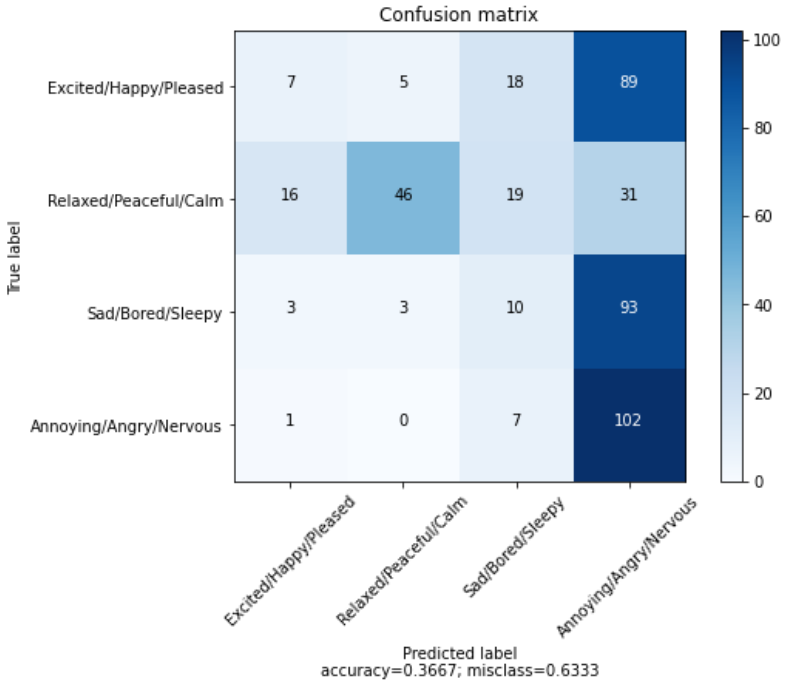


Figure 25: Mood End to End

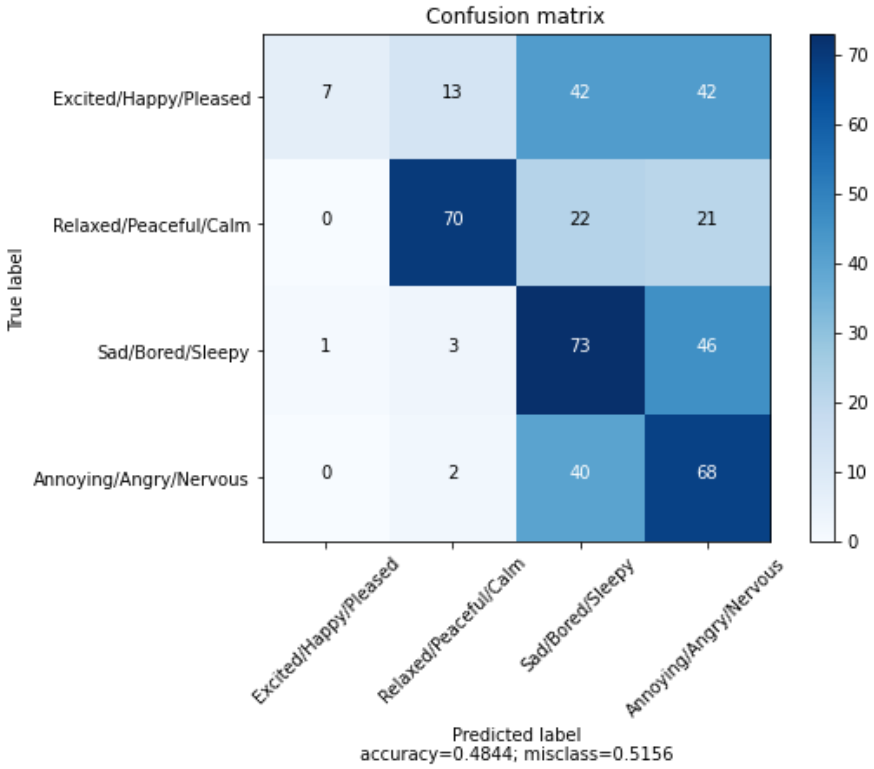


Figure 26: Mood End-to-End GRU