Music Genre Classifier

Ehtan Johnson
Johnson.10404@buckeyemail.osu.edu
Department of Physics, The Ohio State University

Motivation

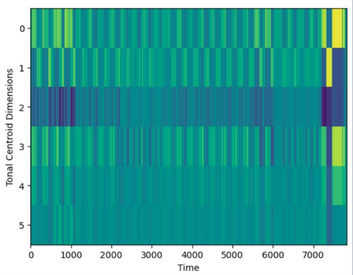
This project focuses on creating a music genre classification model using machine learning, driven by a personal passion for music and the challenge of working with audio. It uses the Jamendo dataset for training, implementing a Keras Fully Convolutional Network (FCN) for multi-label, multi-class classification of the 10 most populated genres. The model achieved an accuracy of 82%, a loss of 1.40, and an AUC of 0.86, though results were based on a smaller subset of 500 audio files due to time constraints. The project aims to be a foundational tool for music recommendation algorithms, analyzing genres and relating similar music. The model processes four unique features (MFCC, Mel Spectrogram, Chroma Vector, and Tonal Centroid) to classify genres from audio files.

Dataset

This project uses the Jamendo dataset, which contains over 55,000 audio tracks in .opus format with 195 tags across genre, instrument, and mood/theme categories. The dataset is split into 90% training and 10% validation, with metadata including audio file ID, artist ID, album ID, duration, genre, instruments, and mood. To process this data, five functions were created to extract features from each audio file, resulting in arrays of length 498 for normalization. Tracks with missing metadata were removed, and the data was split using the train_test_split function from scikit-learn, leaving the final dataset as 72% training, 18% testing, and 10% validation. There were about 10,000 total tracks to work with after the data had been cleaned.

Features :

- **Mel-Frequency Cepstral Coefficients (MFCC)**: A representation of the power spectrum mapped onto the Mel scale, useful for capturing speech and music characteristics.
- **Mel Spectrogram**: A time-frequency representation of the audio signal.
- **Chroma Vector**: A representation of harmonic content, broken into 12 bins corresponding to musical octaves.
- **Tonnetz**: A projection of the Chroma Vector into a 6-dimensional space representing musical harmony, such as the perfect fifth, minor third, and major third.

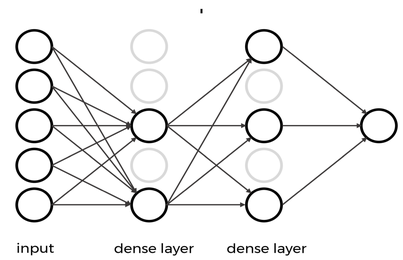


Visual representation of the Tonnetz feature

Methods

A Fully Convolutional Newtork was used. The architecture is as follows :

- Three dense layers (256, 128, 64 filters)
- Dropout layer (0.5 -> Drops the number of nodes used by 50%)



A visual representation of the dropout layer. Note that my model has an additional dense layer

Training :

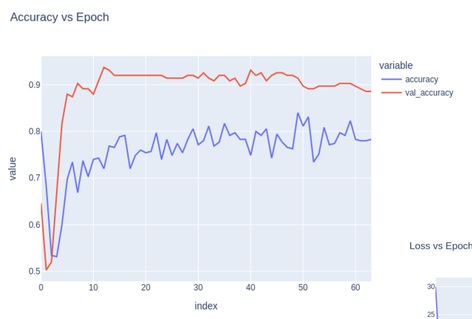
- Learning rate – 0.0001
- Batch size – 32
- Early stopping incorporated to aid against overfitting
- Metrics – Accuracy, AUC (area under curve)

Results

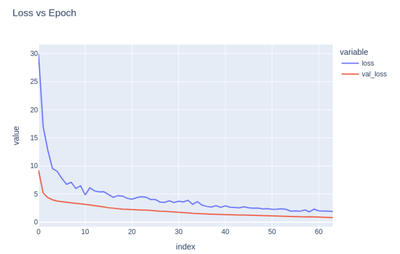
These are the results from evaluating the model :

Accuracy	Loss	AUC	Hamming Loss
0.82	1.40	0.86	0.043

F1 Score	Precision	Recall	
0.68	0.78	0.61	Micro
0.08	0.081	0.078	Macro



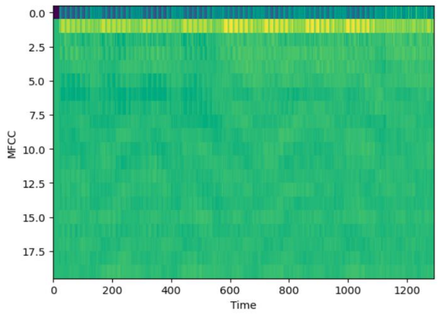
These plots show accuracy and loss as the model learns. They indicate that the model is performing well in conjunction with the values in the charts above.



Discussion

The algorithm's hyperparameters were optimized through trial and error, using a 1-D input array and three hidden layers in a Fully Convolutional Network (FCN). ReLU activation was used for hidden layers, and sigmoid for the output, suitable for multi-label classification. Adam optimizer with a learning rate of 1e-4 and BinaryCrossentropy loss function were selected. The model was trained for 64 epochs with early stopping to prevent overfitting. Data was split using sklearn's train_test_split.

Performance was evaluated using accuracy, AUC (0.86), precision (0.78, 0.087), recall (0.61,0.078), F1 score (0.68. 0.08), and Hamming loss (0.043). The model achieved 82% accuracy, 1.40 loss, indicating good performance. However, due to time constraints I was unable to run on the full dataset, which could have resulted in an underrepresentation of some genres, leading to the low macro scores.



Visual representation of the MFCC feature

Conclusions and Future Work

This project has been a valuable learning experience, as it was my first time creating a project from scratch. By using techniques learned in class and exploring additional methods, I filtered the Jamendo dataset to create a multi-label classifier for 10 music genres. The model, a Keras Fully Convolutional Network (FCN), performs well overall, as shown in the results. While I didn't explore many models due to time and personal constraints, the FCN outperformed the XGBoost model. With more time and a team, I would explore other models (e.g., CNN, Random Forest) and address potential overfitting issues, as well as expand the genres the model can classify. Overall, I am very happy with the results of this project.

References:
Picture in "Methods" was sourced from - https://enccs.github.io/deep-learning-intro/04-networks-are-like-onions/index.html