Music Genre Classification: comparing impact of feature extraction on the model prediction

Deep Learning Project Report

Sulagna Saha (saha23s@mtholyoke.edu)
Mentor: Bálint Gyires-Tóth (toth.b@tmit.bme.hu)

Introduction:

This project proposes the implementation of a deep learning model for music genre classification, mainly to observe how proper feature extraction can impact the model accuracy. The task of music genre classification, as defined by Tzanetakis and Cook [6], aims to accurately predict the genre of a given piece of music. Such a task could enhance various music-related applications, including music recommendation systems and search engines, which rely on precise genre labeling. Accurate genre classification could also assist in music analysis and understanding the characteristics of different music genres. This project was specifically designed to

- Extracting features from scratch and also using a pre-trained model.
- Using current deep learning techniques: LSTM, CNN and Transformer model and comparing their predictions

Previous solutions:

Recent works [7][8][9] heavily focused on CNN models and the combination of CNN with other deep learning techniques. Li et al. 2017 [4] said the variations of musical patterns with a certain transformation such as, Fast Fourier Transform (FFT) and Mel-frequency Cepstral Coefficient (MFCC), are similar to images which work well with CNNs in image classifications. Dong et al 2018 [5] mentioned that CNN models achieved human-level accuracy (70%) in case of music genre classification. Definitely, CNN models are good for classification but not many researchers used it for feature extractions. Moreover, not many of the researches worked by comparing how other techniques such as LSTM and Transformer model works in the similar

dataset. Besides, the VGGish model from Google has a great potential for extracting features from large-scale audio files, which is used in our proposed method.

Dataset:

The GTZAN dataset consists of 1000 audio tracks each 30 seconds long. It contains 10 genres, each represented by 100 tracks. The tracks are all 22050 Hz Mono 16-bit audio files in .wav format.

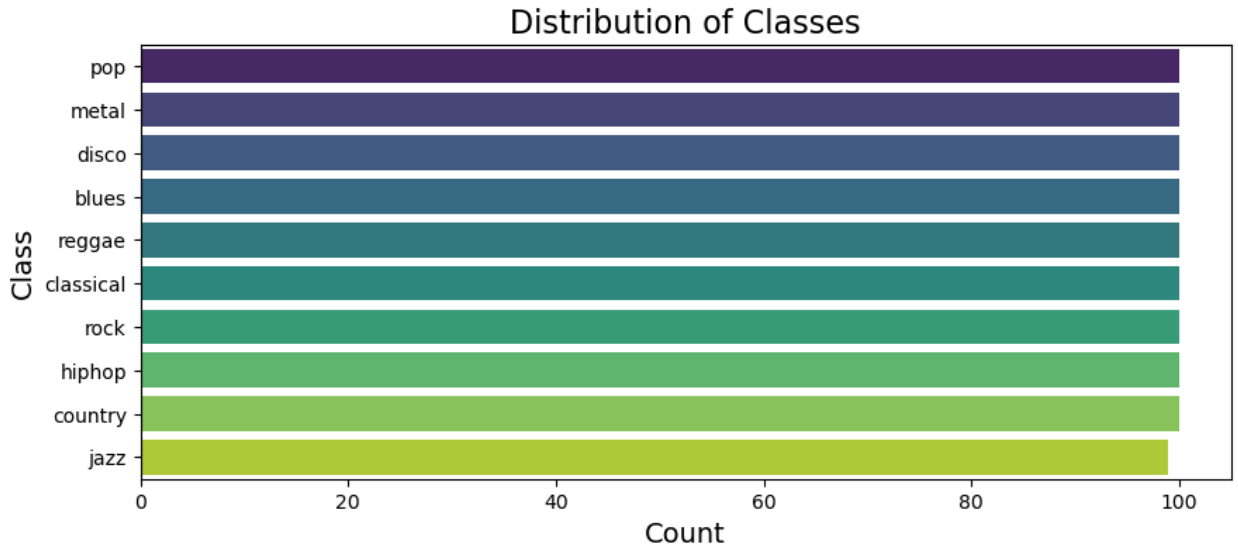

Figure 1: The distribution of 10 genres in GTZAN dataset

Proposed Method:

There's recent work on the same dataset using the VGGish model for feature extraction and then CNN model to classify the genres. Their validation accuracy was 45.60%. I was curious if that accuracy can be increased by using other model architectures - that can emphasize the importance of using the VGGish model to extract features. Here's the visualization of 6 proposed models:

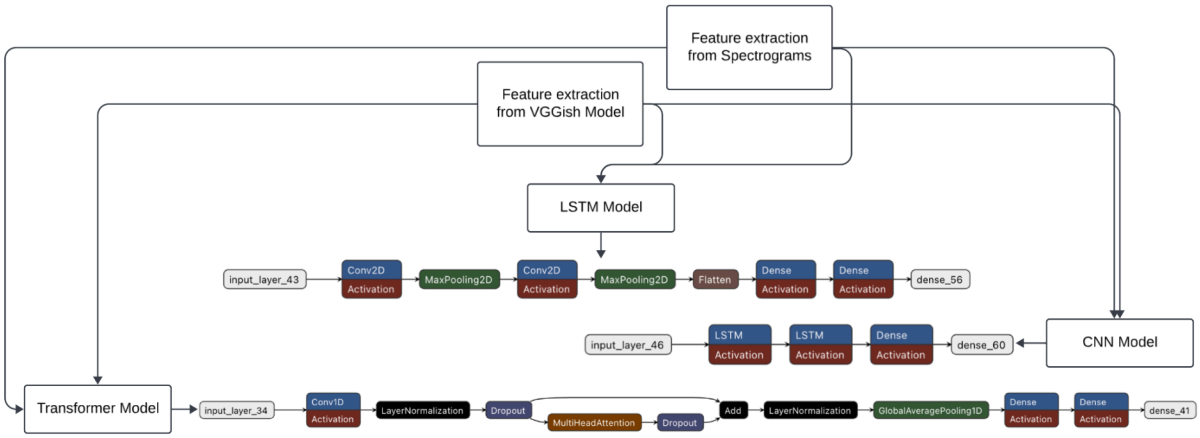

Figure 2: Model architectures and 2 kind of feature extraction

Evaluation Method:

Because it's a multi-class classification problem, I have used these procedures to evaluate:

i) balanced accuracy score,

ii) metrics with the classification_report function,

iii) confusion matrix

Here's the visualization for the models I have tried and the comparison of results of balanced and validation accuracies:

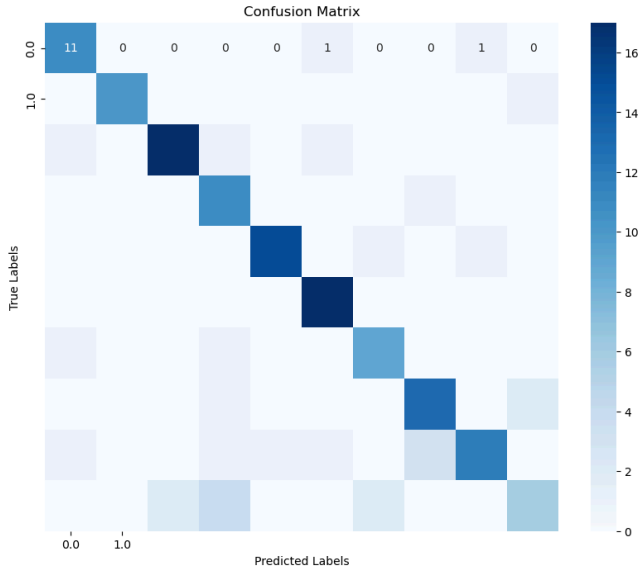

Figure 3: Confusion matrix of transformer model using VGGish as a feature extractor

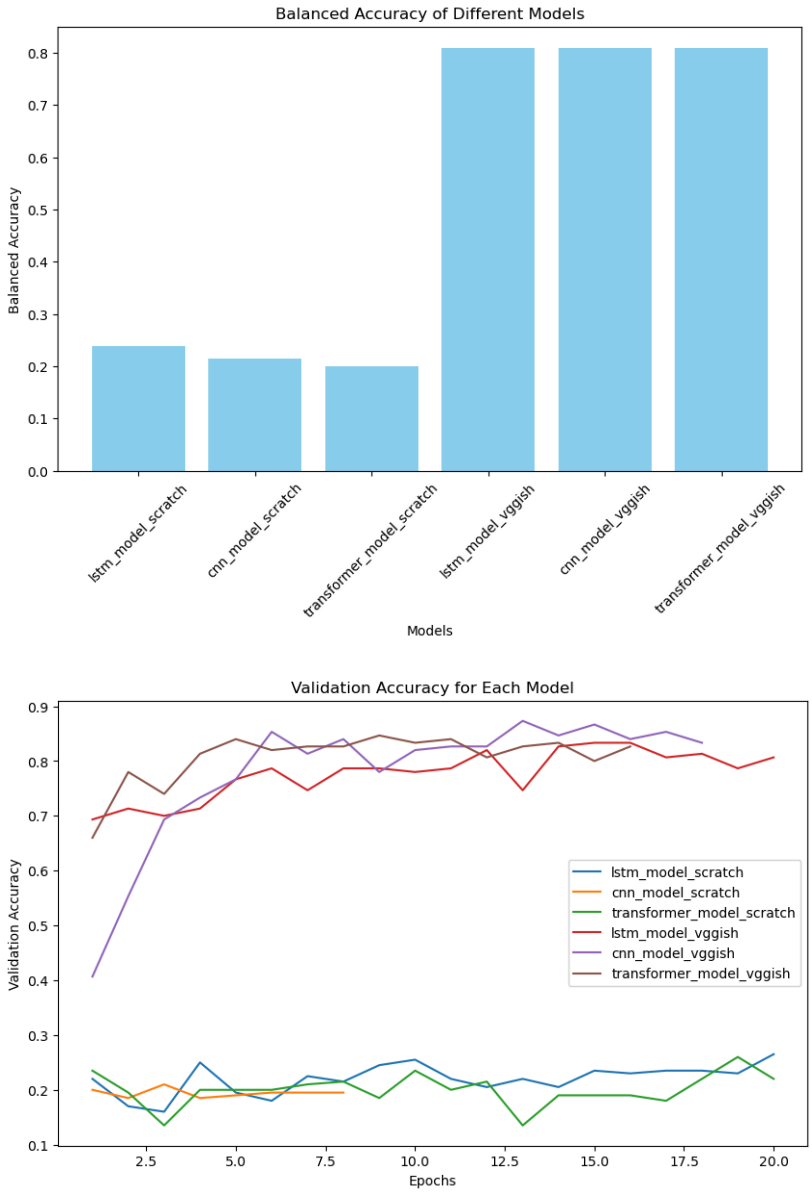

Figure 4: Evaluation methods (balanced accuracies and validation accuracy changes

Results:

The transformer model using the VGGish model as a feature extractor achieved a better validation (80.9%) and test accuracy compared to the previous works using the VGGish model as a feature extractor. (45.60%) [3] The spectrograms had not been a good way to extract features (even though merging it with other information collected from the audios probably can help make the features more meaningful).

References:

[1] GTZAN Dataset (https://datasets.activeloop.ai/docs/ml/datasets/gtzan-genre-dataset/)

[2] VGGish Model (https://apple.github.io/turicreate/docs/userguide/sound_classifier/how-it-works.html)

[3] Bhavesh Mittal. "Music Genre Classification using VGGish CNN." (https://www.kaggle.com/code/bhaveshmittal/music-genre-classification-vggish-cnn/notebook)

[4] T. L. Li, A. B. Chan, and A. Chun, “Automatic musical pattern feature extraction using convolutional neural network,” in Proc. Int. Conf. Data Mining and Applications, 2010

[5] D. Mingwen, “Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification”, 2018

[6] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on speech and audio processing. Scaringella, N., Zoia, G., & Mlynek, D. (2006)

[7] Automatic genre classification of music content: a survey. IEEE Signal Processing Magazine. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017)

[8] Convolutional recurrent neural networks for music classification. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

[9] Dieleman, S., & Schrauwen, B. (2014). End-to-end learning for music audio. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).