MUSIC GENRE CLASSIFICATION & GENERATION USING DEEP LEARNING

Gaurav Dawra (2019039) Subhanshu Bansal (2020135) Hardik Patel (2020507)

Abstract

This research paper investigates the problem of music genre classification, which is a fundamental task in music information retrieval. The paper extensively studies deep learning algorithms for automatic music genre classification, including KNN(K-Nearest Neighbours), CNN (Convolutional Neural Networks) and PRCNN (Convolutional Recurrent Neural Networks). The study employs a large dataset of audio tracks and evaluates the performance of each algorithm using standard classification metrics. The paper also explores audio features, such as spectral features, and compares their effectiveness in genre classification. The study results provide insights into the state-of-the-art music genre classification and offer guidelines for selecting appropriate algorithms and features for different music classification tasks. The findings of this research can have significant implications for music recommendation systems, music streaming platforms, and other music-related applications.

Problem Statement

Music genre classification is a challenging problem in music information retrieval, as it involves identifying the musical style of a given audio recording. Despite significant progress in this area, several issues still need to be addressed. One of the main challenges is the need for a standard methodology for evaluating the performance of different algorithms and features. Moreover, the effectiveness of existing approaches may vary depending on the specific characteristics of the audio dataset and the music genres involved. Another challenge is the need to develop more robust and accurate methods to handle the diversity and complexity of music genres, particularly for emerging or hybrid genres that may not fit into conventional classification schemes. Furthermore, there is a growing demand for music genre classification systems that can adapt to individual preferences and user contexts, which requires developing more sophisticated models that can capture the subjective and contextual aspects of music listening. Therefore, there is a need for further research to address these challenges and advance the state-of-the-art in music genre classification, which can have significant implications for the music industry and music-related applications.

Related Work and Existing Baselines

Music genre classification is a well-established research area in music information retrieval, and there has been a significant amount of prior work in this field. Several approaches have been proposed for automatic music genre classification, including traditional machine learning algorithms such as support vector machines, decision trees, and k-nearest neighbours, as well as deep learning methods such as convolutional neural networks, recurrent neural networks, and hybrid models.

One of the most widely used baselines for music genre classification is the GTZAN dataset, used in several studies as a benchmark for evaluating the performance of different algorithms and features. In a seminal work, Tzanetakis and Cook used the GTZAN dataset to compare the effectiveness of various audio features for music genre classification, including spectral, temporal, and statistical features. They showed that combining multiple features can improve the accuracy of genre classification and that some features are more effective than others for different genres.

Since then, many studies have explored different approaches to music genre classification, such as ensemble methods, feature selection techniques, and transfer learning. For instance, in a recent study, Liu et al. proposed a hybrid deep learning model that combines convolutional neural networks and recurrent neural networks for music genre classification. They achieved state-of-the-art performance on the GTZAN dataset and showed their model could generalise well to other datasets.

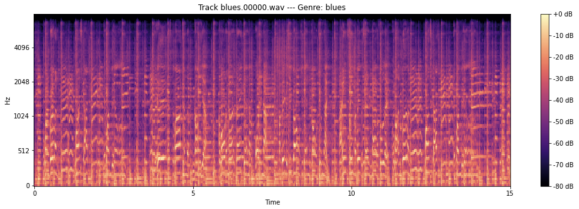
Another line of research has focused on developing more fine-grained music genre classification systems that recognise subgenres and hybrid genres. For example, in a recent study, Huzaifah et al. proposed a hierarchical deep-learning model that can classify music tracks into multiple levels of genres, including

main, subgenres, and fine-grained. They showed that their model could achieve high accuracy on a dataset of Indian classical music with a complex taxonomy of music genres.
Overall, the existing baselines and related work in music genre classification provide valuable insights into the state-of-the-art in this field and offer guidelines for selecting appropriate algorithms and features for different music classification tasks. However, several challenges still need to be addressed, such as improving the robustness and generalisation of classification models, handling the diversity and complexity of music genres, and developing more user-centric and context-aware music recommendation systems.

Dataset Details

The GTZAN dataset is a widely-used music dataset for research in music genre classification, which Tzanetakis and Cook introduced in 2002. The dataset contains 1,000 audio tracks, each 30 seconds long and sampled at a rate of 22.05 kHz. The audio tracks are evenly distributed across 10 music genres, including blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock. Each genre contains 100 audio tracks from different artists and albums, which were selected based on their popularity and representativeness of the genre. The audio tracks are in the WAV format and have been preprocessed to remove silence and normalise the volume level. The dataset also includes metadata for each audio track, such as the artist name, album name, and song title. The GTZAN dataset has been widely used in research on music genre classification, with many studies reporting their results on this dataset. Despite some criticisms of the dataset, such as its limited size and potential biases in genre labelling, it remains a popular benchmark for evaluating music genre classification algorithms and comparing their performance with other studies.

Dataset Pre-Processing



As we all know, Deep Neural Networks need enormous input data to learn robust feature representation. However, the datasets used in our experiments are 1000 song excerpts and 4180 music tracks, respectively. To increase the number of tracks, we cut each song excerpt into shorter music clips with 3 seconds duration and 50 overlaps. Thus, the increased training datasets help our architecture avoid overfitting partly and have better performance on feature extraction. Similarly, we calculate Fast Fourier Transforms (FFTs) on frames of length 1024 at 22050 kHz sampling rate with 50% overlap and use the absolute value of each FFT frame.
We finally construct an STFT spectrogram with 128 frames; each frame is a 513-dimensional vector.

Methodology

While we implemented the baseline modules mentioned, we also implemented our own CNN structure over the mel-spectograms. The layers involved 2 convolutional layers, one of kernel size 3x3 and other of kernel size 5x5. We applied 2d max-pooling after each convolutional layer. After the convolution we ended up with a feature vector of size 16x23x35. We then applied a fully connected ANN over this feature vector and our output layer consisted of 10 neurons as expected. This ANN consists of 3 hidden layers. We got 70.5% accuracy over the GTZAN which exceeds any of the baseline models. Here we have shown the classification report for the same.

Accuracy of the network on the test images: 70.500000 %

	precision	recall	f1-score	support
0	0.71	0.60	0.65	20
1	0.95	1.00	0.98	20
2	0.75	0.60	0.67	20
3	0.50	0.50	0.50	20
4	0.50	0.70	0.58	20
5	0.60	0.90	0.72	20
6	0.89	0.80	0.84	20
7	0.84	0.80	0.82	20
8	0.71	0.50	0.59	20
9	0.76	0.65	0.70	20
accuracy			0.70	200
macro avg	0.72	0.71	0.70	200
weighted avg	0.72	0.70	0.70	200

Observation and Future Work

Observation: On running the KNN, CNN and PRCNN Models (the baseline models), the Classification Report of each of the models was as follows-

	precision	recall	f1-score	support
0	0.45	0.56	0.50	9
1	0.86	0.86	0.86	7
2	0.29	0.44	0.35	9
3	0.45	0.42	0.43	12
4	0.45	0.83	0.59	6
5	0.29	0.25	0.27	8
6	0.80	0.80	0.80	10
7	0.85	0.73	0.79	15
8	0.45	0.42	0.43	12
9	0.80	0.33	0.47	12
accuracy			0.55	100
macro avg	0.57	0.56	0.55	100
weighted avg	0.59	0.55	0.55	100

KNN

	precision	recall	f1-score	support
0	0.40	0.22	0.29	9
1	0.56	0.71	0.63	7
2	0.09	0.11	0.10	9
3	0.62	0.42	0.50	12
4	0.40	0.67	0.50	6
5	0.86	0.75	0.80	8
6	0.56	1.00	0.71	10
7	0.60	0.20	0.30	15
8	0.83	0.83	0.83	12
9	0.33	0.42	0.37	12
accuracy			0.51	100
macro avg	0.53	0.53	0.50	100
weighted avg	0.54	0.51	0.49	100

CNN

	precision	recall	f1-score	support
0	0.60	0.33	0.43	9
1	0.86	0.86	0.86	7
2	0.33	0.22	0.27	9
3	0.38	0.42	0.40	12
4	0.08	0.17	0.11	6
5	0.57	0.50	0.53	8
6	0.25	0.80	0.38	10
7	0.50	0.13	0.21	15
8	0.60	0.50	0.55	12
9	0.00	0.00	0.00	12
accuracy			0.37	100
macro avg	0.42	0.39	0.37	100
weighted avg	0.41	0.37	0.35	100

PRCNN

As you can see, our CNN model performed better overall than baseline KNN, CNN and PRCNN models.

Future Work: Using this Classification model, we will implement a Generative Adversarial Network (GAN) to generate the most accurate music from all the data fed. This model will be able to make music using text inputs provided by the user.

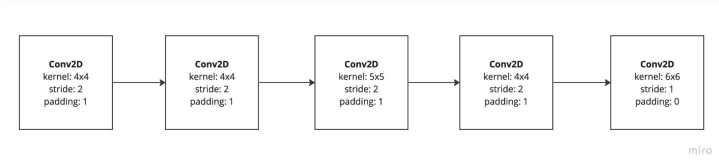
Music Recommendation

One part of the project we worked on was music generation. Although this recommendation might seem basic, but it has a unique way to be implemented. It recommends music on the basis of your mood. The first step of this is to collect data and it is done by storing various facial expressions which show emotions such as happy, sad, angry, punk etc. The data is collected using CV techniques and once the data is collected the model is trained on this data. After successfully training the model we run the model on web based app using python library named streamlit. The prompt there asks us to enter the which language of songs you want to hear along with the name of the artist. Then it opens up the camera and reads your emotion. Once it registers the emotion, it searches for those songs on youtube and presents them to you. The reason to youtube as a mean to show the songs is due to the unavailability of dataset of well classified songs based on emotions.

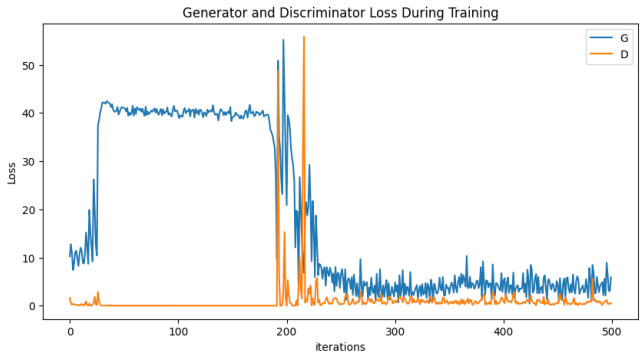
Music Generation

For music generation, we implemented a naive approach. We tried to generate fake MEL spectrograms for the audio files and convert them back to audio. For this purpose, we used GANs. We were able to successfully generate mel spectograms belonging to the distribution provided in GTZAN datasets. Although we were not successfully and completely able to convert these mel spectrograms to audio files. But we were able to achieve a reasonable equilibrium in the GAN model.
We used DCGANs or GANs that involve convolutional layers to generate and discriminate against the fake spectrograms.

The diagram for the discriminator is shown below. The one for the generator is the same in reverse and with Conv2dTranspose layers instead of Conv2d.



We were able to get successful convergence to equilibrium.



The next step would be to apply this to general sized mel spectrograms to obtain viable audio files.

References

G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," in IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293-302, July 2002

Choi, Keunwoo, et al. "Convolutional recurrent neural networks for music classification." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017

Feng, Lin, Shenlan Liu, and Jianing Yao. "Music genre classification with paralleling recurrent convolutional neural network." arXiv preprint arXiv:1712.08370 (2017)

**Github Repository Referenced: link