Music Genre Classification*

1st Justin James
Computer Science and Engineering
PES University
Bangalore, India
justinjames838@gmail.com

2nd Anagha HM
Computer Science and Engineering
PES University
Bangalore, India
anu.hm0520@gmail.com

3rd Hanuraag Ravilla Baskaran
Computer Science and Engineering
PES University
Bangalore, India
rbhanuraag01@gmail.com

***Abstract*—Categorizing music files according to their genre is a challenging task in the area of music information retrieval (MIR). In this study, we compare the performance of two classes of models. We use three methods, the first being an artificial neural network, to establish a baseline model. We then account for overfitting with the previous model, as artificial neural network models are prone to suffering from overfitting. After that we train a convolutional neural network model as our main driver model. We find we get a 74 percent accuracy.**

I. Introduction

With the growth of online music databases and easy access to music content, people find it increasing hard to manage the songs that they listen to. One way to categorize and organize songs is based on the genre, which is identified by some characteristics of the music such as rhythmic structure, harmonic content and instrumentation. Being able to automatically classify and provide tags to the music present in a user's library, based on genre, would be beneficial for audio streaming services such as Spotify and iTunes. We use three methods, an ANN with 5 layers, the same ANN model accounted for overfitting within the model, and a CNN with 6 layers. The rest of this paper is organized as follows.we describe the existing methods in the literature for the task of music genre classification, following an overview of the the dataset used in this study and how it was obtained. Next, we discuss the proposed models and the implementation details. We finally report the results and conclude this paper with a quick overview.

II. Literature Review Summary

In this paper the dataset used is the GTZAN dataset which comprises 10 genres and 100 30 sec audio files. The audio files were each further split into 10 files of 3 sec each. The spectrograms for the audio files were generated. The authors then used a Convolutional Neural Network on the spectrograms to extract 4 feature maps. These features are hierarchical and are extracted by convolutional kernels. It extracts both low level features (e.g onset) and high level features (e.g musical instrument patterns). The authors also use a Support Vector Machine and K Nearest Neighbour model for accuracy comparisons. They conclude that the CNN along with the SVM offer higher accuracy than the KNN model. [1]

This paper uses the GTZAN dataset. Audio segments of 10-100ms are extracted. MFCC is a short term power spectrum of sound and is performed on these audio segments. In this work, R machine learning library 'H2O' Deep Learning is used to implement DNN. For classification of music genres, the dataset is partitioned randomly into three parts: 60% for training, 20% for validation, 20% for testing. This experiment uses 350 hidden layers and 60 epochs .It's followed by rectifier activation. The accuracy is 97.8% [2]

This paper uses the GTZAN dataset for training and testing. There is a 60% to 40% training split between training and testing data of the feature sets. The deep belief model developed based on the principle of restricted Boltzmann Machines is first tested on two class classification, then 3 class classification, and so on. While comparing a neural network(NN) and the deep belief model(DBM), it is seen that while NN and DBM have about 97 to 98 percent accuracy, the DBM is more computationally intensive and the NN is more efficient [3]

This paper uses different modals i.e text (album reviews) , images(album covers) and audio from the songs to generate a multi-modal feature space and perform multi-label classification. They have used the MST-1 Dataset and have used CNN's with softmax Activations for classifications. They have generated spectrograms from the audio and have used tf-idf vectorization on the text to generate representations. Then a multimodal feature space is learned and multi-modal fusion is performed. The CNN based approach applied on the audio spectrograms achieved an accuracy of 88% and it was concluded that multi-modal feature representations performed better than classifications based on individual feature representations. [4]

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

III. Dataset

The dataset that was chosen for our problem statement was GTZAN Genre Dataset. The GTZAN Genre Dataset consists of 1000 audio tracks with a 30-second duration. The dataset is

divided into a total of 10 genres, each with 100 tracks. All the tracks are 22050Hz Mono 16-bit audio files in .wav format.

IV. Proposed Methodology

The main goal of this project is to implement music genre classification. The first part of the approach is preprocessing the dataset, which involves extracting the MFCC of the audio files in the dataset. The second part of the approach involves modelling an Artificial Neural Network. The third part of the approach tries to overcome the shortcomings of the Artificial Neural Network, namely the tendency of ANN models to overfit. The last part of the project involves implementing a Convolution Neural Network model.

V. Model Definition

A. Preprocessing

An important part of the project involves preprocessing the dataset in order to bring it to appropriate form, that could then be subjected to model evaluations. This is done by loading the dataset and performing MFCC over the .wav files. In sound processing, the mel-frequency cepstrum co-efficients are coefficients that are derived from the type if cepstral representation of sound. The mel scale tries to capture small frequency differences that can be identified by humans. The python model librosa has an inbuilt function called MFCC that can be used to implement this.

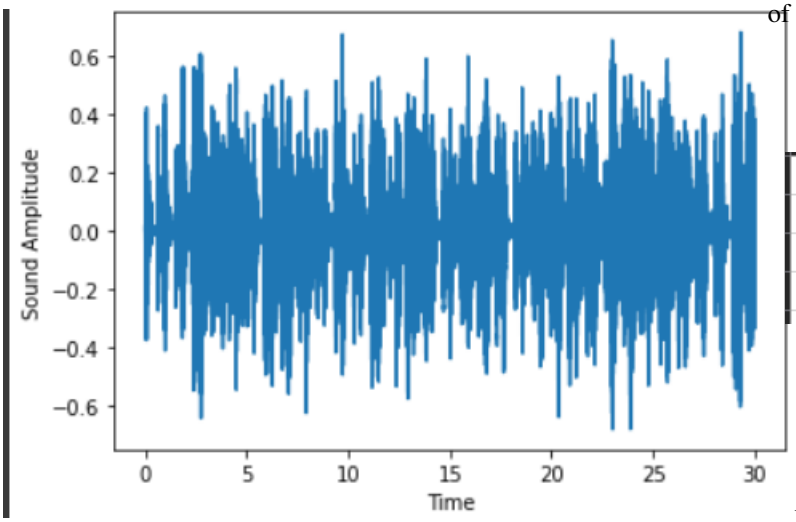

Fig. 1. Spectrogram

B. Artificial Neural Network

Artificial Neural Network is a computational model that mimics the way nerve cells work in the human body. The neural network implemented here has 5 layers. The first layer is Flatten and this is followed by 4 dense layers. This neural network has a total of 1,014,218 parameters. To compile this model we have used the 'Adam' optimizer and the loss function used is "sparse_categorical_crossentropy". This achieves an accuracy of 59%.

C. Overcoming Overfitting of ANN

One of the major shortcomings of ANNs are that they tend to overfit the data. This can be majorly overcome by following these processes: Making architecture less complicated Using augmented data Early stopping of training Adding dropout layers Regularization / Standardization We have implemented Dropout layers and kernel regularizers. Dropout layers set randomly sets input units to 0 with a frequency of rate at each step during training time, which helps prevent overfitting. Kernel regulizers apply a penalty on the layer's kernel, thus overcoming overfitting. This model has 9 layers. The first layer is Flatten followed by alternating dense layers and dropout layers. The last layer of the model is a dense layer. This model has 1,015,978 trainable parameters. To compile this model we have used the 'Adam' optimizer and the loss function used is "sparse_categorical_crossentropy". This achieves an accuracy of 54%.

D. Convolutional Neural Network

A convolutuonal neural network is a class of deep neural network used to visualize imagery. This is implemented using keras layers conv2d, MaxPool2D and BatchNormalization. This model has 16 layers and has 33,242 parameters.It was compiled with 60 epochs. To compile this model we have used the 'Adam' optimizer and the loss function used is "sparse_categorical_crossentropy". This achieves an accuracy of 73%.

VI. Final Results

No	Model Used	Accuracy Percentage
1	ANN	59%
2	ANN w/ overfitting	54%
3	CNN	73%

Fig. 2. Final results

VII. Conclusion

In this paper, we aim to train multiple models to effectively classify music clips into different genres. We propose three methods to solve this. The first method uses an Artificial neural network with 5 layers, with a 59% accuracy. The second method accounts for overfitting in the artificial neural network, and we get a 54% accuracy after accounting for overfitting. The third method uses convolutional neural network, with 6 layers, with which we achieve a 73% accuracy. Hence, it is shown that our CNN model has a better accuracy, and is the best proposed method for classifying music. More work can be done on MFCC preprocessing, to account for the model being more realistic in differentiating between frequency steps as low as 100Hz.

VIII. Acknowledgment

We would like to thank Dr.Uma D for her continued support and engagement, and the Topics in Deep Learning team for their invaluable guidance throughout this experience. We would also like to thank the PES Computer Science Department for this wondrous opportunity and constant support with our research endeavors.

IX. Contributions

- Anagha H M: Convolutional Neural Network
- Hanuraag R Baskaran: Artificial Neural Network
- Justin James:Overcoming Overfitting of ANN

References

[1] Jawaherlalnehru, G. , S. Jothilakshmi. "Music Genre Classification using Deep Neural Networks." International Journal of Scientific Research in Science, Engineering and Technology 4.4 (2018): 935.

[2] Feng, Tao. "Deep learning for music genre classification." private document (2014).

[3] Yang, Xiaohong Chen, Qingcai Zhou, Shusen Wang, Xiaolong. (2011). Deep Belief Networks for Automatic Music Genre Classification.. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2433-2436. 10.21437/Interspeech.2011-633.

[4] Oramas, Sergio, et al. "Multimodal deep learning for music genre classification." Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21. (2018).