Music Genre Classification

ISSN 2395-1621

#1Rohan Mistry, #2Prof. N. G. Bhojane, #3Ashutosh Deshmukh, #4Renukesh Kothalkar,
#5Abhishek Kharmale

1mistryrohan0@gmail.com
2ngbhojane.scoe@sinhgad.edu
3d22.ashutosh@gmail.com
4renukesh.kothalkar@gmail.com
5abhishek9817@gmail.com

#12345Computer Department, Sinhgad College of Engineering,
Pune, Maharashtra, 411041, India.

ABSTRACT

Music genre allows us to categorize the musical items into broader categories that share similar characteristics. A music sample consists of the audio file, cover image and text review. By analysing these attributes, we can predict the genre of given sample which can then be used for applications like music recommendation, customized music playlist creation and for other Music Information Retrieval (MIR) applications. For classification, convolutional neural network deep learning algorithm is used. Multi-modal dataset is used for feature extraction. Auralisation will be implemented for better understanding of extracted features. To improve the performance and better learning of the model, data augmentation technique will be applied on dataset.

Keywords— Convolutional Neural Network, Deep learning, Multi-modal, Data augmentation, Auralisation.

ARTICLE INFO

Article History

Received: 8th March 2020

Received in revised form :

8th March 2020

Accepted: 10th March 2020

Published online :

11th March 2020

I. INTRODUCTION

Music genre can be defined as a categorization method that identifies chunks of music which may belong to shared class or set of conventions. The term genre (e.g., Jazz, Indie, Pop) is a subject to interpretation and it is often the case that genres may very fuzzy in their definition. Despite the lack of a standard criteria for defining genres, the classification of music based on genres is one of the broadest and most widely used. Music genre plays a very significant role in music recommender system. Labels in music genre provides category in organized manner and can classify music, artists and albums in broader group having similar features. Music genre classification is a widely studied problem in the Music Information Research (MIR) community [1].

Music classification is widely based on music genres, and popularly used in online music streaming services. On a daily basis the amount of songs released continues to grow rapidly, and specifically on online platform such as Saavn, Amazon Music and Spotify – a 2016 report suggests that tens of thousands of songs were released every month on Spotify – this certainly creates a need of a proper system which can organize and classify songs when provided with accurate meta- data. The functionality of being able to instantly classify songs according to Genre or playlist or library is immensely useful for purchasing or Streaming services. This in turn increases the capacity for statistical analysis that is more accurate and completes labelling of music

II. RELATED WORK

Multi-modal deep learning approach [2] is used to learn and combine multi-modal data i.e. audio, text, and images for music genre classification. The learned representations from different modalities are aggregated together which improves the accuracy of the classification. Multi-modal data repre- sentations and effect of different learned features and their combinations is proposed by Oramas et al. (2018). Qualitative analysis of results of experiments on different modalities is done. The MSD dataset [3] used for the experiment consists of audio-descriptors and not the actual audio samples.

Oramas et al. (2017) categorizes musical items into multiple labels, using three different data modalities: audio, text, and images. This paper aims on advancing the field of music classification by framing it as multi-label genre rather than single-label genre as these may not be necessarily mutually exclusive. The explanation of combination of different modalities for classification has not been addressed in this paper.

Procedure and results of deconvolution and Auralisation [4] is used to extend the understanding of CNNs in music classification. Auralisation is reconstructing audio signal from deconvolved spectrograms. Choi et al. (2016) explains the trained features of a 5-layer CNN based on deconvolved spectrograms and auralised signals. Auralisation is introduced which is extension of CNNs visualization which enables to understand the mechanism of CNNs that are trained with audio signals. Audio feature extraction by the layers on CNN is explained through experiments. Analysis of learnt features is not done in this paper. After feature extraction, music classification is not explained in this paper.

Deep convolutional neural network architecture [5] which, in combination with a set of audio data augmentations, pro- duces results for environmental sound classification (Salamon et al., 2017). This combination improves the performance and outperforms both the proposed CNN without augmentation with augmentation. Data augmentation technique is used to overcome the problem of data scarcity which significantly influences the performance of proposed CNN architecture. The performance of the model could be improved further by applying class-conditional augmentation.

Pons et al. (2019) proposes a way to evaluate CNN archi- tecture via comparing the obtained classification accuracies by using different randomly weighted CNN architecture as feature extractors [6]. The goal is to compare classification accuracies when using different randomly weighted architectures. Nontrained randomly weighted CNNs are used as feature extractors which in some cases match or outperform the trained CNNs.

III. DATASET

We will use two datasets GTZAN for audio feature extraction and MARD (Multimodal Album Review Dataset) for text and image feature extraction to train our system models.

a. *GTZAN Dataset*

The dataset consists of 1000 audio tracks each 30 seconds long. There are 100 tracks corresponding to each genre so we have a total of 10 genres. The tracks are all 22050Hz Mono 16-bit audio files in .wav format. We are going to use this dataset for training audio feature extractor model. We will first apply data augmentation on this data and create 4000 additional audio tracks by applying time stretching and pitch shifting on each audio track.

b. *MARD (Multimodal Album Review Dataset)*

The dataset contains texts and accompanying metadata originally obtained from a much larger dataset of Amazon customer reviews, which have been enriched with music metadata from MusicBrainz, and audio descriptors from AcousticBrainz. MARD amounts to a total of 65,566 albums and 263,525 customer reviews from 13 different genres. Album cover images and customer reviews corresponding to 10 genres from GTZAN dataset will only be used instead of the available 13 genres for compliance of image and text model with audio model. The album cover images will be downloaded from their URL provided in the dataset and stored according to their classes. These album cover images will be then used to train image feature extractor model. The dataset consists of customer reviews which we will use to train the text feature extractor model

IV. METHODOLOGY

We will initially build the system models by training them using training dataset. The audio feature extractor model will be trained using audio files by applying augmentation and convolutional neural networks. The text feature extractor model will be trained using text reviews of the songs and using feed forward network. The image feature extractor model will be trained using album cover image of the songs and deep residual network. We will train the classification model by concatenating feature vectors of above three models into a single feature vector, which becomes the input to a simple feedforward network, where the input layer is directly connected to the output layer.

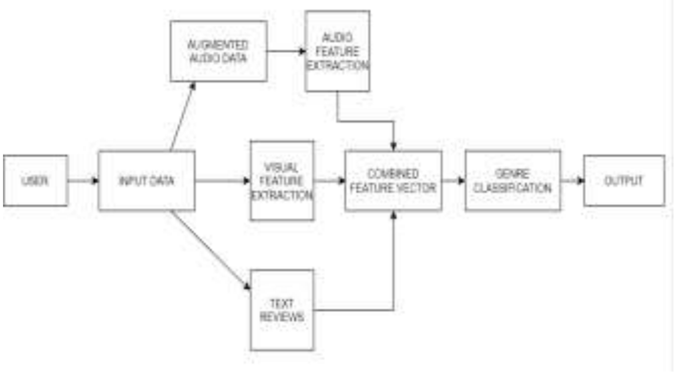

Fig. 1. System Overview Diagram

The system then can take input from the user which consists of audio file of the song, text reviews of the song and album cover image. Audio file is augmented and processed and then the audio features are extracted from it in the form of audio feature vector. Album cover image and text reviews for particular input are used to extract the features and then concatenated together and sent to classification model for classifying the genre.

a. Data Augmentation

Data augmentation is used in machine learning in order to increase the diversity of data for training model and thereby achieving a good performance. There are various data augmentation techniques such as visual augmentation and audio augmentation. In our work we are going to implement the data augmentation on audio files before giving it to the Audio feature extractor. Data Augmentation can be performed by using two methods. First, we will perform time stretching which involves slowing down and speeding up the given audio sample while keeping the pitch untouched. Second, Pitch Shifting which is to raise and lower the pitch of the audio sample while keeping the duration unchanged. We are going to apply both methods on audio files to enrich the data and audio feature extraction.

b. Audio Classification Methodology

Log compressed constant Q transforms (CQT) will be computed for all the available audio files in the dataset. We will sample two 15-second long patches from each track, resulting in the fixed-size input to the CNN. CQT patches are produced for each audio file and are fed to series of convolutional layers with ReLU (Rectified Linear Units) which will implement the activation function. Then to reduce the dimensionality and improve our assumptions on the given features, max pooling layers are used. Output layer is connected to the flattened output of last convolutional layer. The activations of last hidden layer gives us the intermediate audio representation.

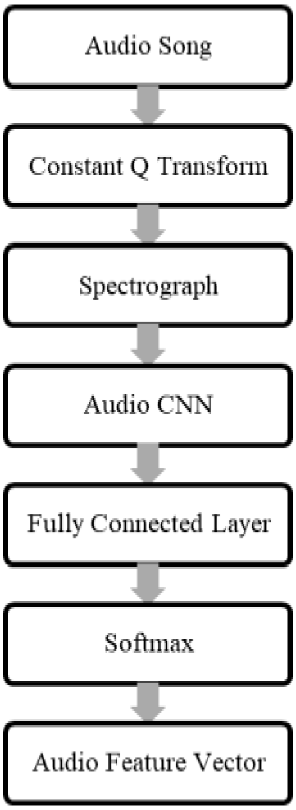

Fig. 2. Audio Feature Extraction

C. *Auralisation*

Spectrograms obtained can be deconvolved, but unlike other images these deconvolved spectrographs do not provide any intuitive explanation. Hence to solve this problem we use auralisation. Auralisation can be defined as the process of reconstructing audio signal from deconvolved images. To understand the functioning of CNNs that are trained with audio samples, we are using auralisation just as an extension to visualization of CNN's. We just need to perform an additional process of inverse transformation of Deconvolved spectrogram.

D. *Text Classification Methodology*

Given a text review of a musical item a process of semantic enrichment will be applied first. To semantically enrich texts, we adopt entity linking. Entity linking consists of assigning a text collection such as artist's name, place etc. with their corresponding entries in the reference Knowledge Base. The Wikipedia categories of entities will be identified in each document and then we will add them at the end of the text as new words. We apply then a VSM(Vector Space Model) with TF-IDF weighting over the enriched texts. From this representation, a feed forward network and a Rectified Linear Unit (ReLU) after each layer is trained to predict the genre labels.

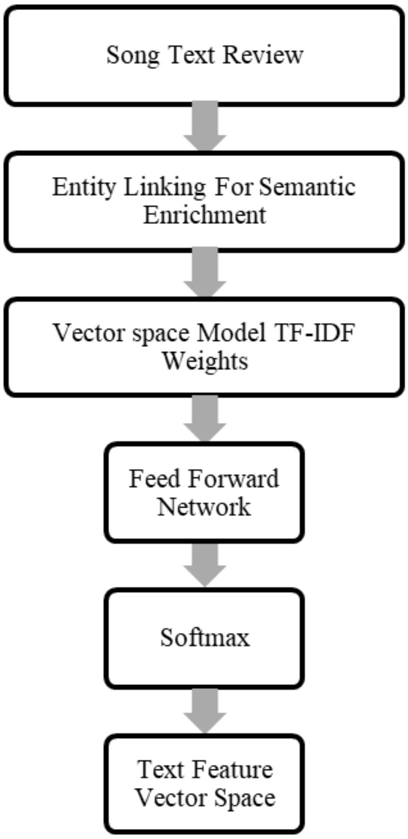
Song Text Review
Entity Linking For Semantic Enrichment
Vector space Model TF-IDF Weights
Feed Forward Network
Softmax
Text Feature Vector Space

Fig. 3. Text Feature Extraction

E. *Visual Classification Methodology*

We will use a Deep Residual Network (ResNet) [7] which is a specific type of CNN that has become one of the best architectures for several image classification tasks. A ResNet is a feedforward CNN with residual learning, which consists of bypassing two or more convolution layers thereby allowing us to add a greater number of layers. This addresses the underfitting problem originated when using a high number of layers, thus we can implement very deep architectures.

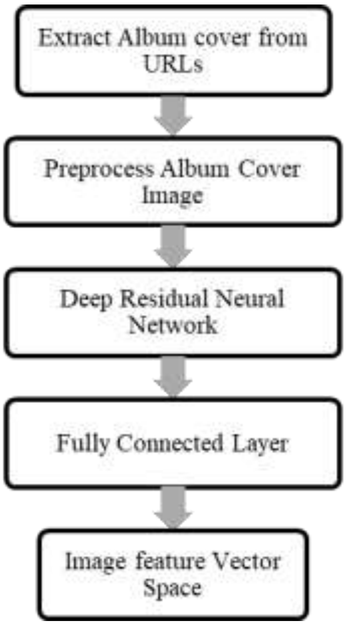
Extract Album cover from URLs
Preprocess Album Cover Image
Deep Residual Neural Network
Fully Connected Layer
Image feature Vector Space

Fig. 4. Visual Feature Extraction

F. *Multimodal Fusion*

Our aim is to combine all the different types of data into one single model which will be used to classify the input file. We will train each data representation on given music genre classification task separately and obtain the internal data representation. The activations of last hidden layer of each modality will be used as feature vector which will then be concatenated into a single feature vector. The resultant feature vector will be the input to a feedforward network. This feedforward network will have no hidden layers and input layer will be directly connected to the output layer. Softmax activation will be applied if we want single label classification else, sigmoid activation is used instead for multi- label classification.

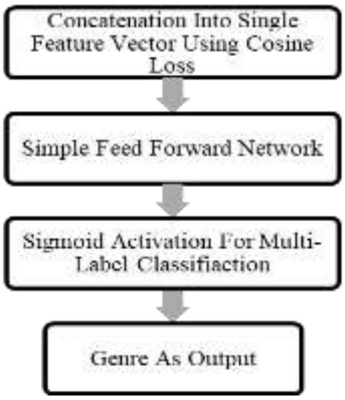
Concatenation Into Single Feature Vector Using Cosine Loss
Simple Feed Forward Network
Sigmoid Activation For Multi-Label Classifiaction
Genre As Output

Fig. 5. Multimodal Fusion and Classification

V. CONCLUSION

System will consist of multi-modal dataset of audio features, text reviews, and album cover images for feature extraction and classification using deep learning algorithms along with auralisation for better understanding of extracted features and data augmentation in order to improve training of audio feature extractor model and generate more accurate result. We are expecting the overall accuracy of our system to be around 75%.

ACKNOWLEGDEMENT

We feel great pleasure in expressing our deepest sense of gratitude and sincere thanks to our guide Prof. N. G. Bhojane for his valuable guidance during the Project work, without which it would have been very difficult task. We felt highly motivated by his continuous encouragement and enthusiasm.

We have no words to express our sincere thanks for valuable guidance, extreme assistance and cooperation extended to all the Staff Members of our Department too. This acknowledgement would be incomplete without expressing our special thanks to Prof. M. P. Wankhade (Head of the Department (Computer Engineering)) for his support during the work.

Last but not least we would like to thank all the Teaching and Non-Teaching staff members of our Computer Engineer Department, our parents and those who helped us directly or indirectly for completing the preliminary work for our project.

REFERENCES

[1] Oramas, Sergio Nieto, Oriol Barbieri, Francesco Serra, Xavier. (2017). Multi-label Music Genre Classification from Audio, Text, and Images Using Deep Features. Music Technology Group, University at Pompeu Fabra – 2017.
[2] Oramas, Sergio Barbieri, Francesco Nieto, Oriol Serra, Xavier. (2018). Multimodal Deep Learning for Music Genre Classification. Transactions of the International Society for Music Information Retrieval. 1. 4-21. 10.5334/tismir.10.
[3] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million-song dataset. In ISMIR, 2011.
[4] Choi, Keunwoo Fazekas, George Sandler, Mark. (2016). Explaining Deep Convolutional Neural Networks on Music Classification.
[5] Salamon, Justin Bello, Juan. (2017). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. IEEE Signal Processing Letters. PP. 10.1109/LSP.2017.2657381.
[6] Randomly Weighted CNNs For (Music) Audio Classification. IEEE - 2019.
[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.