Music Genre Classification/Recognition

Manjula M[1], Mallikarjun Pujeri[2], Naresh Babu R[3], Yashwanth A N[4]
[1, 2, 3, 4] Atria institute of technology

Abstract- *In this paper we talk about the architecture of Neural networks, mostly of which includes CNN (convolutional neural network) and CRNN (convolutional recurrent neural network) which we are planning to implement in our system for a custom music genre classification with our own data-set and GTZAN data-sets. Here we try to develop our system in the case of low computational level and data budget by which we will be able to train with a much larger data-sets. We use multiple strategies for tuning , initializing and optimizing. We also use a multiframe method by which we can almost analyse a n entire song or audio in detail. Basically which is used during training time for producing more samples it is also used during testing time to get an entire an overall summary of the entire song or audio. We finally at the end evaluate the results with both our handmade data-sets and GTZAN data-sets which are used in a lot of work.*

Keywords- Music genre classification, Neural Networks, Multi frames.

I. INTRODUCTION

Music genres are a fixed of descriptive key phrases that carry excessive-degree records approximately a song clip (Pop, Disco, Hip-hop…etc.).Music Genre classification is a project that targets to are looking ahead to track style using the audio signal. Being able to automatize the project of detecting musical tags permit to create exciting content fabric for the character like tune discovery and playlist creations.

Implementing this project calls for extracting acoustic competencies which can be accurate estimators of the type of genres we are interested in, observed via a single or multi-label type or in a few instances, regression degree. Conventionally, feature extraction is based on a signal processing the front-give up in order to compute applicable capabilities from time or frequency domain audio illustration. The capabilities are then used as enter to the ma-chine studying degree. However, it's miles difficult to realise which functions are be the most relevant to perform every undertaking. The recent techniques the use of Deep Neural Networks (DNNs), unify function extraction and selection taking. Thus permit gaining knowledge of the applicable capabilities for every task at the identical time that this project is learning to classify them.

II. CNN AND CRNN USED FOR CLASSIFICATION

In this system for the most of the work, we use melspectogram of the songs signals to provide as an input to our project.

CNN(convolutional neural network):

Convolutional neural network has been widely used in many music category task such as classifying genres [5] [6], latent feature prediction for recommendation [7], and also tagging music [3] [4]. By using Convolutional kernel of CNN we can extract features that resides in different levels of hierarchy. These hierarchical features are then learned to get the results of a given task during a maintained training.

CRNN(convolutional recurrent neural network):

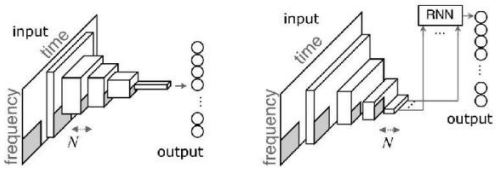
In the recent days CNNs have been incorporated with convolutional recurrent neural network(CRNNs) basically which are mostly used for sequential modelling of data most of which are sequences of words and music or audio signals.

CRNN can be said to be an updated version of the CNN as it replaces the end convolutional layer with RNN(recurrent neural network). Both the CNN and RNN techniques used in CRNNs are used for feature extracting and summarizing. This technique was first introduced in the document classification [9] and image classification [10] and for transcription of music [11].

III. NETWORK ARCHITECTURE

In this project we will first study the architecture of both the CNN and the CRNN in [4] [12] proposed by Choi- et-al.

Let us first start with CNN which has 5 convolutional Layers of 33 kernels max pooling layers As shown in figure 1a. By implementing this network it reduces the feature maps to 1×1 at the last layer. And each and every feature covers the entire input. CNN model permits time and frequency variations in various scales by 2D sub-sampling. It also reduces the no of parameters.



(a) CNN architecture (b) CRNN architecture

Fig.1: Network architecture from [4] and [12]

The CRNN architecture uses a two layer RNN along with a (GRU) gated recurrent units [13]. The 2-dimensional 4 layer CNN is as shown in figure 1b.it is assumed in this model that the temporal patterns can be applied better with RNNs rather than CNNs, But eventually they depend on CNN for input side from feature extraction. In this model RNNs are used to combine the temporal patterns rather than, sub-sampling used in other CNNs.

IV. MULTIFRAME

Multiframe is basically a strategy by using it we can make use of multiple frames, that is extraction of multiple frame.

In our project we propose to use multiframe (mel-spectogram) for each song. For every song first and the last N seconds which are no longer needed are discarded. Then we divide the remaining frames of same time length t. This method as two benefits:

At training time : it is possible to produce more data for training the network unlike in approach of [4], since they are always extracting the mid part part of a song and most likely to be limited by the number of songs, its is of huge help in data augmentation.

AT test time : we can find the mean or average with the result of all other frames to interpret the genre tag for an overall song.

V. GTZAN and HAND-MADE data-set

GTZAN data-set

It is a data-set created by Tzanetakis-et-al. It consists of 1000 music picked out of 30 seconds time duration with 100 examples in every 10 different music genres:

Hip-hop, Jazz, Blues, Country, Metal, Classic, Disco, Rock, Popular, Reggae. All the files for these genres were collected in 2000-2001 from different resources which includes radio stations, personal CDs, recordings from a microphone etc. All these are in 22050Hz Mono 16-bit files of audio in .wav format.

Hand-made data-set

For us to be able to test the performance of the multiframe method, we developed a data-set by using the genres of GTZAN data-set, but we use songs with a longer time duration. Exactly, our data-set consists of 300 music picked up with 30 examples for each and every 10 music genres of GTZAN.

We use whole songs, Hence number of frames per song can vary for every song, approximately leading to 4 frames per song for the shortest time duration songs. Which can vary for songs with longer time duration, may be more than 10.

VI. RESULT

The result will be a confusion matrices. In which each column will be a predicted label and each row will be a true label ranging from 0 to 100%, it is as shown in the figure.
(a) by using each and every frame (b)using mean

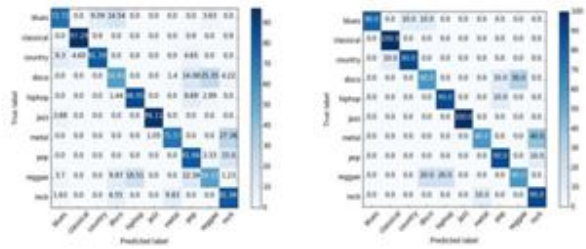


Fig.2: Confusion matrices

As in the figure you can see that both the confusion matrices are diagonal which makes it easier for all 10 genres to work well.

VII. CONCLUSION

We use the applications of neural networks for classifying genres of a song. But these kind of networks usually requires large amount of data to be trained from the beginning. By using multiframe model we can set an average frame for a single song. Our home made data-set is used as an experiment for songs longer than our frame duration to be compounded.

REFERENCES

[1] Sander Dieleman and Benjamin Schrauwen, "Multi-scale approaches to music audio featurelearning," in14th International Society for Music Information Re-trieval Conference (ISMIR-2013). Pontif ́ıcia Universi-dade Catolica´ do Parana,´ 2013, pp. 116–121.

[2] Aaron¨ Van Den Oord, Sander Dieleman, and Ben-jamin Schrauwen, "Transfer learning by supervised pre-training for audio-based music classification," in Con-ference of the International Society for Music Informa-tion Retrieval (ISMIR 2014), 2014.

[3] Sander Dieleman and Benjamin Schrauwen,"End-to-end learning for music audio," in 2014 IEEE Interna-tional Conference on Acoustics, Speech and Signal Pro-cessing (ICASSP). IEEE, 2014, pp. 6964–6968.

[4] Keunwoo Choi, George Fazekas, and Mark Sandler, "Automatic tagging using deep convolutional neural net-works," arXiv preprint arXiv:1606.00298, 2016.

[5] Keunwoo Choi, George Fazekas, Mark Sandler, and Jeonghee Kim, "Auralisation of deep convolutional neu-ral networks: Listening to learned features," in Proceed-ings of the 16th International Society for Music Informa-tion Retrieval Conference, ISMIR, 2015, pp. 26–30.

[6] Paulo Chiliguano and Gyorgy Fazekas, "Hybrid music recommender using content-based and social informa-tion," in 2016 IEEE International Conference on Acous-tics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 2618–2622.

[7] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen, "Deep content-based music recommenda-tion," in Advances in Neural Information Processing Systems, 2013, pp. 2643–2651.

[8] Keunwoo Choi, George Fazekas, and Mark Sandler, "Explaining deep convolutional neural networks on mu-sic classification," arXiv preprint arXiv:1607.02444, 2016.

[9] Duyu Tang, Bing Qin, and Ting Liu, "Document mod-eling with gated recurrent neural network for sentiment classification," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1422–1432.

[10] Zhen Zuo, Bing Shuai, Gang Wang, Xiao Liu, Xingxing Wang, Bing Wang, and Yushi Chen, "Convolutional re-current neural networks:

[11] Learning spatial dependencies for image representation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni-tion Workshops, 2015, pp. 18–26.

[12] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon, "An end-to-end neural network for polyphonic piano music transcription," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 5, pp. 927–939, 2016.

[13] Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho,"Convolutional recurrent neural networks for music classification,"arXiv preprint arXiv:1609.04243, 2016.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hin-ton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[15] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in European Con-ference on Computer Vision. Springer, 2014, pp. 818–833.