JUXTAPOSITION BETWEEN TWO CONVOLUTIONAL NEURAL NETWORK FOR MUSIC GENRE CLASSIFICATION

[1] Dr. J. Shana, [2] Ms. N. Priyadharshini Jayadurga, [3] Pratiba K R, [4] Prakash R
[1,2,3,4] Department of Computing – Artificial Intelligence and Machine Learning,
[1,2,3,4]Coimbatore Institute of Technology, Coimbatore
[1]shana@cit.edu.in, [2]dharshinidurga@cit.edu.in, [3]pratiba.karthi@gmail.com,
[4]prakashr7d@gmail.com

ABSTRACT

Music genre classification is the fundamental step involved in building a strong recommendation system. If music classification has to be carried out manually, then one has to listen to numerous songs and then select the genre. This process is not only time-consuming, it is quite a tedious task. The music industry has seen an excellent flow of latest channels to browse and distribute music. This doesn't return without drawbacks. With the increase in data, manual curation has become a difficult task. Audio files have a plethora of features that could be used to make parts of this process a lot easier. Advancements in technology have made it possible to extract the features of audio files. However, the most effective way to handle these for various tasks is unknown. This paper compared the two deep learning models of convolutional neural networks namely Alex-net and Res-net for the purpose of music genre classification using mel-spectrogram images for training. These aforementioned models were tested on GTZAN datasets. It was found that the results showed 56.0% accuracy for the res-net model which was outperformed by alex-net with an accuracy of 80.5%.

Keywords: *GTZAN Dataset, Alex-net, Res-net, Convolutional Neural Network*

INTRODUCTION

With the increase in technology, the amount of data being available to the public is increasing day by day. The increase of data is rapid to the point where manual curation is becoming infeasible and classification using automated systems are necessary. The music industry is no exception. Automating the process of music tagging would end in higher organization of the information and thereby creating any further development on the data is easier, like making themed playlists or recommending songs to users.

Machine learning is used to find the delicate patterns within the data, which might preferably be tough to explicitly code algorithms for. One such case is deciding what genre a song belongs to, that is the use case this report can cover. Companies (such as Soundcloud, Apple Music, Spotify, Wynk etc) use music classification, either to place recommendations to their customers, or simply as a product (like Shazam). Automatic musical genre classification can assist humans or perhaps replace them in this process and would be of a very valuable addition to music information

retrieval systems. Furthermore to this, automatic classification of music into genres can offer a framework for development and analysis of features for any form of content-based analysis of musical signals The concept of automatic music genre classification has become very highly regarded in recent years as a result of the rapid climb of the digital entertainment industry.

Categorizing music files in line with their genre could be a challenging task within the space of music information retrieval (MIR). As the availability of data increases, the need for categorization of said data becomes necessary in order to make good use of it. There are multiple methods of classifying data. In this study, Convolutional Neural Network (CNN) is used for classifying the data. It is a deep learning approach wherein two CNN models are trained end-to-end, to predict the genre label of an audio signal, solely using its spectrogram. The goals of this project are:

- Developing a machine learning model that classifies music into genres shows that there exists a solution which automatically classifies music into its genres based on various different features, instead of manually entering the genre.
- Achieve a good accuracy so that the model classifies new music into its genre correctly. This model ought to be higher than a minimum of a couple of antecedent models.

RELATED WORKS

Machine learning techniques have been used for music genre classification for decades now. In 2002, G. Tzanetakis and P. Cook [2] used both the mixture of Gaussians model and k-nearest neighbors. They achieved 61% accuracy. As a benchmark, human accuracy averages around 70% for this kind of genre classification work. Tzanetakis and Cook used MFCCs, a close cousin of mel-spectrograms, and primarily all work has followed in their footsteps in reworking their data in this manner.

In the following years, methods such as support vector machines were also applied to the present task, such as in 2003 when C. Xu et al. [3] used multiple layers of SVMs to attain over 90% accuracy on a dataset containing only four genres. Within the past 5-10 years, however, convolutional neural networks have shown to be incredibly accurate music genre classifiers, with excellent results reflecting both the complexity provided by having multiple layers and therefore the ability of convolutional layers to effectively determine patterns within images (which is essentially what mel-spectrograms and MFCCs are). These results have far exceeded human capability for genre classification, with our research finding that current state-of-the-art models perform with an accuracy of around 91% when using the full 30s track length. Several papers applied CNNs compared to their models to alternate ML techniques, including k NN, mixture of Gaussians, and SVMs, and CNNs performed favorably in all cases. Thus we tend to focus our efforts on implementing a high-accuracy CNN, used as a baseline.

METHODOLOGY

The musical data was obtained from the publicly available GTZAN dataset. The GTZAN dataset

is a relatively well-known resource in the field of MIR. It consists of 10 groups of 100 30-second song excerpts, with each group representing one genre, totaling 1000 audio files overall. The dataset comes pre-organized into folders - one per genre - making it fairly straight-forward to navigate, and balanced. The genres are - blues, classical, country, disco, pop, jazz, reggae, rock, metal. Each audio clips are 22050Hz Mono 16-bit files. The dataset incorporates samples from variety of sources like CDs, radios, microphone recordings etc.

To analyze the audio files from the two datasets, a python library called Librosa was used.The preprocessing part involved converting the audio from .au format to .wav format.To make it compatible to python's wave module for reading audio files. Next, we split a single audio file into 10 audio files each of 3 seconds.Now our training examples have become tenfold i.e. each genre has 1000 training examples and total training examples are 10,000. So we increased our dataset and this will be helpful for a deep learning model because it always requires more data. The last preprocessing step is generation of Mel-Spectogram. As we are going to use a Convolutional Neural Network, we need an image as an input, for this we will use the mel spectrograms of audio files and save the spectrograms as an image file (.jpg or .png).

A spectrogram is a visual way to represent signal strength of various frequencies. (time, frequency, and magnitude). The frequencies of spectrogram are mapped to the mel scale. Rationale behind this is because human auditory system does not operate on a linear scale, rather it is more logarithmic. Mel Scale is a Spectrogram with frequencies in Mel-Scale to mimic human auditory sensory. This visual illustration as depicted in Fig:1, will then be viewed and analyzed by pattern recognition tool.

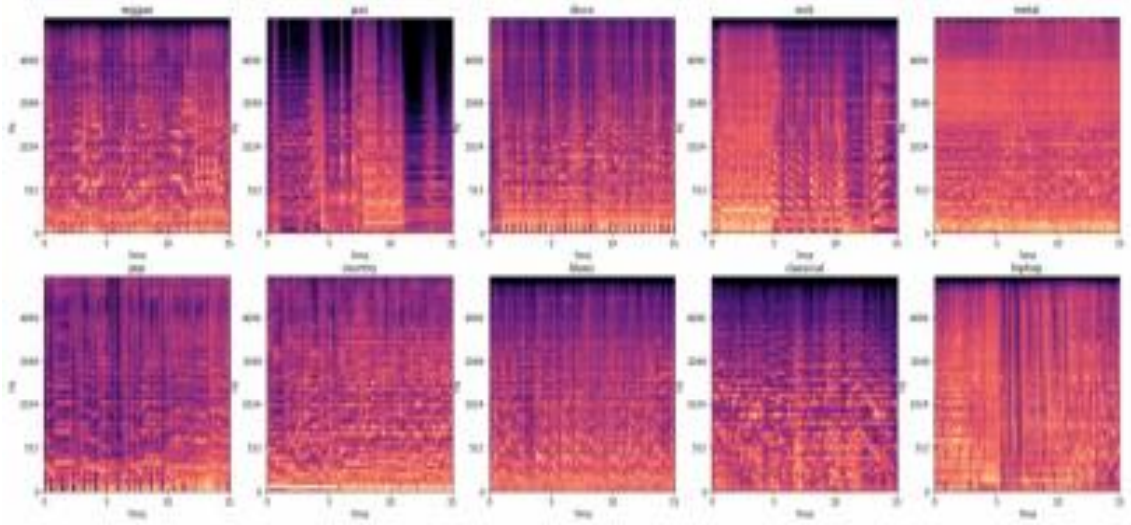

Fig 1: Visual Illustration of the Spectrogram with Frequencies in Mel-Scale

Structure of the Res-Net model

The input layer to the network was given the same number of nodes as features of the data. The output layer, which was activated with Softmax, was given ten units, one for each genre. However,

care must be taken not to overfit the training data; hence, regularisation was used to prevent this from occurring. The best result was achieved with a thirty two network in which the number of nodes decreased as the layers went on. The three hidden Dense layers were, like the input layer, activated with Relu.

Structure of Alex-net model

AlexNet architecture consists of 5 convolutional layers, 3 max-pooling layers, 2 normalization layers, 2 fully connected layers, and 1 softmax layer. The softmax layer is the output layer which has 10 units for each genre. Each convolutional layer consists of convolutional filters and a nonlinear activation function ReLU. The pooling layers are used to perform max pooling. Input size is fixed due to the presence of fully connected layers.

RESULTS AND DISCUSSION

This displays the results in the form of highest accuracies achieved for each experiment, as well as confusion matrices to help visualize classification accuracies for individual genres. All confusion matrix values have been normalized and rounded.

The best performance in terms of accuracy is observed for the Alex-net model that uses only the spectrogram as an input to predict the music genre with a test accuracy of 80.54%. The Res-net model , however robust and complex their design is, does not give a good accuracy even though the regularization metrics are modified or the epochs are increased.

1.Res-Net

The following displays the model accuracy when trained on the GTZAN dataset resulted in a 56% accuracy.

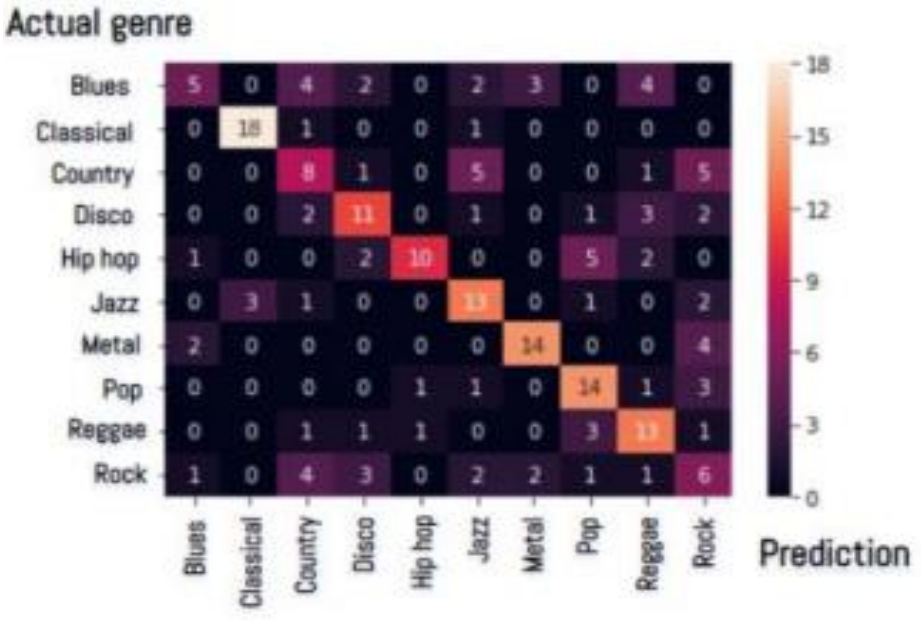

Fig: 2.1 Confusion matrix of the Res-Net model trained on the GTZAN dataset using MelSpectrogram.

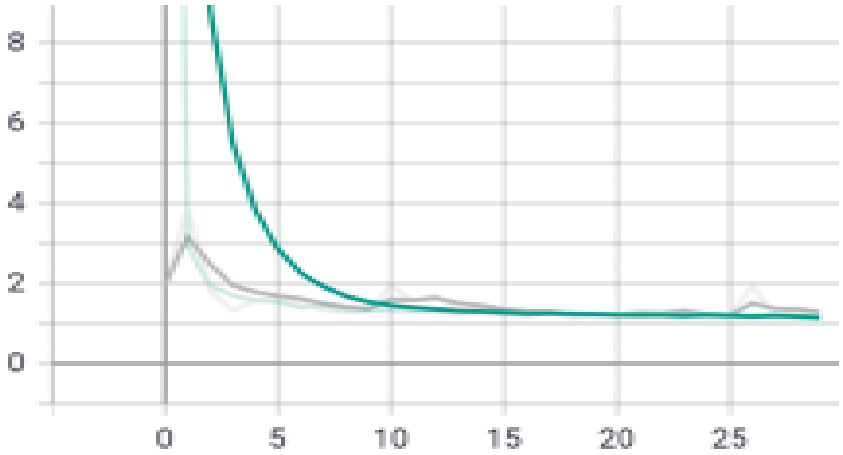

Fig: 2.2 Model accuracy of the Res-net model

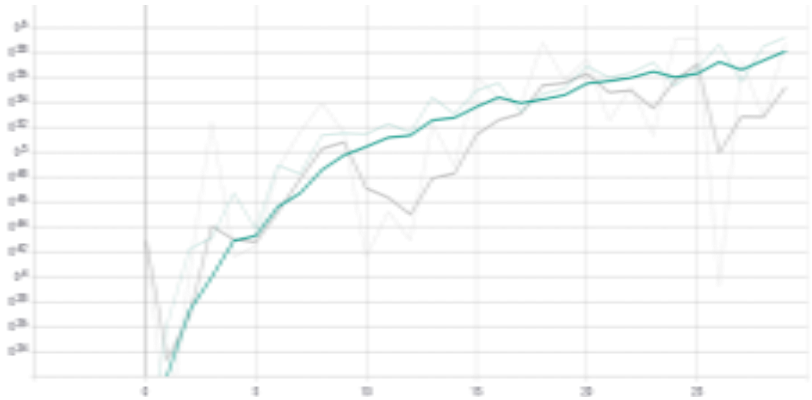

Fig: 2.3 Model loss of the Res-Net model

2.Alex-net

The following displays the model accuracy when trained on the GTZAN dataset resulted in a 80% accuracy.

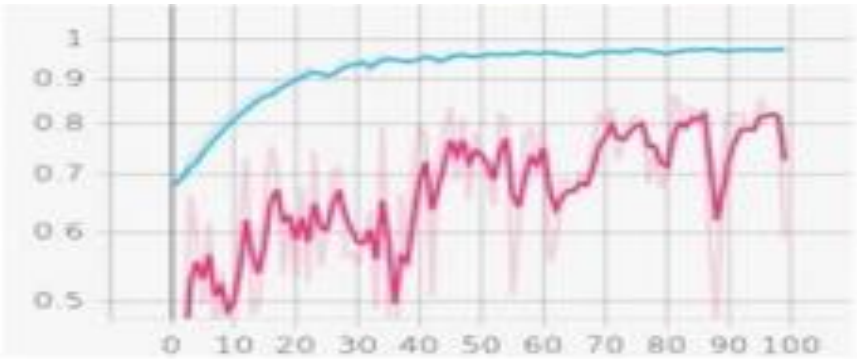

Fig: 3.1 Model Accuracy of the Alex-net Model

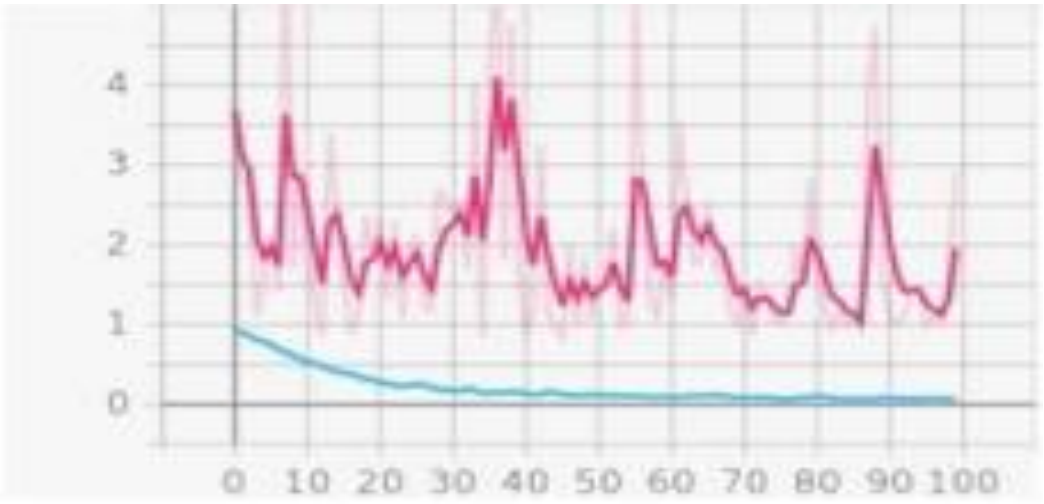

Fig: 3.2 Model Loss of the Alex-Net Model

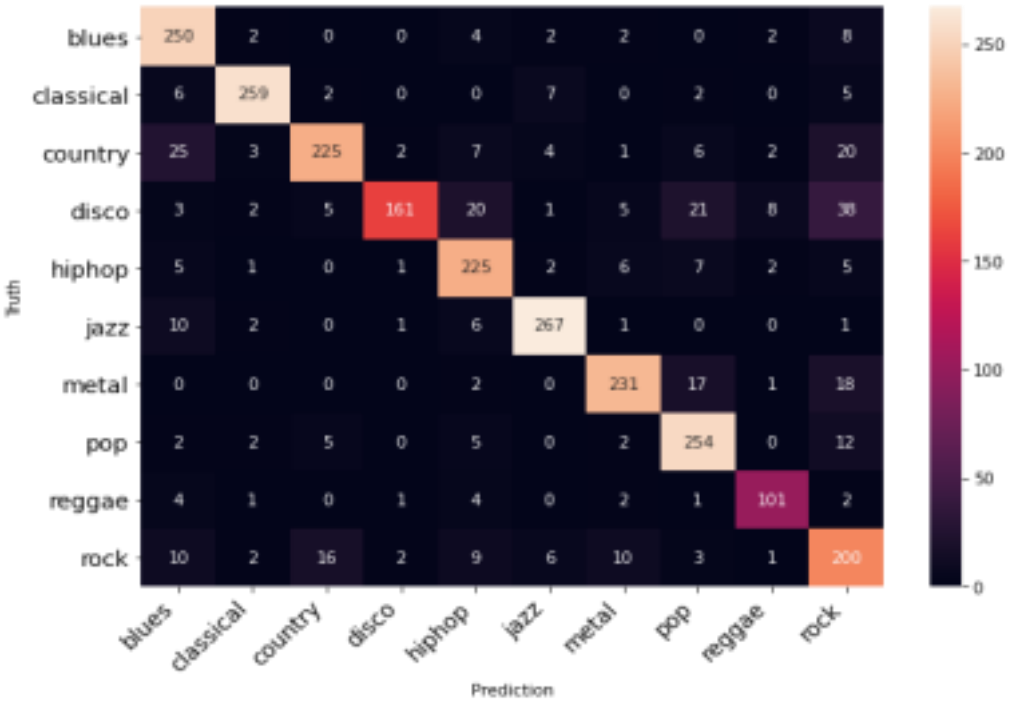

Fig: 3.3 Confusion matrix of the Alex-Net model trained on the GTZAN dataset using MelSpectrogram

CONCLUSION

According to this study, the Alex-Net model outperformed the Res-net and that leads to the conclusion that Alex-Net is better suited for music genre classification. As mentioned in the background, there has been studies done on combining models with each other which might be the better thing to actually analyze. Considering that it also was a lot more consistent in its

classifications shows that it found patterns in the songs easier. In general, it is very difficult to say what a good enough prediction accuracy for an automatic system would be since even humans are not perfect at classifying songs. However, it could be useful for a system where human confirmation is used, as it would decrease the workload.

Both the model fulfils its primary objective of classification of an audio file hello into its genre. The system proves out to be exemplary for performing the classification without human intervention and thereby making it faster and easier process. Of the approaches surveyed, the Alex-net model was shown to be the most promising.

REFERENCES

[1] Hareesh Bahuleyan, Music Genre Classification using Machine Learning Techniques, University of Waterloo, 2018

[2] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10(5):293–302, July 2002.

[3] Changsheng Xu, N. C. Maddage, Xi Shao, Fang Cao, and Qi Tian. Musical genre classification using support vector machines. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03)., Volume 5, April 2003.

[4] François Chollet et al. Keras. https://keras.io, 2015

[5] Y. Panagakis, C. Kotropoulos, and G. R. Arce. Music Genre Classification Via Sparse Representations of auditory temporal modulations. In 2009 17th European Signal Processing Conference, pages 1–5, Aug 2009.

[6] Defferrard, Michaël et al. “FMA: A Dataset for Music Analysis”. In: 18th International Society for Music Information Retrieval Conference. 2017. URL: https://arxiv.org/abs/1612.01840.

[7] Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. Deep Learning. http://www.deeplearningbook.org. MIT Press, 2016.

[8] Thomas Lidy and Alexander Schindler. Parallel Convolutional Neural Networks for Music Genre and Mood Classification. MIREX2016, 2016.

[9] Chathuranga, Y. M., & Jayaratne, K. L. Automatic Music Genre Classification of Audio Signals with Machine Learning Approaches. GSTF International Journal of Computing, 3(2), 2013.