DEEPMUSIC: A DEEP-LEARNING APPROACH TO MUSIC GENRE CLASSIFICATION

Anirban Lahiri — Shivam Raj — Subhrasankar Chatterjee

IIT Kharagpur

1. BACKGROUND

A lot of research has been put into the field of music classification using machine learning as well as deep-learning, for example [1], [2], [3] and [4]. Some of these classification mechanisms have been further exploited for music recommendation systems like [5]. Three major classes of music recommendation systems exist currently. These are content-based, collaborative filtering and hybrid systems. In this work we primarily investigate the use of content-based deep learning approaches for music classification.

The current work attempts to explore a number of deep-learning models for music classification with the aim for providing music recommendation. It is assumed that if a person likes to listen to music of any particular genre then they would like to listen to more songs from the same genre. The GTZan Dataset has been used for this exploration [6]. It consists of 10 music genres namely blues, classical, country, hiphop, disco, jazz, metal, pop, reggae and rock, with 100 songs in each Genre. Although the GTZan Dataset has been criticised by some literature [7], it is a freely available resource and a reasonable starting point for music classification experiments.

2. INTRODUCTION

In order to distinguish between the classes or genres the model needs to learn the characteristics of each genre. A widely known representation for the capturing music characteristics over time are mel-spectrograms, some examples are shown in figure 1. Spectrograms are a representation of the frequency composition of a signal over time. Mel-Spectograms try to capture the nuances of the human hearing by factoring in a logarithmic db(Decibel) scale.

In this work we have explored a number of deep-learning models utilizing RNNs(Recurrent Neural Networks) and CNNs(Convolutional Neural Networks) and have presented the most promising ones from each class. The remaining sections outline the methodology used for building the model, the details of the model, the experiments conducted and the results obtained.

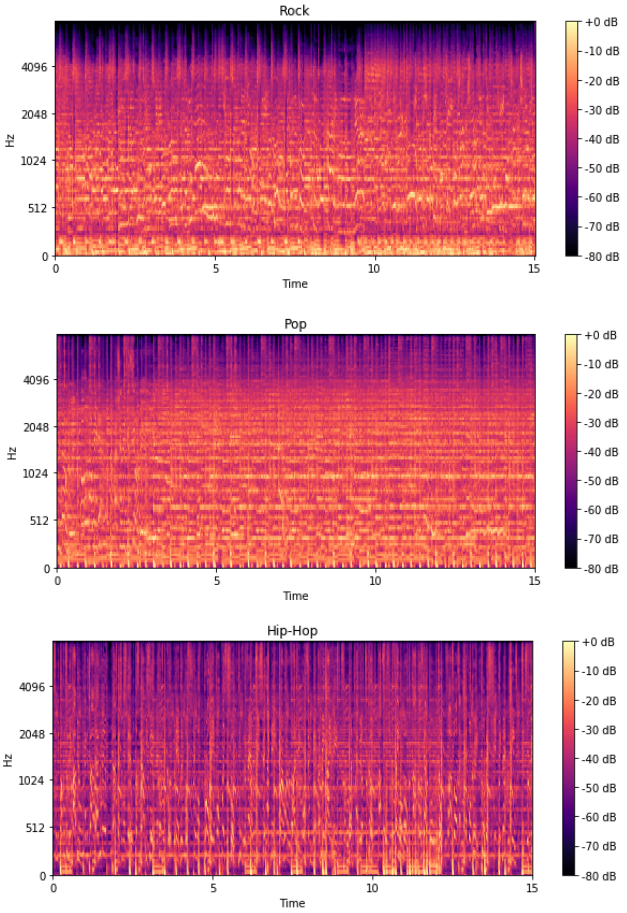


Fig. 1. Example Spectrograms from the Music Dataset

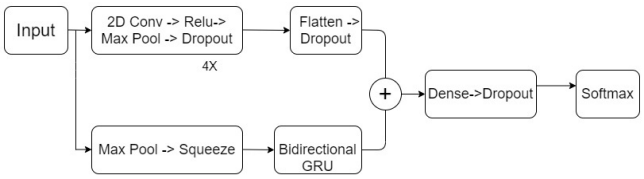
3. METHODOLOGY

CNNs [1], RNNs [8] and CRNNs(Convolutional Recurrent Neural Networks) [3] have been used in the past for music classification and recommendation. The input to these Neural Networks are mel-spectograms. These were computed using the Python librosa library. For this computation the window length over which the spectrum is computed needs to be specified. In this case the window length is kept at 2048 samples since it translates to about 10ms which is the at the lower end of human perception of hearing. The hop length which specifies the distance between two consecutive windows is set at

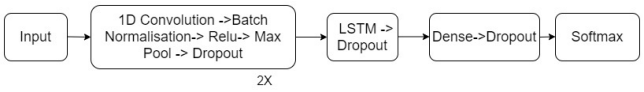
512 samples.

The mel-spectograms exhibit the features characteristic to each genre of music as shown by the examples in figure 1. CNNs have been widely used in the past for extracting characteristic features from images and the feature extraction from mel-spectograms is similar to that from other images. Hence, CNNs are a viable candidate for music classification. CRNNs have been recently introduced as a potential the state of the art architecture [3] in this field. In this exploration, we have run these models and compared their results.

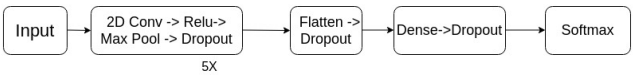
4. MODEL ARCHITECTURES



(a) Parallel CNN GRU



(b) CRNN



(c) CNN

Fig. 2. Model Architecture Block Diagrams

4.1. Parallel CNN GRU :

Inspired by the results in [9] , we explored the Parallel CNN GRU Model. The intution behind passing the input spectrograms through CNN and GRU layers in parallel is that we can understand the layout of the data (through CNN) and obtain the temporal summarization of the data at the same stage. We avoided the use of LSTMs, the popular choice for temporal summarization, as our dataset available was small as per the standards of deep learning.

Each CNN block comprises of a 2D convolution layer, relu activation layer, 2D maxpool layer and a dropout layer. 4 such CNN blocks, with successively increasing number of channels are connected sequentially. We also pass the input image through a bidirectional GRU in parallel. We then pass their concatenation through a fully connected layer and then a dense layer with softmax activation for classification.

4.2. CRNN :

Inspired by the results in [3], we also explored the CRNN model. CNNs are used for understanding the 2D layout of spectogram data and LSTMs are used as RNN layers for capturing temporal information.

We use 1D convolution layers that perform convolutions on spectograms across time, followed by batch normalisation, relu activation, 1D maxpooling layer and dropout layers . We use 2 such blocks and sequentially connect it with the LSTM layer, which aims at capturing short term and long term structure of the song. We pass its output through a dense layer and a softmax layer for classification.

4.3. CNN :

The temporal summarizers LSTMs and GRUs, as discussed above, did not perform reasonably well as each genre in the GTGAN dataset contained only 100 songs(30 seconds each). So, we moved ahead without the RNN models and used blocks of CNNs as feature extraction pipelines.

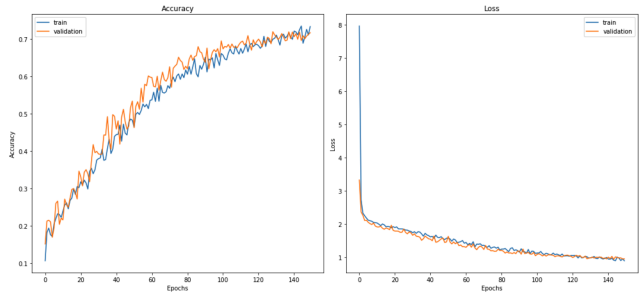
Each block consisted of a 2D convolution layer, followed by relu activation, 2D max pooling, and a dropout layer. We used 5 such blocks serially, each with increasing number of channels. We flattened the final layer and fed it into a dense layer, which was followed by a softmax layer,

5. EXPERIMENTATION DETAILS

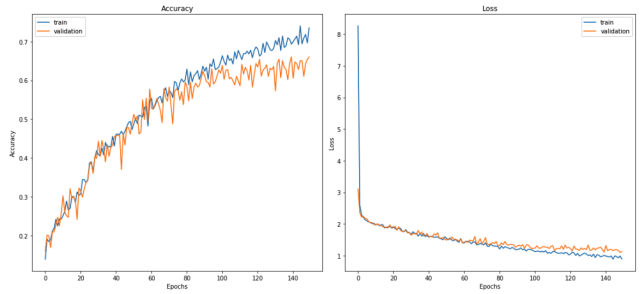
The training dynamics of the used models are shown in figure 3. For each of the models, the training was run for sufficient number of epochs until the model approximates its final steady state behaviour and there was no room for its improvement. For avoiding overfitting, we used standard techinques like dropout, L2 regularisation, batch normalisation, decrease in model size, etc over a range of parameters. Even then, the CNN GRU model overfits the training data, clearly indicating a small dataset. The code for this work can be found at [10] and has been developed on top of previous code repository [11].

6. RESULTS

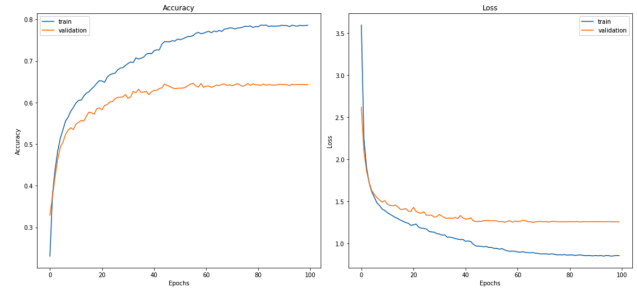
The results are summarised in Table 1. The confusion matrices for the various models are shown in figures 4, 5 and 6. In general the 2D-CNN model outperforms the other two models. The 2D-CNN model performs well for the genres metal, disco, classical, hiphop, pop and reggae.Conversely the 2D



(a)



(b)



(c)

Fig. 3. Plots for training accuracy over time for compared methods (a)2D CNN (b)CRNN (c)CNN-GRU

CRNN Model takes the lead in blues, classical and disco. Finally the CNN-GRU model performs well only for classical, metal and hiphop genres. In summary the 2D-CNN model has a superior performance compared to the other two models for most genres.

7. CONCLUSIONS AND FUTURE WORK

The features generated, like the mel-spectograms try to capture and represent maximum available information from the songs and using feature distillation pipelines like CNN , it is possible to perform well on recognising the genre of music. We used the sequential models like GRUs, LSTMs to capture the temporal dependence of music features . However , as the dataset contained only 100 songs (of 30 seconds each)

Table 1. Summarized Results

Model	Training Acc.	Test Acc
Parallel CNN GRU	0.78	0.61
CRNN	0.75	0.64
CNN	0.78	0.72

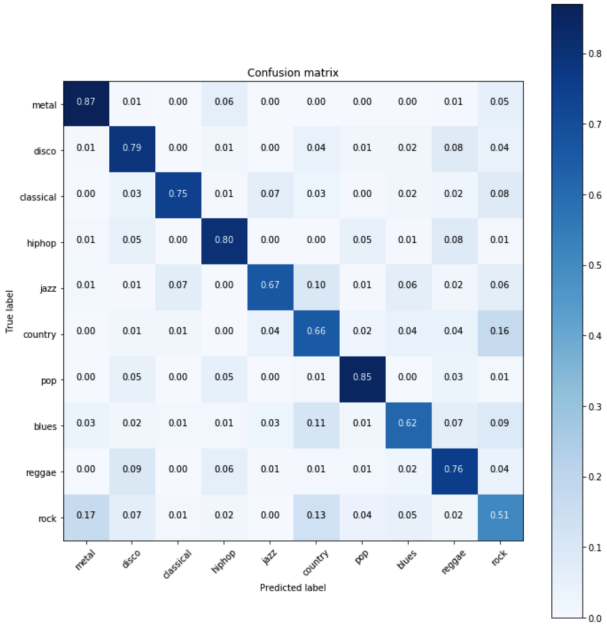


Fig. 4. Confusion Matrix for 2D CNN Model

per genre, training deep networks with very good accuracy was challenging. So, even though deep learning models like CNN, GRU, LSTM, etc have a good potential to perform well on task of music genre classification, they did not do remarkably well on the small GTZAN dataset. Although we looked into applying our model on large music datasets like FMA (Free Music Archive), MSD (Million Song Dataset),we could not do so because of various resource and time constraints. We would like to evaluate the performance of our models on these datasets as a continuation of this project. We would also like to explore the use of transfer learning by initialising the weights from different pretrained models like the resnet or the inception model as explored in [12]. In addition since some of the models perform well only for certain genres. Hence, ensemble of multiple models could be used to obtain a better overall classification.

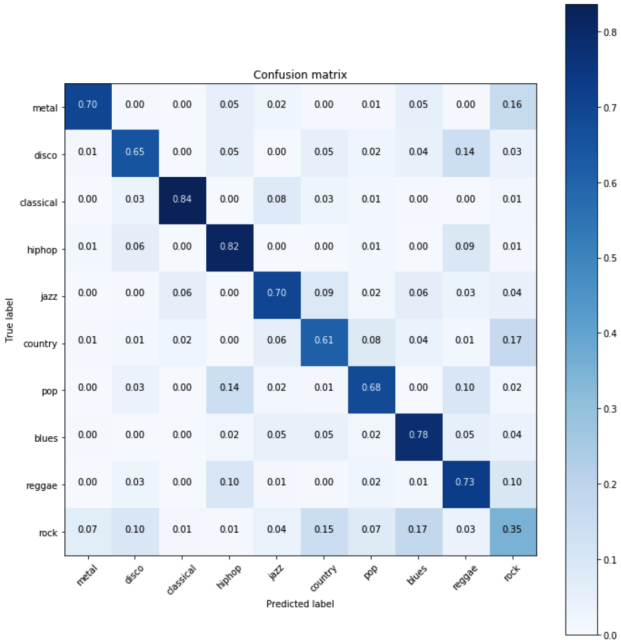


Fig. 5. Confusion Matrix for 2D CRNN Model CRNN

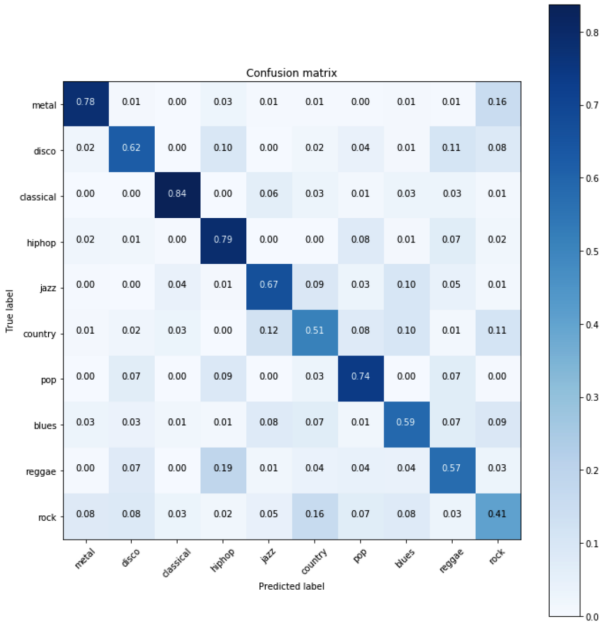


Fig. 6. Confusion Matrix for 2D CRNN Model CRNN

8. REFERENCES

[1] Sander Dieleman, Philémon Brakel, and Benjamin Schrauwen, "Audio-based music classification with a pretrained convolutional network," in *Proceedings of the 12th international society for music information retrieval conference : Proc. ISMIR 2011*, Anssi Klapuri and Colby Leider, Eds. 2011, pp. 669–674, University of Miami.

[2] Aäron van den Oord, Sander Dieleman, and Benjamin Schrauwen, "Transfer learning by supervised pretraining for audio-based music classification," in *Conference of the International Society for Music Information Retrieval, Proceedings*, 2014, p. 6.

[3] Keunwoo Choi, György Fazekas, Mark B. Sandler, and Kyunghyun Cho, "Convolutional recurrent neural networks for music classification," *CoRR*, vol. abs/1609.04243, 2016.

[4] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 4–21, DOI:.

[5] Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra, "A deep multimodal approach for cold-start music recommendation," *CoRR*, vol. abs/1706.09739, 2017.

[6] "Musical genre classification of audio signals " by g. tzanetakis and p," *Cook in IEEE Transactions on Audio and Speech Processing*.

[7] Bob L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," *CoRR*, vol. abs/1306.1461, 2013.

[8] Jeremy Irvin, Elliott Chartock, and Nadav Hollander, "Recurrent neural networks with attention for genre classification," .

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

[10] Anirban Lahiri, Shivam Raj, and Shubhrasankar Chatterjee, ," `https://github.com/anirbanlahiri2017/DL_Music`.

[11] H. Guimares, ," `https://github.com/Hguimaraes/gtzan.keras`.

[12] Grzegorz Gwardys and Daniel Michał Grzywczak, "Deep image features in music information retrieval," *International Journal of Electronics and Telecommunications*, vol. 60, no. 4, pp. 321–326, 2014.