AI based Music Recommendation system using Deep Learning Algorithms

Anand R¹, Sabeenian R.S¹, Deepika Gurang¹, Kirthika R¹, Shaik Rubeena¹

¹Department of ECE, Sona College of Technology, Salem.

E-mail: anand.r@sonatech.ac.in

Abstract. The customized recommendation framework for music should accurately represent private tastes. To obtain tailored feedback for the needs of various viewers, it takes adjustments. To find the better deep learning model for the recommendation may pave a way for a better recommender. Compared to the previous era, with commercial music streaming sites that can be downloaded from mobile devices, digital music availability is currently plentiful. It takes a very long time to figure out all this digital music and induces data exhaustion. It may be helpful to create a music recommendation system that can automatically scan the music libraries and suggest appropriate songs to users. The music provider will anticipate and then give their customers the appropriate songs based on the characteristics of the music previously heard by using the music recommendation system. Our study would like to build a framework for music recommendations that can provide recommendations based on the similarity of audio signal features. This research uses the Convolutional Neural Network (CNN) and the Recurrent Neural Network (RNN). Customized recommendation system for music should effectively represent private preferences. To attain tailored recommendations for the demands of different listeners, it needs changes and therefore, attempting to find a better deep learning model for the recommendation will pave a way for a better recommender.

1. Introduction

CNN was inspired by the visual cortex of the brain and prominently used for object or face detection and classification. CNN have three-dimensional layers height, width and depth and all neurons are not connected to the previous layer rather a layer is only connected to a small portion of neurons in the previous layer. LSTM is also a deep learning algorithm to selectively remember patterns for long duration of time and to overcome long term dependencies problem of RNN. LSTM just does some changes to information by implementing multiplication and addition. In this work, a deep learning prototype is recommended for the automatic classification of different of audio signals. The proposed CNN and LSTM model has an end-to-end architecture with MFCC feature extraction methods.

2. Literature Survey

A Genetic Algorithm (GA) for Music Data Recommendation System. This recommendation system designed by Hyun-Tae Kim, Eungyeong Kim, Jong-Hyun Lee and Chang Wook Ahn is a joint filtering (CF) and genetic algorithm hybrid solution [2]. Gender-based music recommendation framework using Adi-Convolutionaryansjah’s Recurrent Neural Networks, Alexander A S Gunawan and Derwin Suharto [3]. The proposed system uses Mel-spectrograms for audio interpretation and RNNs for highlight extraction. The precision of CNN model is 0.687, CRNN is 0.712 as well as Recall of CNN is 0.781 and CRNN is 0.804. C.Wang et al. expand probabilistic matrix factorization (PMF), prior to the latent factor vectors of the products with a subject model and extend this model to the recommendation of the science article [4]. They used CNN to identify the music and create a log file and used CF to provide suggestions for A D.Shun-
Hao Chang, Ashu Abdul, Jenhui Chen and Hua-Yuan Liao [5] on "A Personalized Music Recommendation System Using Convolutional Neural Networks Approach". There work suggested that use SVN or KNN classifiers reduces the efficiency during complex data [6].

3. Proposed Method
As we all know that CNN understands the data accurately using an image. So, at first, we need to convert our audio data into image. To visualize the data, plot the amplitude versus the time graph but it becomes harder to study the response so the frequency graph utilized for that purpose using fast Fourier transform [7,8] and we notice that data becomes more reluctant at low frequency so further we would down-sample the data (say to maybe 16Khz). Figure 1 & 2 represents the amplitude vs time graph and frequency, respectively.

4. Feature Extraction
MFCC tells us about the overall shape of a spectral envelope using a small set of features (about 10-20) or represents a set of short-term power spectrum characteristics of the sound which is shown in figure 1 [9]. The human perception of sound intensity is logarithmic in nature. So, Mel-scale is used to get a frequency that is close to human hearing capability which uses quasi-logarithmic frequency, and the frequency bands are equally distant. Figure 2 shows us MFCC features of the respiratory sounds. First and foremost, the spectrogram gives us the intensity of frequency over time. And using Mel-scale in place of frequency, Mel-spectrogram is obtained which is shown in figure 3 and is fed to the CNN and LSTM model.

\[
\text{Mel-scale formula: } M(f) = 1125 \ln \left( 1 + \left( \frac{f}{700} \right) \right)
\]

Figure 1: Time series for Music Recommendation system

Figure 2: Mel Frequency Cepstral Coefficients

4.1 Convolutional Neural Network
The recommended CNN model has 3 convolution layers followed by a pooling, each with activation or regularization parameter, which is fed into 3 fully connected layers [10] with activation of ReLU, soft-max output, and categorical loss of cross entropy. In figure 3, the architecture is visually presented in the block diagram.
The input for convNet is a MFCC image which passes through first Convolutional layer with filters [11] having size of 3x3x16 and a stride =1. The image dimensions change from 9x13x3 to 9x13x16. Notice that the 3 becomes 16 and 32 to 128 features which is shown in equation 2.

\[ \frac{n - f + 2p}{s} + 1 \]  

Where, \( n \) = input image size; \( f \) = filter size; \( p \) = padding; \( s \) = stride. Next, second Convolutional layer with 32 feature maps, size 3x3 and a stride = 1. The third and fourth Convolutional layer with 64 feature maps, size 3x3 and a stride = 1 and 128 feature maps, size 3x3 and a stride = 1 respectively. The fifth layer is a max pool layer with filter size 2x2 and a stride = 1. This layer is the same as the previous layer which will be reduced to 4x6x128 and the output of which goes into a dropout, flattening and a series of two fully connected layers. The second fully connected layer [12] is fed into a soft-max classifier with 8 class labels. Table 1 gives us the summary of CNN model.

### 4.2 Recurrent Neural Network

For consistent gradient , RNN classifies by taking various glimpse of the image and also uses back propagation through time . In case of vanishing and exploring gradient , the weight either becomes negligible or increases tremendously . So, to overcome this fluctuation, utilization of LSTM is done . LSTM uses cell state , sigmoid layer and tanh layer which helps LSTM in understanding long term dependencies . There are 3 many to many LSTM and have applied time distributed dense layer to wrap a fully connected Dense layer with 8 outputs, ReLU activation , objected to flattening and fed to dense layer with soft-max activation function to obtain the 8 class layers. In time distributed dense layer, each dense layer is applied on each timestamp i.e., apply a fully connected operation to each of its channel. Table 2 represents LSTM architecture [12].

### 5. Results and Discussion

We perform experiment to detect and Acoustic guitar, bass drum, cello, clarinet, double bass, flute, hi-hat, saxophone, snare drum, and violin or fiddle [13,14]. In fig 5 and 6, we observe that LSTM model is more stable than CNN model as it keeps fluctuating. In CNN [16], the minimum validation loss and highest validation accuracy is in between 10 to 15 epochs [16]. Figure 5 represents the loss vs number of training epochs and accuracy vs number of training epochs graph using CNN model and figure 6 presents the loss vs number of training epochs and accuracy vs number of training epochs graph using LSTM model. The proposed work classifies First and foremost, the proposed CNN model resulted in recall of 98.3, precision of 98.3, \( f \)-score of 99.3 and accuracy of 96.6 whereas in the RNN model - recall of 84.3, precision of 85, \( f \)-score of 84.5 and accuracy of 84.4 in 20 epochs. The recommender system obtains using both CNN and RNN overall has better performance. The system of measurement used to take out show testing of the system are Accuracy score, Precision (P), Recall (R), F-measure which is shown in equation 3,4,5 respectively and Confusion matrix. Precision metric provides the measure of positive analysis that is correct. Recall
defines the measure of actual positives that are correct. F-measure tests accuracy. Confusion matrix is used to measure the performance of the model.

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (3)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (4)
\]

\[
F\text{Measure} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (5)
\]

The performance metrics of CNN as well as RNN model is shown in Table 3.

Table 1: Summary for CNN model

| Layer | Feature maps | Size | Kernel size | Stride | Activation |
|-------|--------------|------|-------------|--------|------------|
| Input | Operation    | 3    | -           | -      | -          |
| 1     | Convolution  | 16   | 9, 13       | 3,3    | 1          | Relu       |
| 2     | Convolution  | 32   | 9, 13       | 3,3    | 1          | Relu       |
| 3     | Convolution  | 64   | 9, 13       | 3,3    | 1          | Relu       |
| 4     | Convolution  | 128  | 9, 13       | 3,3    | 1          | Relu       |
| 5     | Max Pool     | 128  | 4, 6        | 2,2    | 1          | Relu       |
| 6     | Flatten      | -    | 3072        | -      | -          | Relu       |
| 7     | Dense        | -    | 128         | -      | -          | Relu       |
| 8     | Dense        | -    | 64          | -      | -          | Relu       |
| Output| Dense        | -    | 8           | -      | -          | Softmax    |

Table 2: LSTM Architecture

| Layers | Size | Activation |
|--------|------|------------|
| Input  | Image | 9,13       |
| 1      | LSTM  | 9, 128     |
| 2      | LSTM  | 9, 128     |
| 3      | LSTM  | 9, 128     |
| 4      | Dense | 64         | Relu     |
| 5      | Dense | 32         | Relu     |
| 6      | Dense | 16         | Relu     |
| 8      | Dense | 8          | Relu     |
| Output | Dense | 8          | Softmax  |
Table 3: CNN Model and RNN Comparison.

| NAME      | CNN  | RNN  |
|-----------|------|------|
| Recall    | 98.3 | 84.3 |
| Precision | 98.3 | 85.0 |
| F_score   | 99.3 | 84.5 |
| Accuracy  | 96.6 | 84.4 |

6. Conclusion

The experience of this project provided great learning about MFCC, CNN, RNN(LSTM). After completing the model, we conclude that our CNN and RNN definition can operate till the prototype is trained carefully to categorize the audio signals. Although the precision is only sufficient for 10 epochs, it can be improved significantly by training the model. We will remain to focus on training the model to make forecasts more precise.

References

[1] Mayer R, Rauber A. Music Genre Classification by Ensembles Of Audio And Lyrics Features. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (Ismir 2011). Miami, Fl (2011).
[2] Sturm Bl. Classification Accuracy Is Not Enough: On the Evaluation of Music Genre Recognition Systems. J Intell Inform Syst. (2013)
[3] Yang Yh, Chen Hh. Machine Recognition of Music Emotion: A Review. Trans Intell Syst Techn. (2013)
[4] Huq A, Bello Jp, Rowe R. Automated Music Emotion Recognition: A Systematic Evaluation. J New Music Res. (2010–11) 39:227–44.
[5] Knees P, Schedl M. Music Similarity and Retrieval — An Introduction to Audio- And Web-Based Strategies. Berlin; Heidelberg: Springer (2016).
[6] Karydis I, LidaKermanidis K, Sioutas S, Iliadis L. Comparing Content and Context Based Similarity for Musical Data. Neurocomputing. (2013)
[7] Schedl M, Knees P, Mcfee B, Bogdanov D, Kaminskas M. Music Recommender Systems. In: Ricci F, Rokach L, Shapira B, Kantor Pb, Editors. Recommender Systems Handbook, 2nd Ed. Boston, Ma: Springer (2015).
[8] Ricci F, Rokach L, Shapira B, Kantor Pb Editors. Recommender Systems Handbook, 2nd Ed. Boston, Ma: Springer (2015).
[9] Bell Rm, Koren Y. Lessons from The Netflix Prize Challenge. AcmSigkdd Exp Newslett. (2007) 9:75–9. Doi: 10.1145/1345448.1345465
[10] De Gemmis M, Lops P, Musto C, Narducci F, Semeraro G. In: Ricci F, Rokach L, Shapira B, Kantor Pb, Editors. Semantics-Aware Content-Based Recommender Systems, 2nd Ed. Boston, Ma: Springer (2015).
[11] Chen Cw, Lamere P, Schedl M, Zamani H. Recsys Challenge 2018: Automatic Music Playlist Continuation. In: Proceedings of the 12th Acm Conference on Recommender Systems (Recsys 2018). Vancouver, Bc: Acm (2018). P. 527–8.

[12] Batmaz Z, Yurekli A, Bilge A, Kaleli C. A Review on Deep Learning for Recommender Systems: Challenges and Remedies. ArtifIntell Rev. (2018) 52:1–37. Doi: 10.1007/S10462-018-9654-Y

[13] Mcfee B, Lanckriet G. The Natural Language of Playlists. In: Proceedings of the 12th International Society for Music Information Retrieval Conference (Ismir). Miami, Fl (2011).

[14] Chen S, Moore JI, Turnbull D, Joachims T. Playlist Prediction Via Metric Embedding. In: Proceedings of The 18th AcmSigkdd.

[15] Shanthi, T., Sabeenian, R.S. And Anand, R., 2020. Automatic Diagnosis of Skin Diseases Using Convolution Neural Network. Microprocessors and Microsystems, 76, P.103074.

[16] Anand R, Shanthi T, NithishMs, Lakshman S. Face Recognition and Classification Using Googlenet Architecture. Insoft Computing for Problem Solving 2020 (Pp. 261-269). Springer, Singapore.

[17] R. Anand, S. Veni, P. Geetha, S. Rama Subramoniam, Extended morphological profiles analysis of airborne hyperspectral image classification using machine learning algorithms, International Journal of Intelligent Networks, Volume 2, 2021, Pages 1-6, ISSN 2666-6030 https://doi.org/10.1016/j.ijin.2020.12.006.