Music Genre Recognition using Deep Neural Networks and Transfer Learning

Deepanway Ghosal, Maheshkumar H. Kolekar

Indian Institute of Technology Patna, India
deepanwayedu@gmail.com, mkolekar@gmail.com

Abstract

Music genre recognition is a very interesting area of research in the broad scope of music information retrieval and audio signal processing. In this work we propose a novel approach for music genre recognition using an ensemble of convolutional long short term memory based neural networks (CNN LSTM) and a transfer learning model. The neural network models are trained on a diverse set of spectral and rhythmic features whereas the transfer learning model was originally trained on the task of music tagging. We compare our system with a number of recently published works and show that our model outperforms them and achieves new state of the art results.

Index Terms: music genre recognition, music information retrieval, deep learning, transfer learning

1. Introduction and Related Work

Music information retrieval (MIR) is an interdisciplinary field dealing with the analysis of musical content by combining aspects from signal processing, machine learning and music theory. MIR enables computer algorithms to understand and process musical data in an intelligent way. Music genre recognition (MGR) is one of the most important subfields of MIR. Music genre is defined as an expressive music style incorporating instrumental or vocal tones in a structured manner belonging to a set of conventions. Automatic music genre recognition is a very interesting problem in the context of MIR because it enables systems to perform content based music recommendation, organizing musical databases and discovering media collections.

The first significant work on musical genre recognition was performed in [1] by Trantakis and Cook. Timbral texture, rhythmic content & pitch content based features were proposed and classification was done using Gaussian mixture model (GMM) and K-nearest neighbor (K-NN) algorithms. Musical genre recognition using support vector machines were proposed in [2] by Xu et al. In [3] Costa et al. proposed the approach of musical genre recognition using spectrogram features. In [4, 5] specific musical features were used with feature selection techniques. Musical genre classification using deep learning models has been performed in [6, 7, 8]. Survey works performed in [9, 10] gives a comprehensive account of genre classification of musical content and evaluation techniques. Authors in [11] introduced the Million Song Dataset - a collection of audio features and metadata for a million contemporary popular music tracks. A wide range of musical information retrieval systems can be build using this dataset including genre recognition, automatic music tagging, music recommendation, etc [1].

2. Proposed Methodology

In this work we have focused on genre recognition of the songs in the GTZAN dataset [1], which has been widely studied in the area of MGR. The dataset contains songs of ten different genres - blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae & rock. To recognize the genre of a song, we first train our deep neural network models on a set of extracted spectral and rhythmic features. We also utilize a transfer learning system to extract meaningful features from the songs. A multilayer perceptron network is then trained on this transferred features to predict the genres. Finally the predictions of different models are combined using a majority voting ensemble.

2.1. Feature Extraction

2.1.1. Two Dimensional Spectral and Rhythmic Features:

A diverse set of spectral and rhythmic domain features are first extracted from the raw musical raw signals. In the features listed below, ‘Tonnetz’ and ‘Tempogram’ are rhythmic features, the rest are spectral features. The musical data in the GTZAN dataset is sampled at 22050 Hz and are around 30 seconds long resulting in a total of roughly 22020 × 30 = 661500 samples. We compute the features for each sliding window of 2048 samples with shift of 1024 samples. We pad appropriate number of zeros at the end such that there are a total of 661500/1024 = 464 windows and thus each song is represented as a (464, k) dimensional feature matrix. The exact choice of k depends on the feature being computed.

- **Mel Spectrogram**: Mel-frequency cepstrum (MFC) representations introduced in [12] are widely used in automatic speaker and speech recognition. The mel spectrogram produces a time-frequency representation of a sound imitating the biological auditory systems of human beings. We compute the magnitude spectrum from the time series musical data and then map it on to the mel scale. We used k = 128.
- **Mel Cepstral, Delta and Double Delta Coefficients**: Mel cepstral coefficients (MFCCs) are the coefficients that collectively make up a mel-frequency cepstrum. We used k = 20 mel cepstral coefficients.
- **Delta Coefficients**: We used k = 20 delta coefficients (derivative of the mel cepstral coefficients).
- **Double Delta Coefficients**: We used k = 20 double delta coefficients (double derivative of the mel cepstral coefficients).
- **Energy Normalized Chromagram**: Chroma audio features are extensively used in musical signal processing. Chroma features are effective in audio matching and retrieval applications [13, 14] as they capture melodic and harmonic characteristics of music and are robust to changes in instrumentation and timbre. In [15] authors introduced Chroma Energy Normalized Statistics (CENS) features by considering short time statistics over energy distributions within the chroma bands. We took k = 12 as it represents 12 distinct semitones of the musical octave.
- **Constant Q Chromagram**: Constant Q transform [16] constitutes of a bank of filters with logarithmically spaced center frequencies $f_n = f_0 \times 2^\frac{n}{b}$ where $n = 0, 1, \ldots$; central frequency of the lowest filter is denoted by $f_0$ and the number of filters in each octave is denoted by $b$. An appropriate choice of
\( f_0 \) and \( b \) directly corresponds to musical notes. This transform also has increasing time resolution towards higher frequencies resembling the human auditory system. \( k = 12 \) was taken.

- **Short Time Fourier Transform (STFT) Chromagram:** Chromagram of short-time chroma frames are used with \( k = 12 \).
- **Tonnetz:** Tonal centroid features (tonnetz) are computed following works in [17]. Authors show that these features are successful in detecting changes in the harmonic content of musical audio signals, such as chord boundaries in polyphonic audio recordings. We used \( k = 6 \) tonnetz features.
- **Tempogram:** The aspects of tempo and rhythm are very important dimensions of music. In [18], the authors introduced a robust mid-level representation that encodes local tempo information by computing local autocorrelation of the onset strength envelope in music signals. This tempogram feature can act as a very important source of information for MGR, specifically where music reveals significant tempo variations. We used \( k = 128 \) tempogram features.

### 2.1.2. One Dimensional Averaged and Transfer Learning Features

We also compute the following one dimensional vectors as a summary statistic of the whole song.

- **Averaged Signal Vector:** This vector is calculated simply by taking the average of all the extracted two dimensional features listed above. After extracting \((646, k_1)\) dim matrix from mel spectrogram, \((646, k_2)\) dim matrix from mel cepstral coefficients, ..., \((646, k_n)\) features from tempogram, the averaging was performed over these 646 windows. Finally vectors of \(k_1, k_2, ..., k_n\) dimensions were obtained which were then concatenated to obtain the averaged signal vector. Our particular choices of \(k_1, ..., k_n\) led to this vector having dimension of 342.

- **Music Transfer Learning Vector:** Transfer learning is frequently used in computer vision problems. In this kind of systems, generally a deep convolutional net trained on the large scale ImageNet data [19] is used. Although the original network is trained on ImageNet data, it is able to capture a wide variety of visual features which are then used for other recognition tasks. In [7] authors introduce a musical transfer learning system. A deep convolutional neural network is first trained on a large dataset [11] for music tagging. The tags include genre, era, instrumentation, and mood labels. This trained network is then used as a feature extractor for other related tasks. We use the model to extract a 160 dimensional vector for each song.

### 2.2. Models

Convolutional neural networks (CNN) are specially designed neural networks for processing data that has a grid-like topology [20]. Introduced in [21] convolutional neural networks have produced excellent results in a wide variety of problems including computer vision [22, 23, 24], speech recognition [25] and natural language processing [26, 27]. Long short term memory (LSTM) networks [28] are also widely used in sequential time series data to capture long term dependencies.

In this work, we apply variants of CNN and CNN-LSTM models for musical genre prediction. Following [26] we use 1D convolution in our models. Here, the extracted features have dimensions of \((646, k)\) (Section 2.1), and our convolutional filters have dimension of \((3, k)\). The 1D convolution operation is performed by sliding the filters over the 646 windowed time-steps. The operation is denoted as 1D convolution because the convolutional filters and the features have same length and hence the sliding of the filters are performed only over the width (time dimension) of the features.

In total, we apply four different CNN and CNN-LSTM models on all the extracted two dimensional features separately to predict the genre of the song. Structure of these models are outlined in Fig. 2. For the two different kinds of one dimensional vectors we use two separate multilayer perceptron (MLP) models for genre prediction. We briefly describe configurations of these models below.

- **CNN Max Pooling Model:** A two layer deep CNN model is used with 128 and 64 filters (width 3) respectively in the two layers. Between these two layers max pooling is performed with factor two. After the second convolution layer, global max pooling is used to create a representational vector. This vector is then used in a fully connected layer to create the output genre.

- **CNN Max Pooling LSTM Model:** This model is similar to
3. Experiments, Results and Discussion

The GTZAN dataset consists of 1000 audio tracks each being 30 seconds long. All the tracks are 22050Hz mono 16-bit audio files in .wav format. It contains 10 genres of songs - blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae & rock. Each genre is represented by 100 tracks. We evaluate our models in this ten class classification framework. We run our experiments in a 10 fold cross validation setup. We maintain the uniform distribution of musical genres in each fold i.e. there are 80 songs of each genre in the train split and 20 songs of each genre in the validation split for each fold.

The average 10 fold accuracy score of our models are reported in Table. 1. A number of interesting observations can be made from the results. First of all, we observe that the best result is obtained by the multilayer perceptron model when used with music transfer learning features. This result can be expected as the original system was trained on the very large Million Song Dataset[11] containing rich label sets for various aspects of music including mood, era, instrumentations and most importantly genre. Also further fine-tuning was performed on our experimental setup leading it to produce the best results. We also observe that the mel spectrogram features produces best results in CNN Max Pooling and CNN Average Pooling models, whereas mel coefficients produces best results for CNN Max Pooling LSTM and CNN Average Pooling LSTM models.

The introduction of LSTM resulted in improved perfor-
Table 1: Average 10 fold cross validation accuracy scores for different features and models.

<table>
<thead>
<tr>
<th>Features &amp; Models</th>
<th>CNN Max Pooling</th>
<th>CNN Max Pooling LSTM</th>
<th>CNN Average Pooling</th>
<th>CNN Average Pooling LSTM</th>
<th>Multilayer Perceptron</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mel Spectrogram</td>
<td>83.0</td>
<td>73.6</td>
<td>82.5</td>
<td>75.7</td>
<td>-</td>
</tr>
<tr>
<td>Mel Coefficients</td>
<td>80.2</td>
<td>79.0</td>
<td>81.6</td>
<td>80.5</td>
<td>-</td>
</tr>
<tr>
<td>Delta Mel Coefficients</td>
<td>70.4</td>
<td>77.2</td>
<td>74.5</td>
<td>77.0</td>
<td>-</td>
</tr>
<tr>
<td>Double Delta Mel Coefficients</td>
<td>72.1</td>
<td>72.9</td>
<td>72.1</td>
<td>76.5</td>
<td>-</td>
</tr>
<tr>
<td>Energy Normalized Chromagram</td>
<td>45.7</td>
<td>34.5</td>
<td>43.0</td>
<td>36.2</td>
<td>-</td>
</tr>
<tr>
<td>Constant Q Chromagram</td>
<td>60.0</td>
<td>49.4</td>
<td>57.5</td>
<td>45.6</td>
<td>-</td>
</tr>
<tr>
<td>STFT Chromagram</td>
<td>62.8</td>
<td>52.5</td>
<td>63.4</td>
<td>53.7</td>
<td>-</td>
</tr>
<tr>
<td>Tonnetz Features</td>
<td>50.2</td>
<td>53.5</td>
<td>51.0</td>
<td>55.8</td>
<td>-</td>
</tr>
<tr>
<td>Tempogram Features</td>
<td>41.5</td>
<td>42.0</td>
<td>41.6</td>
<td>43.3</td>
<td>-</td>
</tr>
<tr>
<td>Averaged Signal Features</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>77.1</td>
</tr>
<tr>
<td>Transfer Learning Features</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>85.5</td>
</tr>
</tbody>
</table>

Table 2: Ensemble models and comparative results with other state-of-the-art systems.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN Max Pooling &amp; MLP</td>
<td>93.6</td>
</tr>
<tr>
<td>CNN Max Pooling LSTM &amp; MLP</td>
<td>91.5</td>
</tr>
<tr>
<td>CNN Average Pooling &amp; MLP</td>
<td>94.2</td>
</tr>
<tr>
<td>CNN Average Pooling LSTM &amp; MLP</td>
<td>91.4</td>
</tr>
<tr>
<td>Grzegorz and Grzywczak [6]</td>
<td>78.0</td>
</tr>
<tr>
<td>Baniya et al. [32]</td>
<td>87.9</td>
</tr>
<tr>
<td>Choi et al. [7]</td>
<td>89.8</td>
</tr>
<tr>
<td>Arabi and Lu [33]</td>
<td>90.8</td>
</tr>
<tr>
<td>Panagakis et al. [4]</td>
<td>91.0</td>
</tr>
<tr>
<td>Panagakis et al. [5]</td>
<td>93.7</td>
</tr>
<tr>
<td>Proposed System</td>
<td><strong>94.2</strong></td>
</tr>
</tbody>
</table>

With rigorous examples and case studies, it is demonstrated that the perfect system in the GTZAN dataset would not be able to surpass the accuracy score of 94.5% due to the inherent noise in some of the repetitions, mis-labelings and distortions of the songs. Interestingly, our proposed system achieves accuracy of 94.2%, an almost perfect score.

5. Conclusion

In this work we proposed a novel approach for music genre recognition. Firstly variants of CNN and CNN-LSTM based models are trained on a variety of spectral and rhythmic features. Secondly, a MLP network is trained on extracted representational features from a transfer learning system trained for music tagging. Finally, these models are combined in a majority voting ensemble setup. With our experiments we showed that the ensemble model is effective in greatly improving the performance. Our proposed model outperforms the current state-of-the-art systems and achieves a near perfect score for musical genre recognition in the GTZAN dataset.
References


