Literature Review about Music Genre Classification

Lam Hoang
Earlham College
Richmond, Indiana, USA
ldhoang18@earlham.edu

ACM Reference Format:
Lam Hoang. 2018. Literature Review about Music Genre Classification. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Music Genre Classification has been one of the most prolific areas in machine learning, specifically, and in computer science, generally. One of the most popular classification methods for this is the use of deep learning techniques, most notably the Neural Networks (or NN) to process large music datasets to identify the corresponding genre. This literature review will be about how NN has been experimented within major researches. The contents of this literature review include 1) The datasets researchers used in their papers to apply deep learning techniques, and 2) The methods they select to classify music genres. For the "Conclusion" section, I will discuss possible future works in applying NN to differentiate music genres. Although other approaches are mentioned in papers I have found, this literature will concentrate only on Neural Network approaches.

2 DATASET

This section will identify the datasets that are used in research papers concentrating on this topic that I have found. It is important to have an understanding of their descriptions as there are components from those dataset that would be helpful to propose an approach for the research topic. The most used datasets that the majority of computer science papers mentioned are GTZAN and Extended Ballroom.

2.1 GTZAN

The majority of research papers on this topic use GTZAN as their proposed dataset. GTZAN, often found in Kaggle, consists of 1000 music excerpts with a time duration of 30 seconds [4]. This dataset has 10 different genres such as blues, classical, country, disco, hip hop, rock, metal, pop, jazz and disco, which means that there are 100 audio clips for each genre [7]. GTZAN dataset has been used in more than 100 published CS papers of the same topic and it is considered one of the most well-known public datasets available for music genre recognition [8]. Some researchers point out several integrity problems from this dataset, such as replications, mislabeling, and distortion [2]. Specifically, Lau et al determined that there were 50 of the entire dataset were replicas, 22 excerpts were from the similar audio file, and 13 audio pieces were of the similar song but from other recordings [8].

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

2.2 Extended Ballroom

Extended Ballroom is a genre classification dataset that extends from the original Ballroom dataset. It is 6 times more tracks, better audio quality, and more advantageous to implement deep learning techniques than the original version [10]. I have taken a look at both Ballroom and Extended Ballroom datasets, and I found out that the latter provides a md5 hash to test the appropriation of audio and several properties to indicate where there is a repetitive and duplicated versions of the audio track.The duration for each track in this dataset is 30 seconds, and it contains 13 different genres with a total of 4180 songs [10]. Thanks to being a large dataset, Extended Ballroom was an ideal selection for an improved version of CNN as the model eliminated the requirement for pre-training on this dataset [9].

3 METHODS

This section will mention some popular approaches researchers have done in their papers. One method is to use a Recurrent Neural Network and the other is Convolutional Neural Network.

3.1 Recurrent Neural Network (RNN)

A lot of research papers on this topic proposed to apply RNN to classify music genres. This term is defined as networks built for data in sequence [10]. Different from other Neural Network (NN) techniques, RNNs supply time-related context-based information to make decision relying on connections formed in cycle [10]. The connections transfer the activations from the previous temporal step to another [10]. The plain RNN structure could not handle long-term dependencies as the issues relating to vanishing gradient might arise. Therefore, Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) were suggested as they to make another connection state present from successfully updated current activations [10].

Yu et al proposed to build an alternate RNN architecture - a bidirectional RNN (BRNN) - to classify music genres using Gated Recurrent Unit (or GRU). The reason to apply GRU for BRNN architecture is due to its having better performance than Long-Short Term Memory (LSTM) after declining the number of gate categories [10]. This model contained two stacked bidirectional RNN (BRNN) layers, one of them moved forward from the beginning of the sequence to the end and the other went backward [10]. Shortcut connections overlooking a hidden BRNN layer were manifested to detect vanishing gradient without the need for any computational complexity [10]. The spectrogram sequences, derived from

the GTZAN and Extended Ballroom audio inputs, were introduced as they entered the preprocessing that involved presenting backward and forward sequences for each spectrogram components [10]. Results indicated that the accuracy of BRNN on Extended Ballroom was higher than on GTZAN (92.7% to 90%).

Rafi et al proposed an improved RNN architecture for music genre classification - an Independent Recurrent Neural Network (Indrnn) - classification to handle gradient decay [9]. This architecture was introduced first by Wu et al because it performed better long time learning than Long-Short Term Memory (LSTM) and RNN. The construction of IndRNN consisted of a scattering transform that initiated feature extraction as the dataset entered pre-processing stage, a 5-layer IndRNN with labeled data combining with the ReLU function, and a softmax function being responsible for genre classification after training. The architecture was tested on GTZAN and obtained a high accuracy of 96% with only 23 epochs [9].

3.2 Convolutional Neural Network (CNN)

This type of method has been widely used in various research papers about music genre classification. The way CNN works is to take several spectrograms derived from the audio files as inputs and extract their patterns into a 2D convolutional layer with appropriate filter and kernel sizes [9]. The reason why spectrogram is mentioned in CNN is due to the model's effectiveness in recognizing image details [8].

Lau proposed to use a prepocessed GTZAN dataset to implement the Convolutional Neural Network (CNN) model. The dataset included an extracted Mel-Frequency Cepstrum Coefficient (MFCC) spectrogram for each song. Also, the audio excerpts in 3 seconds and 30 seconds were accompanied with their feature descriptions compiled in an additional .csv file [8]. He then built a CNN architecture using Keras that consisted of 5 convolutional blocks. Each block had a convolutional layer with 3x3 filter and 1x1 stride, a max pooling with 2x2 windows size and 2x2 stride, and a Rectifying Linear Unit (ReLU) function to display the probabilities for 10 genre genres, the highest of which would be chosen as a classified label for an input [8]. There were three CNN models trained on spectrograms, 20 MFCCs on 30-second and 3-second music pieces, and a classification test was operated on the test sets after training [8]. There was an issue in training datasets as Lau mentioned that the 3-second one was not in par with the numbers of genres in tha sample. That being said, there were genres that had less or more than the base number of samples (1000) [10].

Yu et al introduced the CNN method utilizing the Short-term Fourier Transform (STFT) spectrograms, consisting of various sequences of spectrogram vectors over time, as inputs [10]. The datasets mentioned in their paper were GTZAN and Extended Ballroom. Yu et al went on to extract each song from both datasets into 18 smaller pieces in 3 seconds with a 50% overlaps, making the data size set 18 times larger than the original for each genre label [10]. The STFT spectrograms were evaluated with size of 513x128, and the train-validate-test ratio was 8:1:1 [10]. In the first few layers of the CNN model, convolution filters and pooling kernels were set up in small sizes in order to capture unique audio features represented in the STFT spectrograms and diminish source loss [10].

Athulya and Sindhu came up with the idea of building a 2D Convolutional Neural Network (CNN). They used the GTZAN dataset to extract the audio files into various types of spectrograms using Librosa library. Those spectrograms were taken as binary inputs of a 2D CNN model, which were formed using Keras library. The layers were also created using TensorFlow library [6]. A 2D convolutional layer was presented at input shape 128x128x1. It contained a 2D NumPy array from the inputs that would be passed to the max-pooling layer, which would then operated a matrix that was half the size of the input layer [6]. There were 5 convolutional layers with a kernel size of 2x2, a stride of 2, and a max-pooling layer. Each output of a layer would then be inserted to a fully-connected layer, which also received a decreasing and flattened matrix size as inputs to carry out the classification [6]. The softmax function appeared at the end of the output layer to produce probabilities output. The architecture achieved an accuracy rate of 94%. Similarly, Nandy and Agrawal proposed a 2D CNN with 1D kernel on spectrograms derived from audio excerpts of Free Music Archive (FMA) datset. The model had an input dimension of 500x1500, resulting in producing an output of a vector with length 5000. The CNN was designed with blocks of convolution layer, batch normalization layer, activation layer, and - if possible - a max-pooling layer. A training/validation/testing ratio of 80/10/10 was applied to the 2D CNN model, along with the dropout parameter of 0.5 [1]. The model performed with an accuracy rate of 76.2% and a logloss rate of 0.7543, which outperformed other models from related research papers. An F1-Score of larger than 0.7 implied that the model was operating gradually well in classifying music genres.

One-dimensional (1D) CNN was also proposed for building the convolutional neural network training model to classify music genres. Allamy and Koerich introduced the 1D Resnet model whose convolutional layers (CLs) were replaced by residual blocks with the aim to prevent the model from degradation and vanishing gradient issues [2]. The blocks comprised two 3x3 (filter x kernel) CLs with stride one, two batch normalization (BN) layers, and an identity shortcut. All components would be activated by LeakyReLU function [2]. To be eligible for the structure of the model, the audio files from GTZAN datasets were extracted into music pieces of same length - 5 seconds - to be integrated with a sliding window of fixed width [2]. Falola and Akinola constructed the 1D CNN that included two separated layers: CNN layers where 1D convolution, activation functions (Rectified Linear Unit - ReLU) and pooling took action, and fully-dense layers. The 1D CNN model comprised 3 hidden CNN and 2 Multi-layer Perceptron (MLP) layers, a kernel size and sub-sampling (pooling) in each layer [3]. The model, after training on audio pieces from a Nigerian songs dataset, performed with an accuracy rate and f1-score of 92.7% and 92.7% for precision, which was the embodiment of a high-quality model. Ghosal and Kolekar introduced four different 1-Dimensional CNN models; a Max-Pooling model, a Max-Pooling with Long-Short Term Memory (LSTM) model, an average pooling model, and an average pooling LSTM model. Both max-pooling methods had an input filter of 128x64 with the kernel length of 3 and a factor of 2, while the LSTM one had hidden dimension of 64 [5]. The result after testing on inputs from GTZAN and Ballroom datasets revealed that Mel-Spectrograms worked best on CNN Max Pooling and Average

Pooling models, while Mel-Coefficients brought positive results for CNN Max-Pooling LSTM and Average-Pooling LSTM designs [5].

4 CONCLUSION

In This literature review, I have discussed how the Neural Network, or NN, works to classify music genres in general. First of all, I have discussed the type of datasets most researchers selected for their papers. As can be seen, the majority of those papers implement genre classification via GTZAN due to its popularity and accessibility. Second of all, I have listed significant Neural Network techniques such as Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). Each researcher has built his/her result by using a wide range of techniques on a regularly proposed dataset; thus, we cannot come to any conclusion about identifying the most optimum architecture. In the near future, NN will be an essential source of research for multiple tries on different music datasets in order to raise objective perspectives on the performance of each NN method.

REFERENCES

[1] Manish Agrawal and Abhilash Nandy. 2020. A Novel Multimodal Music Genre Classifier using Hierarchical Attention and Convolutional Neural Network. *arXiv preprint arXiv:2011.11970* (2020).

[2] Safaa Allamy and Alessandro Lameiras Koerich. 2021. 1D CNN Architectures for Music Genre Classification. *arXiv preprint arXiv:2105.07302* (2021).

[3] Peace Busola Falola and Solomon Olalekan Akinola. 2021. Music Genre Classification Using 1D Convolution Neural Network. (2021).

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

[5] Deepanway Ghosal and MF Kolekar. 2018. Musical genre and style recognition using deep neural networks and transfer learning. In *Proceedings, APSIPA Annual Summit and Conference*, Vol. 2018. 12–15.

[6] Athulya KM et al. 2021. Deep Learning Based Music Genre Classification Using Spectrogram. (2021).

[7] Macharla Vaibhavi P Radha Krishna. 2021. Music Genre Classification using Neural Networks with Data Augmentation. (2021).

[8] Dhevan S Lau and Ritesh Ajoodha. 2021. Music genre classification: A comparative study between deep-learning and traditional machine learning approaches. In *Sixth International Congress on Information and Communication Technology (6th ICICT)*. 1–8.

[9] Quazi Ghulam Rafi, Mohammed Noman, Sadia Zahin Prodhan, Sabrina Alam, and Dip Nandi. 2021. Comparative Analysis of Three Improved Deep Learning Architectures for Music Genre Classification. (2021).

[10] Yang Yu, Sen Luo, Shenglan Liu, Hong Qiao, Yang Liu, and Lin Feng. 2020. Deep attention based music genre classification. *Neurocomputing* 372 (2020), 84–91.