Applying and Comparing 1D-CNN and Bi-RNN Architecture in Music Genre Classification

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

ABSTRACT

Music Genre Classification has been one of the most prolific areas in machine learning. One of the most popular classification methods for this is the use of Neural Networks (or NN) to process large music datasets to identify the corresponding genres. There are various types of NN techniques applied on different music datasets such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). The study will focus on applying and comparing the efficiency between 1D-CNN and Bidirectional-RNN architecture on Free Music Archive (FMA) dataset.

KEYWORDS

Music Genre Classification, neural networks, convolutional neural network, recurrent neural network

ACM Reference Format:
Lam Hoang. 2018. Applying and Comparing 1D-CNN and Bi-RNN Architecture in Music Genre Classification. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Music Genre Classification is the method of classifying music into wide range of groups regarding to the complexity of cultures, musicians, and marketplaces. It has been a significant method among many music information retrieval (MIR) strategies [9]. Neural Network models have been widely applied in many research papers because of its ability to work in large dataset and result in better classification performance [7].

The first Neural Network architecture I would like to introduce is Convolutional Neural Network (CNN). Many research papers focus on building a 2D convolutional model for music genre classification because the audio data is converted into a 2D spectrogram [2]. Recently, there have been many researchers choosing to build a 1D-CNN model in their paper. This model is deemed more effective in extracting features from shorter pieces of the dataset [5].

The second Neural Network mentioned in this section is Recurrent Neural Network (RNN). RNN is a kind of neural network that is applied in many sequence labeling tasks [3]. The original RNN structure could not handle long-term dependencies as the model struggled from vanishing gradients - a factor leading to the failure of a model during its training. Therefore, many research papers has paid attention to building a RNN with Long-Short Term memory (LSTM) model or a Bidirectional RNN with Gated Recurrent Unit (GRU) to handle the issue.

From the listed CNN and RNN descriptions and methods, I plan to implement and evaluate the performance between 1D Convolution Neural Network (CNN) model and Bidirectional Recurrent Neural Network (RNN) model on Free Music Archive (FMA) dataset. Those are the approaches that improve the quality of neural network methods on music genre classification by handling existing issues from traditional architectures. The rest of the proposal will be organized as follows: section 2 - Related Works - will describe several past works from researchers focusing on this topic. Section 3 and 4 introduce the design of my intended methods and how to evaluate their performance. Section 5 - Timeline - will demonstrate the details of my timeline to work on my topic for the senior capstone project.

2 RELATED WORKS

2.1 Recurrent Neural Network (RNN)

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 cyclic connections that transfer the activations from the previous temporal step to another [10]. The original RNN structure could not handle long-term dependencies as the issues relating to vanishing gradient might arise. Vanishing gradient occurred when the information about the initial layer (root) such as weight and biases (gradients) are not updated during traversal from one layer (child) to the initial in each training session. The incident makes RNN decrease its gradients exponentially, resulting in an inaccurate training result [8] Therefore, many research papers has paid attention to building a Bidirectional RNN with Gated Recurrent Unit (GRU) to handle the issue.

Yu et al proposed to build a bidirectional RNN (BRNN) to classify music genres using Gated Recurrent Unit (or GRU). The reason to apply GRU for BRNN architecture is that the unit made recurrent unit keep track of dependencies of different time frames flexibly [10]. This model contained two stacked bidirectional RNN (BRNN) layers and a shortcut connection to detect vanishing gradient without the need for any computational complexity. The spectrogram sequences were created from the GTZAN and Extended Ballroom audio inputs[10]. Results indicated that the accuracy of BRNN on Extended Ballroom and on GTZAN were 92.7% and 90%, respectively.

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Feng et al introduced the Bidirectional Gated Recurrent Block (BGRU) as a parallel design with a CNN to extract features within time-related sequences. The input was a Short-term Fourier Transform (STFT) spectrogram of size 128x512, whose features were extracted in the next 7 layers including the two-stacked BRGU unit. The components from this unit eventually turned into a 256-dimension output. Adding the BGRU unit aimed at improving the extraction performance of the model, and the result showed an impressive test accuracy of 88% to 92% when combining with a different model on GTZAN dataset.

2.2 Convolutional Neural Network (CNN)

The way Convolutional Neural Network (CNN) works is to transform audio inputs into data visual representation and extract their patterns at convolutional layers with corresponding filter and kernel sizes. The input will then be fed to a classification layer that produces an output, which is a genre prediction of a music track. CNN has been applied in various computer vision tasks that involves processing and recognizing images and videos [1]. Recent studies have suggested applying a 1D architecture for this topic because of its shallow design, easy training and operation, and less computational demands [5].

Allamy and Koerich introduced the 1D-CNN Resnet model whose convolutional layers (CLs) aimed at preventing the model from degradation and vanishing gradient issues [1]. To be eligible for the structure of the model, the audio files from GTZAN datasets were extracted into music pieces of same length. 1D Resnet was tested on raw audio samples and artificial audio samples derived from GTZAN dataset, and it achieved a fairly high accuracy of 76% and nearly 81%, respectively.

Falola and Akinola constructed the 1D CNN that included CNN layers and a layer that connected the input to the output. The audio file inputs were taken from a dataset consisting of 1000 Nigerian songs with four genres. The features in each file were extracted at the beginning, the middle, and the end with a total of 30 seconds using Librosa library. The model performed with an accuracy rate and f1-score of 92.7% and 92.7% for precision, which demonstrated a high efficacy of the model [5].

Bian et al proposed a 1D CNN model that included 5 convolutional layers with kernel size of 4. The dimension of the spectrogram input decreased at each layer. After that, the input would be taken into a genre prediction layer to produce an output reflecting the genre of a song. The model performed at 64.7% on FMA small dataset and 84% on GTZAN. The advantage of 1D CNN in their work is the model's ability to traverse throughout the entire frequency at once [2].

3 DESIGN

For the proposal, I will concentrate on comparing the performance of two neural network methods: the 1D convolutional neural network (CNN) and the Bidirectional recurrent neural network (RNN). As both CNN and RNN are deep learning techniques, a spectrogram is proposed as an input for both models due to deep learning's success in various Computer Vision tasks such as image classification, object detection, facial expression recognition, and more [9]. The goal is to evaluate these neural network methods in classifying music genres based on the accuracy using a music dataset.

3.1 Dataset

A medium version of Free Music Archive (FMA) will be used for this project. This version contains 25,000 tracks from 16 unbalanced genres, each has a length of 30 seconds. The FMA dataset, built in 2017, is suitable for music genre classification thanks to providing clear genre information (for example, specified sub-genre in a track), having a music genre hierarchy, and being updated with high audio quality [4].

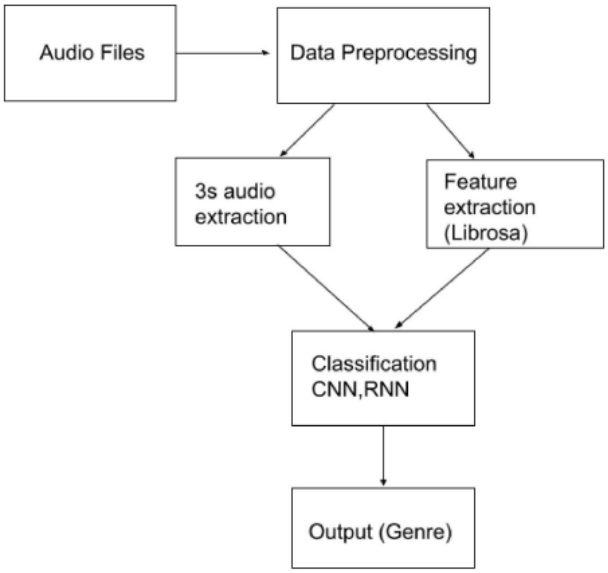

Figure 1: Proposed diagram of the whole design

3.2 Convolutional Neural Network

The proposed Convolution Neural Network design is based on the 1D-CNN architecture developed by Bian et al. Spectrograms created from preprocessing and feature extraction on FMA dataset (medium) will be set up as input layers of size 128 x 128. This architecture consists of 5 convolution layers with a consistent kernel size of 4 throughout each layer. After each convolution layer, a max pooling layer will appear with filter size of 4, making the output size divided by 4 per layer. Batch normalization and rectified linear unit (ReLU) are embedded at each convolution layer as well. After the third convolution layer, the filter size for max pooling decreases to 2 and 1 sequentially. The final layer, replacing a fully connected layer, has a filter size of 1, and it will be fed into a Support Vector Machine (SVM) for music genre prediction. After that, there are two dense layers with 1024 and 8 hidden units to process the prediction information, which eventually produce a softmax output layer. Apart from other CNN models, this model works much better in analyzing data throughout the full length sequence [5] and being able to discover audio frequency patterns [1]. As Bian et al. have

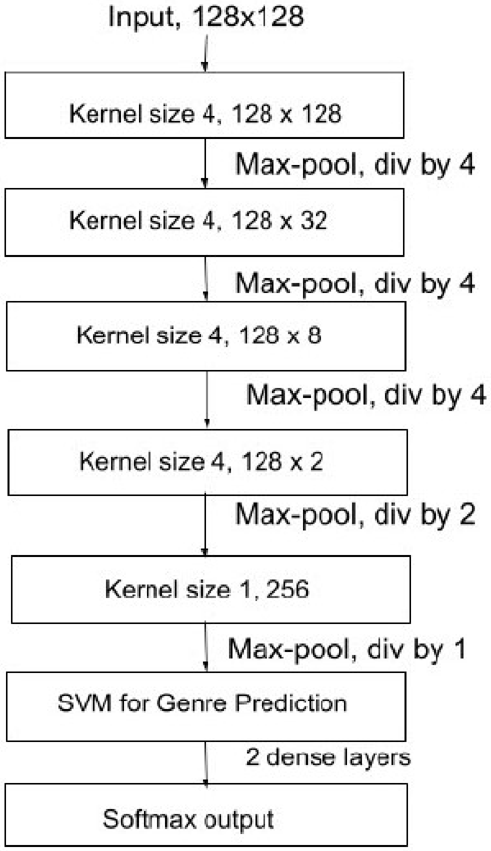

Figure 2: Proposed diagram of the 1D-CNN

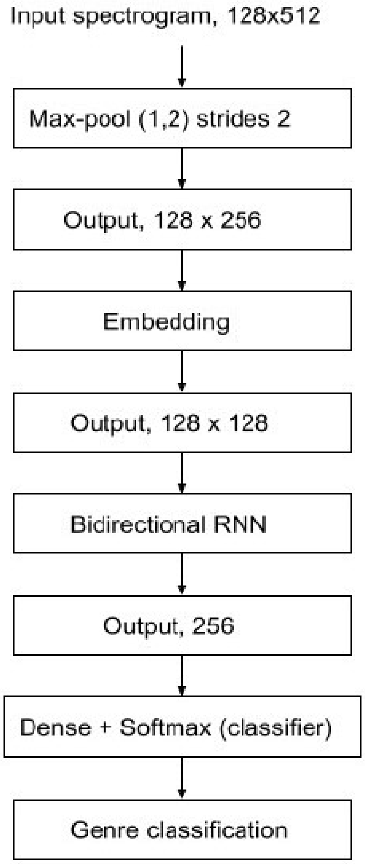

Figure 3: Proposed diagram of the BiRNN-GRU

implemented this model on a small version of FMA, I would want to apply the same method on the medium-sized FMA dataset to see if my proposed 1D-CNN could do much better than in the research paper.

3.3 Recurrent Neural Network

The proposed design will be setting up a Bidirectional Recurrent Neural Network (BRNN) using Gated Recurrent Units (GRU) that was presented by Feng et al. An input layer, containing a spectrogram of size 128 x 512, will be processed by a max-pooling layer to decrease the dimension of the input to 128 x 256. Due to having a complicated design, an embedding layer is generated to reduce the number of parameters needed for dimension reduction. Then, a 128 x 128 feature layer will be taken into a 1-layer Bidirectional RNN with GRU units for feature extraction, producing one 256-Dimensional feature vector as the output. The GRU is deployed in order to receive the information from the sequences in the block, extract more temporal correlations from music samples, and handle vanishing and exploding gradients in the long run. Another benefit from GRU's appearance is making the model more concise, reducing the possibility that the overall model leans towards insignificant data, and operating faster than other RNNs with few frameworks [6]. I choose this model because I would like to experiment on a different music dataset as well as figure out if BiRNN works better if it is not paired with another neural network model.

4 TIMELINE

This section features a detailed schedule table reflecting the completion process of my project in CS488. As can be seen from the table, the project will be done as a part of Senior Capstone Experience, which takes place from early-October to late-November. The project starts with collecting and analyzing data from my proposed dataset and ends with presenting the software as well as publishing the paper. The majority of work for this project involves building the software and writing the paper.

Table 1: Detailed Schedule for CS488

Date	Tasks
Break	Collect and Process data
4-10 Oct	Implement proposed CNN and RNN designs
11-17 Oct	Design evaluation and first draft of paper due
18-24 Oct	Finish up and review the software for first release
25 Oct - 7 Nov	Working on the paper and modifying the software
8-14 Nov	Second paper draft and software due
15-26 Nov	Final paper and software due

5 EVALUATION METHOD

5.1 Accuracy

The accuracy evaluation metric indicates the ratio between precisely predicted outcomes and the aggregated number of predictions [5].

$$Accuracy = \frac{NumberofAccuratePredictions}{TotalNumberofPredictions}$$

5.2 Precision

This evaluation metric will identify whether the model predicts positively correct or incorrect [5]

$$precision = \frac{NumberofCorrectPositives}{TotalNumberofPositivePredictions}$$

5.3 Recall

The recall metric determines the ratio between true (correct) positives and the total number of actual positives [5]

$$Recall = \frac{NumberofTruePositives}{TotalNumberofActualPositives}$$

6 ACKNOWLEDGEMENT

I sincerely am thankful for the help of David Barbella, Assistant Professor of Computer Science and the instructor of CS388. He has given me many useful feedbacks and suggestions that are helpful to enhance the quality of my proposal.

REFERENCES

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

[2] Wenhao Bian, Jie Wang, Bojin Zhuang, Jiankui Yang, Shaojun Wang, and Jing Xiao. 2019. Audio-based music classification with DenseNet and data augmentation. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 56–65.

[3] Jia Dai, Shan Liang, Wei Xue, Chongjia Ni, and Wenju Liu. 2016. Long short-term memory recurrent neural network based segment features for music genre classification. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 1–5.

[4] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2016. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840* (2016).

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

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

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

[8] Phong Le and Willem Zuidema. 2016. Quantifying the vanishing gradient and long distance dependency problem in recursive neural networks and recursive LSTMs. *arXiv preprint arXiv:1603.00423* (2016).

[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.