Applying and Comparing 1D-CNN Architectures 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 Artificial Intelligence, especially in the field of deep 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, one of which is the Convolutional Neural Network (CNN). Therefore, in this paper, I will focus on the process of applying 1D-CNN in Python and how the models' performance are evaluated using metrics. My approach will be implemented on the medium-sized Free Music Archive dataset due to its containing a large amount of songs and genres. Based on what I have done, the model performs at an accuracy of 35% on the medium-sized Free Music Archives dataset, which contained 16 different and unbalanced genres. The result was obtained after a set of training and testing with 8 out of 16 genres selected.

KEYWORDS

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

ACM Reference Format:
Lam Hoang. 2018. Applying and Comparing 1D-CNN Architectures in 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 AND RELATED WORKS

With the rise of the Internet, music has become more accessible to everyone around the world. Therefore, music information retrieval (MIR), concentrating on research and developing the computation design to give researchers more depth about the music information such as genres and length, exists to help listeners listen to their favorite songs from a lot of options [5] and discover different music backgrounds [2]. Of all the methods, Music Genre Classification, the method of classifying music into wide range of groups regarding to the complexity of cultures, musicians, and marketplaces, has been a significant method among many music information retrieval (MIR) strategies [11].

Recently, Neural Network models have been widely applied in many research papers because of its ability to work in large dataset and perform outstandingly in genre classification [8]. One of the models that is chosen to be presented in this paper is the Convolutional Neural Network (CNN), which has been experimented in many current studies [2] due to its end-to-end training architectures which combine feature extraction with music classification in one stage [6]. Having been applied in various computer vision tasks that involves processing and recognizing images and videos [2], the way CNN works is to transform a 2D spectrogram from audio inputs and extract its features to classify the genre through 2-dimensional CNN layers. A reason why a spectrogram is presented as an input for the CNN model is due to its success in various Computer Vision tasks such as image classification, object detection, facial expression recognition, and more [11]. Also, due to the high sampling rate in each audio, representing audio in spectrogram could help reduce the complexity of the model more than in raw waveform featuring time domain only [12]. Therefore, the sound waves can be represented as spectrograms, which will later be processed as images and let researchers regard the CNN as a model that performs its visual approaches. Athulya and Sindhu developed their 2DCNN with five convolutional layers in order to validate their recommendation system [8]. Nandi and Agrawal built the 2DCNN model along with the hierachical attention network to compare the performance between those models and to examine whether processing audio inputs into time-frequency axis or natural language processing is more effective [1].

Although the approach of applying 2D CNN reached its highly developed performance and was considered effective for feature extraction, this model did not seem to perform well on frequency axis [3]. Hence, 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]. The CNN features 1D layers that are deemed more effective in extracting features from shorter pieces of the dataset [5]. Several researchers have applied 1D CNN model in a more complex way. For example, Allamy and Koerich introduced the 1D-CNN Resnet model whose convolutional layers (CLs) aimed at preventing the model from degradation and vanishing gradient issues [2]. Li et al proposed a Dilated 1D CNN that differs from the traditional CNN in producing gaps between the layer's filter for every input unit [9]. Bian et al introduced 1D CNN models including the Resnet blocks that improve the functions of the original 1D CNN.

In this final paper, I concentrated on implementing the 1D CNN and comparing its performance with other works done by researchers. The goal is to evaluate these neural network methods in classifying music genres. From what I have studied, a lot of the researches

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

used GTZAN dataset as the primary music dataset to be experimented with the 1DCNN; therefore, what I intended to do was to experiment the 1D CNN model on a different dataset. I selected the FMA dataset because it contains different dataset sizes (large, medium, small), which encourages me to explore more and compare the performance of my model against those from various research papers.

2 DESIGN AND IMPLEMENTATION

The Design and Implementation introduce the sketch of my design and how I implemented them using software.

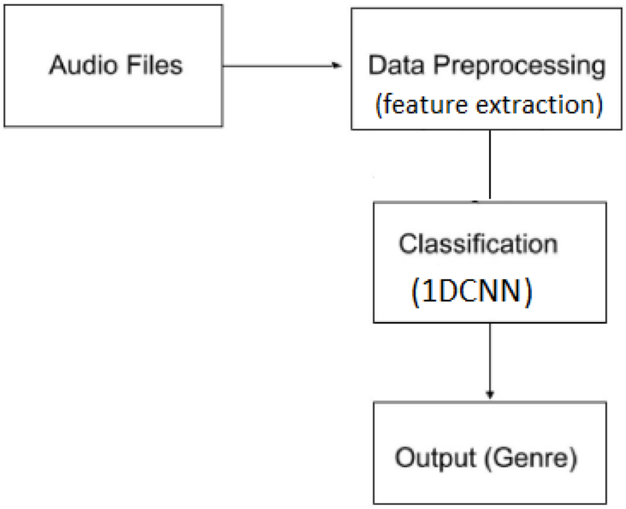


Figure 1: Proposed diagram of the whole design

2.1 Dataset

A medium version of Free Music Archive (FMA) was used for this project. This version contains 25,000 tracks from 16 unbalanced genres, each has a length of 30 seconds [4]. 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].

2.2 Preproccessing and Feature Extraction

For this stage, a Python library called Librosa [10] was deployed to take the audio file with its corresponding track id from the csv file in order to create a mel-spectrogram. Then, the features represented by the spectrogram were extracted and later fed into the 1dCNN as an input. The tracks were set with a sample rate of 22050 Hz, forming 661500 samples from 30 seconds each. Then, with a sliding window of 1024 samples, the mel-spectrogram was created with 661500/1024 = 646 windows and a fixed height of 128 pixels. The genres were also labeled as numerical values from 0 to 7 to represent 8 most popular genres out of 16.

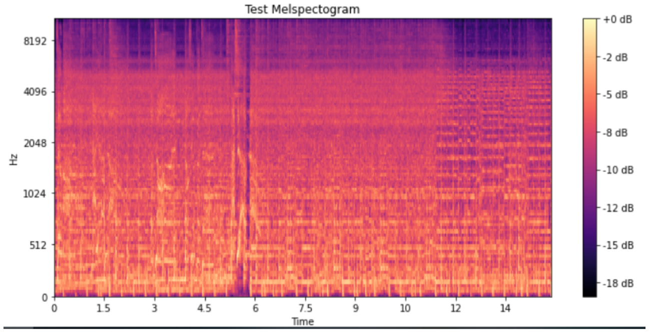


Figure 2: Spectrogram of the audio sample

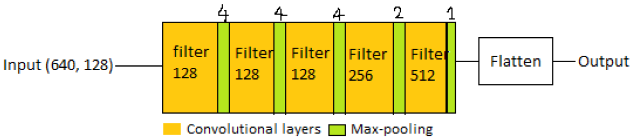


Figure 3: 1DCNN Diagram

2.3 1D Convolutional Neural Network

The 1D Convolutional Neural Network (CNN) model was built based on the architecture developed by Bian et al and it was demonstrated by Figure 3. The difference from Bian's model is that this model used a colored spectrogram with dimensions (640, 128) as an input rather than a grayscale spectrogram (128, 128). This is because the colored spectrogram was apparent enough to spot the intensity of amplitudes and frequencies in tracks. The tool to initiate the model is a Python library called Keras [7] and it was done on Jupyter Notebook. A spectrogram of dimension (646, 128) served as an input and it was passed along 5 convolutional layers, each of which had a kernel size of 3 and a stride size 1, to begin the classification process. Several max-pooling layers were embedded between the output of one 1D CNN layer to another. The max-pooling layers with sizes 4 were placed between each of the first 3 1D layers. A max-pooling layer of size 2 was placed after the fourth convolutional layer (CL), and the final layer of size 1 was placed after the last CL. After the classification process, a flatten layer processed the prediction information, which were used to produce a softmax output layer.

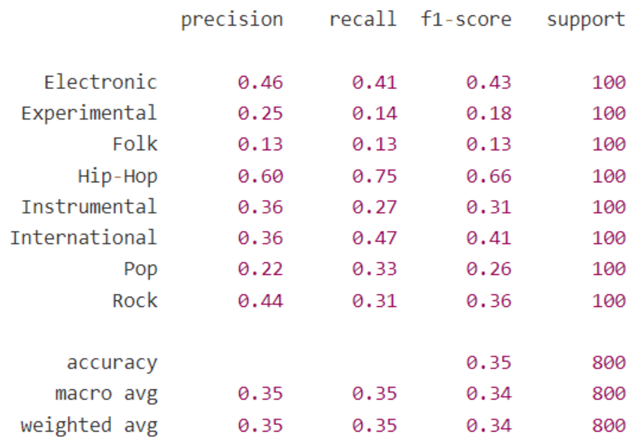
2.4 Classification

The model was initiated by taking Adam as an optimizer function. ReduceLRonPlateau was used to decrease the learning rate when metrics "accuracy" or "val-accuracy" cease improving. To make models work, a batch size needed to be defined in order to bring the input at every epoch of training. An epoch is understood to be an iteration along the whole dataset and, for this experiment, the epoch was set to 50 so that the computer would be less struggling to proceed the training model than with lesser epoch [5]. After the training, the model was saved using Keras function. All training process was done by using Python library Tensorflow [13].

Table 1: 1DCNN Performance Comparisons

Model	Author(s)	Result
Allamy-Koerich	1DCNN Resnet	80.93%
Falola-Akinola	1DCNN	92.5%
Li-Xue-Jiang	Dilated CNN (2 layers)	87%
Li-Xue-Jiang	Dilated CNN (3 layers)	79%
Bian	1DCNN	64.7% (FMA-small)
Bian	1DCNN	84% (GTZAN)
Lam	1DCNN	35%

3 RESULTS



```
              precision    recall  f1-score   support

   Electronic       0.46      0.41      0.43       100
 Experimental       0.25      0.14      0.18       100
         Folk       0.13      0.13      0.13       100
      Hip-Hop       0.60      0.75      0.66       100
 Instrumental       0.36      0.27      0.31       100
International       0.36      0.47      0.41       100
          Pop       0.22      0.33      0.26       100
         Rock       0.44      0.31      0.36       100

     accuracy                           0.35       800
    macro avg       0.35      0.35      0.34       800
 weighted avg       0.35      0.35      0.34       800
```

Figure 4: Test Evaluation for 1DCNN model

This figure displayed the performance of the 1DCNN model based on the precision, recall, f1-score, and accuracy. The features were extracted from the spectrogram of each song. As can be seen from the figure, the metrics were applied to evaluate the performance of the model to identify the genres Electronic, Experimental, Folk, Hip-Hop, Instrumental, International, Pop, Folk from the dataset.

In general, the 1DCNN introduced in this paper did not perform as impressive as other models. From this figure, we can see that the model performs at an accuracy score of 0.35, or 35% accuracy. The majority of the models performed at an accuracy rate of around 80% to more than 90%. Specifically, Bian's basic 1DCNN model performed at an accuracy rate of 64% with the small version of the FMA dataset. A reason that could lead to this result may be due to the fact that the filter of the dimension input increased gradually after each convolution layers from size 128 at the beginning to 512 at the end. This justifies the claim made by Allamy and Koerich while comparing his 1DCNN Resnet against other variations of 1DCNN models that an architecture that featured large convolutional layers gained the worst average accuracy [2].

4 CONCLUSION AND FUTURE WORK

Music genre classification has been an important method to support listeners to expand their list of favorite songs and music genres. This paper has introduced a classification technique of using the 1-dimensional Convolutional Neural Network (CNN) model and provided details about how the model is being designed, implemented, and evaluated with other 1DCNN models from previous research papers. Overall, the average accuracy of my 1DCNN model did not perform as impressive as other 1DCNN models that researchers have implemented in the past, which encourages me to explore more on how to make a better model from now on. In the future, I will focus on trying different music genre classification methods that are Neural Network related such as building a 2DCNN or pairing a 1DCNN-RNN model on the same dataset.

5 ACKNOWLEDGEMENT

I sincerely am thankful for the help of David Barbella, Assistant Professor of Computer Science and the instructor of CS488 during the January Term. He has given me many useful feedbacks and suggestions that are helpful to enhance the quality of my proposal. Also, I would give credits to a Github project that focused on the same topic as mine because it brought an inspiration and resources to help me complete my project. The project link can be found here: https://github.com/priya-dwivedi/Music_Genre_Classification

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] 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.

[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] Nikhil Ketkar. 2017. Introduction to keras. In *Deep learning with Python*. Springer, 97–111.

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

[9] Haojun Li, Siqi Xue, and Jialun Zhang. 2018. Combining CNN and Classical Algorithms for Music Genre Classification. (2018).

[10] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8. Citeseer, 18–25.

[11] 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).

[12] Lonce Wyse. 2017. Audio spectrogram representations for processing with convolutional neural networks. *arXiv preprint arXiv:1706.09559* (2017).

[13] Giancarlo Zaccone, Md Rezaul Karim, and Ahmed Menshawy. 2017. *Deep learning with TensorFlow*. Packt Publishing Ltd.