

Automatic Music Genre Classification based on CNN and LSTM

Xining Luo *

Department of Math and Physical Sciences, University of Toronto, Toronto, Canada

* Corresponding Author Email: cynth.luo@mail.utoronto.ca

Abstract. Various applications of machine learning are discovered and receiving more and more attention contemporarily. The music industry has benefited from the incorporation of artificial intelligence, especially the field of music classification, as machines are able to organize big data in a more efficient manner than the traditional human expertises. This paper compares two machine learning models, the Convolutional Neural Network model (CNN), and the Long Short Term Memory model (LSTM), from their architectures, functionality, to classification accuracy based on empirical data. The models were trained on two datasets, GTZAN and FMA. The result indicates that the CNN model achieved a 56.0% and 50.5% accuracy for the two datasets respectively, outperforming the LSTM model, which had a 42.0% and 33.5% accuracy. The paper aims to analyze the two models' capability for music classification and determine which model is better suited for the task. These results shed light on guiding further exploration of computer music.

Keywords: Music Genre Classification; CNN; LSTM; GTZAN; FMA.

1. Introduction

The history of machine learning dates back to the 1950s. The earliest machine learning model is usually believed to be proposed by psychologist Frank Rosenblatt. Based on ideas about the biological nervous system, Rosenblatt created a group that built a machine named “perceptron” for recognizing the letters of the alphabet. This machine then became the prototype for artificial neural networks and inspired more research. In the early 1960s, more groups were engaged in the design and testing of learning cognition systems, and experimented and developed different approaches for machine learning. However, The book *Perceptrons: An Introduction to Computational Geometry* by M. Minsky and S. Papert, published in 1969, pointed out many limitations of perceptrons. Its emergence caused an enormous reduction of funding in artificial intelligence research until the 1980s. Although slightly more studies of deep learning and artificial neural networks were conducted since 1980, practical progress made by these studies was below expectation and the investment in the field of artificial intelligence decreased again around the 1990s. Finally, a real advance was made in 1995, where a new algorithm was introduced by Cortes and Vapnik under the name support-vector networks. This new model was widely used and proven to have high effectiveness; its success had a great impact on both theoretical and applied machine learning and inspired further studies. Then arrived the twenty-first century, a historical pivoting point for machine learning research. The rise of Big Data truly gave birth to more new approaches and perspectives, since it provided the necessary data required for practical experimentations [1-6].

The application of machine learning in the field of music was one of the numerous new aspects that were brought to life in this era. Currently, the two most commonly mentioned and used models are CNN and LSTM. Numerous experiments regarding the two models were conducted. Bahuleyan explored the CNN model's capability for music genre classification [2]. The group trained the model with features extracted in the time and frequency domain of audio signals, and analyzed the results. O'Shea and Nash broke down the CNN model's basic structure and functionality and demonstrated how it classifies a song [7]. Tibadiya discussed the LSTM model's structure and music classification ability, explained the structure and functionality of the model, and trained the model similar to the procedures done on CNN [8]. However, the LSTM model was used in areas other than music classification. Hewahi, AlSaigal, and AlJanahi proposed that neural networks can be used for music generation as well and implemented a computational music generator based on the LSTM model [5].

In the era of Big Data, the need for accurate data organization is becoming more and more crucial. Traditionally, music classification is done by human expertise, and such manual process is labour intensive and time consuming. The application of new deep learning models in this area is beneficial, as it reduces the labour cost and improves the efficiency of the classification process. This paper will discuss the two most commonly used approaches for music genre classification mentioned earlier, which are the CNN model and the LSTM model, and compare their capability regarding this task.

2. CNN

The CNN model relies on artificial neural networks—a computational system inspired by biological nervous systems, consisting of numerous interconnected computational nodes. Traditionally, neuron layers within artificial neural networks only have height and width. However, neurons within CNN are organized into three-dimensions. The added dimension, depth, represents the activation volume, which condenses the output file. Neurons in a CNN within any given layer only connect to a small region of the preceding layer [7].

The CNN model mainly consists of three layers. The first layer is the input layer, which holds the pixel values of the spectrogram image. The second layer, the convolutional layer, determines the output of the neurons of which are connected to local regions of the input. The pooling layer comes next, which performs downsampling on the given input and reduces the number of parameters within activation. Lastly, the fully-connected layer, is the layer where the final classification process is performed [7].

This model specializes in image recognition tasks. When processing a sound input, it first splits the audio into small chunks, then visualizes them into spectrograms, two-dimensional representations of signals, through functions from Librosa python library [2, 6]. The Librosa library contains two domains; The audio signals are analyzed in the time domain using Fast Fourier Transform, and are converted into spectrograms in the frequency domain. Afterwards, it extracts features that can help with the music genre classification from the spectrograms, these include the spectral centroid, the spectral flux, and the spectral contrast [9, 10]. The structure of a CNN is sketched in Fig. 1.

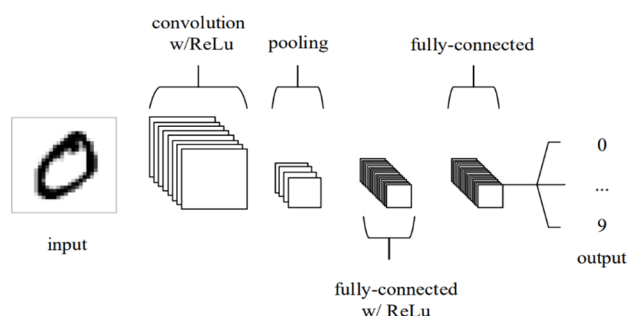


Fig 1. The procedure of CNN.

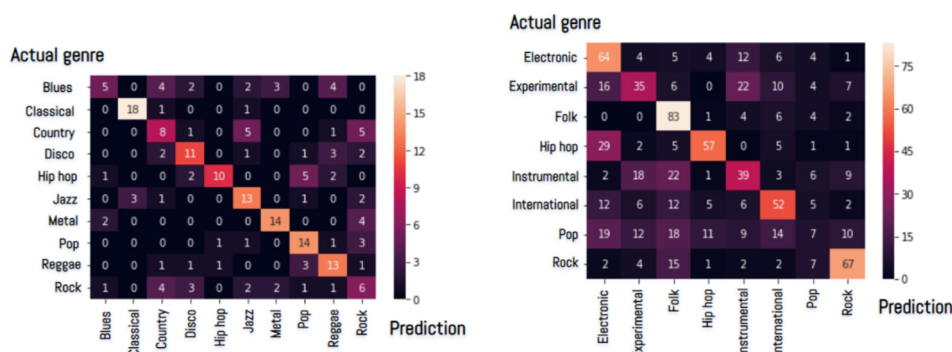


Fig 2. The results of GTZAN (left panel) and FMA (right panel) for CNN.

The GTZAN dataset consisted of 1000 audio tracks with each being about 30 seconds long [4]. The tracks were categorized into 10 genres: blues, classical, country, disco, hip hop, jazz, metal, pop, reggae, and rock. Each genre contains 100 tracks. The CNN model resulted in a 56.0% accuracy when tested by this dataset and managed to classify at least 50% of the songs correctly in 6 out of the 10 genres. It had an easy time classifying classical songs, but had difficulty with blues, country, and rock, as shown in Fig. 2. The FMA dataset contains 106,574 untrimmed tracks. When selecting input, the set returns 8,000 tracks with 8 balanced genres: electronic, experimental, folk, hip hop, instrumental, international, pop and rock. The model had a 50.5% accuracy when trained on this dataset and managed to get an accuracy of above 50% for 5 out of the 10 genres. It had an exceptionally hard time classifying pop songs, where it only got 7 out of 100 songs correct, as shown by the table below. On the other hand, it rarely had difficulty with folk, where it got 83 out of 100 correct [4].

3. LSTM

The LSTM model also uses artificial neural networks. However, the network it relies on is slightly more complex than regular artificial neural networks. It is called the recurrent neural networks, which builds connections between previous and current states in the neural network. This offers the network a kind of short-term memory.

The LSTM model, in addition to the basic architecture of recurrent neural networks, adds a special type of cells called memory cells. These cells effectively provide the network with long-term memory, allowing the network to read, write, or delete stored information. Memory cells consist of three gates: the input gate, the output gate, and the forget gate. The input gate controls the amount of data being loaded into memory, the output gate controls the data being passed onto the next layer, and the forget gate controls the loss within memory storage. The structure of a memory cell is shown in the image Fig. 3.

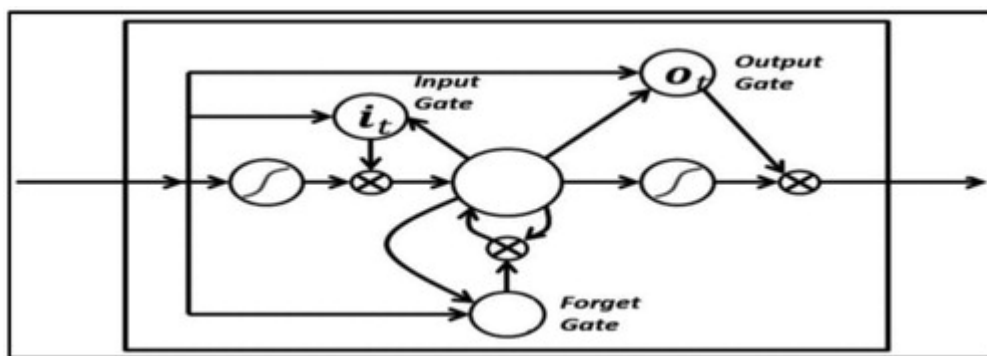


Fig 3. A sketch of LSTM.

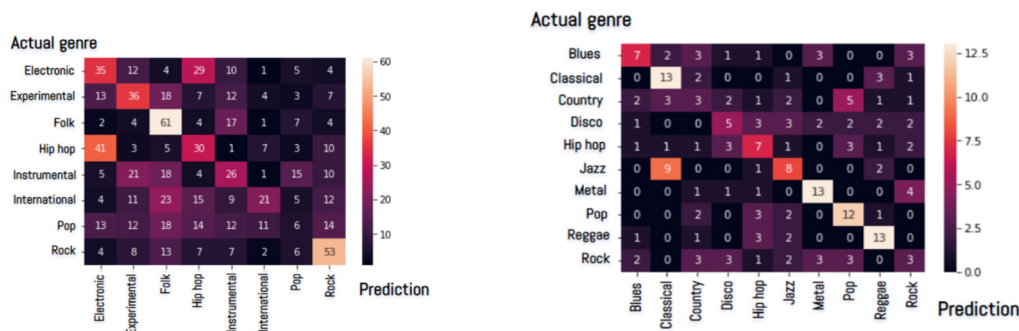


Fig 4. The results of GTZAN (left panel) and FMA (right panel) for CNN

This model specializes in processing sequential data like text and speech. Its music genre classification process relies on mel frequency spectral coefficient, which is a feature commonly used for speech recognition, genre classification, and audio signal similarity measurement. This feature is directly extracted from audio tracks by using functions from the Librosa library [8]. The same two sets of data that were used to test the accuracy of the CNN model—the GTZAN dataset and the FMA dataset—were also used to test that of the LSTM model. When trained on the GTZAN dataset, the model resulted in a 42.0% accuracy. It had an over 50% accuracy for 4 out of the 10 genres, and none of the genres appears to be exceptionally easy or difficult to classify. The result is shown by the table below. The model had a 33.5% accuracy when tested by the FMA dataset. The data reveals that the model is struggling with classifying pop songs, where it only predicted 6 out of 100 songs correct. Furthermore, it only managed to get an accuracy of above 50% for 2 of the genres, which were rock and folk [4]. The results are illustrated in Fig. 4.

4. Comparison, Analysis & Suggestions

Both the CNN and LSTM model make some modification to the basic architecture of regular neural networks to accommodate their intended specialization. CNN adds another dimension to the network, allowing it to compress the input, reducing the number of parameters required to train, and ultimately producing better image analysis ability. The LSTM model, instead of adding more dimensions, constructs a new type of cell within the existing network, and the improved capability of storage enhances its speech processing ability.

Despite both relying on functions from the python Librosa library for feature extraction, the two models differ in their procedure and the type of feature they target. The CNN model first converts the audio input into spectrograms, then extract image-based features from these spectrograms. On the other hand, the LSTM model directly extracts a sound-based feature from the audio tracks.

Regarding the classification accuracy test, the CNN model had a 56.0% accuracy when trained on the GTZAN dataset, as compared to the 42.0% accuracy of the LSTM model. While the former struggled greatly with the prediction of blues, country, and rock songs and had an extremely easy time with that of classical songs, the latter, despite having a lower overall accuracy, did not appear to have an exceptionally difficult or easy time with the classification of any of the genres. When trained on the FMA dataset, the CNN model had a 50.5% accuracy and the LSTM model had a 33.5% accuracy. Both models were visibly having a difficult time classifying pop songs. Nevertheless, the CNN model still managed to get predictions with high accuracy for folk songs, while the LSTM model did not have any genres which it had a noticeably easy time with.

Based on the statistical results of the music genre classification experiments, it is evident that the CNN model outperforms the LSTM model in terms of overall classification accuracy. It is also interesting that certain genres were exceptionally easy or hard for the models to classify. Another issue is that the classification accuracy of the same genres from the two datasets sometimes differ greatly.

The structure and method used for data processing be taken into consideration regarding why the CNN model had a more consistent performance in the classification accuracy test. The CNN model contains a pooling layer, which reduces the number of parameters required for the classification stage. This can possibly make the CNN model identifying patterns more easily than the LSTM model.

The two models were often similar in terms of the genres they had an easy time or struggled with. Both models appear to be having little difficulty with classifying classical songs from the GTZAN set. One reason may be due to the distinctive instruments used in classical music. As compared to most genres involved in the dataset, which are mostly electronic music genres, classical music almost solely features acoustic instruments. Another reason may be the lack of drum beats in classical music relative to other genres.

Folk music from the FMA dataset is a similar case, which the two models rarely trouble with. This might be due to the genre having a distinctive sound like that of classical music, as discussed in the

previous paragraph. The lack of electronic music and unique acoustic instruments used in folk music might have made it stand out from the other genres.

It is very intriguing that both models were having a much more difficult time classifying rock songs from the GTZAN dataset compared to those from the FMA dataset. The CNN model only predicted 6 out of 20 correct for the GTZAN dataset, in comparison to getting 67 out of 100 correct for the FMA dataset. By the same token, the LSTM model only got 3 out of 20 right when trained on the GTZAN dataset, while 53 out of 100 in the case of the FMA dataset. An explanation can be that rock has many sub-genres. The GTZAN dataset most possibly contains rock songs of a much more diverse sub-genres than those from the FMA dataset, which may have resulted in these songs sounding like other genres from the dataset.

The classification pop genre is identical by nature. This time, the two models had a harder time predicting songs from the FMA dataset than songs from the GTZAN dataset. The same reason can be used to explain such differences in the classification accuracy of pop songs from the two dataset, since pop is also a genre that has various sub-genres.

The accuracy test produced discrepancy in the result due to the two dataset having different genres and songs that were collected differently, which may affect the analysis of the models' classification ability. To make future testing more consistent, it is suggested to have one single large dataset that contains songs that were collected through various means. This will eliminate the disparity between datasets and produce more consistent results.

The experiment also shows that the CNN model outperforms the LSTM model in spite of the dataset it was trained on. This may lead to the conclusion that the CNN model is better suited for music genre classification. However, this result is not sufficient enough to dismiss the capability of the LSTM model in the area of music classification. Firstly, it is difficult to determine what a good prediction of songs from machines is, since it is evaluated from a human's perspective and humans are not absolutely perfect at classifying songs. Secondly, there are aspects that are yet to be tested. For example, classifying songs based on emotions. It is recommended to also conduct studies on different types of classification using the two models. This will provide a clear understanding of which model performs better when given with certain different pieces of information.

5. Limitation & Prospects

One limitation of the paper is the insufficiency of data used in the analysis. Only two data sets are involved in the classification accuracy comparison. Moreover, all four tests were done by the same group in one experiment. Having such a small sample size can have a critical impact on both the accuracy of the comparison and scope of the conclusion. Solely based on the single experiment, the LSTM model is shown to have a much lower accuracy than the CNN model. Nevertheless, it is possible that the former may turn out to outperform the latter in experiments conducted by other groups. The second lamination of the paper is its lack of analysis on the datasets, because factors such as how sample audio tracks were collected and the evaluation criteria for genre classification accuracy can all affect the result.

The paper would have been improved if results from more tests are included and compared, especially those done with different datasets. As discussed in the previous section, all audio tracks are pre-categorized by human expertise and there is subjectivity in manual classification, which makes it difficult to set the standard for an accuracy classification done by machines. Gathering data from a variety of experiments and considering the contents of their datasets will provide a more thorough understanding of the two models. The paper will be able to analyze each model along with the different characteristics of the datasets it was trained on. Doing so offers more aspects that can be taken into consideration when making the final comparison.

6. Summary

In conclusion, this paper investigates the CNN model and the LSTM model and discusses the difference in their capability in music genre classification based on training on the GTZAN dataset and the FMA dataset. According to the analysis, both models rely on functions from the Librosa python library to extract specific features from the audio track when attempting to classify its genre. However, their procedures differ. The CNN model breaks down the audio into small chunks, visualizes them into spectrograms, then tracts the target features. On the other hand, the LSTM model directly extracts features from the audio track. The CNN model turns out to outperform the LSTM model regarding classification accuracy, which is believed to be a result of it having a pooling layer that reduces the parameters required for activation. Nevertheless, the capability of the LSTM in the field of music still cannot be dismissed due to the limited data used in the analysis. In the future, studies on other aspects of the two models such as classification based on mood instead of genre can be conducted. Overall, these results offer a guideline for those looking for a comparison between the CNN and the LSTM model's structure, functionality, and ability for music genres classification.

References

- [1] Kirk A, Miranda E R. Guide to Computing for Expressive Music Performance. Springer Publishing Company, Incorporated, 2012.
- [2] Bahuleyan Hareesh. Music genre classification using machine learning techniques. University of Waterloo, n.d.
- [3] Fradkov Alexander. Early history of machine learning. Institute for Problems in Mechanical Engineering, 2021.
- [4] Gessel Gabriel, Akesson Simon. A comparative analysis of CNN and LSTM for music genre classification. KTH Royal Institute of Technology, 2019.
- [5] Hewahi Nabil, AlSaigal Salman, and AlJanahi Sulaiman. Generation of music pieces using machine learning:
- [6] M Athulya, S Sindhu. Deep learning based music genre classification using spectrograms. NSS College of Engineering, 2021.
- [7] O'Shea Keiron, Nash Ryan. An introduction to convolutional neural networks. Aberystwyth University & Lancaster University, n.d.
- [8] Tibadiya Prem. Music genre classification using RNN-LSTM. 2020.
- [9] Zhang Jingwen. Music feature extraction and classification algorithm based on deep learning. Xi'an Peihua University, 2021.
- [10] long short-term memory neural networks approach. University of Bahrain, 2019.
- [11] long-short term memory neural network approach. University of Bahrain, 2019.