

Music Similarity Estimation based on Feature Extraction and Deep Learning

Yiyi Niu *

SWJTU-Leeds Joint School, Southwest Jiaotong University, Chengdu, China

* Corresponding author email: sc20yn@leeds.ac.uk

Abstract. The field of music information retrieval (MIR) has developed rapidly in recent years, which provides a lot of convenience for people's entertainment life after combining with big data and deep learning. This paper studies music similarity estimation in the field of music information retrieval. To be specific, based on the deep learning of music feature data, the music similarity model (a classifier) is obtained to judge whether the music similarity reaches more than 50% (i.e., random model). According to the experiment, using the Siamese neural network to train the model with chroma features as input can achieve high accuracy. It will make it more convenient for music recommendation, humming recognition and plagiarism identification. In the future, more researches ought to focus on the real value instead of classifier to judge the similarity or at least the clustering model (multiple thresholds instead of two). Anyway, these results shed light on guiding further exploration focusing on similarity extraction of music in the context of the state-of-art deep learning scenarios.

Keywords: Music Similarity; Deep Learning; Euclidean Distance; The Siamese Neural Network.

1. Introduction

With the continuous development of big data and the continuous improvement of requirements for entertainment life, the field of multimedia information retrieval has developed rapidly. The field of music information retrieval (MIR), to which music similarity analysis belongs, is part of the multimedia information retrieval field. Since the mid-1990s, the rapid development of the Internet, MP3, semiconductor technology has made traditional music media such as vinyl records, tapes, and CD-ROMs almost disappear, and the digital music is rife. This has promoted the emergence of music information retrieval (MIR) technology, and its connotation has long been extended from the initial narrow music retrieval to a large-scale scientific research field that uses computational means to analyze and understand the content of digital music, including dozens of research topics. The establishment of the 2000 International Society for Music Information Retrieval Conference (ISMIR) can be seen as the official creation of this field. The application of music similarity assessment is very extensive, including music recommendation, cover recognition, music plagiarism judgment, etc.

The similarity of music can be analyzed from many aspects, including style, rhythm, melody, emotion, instrument type, music background, etc. Musical similarity can be analyzed through the physical features of the signal, from the frequency domain angle, time domain angle, etc. Through the processed signal, the pitch, the note time value, the score and so on can be obtained, using this, the similarities between different pieces can be analyzed from a music theory perspective. This is currently more researched in the field of music similarity. In some studies, the Mel-frequency cepstral coefficients (MFCC) and chroma feature are extracted from audio, and the convolutional neural network is used for semantic segmentation. The main melody of music is extracted through the bidirectional long and short-term memory network and the attention mechanism [1]. In another study, a self-organizing neural network (named SARDNET) and a multi-layer perceptron are the two artificial neural networks that are used in system architecture. These networks convert a sequence of temporal intervals (a rhythm pattern that has been produced) as input into a preset group of archetypal rhythm patterns. The similarity of musical rhythms was estimated [2]. Seyerlehner et al. employed a collection of block-level features for three distinct tasks, including music similarity estimates, tag classification, and genre classification [3]. Li et al., obtained six decisive music features by

calculating Pearson correlation coefficient, and the feature vectors were constructed with these features and the cosine similarity was calculated to obtain the music similarity. It is also possible to also to classify music by analyzing users and their favorite songs, or by analyzing the lyrics content, or by analyzing the music background and the musician [4]. There's another study that assesses the similarity of music makes use of three key areas: text retrieval, co-occurrence, and user ratings. [5].

In order to develop a convenient method to determine whether different songs are similar enough for music software to recommend similar songs, this paper used chromaticity features of music to train Siamese neural network based on convolutional neural network (CNN) to judge whether the similarity of different music is higher than 50%. The rest part of the paper is organized as follows. The Sec. 2 will introduce the dataset this paper used, data processing, neural network used and experimental process. The Sec. 3 will introduce the neural network used in this paper in detail, show the model training results and applications. The Sec. 4 is the limitation of this experiment and the prospect of the future. Eventually, a brief summary will be given in Sec. 5.

2. Data & Method

2.1 Data

The dataset used in this paper is MillionSongDataset(MSD) subset [6]. It includes the basic information of songs (e.g., author, author tag, Song TAG, song duration), and also includes the feature data obtained after the signal processing (e.g., energy, chroma feature, Mel-frequency cepstral coefficients (MFCC) feature), but does not contain audio files. Therefore, this paper will directly use the feature data provided in the dataset to train the model. In addition, the additional dataset last.fm subset will also be used at the same time, which contains the similarity between the song and other songs and the tags of the song. It needs to be used in combination with the basic data set.

Due to the large amount of data contained in MSD, the files are all in H5 format. The MSD website provides relevant codes to read the H5 file information. By using these codes and modifying them, the chroma feature data of a total of 10,000 songs are stored in Excel files. The chroma feature is used because it represents the tonal information of the song. The similarity between the song and the other songs in the dataset last.fm Subset is also stored in Excel files. Because the base dataset used in this paper is MSD Subset, some of the songs in Last.fm that provide similarity cannot be found in MSD Subset, hence the number of contrast songs for each song is small.

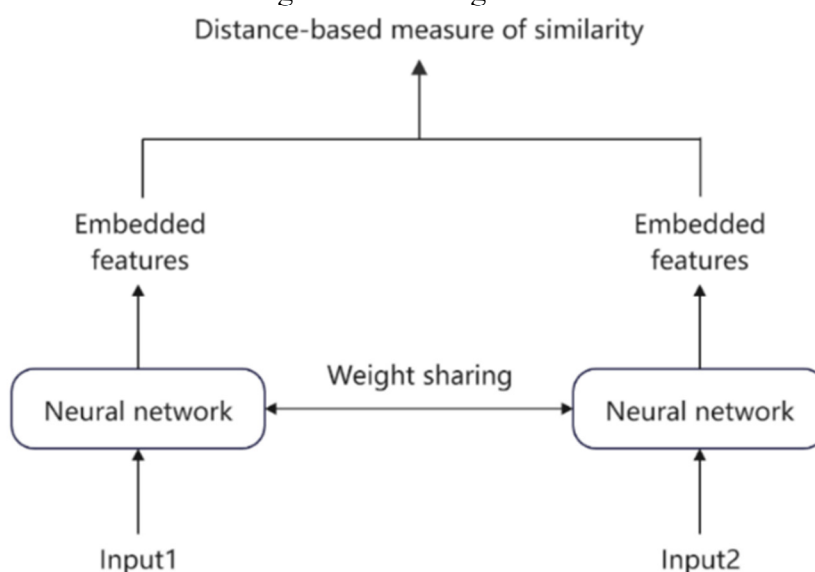


Fig 1. Siamese neural network.

2.2 Model

The model selected in this paper is the model trained by the Siamese neural network constructed based on convolutional neural network (CNN). This network is created by joining two identical sub-networks together at their outputs. The joining neuron keeps track of how far apart the two feature vectors are throughout training as the two sub-networks extract features from the two signatures. Verification is the process of comparing an extracted feature vector to one that has been saved. Unless they are closer to this recorded representation than a chosen threshold, all other signatures are discarded as forgeries [7]. The structure of the Siamese neural network is shown in Fig. 1.

80% of the data will be randomly chosen as the train set, and 20% will be randomly selected as the test set. If the similarity between two songs is greater than 0.5, the label will be 1, and if it is less than 0.5, the label will be 0.

$$D_w(\vec{x}_1, \vec{x}_2) = \|G_w(\vec{x}_1) - G_w(\vec{x}_2)\|_2 \quad (1)$$

Firstly, the Euclidean distance of the two vectors obtained by the model is calculated through this formula, which reflects the correlation of the two inputs. Then Contrastive Loss can be calculated through the formula:

$$L(\vec{x}_a, \vec{x}_b) = (1 - Y) \frac{1}{2} D_w^2 + Y \frac{1}{2} \{ \max(0, m - D_w) \}^2 \quad (2)$$

Here, Y is 0(not match) or 1(match), it indicates whether the two inputs match. m is the set threshold, when the distance between $(\vec{x}_1, \vec{x}_2)$ is less than m, Contrastive Loss will become 0. Since this experiment uses Keras to train the model, and the Binary Cross-Entropy loss function is provided by Keras, it is convenient to use. Instead of using that loss function, this paper used Binary Cross-Entropy. The formula is as follows:

$$loss = -\frac{1}{N} \sum_{i=1}^N y_i \log(p(y_i)) + (1 - y_i) \log(1 - p(y_i)) \quad (3)$$

where y_i is the label with a value of 0 or 1. $p(y_i)$ is the probability of getting y. The loss function is to evaluate the prediction results of the model. Therefore, when $y_i = 0$, $loss = -\log(1 - p(y_i))$. The closer the predicted value of y is to 0, the smaller the loss value is. The closer the predicted value of y is to 1, the greater the loss value is.

2.3 Procedure

First, one needs to download the dataset to computer. Then, the required data in the data set is extracted and saved in an Excel file. Afterwards, one needs to place the chroma features and corresponding labels of the two audio recordings into three arrays, respectively. The array is split, with 80% as the training set and 20% as the test set. The network nervous system was initialized, the model was trained with the processed data set, and loss and accuracy curves were obtained. Finally, the accuracy was obtained by testing the model with the test set.

3. Results & Discussion

3.1 Model Training

The estimation of music similarity by the Siamese neural network is realized by Python language and Keras package. Python has a distinct advantage over other languages in that it has a large number of third-party libraries for data processing and analysis [9]. After processing the data using packages such as Numpy and Pandas, initialize the model. The first layer is a one-dimensional CNN layer, which takes chroma features as parameters. The second layer is also a one-dimensional CNN layer, which can be used to obtain more accurate results. These two layers are CNN encoders, which encode the input chroma features. The two results obtained from the two inputs are subtracted in the subtraction layer to obtain the similarity, and then the output layer activated using sigmoid is created. The optimizer is Adam and the learning rate is set to 0.002. After everything is ready, the paired

chroma features and their labels are used to train the model, epoch is 30, and loss and accuracy curves are obtained. The accuracy of the model is tested using the prepared test set.

3.2 Performances

The loss curve and accuracy curve obtained after training the model with 4800 examples are shown in the Fig. 2 and Fig. 3. It can be seen that both loss and accuracy tend to be stable after the 15th iteration. The loss function used in this model is Binary Cross-Entropy loss. No matter whether the label is 0 or 1, the smaller the loss value, the better the model is. It can be seen from Fig. 2 that after the loss becomes stable (i.e., the value remains between 0 and 0.1), the model can be treated as is an ideal result. As illustrtaed in Fig. 3, accuracy keeps above 0.98 after it becomes stable, indicating a high accuracy of the model. After the model training is completed, the test data is used to test the accuracy of the model, and the accuracy of the model is 0.9026504755020142. However, such high accuracy is caused by too many data with label equals to 0 in the dataset, which cannot completely represent the accuracy of the model.

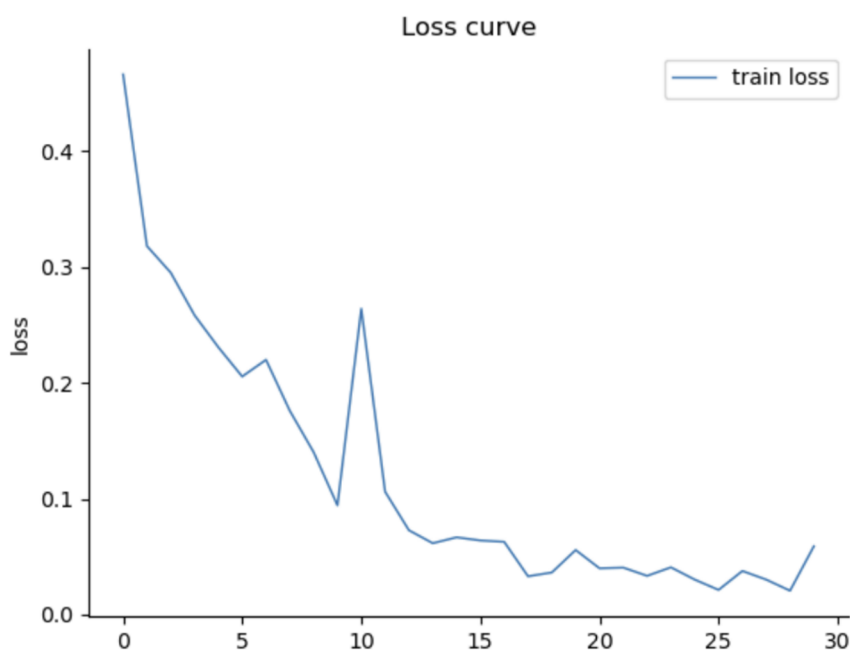


Fig 2. Loss curve as a function of epoch (the x label).

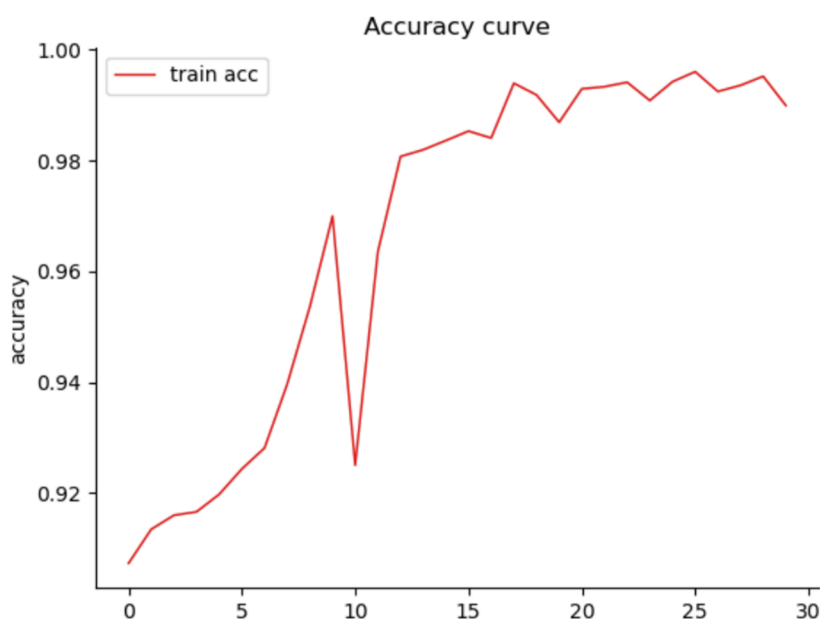


Fig 3. Accuracy as a function of epoch (the x label).

3.3 Applications

The technology and model studied in this paper can be applied to music software to recommend songs for users. After knowing the music that users like, the music with high similarity to the music can be calculated and recommended to the target users. After obtaining the chroma features of music, the experimental model can be used to estimate the similarity. When the label is 1, the two songs are relatively similar and suitable for recommendation. However, a more suitable recommendation systems ought to put the songs in order, i.e., a specific similarity value should be given instead of a Bool value.

4. Limitations & Prospect

Nevertheless, this study has shortcomings and drawbacks. Primarily, this paper only judges whether the similarity of two songs is higher than 50%, without giving a specific value, which is not accurate enough. The similarity data in last.fm dataset used in this paper is not accurate enough, and the similarity of two songs is not uniform in different files. This will adversely affect the training of the model. Moreover, since the selected base dataset is MillionSongDataset Subset, a large proportion of the songs in last.fm for which similarity was calculated could not be found. This leads to a sharp drop in the amount of data. The training speed and accuracy of the model are affected. There are many contents with label 0 in the dataset, which affects the judgment of accuracy and makes the accuracy of the model falsely high.

The field of music information retrieval(MIR) which Music similarity estimation belongs to, will provide great help for people's entertainment life. With the improvement of the accuracy of music similarity estimation, the recommendation function of music software will be more accurate, and it is easier for people to hear the songs they like. Meanwhile, the detection of music creation infringement will be more convenient, and the identification of infringement will become easier. At the same time, it is also of great help to the field of humming recognition. By comparing the similarity between humming songs and songs in the music library, the song with the highest similarity is selected as the result of humming recognition.

5. Conclusion

Through deep learning of music feature data, a model is trained to estimate whether the musical similarity is higher than 50%. In this paper, the Siamese neural network is adopted for audio data processing, but the accuracy still needs to be improved, and the dataset also needs to be expanded. This paper abandoned the traditional method of extracting music theme through deep learning and then performing similarity estimation, and chose to directly use music feature data to train the similarity model, which reduced the operation steps and the number of training models, and improved the efficiency. This model is not accurate enough for music plagiarism detection, but it has high application in similar music recommendation. In the future, datasets can be built with more complete, such as building data sets by genre. It can distinguish the music genre in a specific field from the general genre, and promote the diversified establishment of music genre data sets by constructing music genre data sets suitable for specific fields [10]. Overall, these results offer a guideline for estimate music similarity using the Siamese neural network and audio features.

References

- [1] Hao Liu. Research on music Similarity Detection Method Based on Deep Learning. North China University of Technology,2021.
- [2] Fouloulis Athanasios, et al. Estimating Similarity of Musical Rhythm Patterns through the use of a Neural Network Model. Researchgate, 2010.

- [3] Klaus Seyerlehner, Markus Schedl, Tim Pohle and Peter Knees. USING BLOCK-LEVEL FEATURES FOR GENRE CLASSIFICATION, TAG CLASSIFICATION AND MUSIC SIMILARITY ESTIMATION, 2019.
- [4] Yangqi Li, et al. Research on Cosine Similarity Algorithm in music Recognition. China Science and Technology Information, 2022, (12): 110-111.
- [5] Knees P. and Schedl M. A survey on music similarity and recommendation from music context data. ACM Trans. Multi-media Comput. Commun. Appl. 10, 1, Article 2 (December 2013), 21 pages.
- [6] Thierry Bertin-Mahieux, et al. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.
- [7] Bromley J., et al. Signature verification using a " siamese" time delay neural network. In Advances in neural information processing systems, 1994, pp. 737-744.
- [8] Hadsell R, Chopra S, Lecun Y. Dimensionality Reduction by Learning an Invariant Mapping. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). IEEE, 2006.
- [9] Shuaijing Li. Research and Application of Recommendation System Based on Deep Learning. Northern University for Nationalities, 2021.
- [10] Yang Song, Hailong Wang, Lin Liu et al. Review of music genre classification in the field of music information retrieval. Journal of Inner Mongolia Normal University (Natural Science Chinese Version), 2022, Vol. 51 (4):418-425.