Applications of deep neural networks on music emotion recognition

Yi Du

Boston University, Boston, MA, 02215,United States

duyi233@bu.edu

Abstract. Music Emotion Recognition (MER) is a subfield of Music Information Retrieval (MIR) that focuses on finding a relationship between music and human emotions by applying machine learning and signal processing techniques. In recent years, neural networks have achieved great success in a large number of areas, such as speech recognition and image processing, sparking many attempts to utilize neural networks in the MER task. Although new models for MER are constantly emerging, there are few systematic reviews in this field that involves the latest models and datasets. Therefore, in this paper, we provide a detailed review of this task. Our work first expounds on the practical significance and research status of the MER problem. Then, we encapsulate the background research and contributions of predecessors in both machine learning and psychology fields. Our work also includes a thorough analysis of several important datasets and their mathematical principles. Finally, we summarize four novel models from an application perspective and conclude a few potential challenges for the task.

Keywords: Music Emotion Recognition, Neural Networks, Nature Language Processing

1. Introduction

The processing of the acoustic signal, including speech recognition, voice detection, and music transcription, has long been a hot topic in machine learning. As a very useful yet especially challenging task, Music Emotion Recognition is gaining more and more attention in recent years. Music, being created by human artists, is a natural carrier of human emotions. The elements of music, such as rhythm, melody, pitch, timbre, or harmony, give us different feelings such as joy, anger, fear, or sadness. Accordingly, scientists believe that computers can recognize the emotion of music by analyzing it, especially with the help of machine learning algorithms [1]. This idea leads to the extensive research on the MER problem, a task of finding the relationship between music and human emotions by applying machine learning and signal processing techniques.

With the development of electronic devices, listening to music has become a way of life for modern people. Music software today is so powerful that people can find nearly whatever they want on a single small smartphone. However, researchers did not stop there. Software developers have already realized their users' demand for an intelligent recommendation system. People want their music APP to be familiar with their taste and automatically push songs according to their sentiments and emotions. Moreover, the rapid emergence of new music works also puts forward unprecedented requirements for search engines. People want to search for music through more abstract information, such as feelings and moods, instead of specific names and authors. Additionally, developers are also seeking to provide users with unique visual expressions for each music work by digging its inner emotions, so that users can enjoy a more immersive multisensory experience of music. As can be seen from all the above points, developing efficient computational music emotion recognition systems has become a prominent and demanded research topic with huge value both socially and commercially.

With its function of conveying emotions and arousing resonance, music is said to be a common language of all mankind. Yet recognizing the emotions of music remains difficult for many reasons. First, the amount of available data is relatively limited. Creating a complete music emotion dataset requires large amounts of time and manpower, because emotion labels usually have to be determined by multiple experienced experts through careful work. Some approaches of MER combine audio processing with lyrics, which requires datasets for songs of different languages to be created separately, thus further enhancing the insufficiency of available data. Moreover, the existence of different emotions in music works is very uneven, causing many datasets not ideal for training machine learning models. Due to the variety of data formats across datasets and the lack of unified metrics for result evaluation, it is difficult to compare results obtained by different researchers, which also leads to an absence of comprehensive reviews of multiple works in the field of MER. Furthermore, when we look deeper into the intrinsic character of this problem, we should realize that even we human beings find it difficult to describe our own emotions, let alone machines. This is because emotions are very complex, subjective feelings and there’s no precise measurement of a person’s emotion. Human emotions often exist in a mixed manner, like being sad at the same time feeling a little regret. In addition, studies in psychology and acoustics are not complete yet. There’s no concrete theory informing us which specific feature or set of features in audio is directly related to human emotion.

In the past decade, much research on MER has been done by academy and industry. Through people’s unremitting efforts, a batch of new techniques has emerged to address specific problems. However, we seldom see high-quality reviews that systematically conclude these new techniques and other potential problems in the MER field. Therefore, our work means to fill this long-existing blank by involving the newly proposed methods as well as recently constructed publicly available datasets for this task. Precisely, our work first introduces the task and its applications on fine-grained categories. Then, we demonstrate several benchmark datasets as well as the metrics for the task. In addition, we also illustrate several proposed methods and the details of the neural models. Eventually, we conclude this paper and raise several potential future research directions. Our paper can benefit not only related researchers but also researchers from other fields.

2. Background

2.1. Two representations of emotions

The psychological community has studied the formal representations of emotions for a long time. At present, there are generally two kinds of models for emotions, which can be mathematically described as discrete representation and continuous representation.

The discrete representation is also known as the categorical approach for representing emotions, because it divides human emotion into several main categories, such as anger, joy, fear, sadness, etc. An important study involving this approach is conducted by Hevner in 1936, in which he 66 adjectives were arranged into 8 groups for volunteers to choose. The discrete representation corresponds to the classification problem in machine learning. The advantage of this approach is that it is closer to our daily language and thus easier for users to understand. However, it is also blamed for being not precise enough and lacking consensus standards on the setting of categories.

The continuous approach usually contains two or three dimensions, each representing an independent aspect of emotion. The most famous model is the one proposed by Russell in 1980. This model describes emotions in a two-dimensional vector space. The horizontal axis measures arousal, which shows emotions from calm to excited, while the vertical axis is for valence that gives the measure of displeasure vs. pleasure. Due to the continuous nature of this representation of emotions, it is generally related to the regression problem in machine learning, usually by managing the two dimensions separately. Studies show that the discrete and continuous representations of emotions are actually interrelated and can be interchanged under certain conditions.

2.2. Datasets

2.2.1. MediaEval database for emotional analysis in music (DEAM). DEAM is a widely accepted dataset for MER tasks. It includes more than 1800 songs annotated with arousal and valence values, continuously (every second) and over the whole composition. Specifically, there are 1744 45-second-long excerpts randomly extracted from songs in 2013 and 2014, and 58 full songs from 2015. The primary source of this dataset is royalty-free music from freemusicarchive.org (FMA), jamendo.com, and the medleyDB dataset.

2.2.2. Synchronized lyrics emotion dataset. This dataset is established for synchronized studies of lyrics and audio. It extracts music excerpts in which certain lyrics lines are sung and stored the precise information of the starting and ending time of lyrics in each sample so that researchers can build a direct connection between text and audio. The data are collected from the Musixmatch platform. Each sample in the dataset is annotated with one of the five common human emotions (sadness, joy, fear, anger, disgust) in a discrete manner. However, a slight drawback is that sadness and joy are too dominant in the data, causing a relatively uneven distribution of tags.

2.2.3. Jacek grekow's dataset. Grekow et al. [2] contributes a new dataset to train and examine his new Recurrent Neural Networks (RNN) based approach of MER. Music genres and emotions are well-distributed in this dataset. The annotation was determined by five experts with university music education and a lot of relevant experiences to ensure the quality. The two-dimensional arousal-valence model is used to measure emotions in music [3]. There are altogether 324 6-second music fragments in this dataset. The data are collected from the publicly available GTZAN1 data collection.

2.2.4. Turkish emotional music database (TEM). TEM [4] is a new database dedicated to Turkish traditional music for MER tasks. It contains 124 music excerpts with a duration of 30 seconds, annotated by 21 university students proficient in Turkish music using Russell's two-dimensional arousal-valence representation.

2.3. Metrics

Numbers of metrics are widely utilized in the MER field to evaluate the model's output. We will briefly introduce some of the most typical ones as a reference for future researchers. Multi-label cross entropy loss is applied to evaluate the performance of both lyrics and audio based models:

$$L(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{C} y_i^n \log x_i^n \qquad (1)$$

where N is the number of examples, C is the number of classes, y^n is the one-hot vector representing the true label and x^n is the model prediction vector using the softmax function.

In researches involving continuous representations and regression algorithms, the coefficient of determination (R^2) and mean absolute error (MAE) are usually used to assess model efficiency:

$$R^2 = 1 - \frac{RSS}{TSS} \qquad (2)$$

where RSS is for the sum of squares of residuals, and TSS is for the total sum of squares;

$$MAE = \frac{\sum_{i=1}^{N}|y_i - x_i|}{N} \qquad (3)$$

where y_i is the predicted value, x_i is the true value, and n is the total number of data points.
In some other researches on the 2D vector space, Pearson's correlation coefcient (PCC) and root mean square error (RMSE) are also used for evaluation:

$$PCC = \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}} \tag{4}$$

where x_i is the values of the x-variable in a sample, $\overline{\mathrm{x}}$ is the mean of values of the x-variable, which is the same for y;

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{x_i}-\widehat{\mathrm{x_i}})^2}{\mathrm{N}}} \tag{5}$$

where x_i is the actual observations time series and $\widehat{\mathrm{x_i}}$ is the estimated time series.

Other regular metrics like accuracy, recall, precision and f-measure also play an important role in many papers. Quite a few papers implement 10-fold cross validation to maximize the use of limited data.

3. Recent research

3.1. Research 1

This paper [5] analyses the emotion of music from two aspects: lyrics and audio. The researchers established a new dataset, in which lyrics and audio are put together and studied respectively in a synchronized manner. Instead of randomly selecting 30 seconds from a song, this dataset carefully ignores the intro and outro portion, leaving only the part with vocals, and holds the lyrics complete while keeping the lengths of selected pieces roughly the same, in order to control the consistency of dimension for audio features. Emotions are represented using Hevner's method of discrete categories, with each sample in the dataset labeled with one of five common human emotions.

Then, models are designed respectively for lyrics and audio. In the lyrics part, there are basically two steps: (i) making use of pre-trained word embeddings to transform the text into vectors; (ii) feeding the word representations to Deep Neural Network models. Three different pre-trained word embeddings, fastText, ELMo and BERT, are leveraged. While for acoustic part, the researchers make use of Mel-Spectrogram to model it and leverage Convolutional Neural Networks (CNN) to locally learn its emotion-specific features.

For lyrics, seven experiments on various combinations of word embedding techniques and classification models were conducted. Over all cases, the non-contextual embedding fastText with classifier LSTM outperformed other approaches. Two experiments on audio were conducted. In the first experiment, the original audio was directly converted into Mel-Spectrogram and fed to the CNN; while in the second one, researchers first extracted human vocals from the audio samples, using Wave-U-Net, and then fed the CNN only with information of the vocals. Though the results of prediction were relatively poor compared to lyrics part, the finding that using vocals independently could improve the performance of classification is quite meaningful.

3.2. Research 2

Music Emotion Recognition tasks are traditionally divided into two categories: static and dynamic [6]. The former refers to analyzing longer music clips, usually 15 to 60s, while the latter usually contains only 0.5 to 1s and reflects the changes in emotions over time. In this paper, researchers creatively studied static MER using samples of shorter lengths (6s), which allows them to detect changes in emotions throughout the whole composition, like in the case of dynamic MER. Due to the sequential property of the task, this paper proposed a model based on RNN and described the work of preparing data for RNN in detail. Several experiments were conducted, and a method of using pretrained model to improve the performance was presented. Continuous figures of two dimensions (arousal and valence) are used to depict emotions in this paper.

Before being inputted to RNN, each music file must be converted into a series of numbers readable by the network, i.e., feature vectors. In this paper, two tools are used: Marsyas, mainly for extracting Mel-Frequency Cepstral Coefficients (MFCC) and chroma features of music, and Essentia, which provides more higher-level features like harmony. The researchers cut each sample (6s) into small, possibly overlapping pieces, making them sequential data that RNN can process.

In the experiments, the authors observe that the results of MFCC are far better than that of chroma and are very close to that achieved from all features. Simple linear regression and SMOreg were outperformed by RNN models both when using MFCC and using all features. Using higher-level features leads to better results than the MFCC and the chroma.

To further optimize the results, researchers use pretrained model to select input features. The pretrained model is connected in front of the whole structure. Experiments showed that it improves the results of both arousal and valence detection. After all, the prediction of arousal has higher accuracy than that of valence in all models. For arousal, quite good results can be gotten out of only few features (MFCC); while for valence, adding higher-level features did make a difference.

3.3. Research 3

Though actively studied for years, Music Emotion Recognition remains a challenging task due to the subjective and personal nature of music-induced emotions and the fact that there isn't any apparent low-level audio feature that is directly related to emotion, which is not the case in other audio processing tasks [7]. As a result, most previous studies in MER either indiscriminately fed large amounts of features to the classifier or went through a hand-engineered pre-processing procedure to select features, ignoring the lack of solid theory in the auditory field about which features are emotion-related. Subsequently, their results turn out to be not quite satisfiable, especially in the prediction of valence component of emotion.

In this paper, researchers broke through this limitation by proposing an end-to-end approach that uses CNN to directly handle raw audio data and extract features from it for later use. Specifically, this algorithm can be described as two parts. In the front-end, a simple CNN layer is connected with a time-distributive fully connected layer (FC), serving as the feature extractor. In the back-end part, researchers used Bi-directional Gated Recurrent Unit (BiGRU) and a maxout fully connected layer consisting of two output units that outputs the final results.

Another significant contribution of this paper is the design of Iterative Reconstruction, which takes the place of FC layer that connects CNN and BiGRU layers. Because of its ability to enhance the most dominant features and suppress the weak ones, IR gradually makes data converge to a stable stage, i.e., discrete values {-1, 0, 1}, and thus eliminate the noisy features and improves the overall accuracy of regression.

In this paper, the authors find that the proposed systems not only gave better or very close results in the arousal dimension, but also outperformed other solutions in the valence dimension, especially in terms of PCC. The use of IR leads to improvement of results in all aspects.

3.4. Research 4

Databases of various languages for Music Emotion Recognition appeared rapidly in recent years, with Turkish music being less studied in comparison with others [8]. To fill this gap, researchers of this paper established a Turkish music database consisting of 124 music segments of 30 seconds long, selecting the most representative part of each song, labeled by 21 university students proficient in Turkish music using the dimensional (valence and arousal) representation. Based on this dataset, several experiments were conducted on an MER system proposed in this paper, giving a gratifying result of 99.19% accuracy.

Specifically, a complete MER system usually involves two parts: feature extraction and prediction. For the feature extraction part, one can either utilize the various publicly available tools, or apply deep learning approaches to automatically learn helpful features from audio data (either raw audio signal, spectrogram or gammatone filterbanks). In this paper, both paths were implemented. Researchers chose three standard audio feature extractors, MIRtoolbox, OpenSMILE, and jAudio, extracting 7368 features altogether [9]. Correspondingly, a one-dimensional CNN, consisting of one or more convolutional layers, pooling layers, and flatten layers, was also used for feature extraction. To feed CNN with a sufficient amount of data, 124 music samples in the database were each divided into 3 pieces of equal length. After that, MFCCs and log-mel filterbank energies were calculated out of these data and inputted into the CNN, which transformed them into 256-dimension feature vectors that were then joint together into vectors of length 768 to prevent overfitting in later stage. To sum up, now we have 768 CNN-based MFCC features, 768 CNN-based log-mel filterbank energy features, and 7368 standard audio features for each 30-second music. Then, Correlation-based Feature Selection (CFS), which is a general method in statistics, is applied to diminish feature size and select the most salient features. For prediction, a convolutional long short term memory deep neural network (CLDNN) architecture is designed to handle the features extracted by the above methods.

Experiments showed that this LSTM & DNN approach performs better than other compared approaches both on standard feature set and combined feature set. Additionally, compared to CNN based feature sets and standard feature set, combined feature set gives the best result when same classification method is used. Experiments also found that the results could be remarkably improved when CFS is applied.

4. Long-short term memory neural network

Long Short Term Memory (LSTM) neural network is a variant of RNN [10], capable of learning long-term dependencies inner sequential data. The LSTM is explicitly leveraged to avoid the long-term dependency problem.

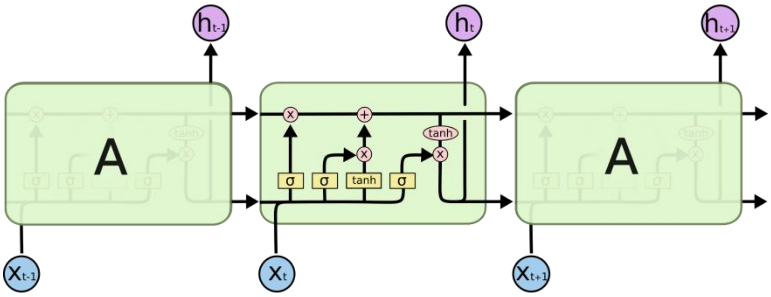

Figure 1. The architecture of the LSTM.

The first step in the LSTM is to decide what information should be absorbed from the cell state. This decision is made by the “forget gate”, which is calculated by:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{6}$$

After that, the LSTM decide what new information should be stored in the cell state. This contains two parts: a sigmoid layer and a tanh layer. The process is as follows:

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \end{aligned} \tag{7}$$

Then, the LSTM multiplies the old state, forgetting the things that are decided to forget earlier. Then we add $i_t * \tilde{C}_t$. This is the new candidate values, scaled by how much to update each state value:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{8}$$

Eventually, the LSTM need to decide what should be output. This output will be based on the cell state, but will be a filtered version:

$$\begin{aligned} o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(C_t) \end{aligned} \tag{9}$$

5. Conclusion

In this paper, we first introduced the MER problem by discussing its application and research situation, especially some existing challenges. Then, we provided a detailed review of the research background of MER problem focusing on two models for emotion representation, four typical current datasets, and several metrics for result evaluation. Moreover, we summarized four leading novel works in the MER field. In the first work, researchers proposed a synchronized method by analyzing both audio data and lyrics data; in the second work, an MER system based on RNN was constructed, enhanced by pre-trained models for feature extraction; the third work proposed an end-to-end MER approach that does not require any form of data pre-processing and designed an Iterative Reconstruction structure to optimize the performance of neural networks; the forth work contributed a Turkish music dataset and conducted several experiments using an LSTM-based MER system. Some of the works obtained very ideal overall accuracy, while some others still have space for future improvement. We believe all four works we presented are of great value to the MER community in terms of innovation and practicality.

For future works, we can see that various targeted datasets are still to be developed, including more languages and music genres and more high-quality samples. Also, we are confident that innovation in neural network architectures will bring new progress to the MER problem, as seen in the presented works. Moreover, new research achievements on human emotions in the psychology community may bring unexpected breakthroughs to the MER task. We also look forward to seeing more systematic reviews that conclude the newest progress in this field.

References

[1] Loreto Parisi, Simone Francia, Silvio Olivastri, Maria Stella Tavella. Exploiting Synchronized Lyrics And Vocal Features For Music Emotion Detection. arXiv:1901.04831, 2019.

[2] Jacek Grekow. Music emotion recognition using recurrent neural networks and pretrained models. Journal of Intelligent Information Systems (2021) 57:531–546.

[3] Richard Orjesek, Roman Jarina, Michal Chmulik. End to end music emotion variation detection using iteratively reconstructed deep features. Multimedia Tools and Applications (2022) 81:5017–5031.

[4] Serhat Hizlisoy, Serdar Yildirim, Zekeriya Tufekci. Music emotion recognition using convolutional long short term memory deep neural networks. Engineering Science and Technology, an International Journal 24 (2021) 760–767.

[5] DEAM dataset - The MediaEval database for emotional analysis of music. http://cvml.unige.ch/datab ases/DEAM. Accessed 30 Sept 2020

[6] Fika Hastarita Rachman, Riyanarto Sarno, Chastine Fatichah. Music Emotion Classification based on Lyrics-Audio using Corpus based Emotion. International Journal of Electrical and Computer Engineering (IJECE), Vol. 8, No. 3, June 2018, pp. 1720~1730.

[7] Russell, J. A. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161–1178, 1980.

[8] K. Hevner. Experimental studies of the elements of expression in music. The American Journal of Psychology, pages 246–268, 1936.

[9] Gers, F. A., Schmidhuber, J., & Cummins, F.A. Learning to forget: Continual prediction with LSTM. Neural Computation, 12, 2451–2471, 2000.

[10] Witten, I. H., Frank, E., Hall, M. A., & Pal, C.J. Data mining fourth edition: Practical machine learning tools and techniques, 4th edn. USA: Morgan Kaufmann Publishers Inc, 2016.