Music Emotion Recognition Using Deep Convolutional Neural Networks

Xiao Quan, Siqi Dong

December 16, 2020

1. Abstract

Music is an important vehicle in communicating emotions. The ability to automatically recognize emotional content of recorded music is an important task in the Music Information Retrieval research community as well as commercial platforms. Music Emotion Recognition (MER) is crucial for effective music organization, user query, playlist generation and more. In our work, we implemented a Deep Convolutional Neural Network to predict music's emotional values on a valence-arousal scale based on mel-spectrograms of input audio files. The results are evaluated based on the R^2 scores of the predicted output. We hope that this work will help future researchers establish a simple working model for MER-related tasks.

2. Introduction

One of music's unique attributes is its ability to directly affect one's emotions. This can be observed in not only traditional recorded music, but also in other media forms such as films or visual installations. In the age of big data, where we have the opportunity to store millions of songs in one catalog, the ability to label music at different levels is crucial for any efficient implementations of applications. Of those musical features, emotion recognition is perhaps the most directly relevant as it often is the starting point when a user searches for what to play. It also has direct implications for tasks such as genre, artist identification, and music recommendation. There are many challenges to the task of music emotion recognition, the most obvious being the task of creating ground-truth emotion labels. In the following section, we offer a brief overview of challenges associated with MER and related works that offers a workaround.

3. Related Work

In the current literature, in order to effectively label the emotional quality of a given music file, two predominant methods are proposed: one is using multiple tags to label the input, the other is to use Russell's two-dimensional valence-arousal space (Yang et al. 2018; Tan et al. 2019). Over recent years, the latter is becoming more frequently used when training machine learning models using larger datasets (Chen et al. 2015). In many deep-learning approaches, tag-like labels are being embedded and converted to valence-arousal scores for effective computation (Delbouys et. al. 2018).

Another challenge for this task is the relative scarcity of labeled datasets. Expert opinions often do not exist in the form of valence-arousal scores, while individually labeled sets suffer from subjectivity. In Soleymani et al.'s work, a continuous valence-arousal scores were recorded and were used to generate the "1000 Songs Database"(Soleymani et al., 2014). Lastly, from an implementation point of view, it has yet to be tested what combinations of input features and machine learning frameworks work best for this given task. For this project, we used a combination of "Deezer Dataset"(18644 tracks) (Delbouys et. al. 2018), which is an annotated subset of the "Million Song Dataset", and Soleymani et al.'s "1000 Songs Database.^{a}s our dataset. The proposed methodology will be discussed in the following section.

4. Method

We first obtained our mp3 files through the Deezer API, and downloaded the "1000 Song Database" at its host website. As a result, we have a total of 14327 annotated 30-second preview mp3 files from Deezer, and 744 annotated 45-second mp3 files from the ”1000 Songs Dataset”. Because the latter dataset had each track labeled on a continuous valence-arousal scale at 2Hz, we averaged the values of each track to obtain a single pair of valence-arousal labels per track. We then converted mp3 files to mel-spectrograms and divided the dataset to a 60-20-20 ratio training, validation and test sets. The mel-spectrograms are generated using librosa's melspectrogram algorithm, with a nfft size of 1024 and no overlap. The output dimensions of each mel-spectrogram is 1292 x 40. This is the input of our ConvNet. For this iteration of the project, we did not perform any data augmentation. The total number of input tracks we have is 15061.

CNN Architecture:

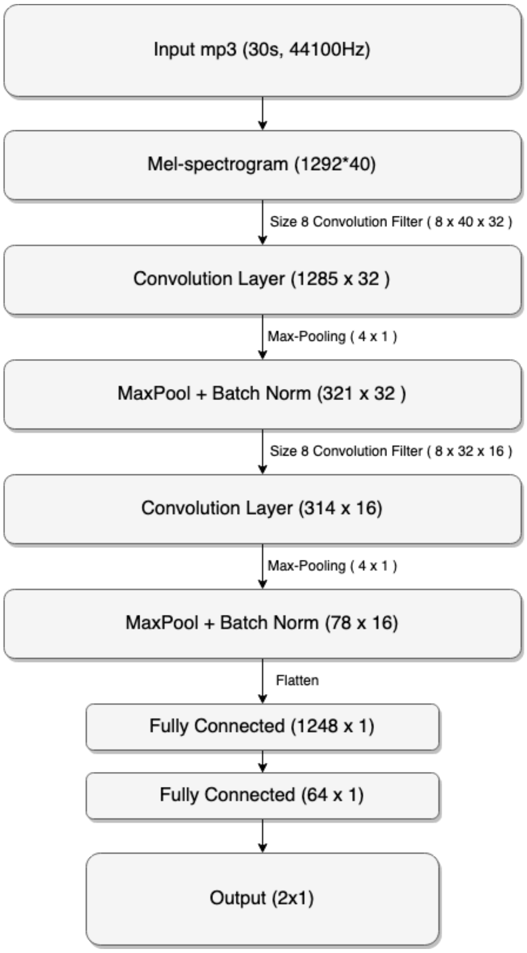
Input mp3 (30s, 44100Hz)
Mel-spectrogram (1292*40)
Size 8 Convolution Filter (8 x 40 x 32)
Convolution Layer (1285 x 32)
Max-Pooling (4 x 1)
MaxPool + Batch Norm (321 x 32)
Size 8 Convolution Filter (8 x 32 x 16)
Convolution Layer (314 x 16)
Max-Pooling (4 x 1)
MaxPool + Batch Norm (78 x 16)
Flatten
Fully Connected (1248 x 1)
Fully Connected (64 x 1)
Output (2x1)

Fig. 1 Architecture of CNN Model

The input mel-spectrogram is first convoluted with 32 size 8 filters with a depth of 40 and a stride of 1, then a size 4 max-pooling layer and stride of 4. The second layer consist of another size 8, stride 1 convolution using 16 filters, and a max-pooling layer of size 4, stride 4. Finally, we have two fully connected layers with a hidden layer of 64 units. The final output is a 2 x 1 vector corresponding to the valence-arousal scores[fig.1].

We also setup a similar architecture with deeper convolution layers to reduce trainable parameters. In this model, we have total of 6 pairs of convolution and max-pooling layers. The filter sizes of the convolution layers is now 4 and stride 1. The max-pooling layer is reduced to size 2. The number of filters for the 6 layers are 32, 32, 16, 16, 8, 8, respectively, before reaching the fully connected layer, which 64 hidden units.

5. Evaluation

We evaluate our model based on the R^2 scores of the predicted labels against reference labels. We ran 100 epochs using an Adam optimizer with a learning rate of 0.001. The results are as follows:

$R^2 Scores$		
Model	Valence	Arousal
Model1	-1.12	0.76
Model2	0.981	0.173
Deezer	0.179	0.235

Table 1. Valence-Arousal R^2 Scores

During the process of training, we experienced very fluctuating results. But the general trend is that the models seems to better predicting arousal values, compared to valence values. This is to be expected as audio data is more adept at predicting the energy of a piece of music, while the lyrics tend to help with valence scores (Delbouys et. al. 2018). The loss for the validation set for both models we used tend to stop improving after 0.11 after 30 epochs. No meaningful results was gained after around 30 epochs as the validation loss does not go down. Using the dataset we created, we were able to achieve the best result using (). This shows some lack in our project, noticeably the models we used is overfitting the data we have. In the original paper by Delbouys et al., there was a significant data-augmentation step that increased the train data 17-fold. We do not have this step. The result is that the valence predictions made by our model often does not show correlation with the test labels.

6. Future Work

There's a lot of work that can be done further in this project. First, data-augmentation techniques should be researched and be applied to the training dataset. Second, the audio mp3 that we used to extract MFCCs are 30s previews from the Deezer API. In the original paper, the files were full length songs, and 7 30s segments were extracted randomly from each song. More careful examination of the data is needed to ensure clearer understanding of the results. Third, the pre-processing of data can be explored further. Currently, we have only converted audio signal to mel-spectrograms using 1024, non-overlapping hanning windows. A smaller or larger windows could prove more suited for the task of Music Emotion Recognition. In conclusion, a more systematic exploration of the different links of data flow need to be designed and experimented from data-preprocessing, augmentation, normalization, model architecture, activation functions and so forth. This would require a significant amount of time that's beyond the scope of our project.

7. References

Brotzer, J. M., Mosqueda, E. R., Gorro, K. (2019). Predicting emotion in music through audio pattern analysis. MSE, 482(1), 012021.

Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., Moussallam, M. (2018). Music mood detection based on audio and lyrics with deep neural net. arXiv preprint arXiv:1809.07276.

Grekow, J. (2017, July). Audio features dedicated to the detection of arousal and valence in music recordings. In 2017 IEEE international conference on innovations in intelligent systems and applications (INISTA) (pp. 40-44). IEEE.

Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C. Y., Yang, Y. H. (2013, October). 1000 songs for emotional analysis of music. In Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia (pp. 1-6).

Tan, K. R., Villarino, M. L., Maderazo, C. (2019). Automatic music mood recognition using Russell's twodimensional valence-arousal space from audio and lyrical data as classified using SVM and Naïve Bayes. MSE, 482(1), 012019.

Yang, Y. H., Chen, H. H. (2012). Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology (TIST), 3(3), 1-30.