Music Emotion Recognition using audio features and deep neural network.

Sylvie Siqi Dong, Xiao Quan

Context and problem:

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 current literature, two predominant methods are proposed: one is using multiple tags to label an input, the other is to use Russell's two-dimensional valence-arousal space (Yang et al. 2018). Over recent years, the latter is more frequently used when training machine learning models using larger datasets (Delbouys et. al. 2018; Chen et al. 2015). Another challenge for this task is the relative scarcity of labeled datasets.

Lastly, it has yet to be tested what combinations of input features and machine learning frameworks work best for this given task. For our final project, we propose using the "Deezer Dataset" (Delbouys et. al. 2018), which is an annotated subset of the "Million Song Dataset", and others (to be researched), as our dataset. The proposed methodology will be discussed below.

Proposed Methodology / Plan:

Usually, there are three steps for music emotion recognition (Yang et al. 2018):

1. Choose the formats of music record and emotion models.
2. Extract music features and ground-truth data.
3. Train the (machine learning) model.

Based on empirical studies (Yang et al. 2018), there are two types of emotion models:

1. Categorical models (such as the typical Hever model) and the Updated Hever Model (UHM), which operates according to the definition of adjectives.
2. Dimensional models (such as Russell's valence-arousal model, Thayer model, and the famous PAD model), which operate according to the definition of dimensions.

Our methodology for the task is currently as follows:

- Curate a set of valence-arousal datasets as well as the associated MFCCs.
- Divide the dataset into appropriate train/test/validate sets.
- Design a deep neural network for training and testing the datasets. Evaluate results.
- Analyze and repeat previous steps while adjusting **input features** and/or **network models,** (i.e. KNN, W-D-KNN, CNN, SVM), and compare performance.

More details need to be fleshed out pending research, but this will be our starting point. We will be dividing the work equally, holding joint research, coding, and writing sessions when working on the project.

References:

Chen, Y. A., Yang, Y. H., Wang, J. C., & Chen, H. (2015, April). The AMG1608 dataset for music emotion recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 693-697). IEEE.
(Deprecated)

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*.

Yang, X., Dong, Y., & Li, J. (2018). Review of data features-based music emotion recognition methods. *Multimedia systems*, *24*(4), 365-389.

Potential Datasets to look into:

https://github.com/ismir/mir-datasets/blob/master/mir-datasets.yaml

https://github.com/deezer/deezer_mood_detection_dataset