COMP4801 Final Year Project

Group: FYP23018

Music Recognition with fragmented input through the use of Machine Learning

Detailed Project Plan

Supervisor

Prof. Xu Dong

Student

Name: Cheng Ho Kin (BEng CS)

UID: 303577961

Email: u3577961@connect.hku.hk

Table of Contents

1. Background

From time to time, people may hear music anywhere like retail shops and restaurants. While they are exposed to the music, they may only remember parts of the rhythm, pitch, or lyrics for that music. As a result, it is hard for those people to find the name of it or know all the details of it. It could be annoying to them if they want to replay the music or use the music in other occasions.

Due to the advancement of machine learning in the past decades, audio signal processing has been possible [1]. This enables many practical applications including speech recognition, environmental sound detection and music information retrieval. Nowadays, many applications on the market have applied different machine learning techniques to do the music recognition. For example, Shazam identifies music by creating a digital fingerprint of the audio to match against its database [2] [3], and Google applies deep neural network that trained with pairs of input (e.g., sung audio with recorded audio) in the machine learning setup to help search the music [4] [5]. Although many of these applications are workable, only a few of them try to merge the traditional searching methods (e.g., search by lyrics, country) with the audio recognition methods mentioned above. Therefore, this project will explore these methods and suggest a solution to perform music recognition that take audio and traditional searching parameters (e.g., lyrics, country) as input. It is expected that the result of this project can help public and content creators to find their desired music and get the music information for different usages (e.g., detection for copyrighted music pieces).

1.1. Audio recognition using deep learning techniques

According to recent researches on machine learning related to audio and music recognition, there are many deep learning techniques being commonly used. Some of these techniques are listed below.

- Convolutional neural network (CNN) is widely applied in image recognition and classification due to its architecture. It can also be used in audio recognition. For example, the one-dimensional audio samples can be converted into two-dimensional data like images. The two-dimensional data, containing matrix of Mel-Frequency Cepstral Coefficients (MFCCs) that are extracted from the audio samples and the number of windows used in the extraction, is then used to train the model [1]. In Luhach et al.'s research [6], the one-dimensional audio sample, converted in 39 x 39 two-dimensional array, is framed with window size of 25ms with an overlap of 15ms. Then the modified dataset was used to train the CNN model give three different convolution layer architectures. In this project, CNN will be used for recognizing the audio input.

- Recurrent neural networks (RNNs) are used as acoustic and linguistic models. Since RNN is inherently deep in time (because of its hidden state is a function of all previous hidden states), it is considered to be a possible alternative [1] [7].

- Long short-term memory (LSTM) is used to extend the RNN for alleviating the vanishing/exploding gradients problem during training. It makes use of a gating mechanism and memory cells to reduce the flow of information in the network [1] [7].

1.2. Motivation and aims of the project

The motivation and goal this project is to predict the music based on the information fragments of the music. The fragments include audio clip, lyrics, genres, and artist. Quite a number of applications for music recognition on the market rely on audio processing and distract from the traditional searching methods like searching with lyrics and genres. Sometimes people may not be able to find the targeted music because of the low-quality audio input (e.g., off-key, mispronounced words). It is possible that they can dig out the target music in the database after spending a lot of time searching through the music with the same genre. Thus, this project should explore on methods including audio processing with machine learning and traditional searching, and suggest a solution that merge these methods to predict the targeted music. The solution may improve the stability in music recognition and may be more convenient to the user.

2. Objective and Scope

The scope of this project is to construct the music predictor. It mainly focuses on the recognition and prediction of the music. It will not cover the generation of the entire music library nor the creation of music dataset for training the audio recognition model.

.

To construct the music predictor, there are 3 main objectives to follow and achieve. They are speech and melody recognition, music classification, and music prediction.

1. **Speech and Melody Recognition**
   1.1. To recognize music pieces according to the audio input
   1.2. To provide a list of candidates based on the percentage of similarity in recognition

2. **Music Classification**
   2.1. To classify the music pieces by criteria including genre and artist

3. **Music Prediction**
   3.1. To provide a list of predicted music pieces according to the user input Including audio, lyrics, genre, and artist.
   3.2. To explore the possibility of generating music piece based on the audio input (when no possible prediction provided)

3. Methodology

To construct the music predictor, the works are identified and have to be done in order. They are grouped into 3 phases. The first phase (Section 3.1) should focus on building model for recognizing speech and melody. The second phase (Section 3.2) should focus on classifying the music pieces by different criteria. In particular, a method should be provided to classify the unprocessed music pieces based on the corresponding genres. The third phase (Section 3.3) should focus on building a model for predicting the music piece given the information fragments of the music. It should make use of the model from the first phase, the classified music datasets, and the lyrics of the music pieces (if available). Testing, analysis and optimization should be performed for all three phases. Noted that this is the preliminary plan of the project. The exact methodology might be subject to change during testing.

3.1. Speech and Melody Recognition

To build a model for recognizing speech and melody, music dataset and deep learning method are required. Since the creation of dataset is hard and time consuming, the model will be trained on available datasets online. The Million Song Dataset (MSD) [8] consists of feature analysis and metadata for a million songs. Each track in the database contains information including the title, artist, release year, etc. Although the dataset does not include any audio, the sample audio can be fetched from external services like 7digital [9]. Regarding the deep learning method, CNN will be used. The researches in [3] [6] will be used as baselines to build the models. The optimal architecture will be selected.

3.2. Music Classification

The music pieces will be classified by different criteria, including genre and artist. The artist of the music pieces should be provided in the dataset or recorded manually. For classifying the genre, CNN will be applied, and the model introduced from [10] will be explored using MSD instead of GTZAN dataset [11]. The researches in [12] will also be referenced, which apply similar approach of transforming audio signals to two-dimensional data with MFCC feature extraction technique for training the music genre classification model using CNN. Noted that not all music pieces in MSD have their genres identified. The model will be used to classify the genres of those music pieces and the performance of the model will be analyzed. Support Vector Machine (SVM), which is a supervised machine learning model, will be considered if the performance of CNN is unsatisfactory, as some researches [10] [13] have analyzed and show that SVM can achieve an acceptable accuracy. For instance, Changsheng Xu et al. [14] used a multi-layer classifier based on SVM and achieved an accuracy of 93.14%. Kyaw and Renu [15] used multi-layer SVM and achieved an accuracy of 93%. This suggest SVM could be a possible alternative.

3.3. Music Prediction

The music predictor will be build based on the work done in Section 3.1 and 3.2. It will use the model for speech and melody recognition, the classified music datasets, and the lyrics of the music pieces to make the prediction. The predictor should first take the input from user. The model in Section 3.1 will process the audio input and provide the first list of candidates based on the percentage of similarity in recognition. The second list of candidates will be provided based on the genre and artist information. It will rely on searching in the classified music datasets. The third list of candidates will be provided based on the input lyrics. It will apply partial lyrics search on the lyrics database. Elasticsearch [16], which is capable of finding relevant documents even the lyrics do not exactly match the data in the lyrics database, will be considered to be the search engine. These three lists will be weighted to give a final list of prediction. The weighting method adopted will undergo testing and optimization.

4. Schedule and Milestones

Date	Milestone	Estimated number of learning hours
Sep 2023	**Research and Preparation of Detailed Project Plan** • Research on current applications for music recognition • Research on current approach in audio recognition and music genre classification • Research on other relevant information for writing the detailed project plan **Update on Project Webpage** • Detailed Project Plan upload • UI design	25 + 5
1 Oct 2023	**Deliverables of Phase 1** • Detailed project plan • Project web page	N/A
Oct-Dec 2023	**Speech and Melody Recognition** • Retrieving music dataset for recognition • Exploration on data type conversion • Exploration on architectures using CNN • Model building • Analysis and optimization **Music genre classification** • Exploration on data type conversion • Exploration on architectures using CNN • Model building	70 + 35
Nov-Jan 2024	**Preparation of First Presentation and Interim Report** • Interim report writeup • Preparation of ppt and relevant aids	25 + 15

8-12 Jan 2024	**First presentation**	N/A
21 Jan 2024	**Deliverables of Phase 2** • Preliminary implementation • Detailed interim report	N/A
Jan-Mar 2024	**Music genre classification (Cont.)** • Model building • Analysis and optimization **Music Prediction** • Integration of the models and search engine • explore the possibility of generating music piece based on the audio input • Testing, analysis and optimization • Application polishing	25 + 70
Mar-Apr 2024	**Preparation of Final Presentation and Final Report** • Final report writeup • Preparation of ppt and relevant aids	35 + 20
15-19 Apr 2024	**Final presentation**	N/A
23 Apr 2024	**Deliverables of Phase 3** • Finalized tested implementation • Final report	N/A
26 Apr 2024	**Project exhibition**	N/A

5. Reference

[1] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. -Y. Chang and T. Sainath, "Deep Learning for Audio Signal Processing," in IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 206-219, May 2019. doi: 10.1109/JSTSP.2019.2908700

[2] “Company - Shazam”. Shazam. Accessed Sep. 22, 2023. [Online] Available: https://www.shazam.com/company

[3] J. Jovanovic. “How does Shazam work? Music Recognition Algorithms, Fingerprinting, and Processing”. Toptal Engineering Blog. Accessed Sep. 22, 2023. [Online] Available: https://www.toptal.com/algorithms/shazam-it-music-processing-fingerprinting-and-recognition

[4] C. Frank. “The Machine Learning Behind Hum to Search”. Google Research Blog. Accessed Sep. 22, 2023. [Online] Available: https://blog.research.google/2020/11/the-machine-learning-behind-hum-to.html

[5] J. Lyon. “Google’s Next Generation Music Recognition”. Google Research Blog. Accessed Sep. 22, 2023. [Online] Available: https://blog.research.google/2018/09/googles-next-generation-music.html?m=1

[6] Luhach, A. K., Kosa, J. A., Poonia, R. C., Gao, X.-Z., & Singh, D. “Experimental Evaluation of CNN Architecture for Speech Recognition”. First International Conference on Sustainable Technologies for Computational Intelligence, vol. 1045, pp. 507–514, 2019. Springer Singapore Pte. Limited. doi: 10.1007/978-981-15-0029-9_40

[7] A. Graves, A. -r. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks". 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, pp. 6645-6649, 2013. doi: 10.1109/ICASSP.2013.6638947

[8] T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman, and P. Lamere. “The Million Song Dataset”. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011. Accessed Sep. 23, 2023. [Online] Available: http://millionsongdataset.com/

[9] “7digital for Music Services”. 7digital. Accessed Sep. 23, 2023. [Online] Available: https://www.7digital.com/music-services/

[10] Gotlur, K., Kulkarni, T., Royyapally, T., Das, B., & Vadlakonda, N. “Music genre classification using machine learning”. AIP Conference Proceedings, 2754(1), 2023. doi: 10.1063/5.0163540

[11] Andrada. “GTZAN Dataset - Music Genre Classification”. Accessed Sep. 24, 2023. [Online] Available: https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification

[12] Sharma, D. K., Peng, S.-L., Sharma, R., & Zaitsev, D. A. “Music Genre Classification Using CNN and RNN-LSTM. In Micro-Electronics and Telecommunication Engineering”, vol. 373, pp. 729–745, 2022. Springer Singapore Pte. Limited. doi: 10.1007/978-981-16-8721-1_67

[13] Smys, S., Bestak, R., Palanisamy, R., & Kotuliak, I. “Automatic Classification of Music Genre Using SVM”. Computer Networks and Inventive Communication Technologies, vol. 75, pp. 439–449, 2022. Springer Singapore Pte. Limited. doi: https://doi.org/10.1007/978-981-16-3728-5_33

[14] Changsheng Xu, N. C. Maddage, Xi Shao, Fang Cao and Qi Tian, "Musical genre classification using support vector machines". 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., pp. V-429, 2003. doi: 10.1109/ICASSP.2003.1199998.

[15] Kyaw, L. Y., Renu. “USING SUPPORT VECTOR MACHINE FOR MUSIC GENRE CLASSIFICATION”. Doctoral dissertation, MERAL Portal, 2009. Accessed Sep. 26, 2023. [Online] Available: https://meral.edu.mm/records/4026

[16] “SEARCH APIS & DEVELOPER TOOLS”. Elastic. Accessed Sep. 26, 2023. [Online] Available: https://www.elastic.co/enterprise-search/search-applications