Music Recommendation System based on Fusion Deep Learning Models

Yuchen Dong†*, Xiaotong Guo†† and Yuchen Gu††

1 International School, Beijing University of Posts and Telecommunications, Beijing, 100876, China

*Corresponding author’s e-mail: dongyuchen@bupt.edu.cn
††These authors contributed equally.

Abstract. With the growth of user popularity of online music platforms, the amount of music data and the demands on accurate recommendation system for music also increases. Therefore, an accurate and efficient recommendation system is urgently needed. In this paper, a music recommendation system based on fusion deep learning models is proposed. Both of the audio information and lyrics data are originally exploited. A supervised training method is applied to train this model with annotated data. In this paper, the result of the test dataset shows that the proposed model can achieve more than 90% accuracy performance on average.

1. Introduction
The recommendation system for music is a more and more significant tool for attracting new users. And the practice of music recommendation system is prevailing among mainstream music playing platforms[1]. In the information recommendation system field, some pioneer researchers have proposed their works over the last decades[2]. Nowadays, a new generation recommendation system for music is demanded to save users from information overloading problems and help them to find music that suits their tastes.

Another important reason for the development of the music recommendation system is the profitability of music platforms. The music copyright owners prefer to distribute the newest songs to their users so that they can obtain the profits at most[3]. Nevertheless, as for those old-fashion or niche songs, unless searched by users, they are hardly recommended by the music platforms, especially when the songs contain less commercial value. In this paper, we start from the music itself to design the model for a more accurate and efficient music recommendation system. A fine recommendation system is crucial to a commercial music player media and provides a large development space and market prospects.

2. Related works
In the current academic community, there are two main-stream methods for recommendation systems, i.e. the content-based approach, and the collaborative approach. The content-based method first analyses the content in the entity that users have viewed or posted in the past and then the system provides another entity with similar content. The collaborative approach recommends entities that the other users have liked in the same user group. The users are grouped by the shared inclinations, i.e. the tags of a song. If a user is in the group that likes pop music, this user is more likely to be recommended with pop music by the collaborative method.
There are some works investigated the music recommendation systems. In [4], authors derived some perceptual properties from the music objects, such as pitch, loudness, and duration. With these properties, the music objects are grouped by clustering algorithms. Meanwhile, the users’ interests are also derived from history and the users’ profiles. By considering these two sources of information, their system provides recommendations to the users. In another previous work [5], a content-based method was proposed using Bayesian networks and utility theory. In which paper, the authors exploited the fuzzy system to process with source information and used a Bayesian network to infer the context of the music object.

3. Data collection and data pre-processing
In this section, how we collected the data, both of the music objects database and the data of users listening history, and how we pre-processed the data are introduced.

3.1. Data collection
The strategy of our system is data-driven, therefore, the performance of the system depends on the quality of data. We downloaded an open-source dataset [6], which provides the feature analysis and metadata of audio information and lyrics data for one million songs and full listening history users. Note that the songs have been already classified into three groups (pop music, classic music, and country music) by the music style tags.

3.2. Audio data pre-processing.
The audio information of the music is stored as a vocal file in the database. The file was too big for fast processing and the time of songs is usually lasting for several minutes. Therefore, to make the audio file accessible to the deep learning model, we convert it to the spectrogram image [7]. In this way, the audio files are deformed to pictures, while the information of the songs remains. figure 1 shows one of the spectrogram images, converted from the audio files in the music objects database.

3.3. Lyric data pre-processing.
The lyric data has the same problem as the audio data for model processing, that the length of the lyric was too long to be efficiently utilized. Therefore, to reduce the length of the lyric while keeping the emotional expression, we wrote a script to analyze the lyric and find out the word that can describe the emotion at the best. In the experiment, we set six different emotions, happy, sad, creepy, exciting, energetic, melancholic, for describing the song’s emotions.

4. Fusion deep learning model
The proposed music recommendation system is a fusion deep learning model. As shown in figure 2, the model mainly consists of four blocks, namely, convolutional neural network (CNN) based spectrogram feature extractor; lyric emotion extractor; recommendation result predictor. The lyric emotion extractor
has been introduced in Section III, therefore, in this section, another two blocks would be explained in detail.

Figure 2. Structure of the proposed music recommendation system.

4.1. CNN based model for spectrogram feature extractor
The CNN based model is the core block of the proposed fusion model. In Section III, the spectrogram is obtained from audio file deformation. The spectrogram is then processed by a CNN model to extract the feature map. The output feature map is used to reflect the rhyme insight of the input songs. We use a simplified and truncated PSPNet\[8] for this task. The structure of this CNN model is shown in figure 3. After the upsampling process, the output feature map size is 4×4, we think 16 digits is enough for representing information of a song.

Figure 3. The detailed structure of the CNN based model.

4.2. Recommendation result predictor
This is the last processing block of the proposed model, the job for this block is to combine the processed and output information from the CNN model and the lyric emotion extractor with the input time-series listening history data. Note that the output feature map from the CNN model has been flattened before inputting to this predictor. For this block, we use logistic regression for its simplicity and efficiency for prediction job[9].

5. Experiment results
Since there are not many publicized music recommendation systems, and the authors did not make their code open-sourced, we only ran the proposed system by the datasets with 10-fold testify scheme. The music styles are pop music, classic music, and country music. We set the number of iterations as 9000 and recorded the accuracy performance at every 3000 iterations. The experimental results are exposed in table 1 show that the accuracy increases with the number of iterations. When the number of iteration is 9000, the overall accuracy is the best, at about 90%. When the iteration number is less than 3000, the performance was not satisfying, because the model is not convergent. The system is built with TensorFlow[10] platform.
### Table 1. Effect of iteration on accuracy.

| Iteration | Pop music | Classic music | Country music | Overall  |
|-----------|-----------|---------------|---------------|----------|
| 3000      | 75.3%     | 62.9%         | 70.6%         | 69.6%    |
| 6000      | 94.3%     | 83.4%         | 90.5%         | 89.4%    |
| 9000      | 95.3%     | 85.1%         | 90.1%         | 90.2%    |

### 6. Conclusion and future work

In this paper, we propose a new music recommendation system which is a fusion deep learning model. For a large public dataset, our model can achieve an overall 90.2% accuracy. Our contribution is to combine the audio information and lyric information with the listening history data of users. This consideration can provide a more comprehensive forecast of what should be the user’s favourite song. The proposed music recommendation system can be regarded as a prototype, where the blocks in the model can be alternated by other models for better performance. Especially the final prediction block, which is too simple for now.

The future work on this model includes improving the recommendation performance on classic music. Among the three music styles, the accuracy of classic music was the worst. Moreover, the recommendation system can be refined to satisfy other users’ requirements in different fields.

### References

[1] Tanaka, A. (2004, June). Mobile music making. In Proceedings of the 2004 conference on New interfaces for musical expression (pp. 154-156). National University of Singapore.

[2] Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In Recommender systems handbook (pp. 1-35). Springer, Boston, MA.

[3] Richardson, J. H. (2014). The Spotify paradox: How the creation of a compulsory license scheme for streaming on-demand music platforms can save the music industry. UCLA Ent. L. Rev., 22, 45.

[4] Chen, H. C., & Chen, A. L. (2005). A music recommendation system based on music and user grouping. Journal of Intelligent Information Systems, 24(2-3), 113-132.

[5] Park, H. S., Yoo, J. O., & Cho, S. B. (2006, September). A context-aware music recommendation system using fuzzy bayesian networks with utility theory. In International conference on Fuzzy systems and knowledge discovery (pp. 970-979). Springer, Berlin, Heidelberg.

[6] Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The million song dataset.

[7] Zhang, X., Zhu, B., Li, L., Li, W., Li, X., Wang, W., ... & Zhang, W. (2015). SIFT-based local spectrogram image descriptor: a novel feature for robust music identification. EURASIP Journal on Audio, Speech, and Music Processing, 2015(1), 6.

[8] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2881-2890).

[9] Larasati, A., DeYong, C., & Slevitch, L. (2012). The application of neural network and logistics regression models on predicting customer satisfaction in a student-operated restaurant. Procedia-Social and Behavioral Sciences, 65, 94-99.

[10] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In 12th Symposium on Operating Systems Design and Implementation (16) (pp. 265-283).