EAMR: An Emotion-aware Music Recommender Method via Mel Spectrogram and Arousal-Valence Model

Zixun Fu
Faculty of Artificial Intelligence in Education,
Central China Normal University,
Wuhan, China
eranonfzx@gmail.com

Zhen Zhang*
[1]School of Sports Engineering and Information Technology, Wuhan Sports University, Wuhan, China
[2]School of Information Management, Wuhan University, Wuhan, China
zhangzhen@whsu.edu.cn

Jie Zheng
School of Mathematics and Statistics,
Hubei University of Arts and Science,
Xiangyang, China
kozhengj@gmail.com

Ke Lin
Department of Control Science and Engineering,
Harbin Institute of Technology (Shenzhen),
Shenzhen, China
sheldon.chris.lin@gmail.com

Duantengchuan Li
School of Computer Science,
Wuhan University,
Wuhan, China
dtclee1222@gmail.com

***Abstract*—For music platforms, the cold start of new songs is a big challenge for recommendation systems. Most existing methods adopt music content (i.e., the audio signal) to alleviate this problem. But these approaches typically overlook the emotional characteristics embedded in the music, which is particularly important for modeling the characteristics of the songs. Considering the existence of negative and positive classifications of songs and the existence of alternating processes of calmness and excitement in song melodies. We propose an emotion-aware music recommender method (EAMR) via Mel spectrogram and the arousal-valence model. First, each audio clip is transformed into a log-Mel spectrogram, and a deep neural network extracts the emotional features. Meanwhile, each audio clip is dimensionally reduced to a 16-dimensional feature vector normalized to obtain each sentiment feature vector of songs. Second, the Euclidean distance of the song feature vector is calculated to recommend the song with the highest similarity to the user's current favorite. Finally, taking into account the impact of features extracted by different convolutional kernels on the model, the ability of model feature extraction is further improved by introducing an attention mechanism called Att-EAMR. The proposed methods are tested on the public Spotify Million Playlist dataset and the free music archive dataset FMA. Experimental results demonstrate that the proposed methods outperform the other comparison approaches when the system faces a cold start.**

Keywords-music recommender; mel spectrogram; arousal-valence model; deep neural network; sentiment feature

I. INTRODUCTION

With the booming of the digital music industry, hundreds of musical works are uploaded to various music platforms every day, which brings a great challenge for people to find suitable music from music platforms [1]. Music recommendation systems, which can provide personalized music recommendation services to users, were born on demand to meet this challenge [2]. A popular recommendation method is collaborative filtering algorithms [3], which are based on historical data of user-music interactions and identify patterns from them to make recommendations. However, these approaches simply consider the user-music interaction data and thus cannot recommend a new piece of music [4] (cold start problem). In addition, the recommendation is affected by the popularity of the music, which means that the more popular music is more likely to be recommended (long-tail effect problem). Another popular recommendation method is the content-based recommendation method [5], [6], which recommends similar music to users based on their previous preferences. Although the content-based recommendation methods can alleviate cold starts to some extent, this approach requires content analysis and the definition of a suitable similarity calculation function.

Considering the rich emotional characteristics embedded in the songs, not only are there positive and negative emotions in the songs. But there are also alternating processes of calm and excitement in the melodies of the songs. And this characteristic in music is consistent with the arousal-valence model in psychology. Hence, we propose an emotion-aware music recommender method (EAMR) via Mel spectrogram and the arousal-valence model, which does not rely on historical data but only uses a convolutional neural network to calculate the emotional similarity between music. To this end, the arousal-valence model in psychology [7] is introduced to classify the emotions expressed by music clips. The arousal and valence dimensions are classified into four categories in terms of strong and weak, positive and negative aspects of music. Subsequently, the emotions of music are classified into sixteen categories according to the different types of arousal and valence. The log-Mel spectrogram of music is then divided into its emotional categories by convolutional neural networks. In this way, when a new work by a new generation of artists appears, it can be recommended to users who might like these works because it has similar emotions with some other artists' works. When a new user appears, the user can also select the current emotional state by himself, and the system can recommend music corresponding to the emotion for the user. The major contributions of this article are summarized in three aspects.

2022 International Conference on Frontiers of Artificial Intelligence and Machine Learning (FAIML) | 978-1-6654-7364-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/FAIML57028.2022.00021

978-1-6654-7364-4/22/$31.00 ©2022 IEEE
DOI 10.1109/FAIML57028.2022.00021

- As the emotion contained in music can reflect the current emotional state of the user, we construct an emotion-aware music recommender method (EAMR). Based on the arousal-valence theory, the EAMR method classifies the emotion of the music by log-Mel spectrogram and convolutional neural network to obtain the emotional characteristics of each music. Then, we utilize the Euclidean distance to express the emotional similarity of two music to achieve music recommendations. Further, we propose an attention-based mechanism (Att-EAMR) to improve the model feature extraction capability.
- We propose a new method to calculate the proportion of different emotions in a piece of music. Based on the arousal-valence theory, we further classified arousal and valence into four categories according to their positive and negative, strength and weakness, respectively. Thus, the basic emotions of music are divided into sixteen categories according to the different categories of arousal-valence.
- The method proposed in this paper relies only on the sentiment expressed in the songs themselves rather than on historical data, which can alleviate the cold-start problem and the long-tail effect in music recommendation to some extent. The results demonstrate that our methods achieve significant improvements for music recommendation tasks.

The rest of this article is organized as follows. Section II presents the related work. Section III elaborates the proposed EAMR model in detail. Section IV introduces the datasets and model parameters setting. Section V presents the experimental results and analysis on public datasets. Finally, Section VI concludes this article.

II. Related Work

A. Arousal-Valence Model

The dimensional theory of emotion advocates the use of continuous indicators for the assessment of emotions. The dimensional theory of emotion suggests that human core affect is continuous in the brain and consists of a mixture of two dimensions: arousal and valence [7]. In addition, emotions fundamentally arise from the different activation of the desire and defense motivational systems [8]. In the arousal-valence model, arousal indicates the degree of activation of each motivational system, and valence indicates which motivational system is activated by the emotional stimulus. In recent years, the arousal-valence model has been widely applied to emotion recognition tasks. Yan *et al.* [9] proposed a nonlinear dynamic analysis method for identifying emotions in an electroencephalogram by combining the arousal-valence model. He *et al.* [10] proposed a joint temporal convolutional network and adversarial discriminant domain for the electroencephalogram-based cross-subject emotion recognition method combined with the arousal-valence model. Moreover, the arousal-valence model has been applied to generative tasks, Sulun *et al.* [11] proposed a symbolic music generation method based on the continuous value in combination with the arousal-valence model. In these works, the experimental results demonstrate that the arousal-valence model works well for modeling emotions.

B. Music Recommendation

The existing algorithms for recommendation are mainly divided into two categories: collaborative filtering recommendation algorithms [3], and content-based recommendation algorithms [5], [6].

Collaborative filtering is based on the user's historical data to make recommendations by discovering the similarities in historical interaction data [12]. The advantage of this algorithm is that it is simple and effective [13], does not require much domain knowledge, and only requires historical data to predict users' future behavior. However, sparse data can affect the accuracy of the model, and there is a cold start problem for new users or new items [3]. The core of a content-based recommendation algorithm is to make recommendations by calculating the similarity between items. Unlike collaborative filtering algorithms, this approach requires more expertise in the corresponding domain to define the similarity between items without relying on historical data. However, the performance of the algorithm may be limited by the definition of similarity between entries. For music recommendations, a content-based recommendation approach is an effective means to mitigate cold starts [14], [15].

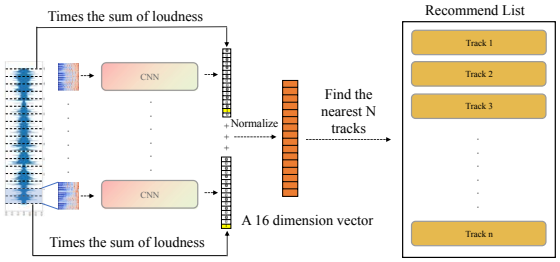

Figure 1. The overall structure of EAMR. The input of the model is the audio of songs, and outputs the most similar tracks.

III. Proposed EAMR Method

Music recommender systems aim to add songs that users prefer to their playlist. However, similar to the traditional recommender system, the cold start problem is still an inevitable challenge. Most of the literature [16] is based on song characteristics and user similarity to generate an ordered list of music. By this way, music feature modeling can effectively increase the exposure of tail songs in the long-tail distribution.

Given that our task is to find the proper song for users' playlists, the feature similarity of songs is critical. For example, a user may prefer classical and punk rock music but usually would not place songs of both genres into the same playlist. In our paper, Fourier transform transforms each song's audio into a spectrum. The embedded representation of emotion in each song can be obtained by feature extraction based on CNN network model. Finally, the similarity of embeddings will be evaluated in the whole set, and the top-N songs were selected to append the playlist. The overall structure of the model is shown in Figure 1.

A. Audio Feature Extraction

For each audio $\boldsymbol{s}$, slice it into a series of segments with a 5-second length, and the set of this audio segments series is noted as $\boldsymbol{A}_S$. Since the Fourier transform always get the best effect on the middle part, the slicing has an overlap of 2.5

seconds between two neighboring segments. For each fragment obtained by slicing, the spectrum $\boldsymbol{E}$ is first calculated using the fast Fourier transform, and then the power spectrum is obtained by the following formula:

$$p = 2 * log(|\boldsymbol{E}|). \quad (1)$$

To obtain the set of spectral set $\boldsymbol{P}_s$ (as shown in Figure 2), the log-Mel spectrogram is multiplied by $\boldsymbol{E}$:

$$p = 2 * log(|\boldsymbol{E}|), \quad (2)$$

where $\boldsymbol{M}$ represents the Mel-scale of the matrix:

$$\boldsymbol{M} = 2595 \times log_{10}(1 + \frac{f}{700}) . \quad (3)$$

The loudness values of each segment in $\boldsymbol{A}_s$ is summed to obtain the set of segment loudness sums $\boldsymbol{D}_s$.

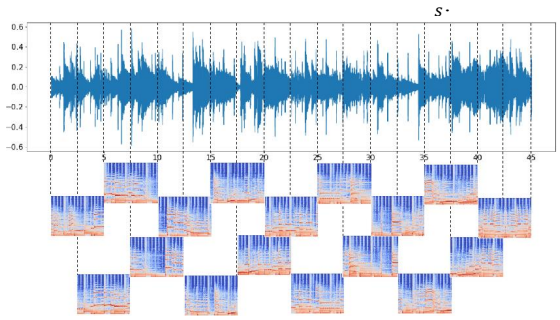

Figure 2. The generation of the spectral set from audio. For each audio s, slice it into a series of segments with 5 seconds in length.

B. Emotion classification method

We introduce the arousal-valence model in psychology to classify the emotions of music clips. Psychologist Russell proposed this model in 2003 [7], which measures emotions through two dimensions: arousal and valence. In Figure 3, the negative half-axis of the arousal dimension represents smooth emotions, and the positive half-axis represents strong emotions. The negative half-axis of the valence dimension represents negative emotions, and the positive half-axis represents positive emotions. On this basis, we divide the arousal dimension into four parts with a critical value L:

$$\begin{cases} Strong\ Positive\ Arousal\ type\ 3, & x>L \\ Weak\ Positive\ Arousal\ type\ 2, & 0\le x\le L \\ Weak\ Negative\ Arousal\ type\ 1, & -L\le x<0 \\ Strong\ Negative\ Arousal\ type\ 0, & x<-L \end{cases} \quad (4)$$

Similarly, the division of valence dimension can be obtained:

$$\begin{cases} Strong\ Positive\ Valence\ type\ 3, & x>L \\ Weak\ Positive\ Valence\ type\ 2, & 0\le x\le L \\ Weak\ Negative\ Valence\ type\ 1, & -L\le x<0 \\ Strong\ Negative\ Valence\ type\ 0, & x<-L \end{cases} \quad (5)$$

The i represents the arousal intensity of a segment, and j represents valence intensity. Thus, the emotional categorization of a segment is calculated as follows

$$T = 4 * i + j + 1 . \quad (6)$$

The classification diagram is shown in Figure 3. We can observe that the spectrogram changes faster when the valence is stronger. The more high-frequency signals when the arousal is stronger.

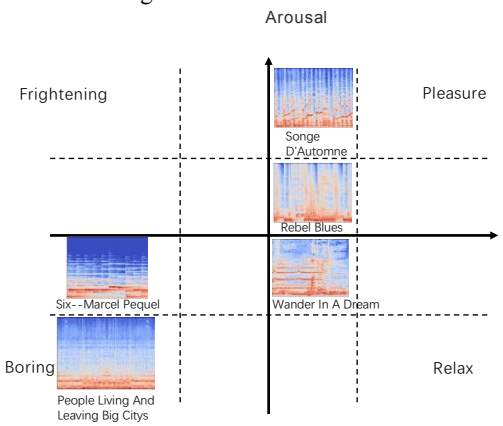

Figure 3. The classification diagram of the arousal and valence.

C. Audio clip emotion classification

For a spectrogram in $\boldsymbol{P}_s$, a convolutional neural network is used to classify sentiment categories. The output category is transformed into a one-hot vector to represent the sentiment of the music in this clip. After performing the above operation on all images, a collection of one-hot vectors $\boldsymbol{T}_s$ is obtained. We design two convolutional neural networks to perform the classification task in this work. The overall model structure is shown in Figure 4 and Figure 5.

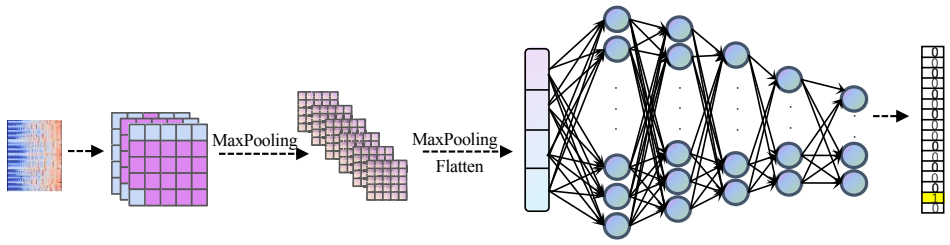

Figure 4. Outline of EAMR framework. The emotion classification model of the audio clip. The output category is transformed into a one-hot vector to represent the sentiment of the music in this clip.

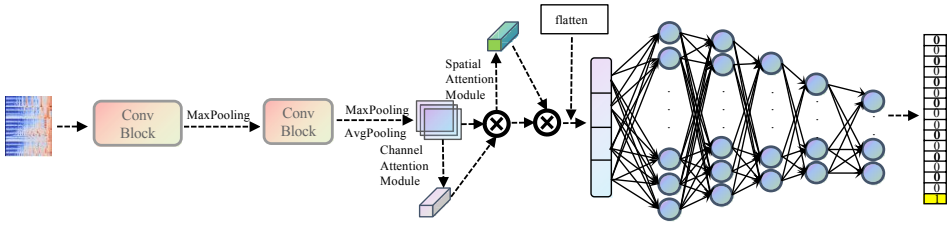

Figure 5. Outline of Att-EAMR framework. The attention emotion classification model of the audio clip.

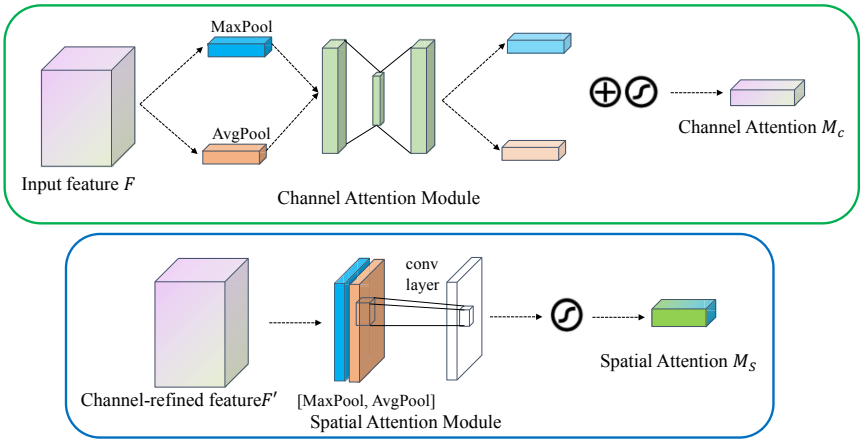

Figure 6. The detailed description of the channel attention module.

Among them, the Conv Block of Att-EAMR model has the same as the Block structure of GoogLeNetv1 [17]. Compared with Figure 4, we use a channel attention module and a spatial attention module in Figure 5. The channel attention module is calculated as the following formula:

$$M_S(F) = \sigma(MLP(AvgP(F)) + MLP(MaxP(F))) \quad (7)$$

where σ denotes the sigmoid function, AvgP denotes AvgPool and MaxP denotes MaxPooling. The spatial attention is calculated as the following formula:

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F)]; MaxPool(F))) \quad (8)$$

where σ denotes the sigmoid function and $f^{7\times7}$ represents a convolution operation with a filter size of 7×7. The specific structure of the attention module is shown in Figure 6.

D. Audio Emotional Component Metric

Considering that the music segment with stronger emotion is always louder than the segment with weaker emotion. Each vector in $\boldsymbol{T}_s$ is multiplied by the loudness sum of the original clip, and then all the vectors in $\boldsymbol{T}_s$ are summed to obtain the vector $\boldsymbol{e}_0$ of the audio emotion. Given the difference in length and loudness of the audio, it is not easy to calculate the difference between audio emotions using only this vector. So, the final audio sentiment component $\boldsymbol{e}$ is obtained by applying regularization to $\boldsymbol{e}_0$.

E. Recommendation List Generation

In this paper, we use the Euclidean distance of the sentiment component vector $\boldsymbol{e}$ of the two audio tracks to express their similarity. To generate the recommendation list, the similarity of any two audio tracks is calculated, and the similarity matrix $\boldsymbol{M}_s$ is obtained, and the Top-N method is used to find the top-N audio tracks with the smallest $\boldsymbol{s}$ from a matrix row to obtain the recommendation list corresponding to this audio track.

IV. Experiment Setting

A. Dataset

The data for this paper comes from three datasets: the music sentiment analysis dataset DEAM [19], the Spotify Million Playlists dataset [20], and the Free Music Archive dataset FMA [21]. Their metadata is shown in Table 1.

TABLE I. The Statistic of Datasets

	FMA	Spotify	DEAM
Tracks	106,574	2,262,292	1,802
Artists	16,371	295,860	
Albums	14,854	734,684	

The DEAM dataset is a media evaluation database for sentiment analysis in music, which contains 1802 tracks from various copyright-free music datasets (websites) such as jamendo.com, medleyDB dataset, etc. The DEAM dataset consists of data in two areas: 1) audio files in Mp3 format with a length of 45s. 2) Audio sentiment tags with an arousal-potency index of 2Hz. The data are only available for the last 30s because of the unstable arousal-valence labeling in the first 15s.

Spotify Million Playlists dataset is a dataset consisting of millions of real existing playlists on Spotify. Each playlist contains metadata of dozens of songs and data such as the number of this playlist, metadata including artist, artist id, music name, music id, album name, album id, and audio length (ms).

The FMA dataset is a dataset for evaluating multiple tasks in music information retrieval and consists of 106,574 tracks from 16,341 artists and14,854 albums, of which the fma_full.zip archive provides the full length and high quality of all of the above audio.

In this paper, we use the audio in the DEAM dataset as training data and the intersection of the million playlists dataset and the FMA dataset as test data.

B. Data Preprocessing

Each audio in DEAM is sliced into a series of segments with 5s in length, and then a crossover with 2.5s in length is made between two adjacent segments. Finally, the log-Mel spectrogram of each segment is drawn. The ten consecutive arousal-valence data were averaged as the arousal-valence values within 5s. 0.1, 0.2, 0.3, 0.4, and 0.5 were utilized as the threshold values for strong and weak arousal (valence), respectively. And five sets of labels were given to each drawn log-Mel spectrogram.

For the data in the playlist dataset, all the music data are extracted from the playlist according to the music ids. Each id record corresponds to the artist, the music name, and the number of the playlist in which it appears. The metadata of all records and the FMA dataset were scanned to filter out the music that emerged in both datasets simultaneously as the test dataset. For the filtered music, a Boolean adjacency matrix **D** is calculated, where $\boldsymbol{d}_{ij}$ is 1, indicating that the *i*-th music and the *j*-th music appear in at least one playlist at the same time.

C. Convolution Neural Network Structure

This paper uses four convolutional neural networks for the spectrogram classification task. Detailed model settings and the number of parameters are shown in Table 2 and Table 3.

TABLE II. THE STRUCTURE OF PARAMETERS IN EAMR

Layer(type)	Patch size/stride	Output size	Activation	#3x3	Params
Conv2d_1	3x3/2	189x249x64	ReLU	64	1,792
Maxpooling2d_1	2x2/1	93x124x64			0
Conv2d_2	3x3/2	46x61x128	ReLU	128	73,856
Maxpooling2d_2	2x2/1	23x30x128			0
Dropout(15%)		23x30x128			0
Flatten		88320	ReLU		0
Dense_1		256	ReLU		22.6M
Dense_2		128	ReLU		32,896
Dense_3		64	ReLU		8,256
Dense_4		32	ReLU		2,080
Dense_5		16	Softmax		528

TABLE III. THE STRUCTURE OF PARAMETERS IN ATTENTION EAMR

Layer(type)	Patch size/stride	Output size	Activation	#1x1	#3x3	#5x5	Params
Block_1		374x500x64	ReLU	52	24	16	8,872
Maxpooling2d_1	3x3	Multiple					0
Block_2		124x166x72	ReLU	60	24	16	12,596
Maxpooling2d_2	3x3	Multiple					0
AvgPool2d	5x5	Multiple					0
Dropout(15%)		Multiple					0
Flatten		Multiple	ReLU				0
Dense_1		Multiple	ReLU				1,622,272
Dense_2		Multiple	ReLU				32,896
Dense_3		Multiple	ReLU				8,256
Dense_4		Multiple	ReLU				2,080
Dense_5		Multiple	Softmax				528

D. Training Convolutional Neural Networks

The training data DEAM dataset was preprocessed to obtain the log-Mel spectrogram in the ratio of 8:2 to divide the training and test sets. The training set was used as the input to the network, and the emotion types at 0.1, 0.2, 0.3, 0.4, and 0.5 as the threshold values of strong and weak arousal (valence) were used as labels, respectively, and input to the above three networks for training, followed by prediction of the data in the test set.

E. Calculating the Emotional Feature Vector of Music

The trained convolutional neural network is used to classify the preprocessed log-Mel spectrograms. The classification result is transformed into a one-hot vector. Then, the vector is multiplied by the loudness sum of each clip. Finally, all the vectors of each audio are summed and normalized to obtain a vector representing the mood of the music.

F. Music Recommendations

For all the audios in the test data, an adjacency matrix **M** is computed, where m_{ij} denotes the Euclidean distance of the sentiment vector e of the *i*-th audio and the *j*-th audio to represent their similarity. To fairly compare the performance under each parameter, Top-N results are calculated in this paper when *N*=5, 10, 20, 50, and 100, respectively.

$$m_{ij} = \sqrt{(e_i - e_{j)}{}^2}\,. \tag{9}$$

V. Evaluation of Experiment Results

The EAMR model recommended N music for a playlist to the user with the most similar emotion, based on the center of the emotional feature space of one or several songs. In evaluating model effectiveness, we used the same training and test data for each model and calculated the corresponding evaluation index to ensure realistic and valid evaluation results. We choose Precision, Recall, and mAP as metrics for evaluating the results. In the evaluation process, if two music tracks appear in the same playlist, we consider that they are related.

Precision is calculated as the following formula:

$$precision = \frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|R(u)|} \tag{10}$$

where R(u) denotes the list of music recommendations generated by the recommendation system for each music on the test set, and T(u) denotes all relevant music for each music on the test set.

The recall is calculated as the following formula:

$$recall = \frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|} \tag{11}$$

The mAP is calculated as the following formula:

$$mAP = \frac{1}{|U|}\sum_{u\in U}\frac{1}{m}\sum_{k=1}^{m}\frac{P_u(k)}{k} \tag{12}$$

where U denotes the set of all music in the test set, m denotes the length of the recommendation list. And if the k-th music in the recommendation list is related to the music corresponding to this list, then $P_u(k)$ is 1, otherwise $P_u(k)$ is set to 0.

A. Accuracy Analysis

All the methods are executed on our datasets. The comparisons of results are summarized in Figure 7. GoogLeNetv1 [17] and GoogLeNetv4 [18] is the most dominant deep neural network models nowadays. Specifically, GoogLeNetv1 increases the network depth and width by using 1x1 convolution to achieve channel reduction and dimensionality increase. GoogLeNetv4 proposes a faster training network by combining the backbone network with the residual network. By comparing different network model architectures, the experimental results show that the proposed model in this paper achieves the best results in Recall, Precision, and mAP metrics for recommendation lists of 5, 10, 20, 50, and 100. Although GoogLeNetv1 [17] and GoogLeNetv4 [18] can achieve good results for computer vision tasks, the models proposed in this paper tended to outperform the two GoogLeNet models. The reason might be that the task was created with a different original intent than the GoogLeNet model, and the training sample size was insufficient. As there are only 1802 audios of length 45s in the DEAM dataset, and the arousal-valence labels are unstable at the beginning 15s of each segment, only 11 Mel spectrograms per audio can be used after dividing the audio using a window of length 5s with a step size of 2.5s, for a total of 19,822.

B. Sensitivity Analysis of Parameters

To reveal the effect of different threshold values of strong and weak arousal (valence) on model performance. we execute some comparison experiments, in which parameters threshold values of strong and weak arousal (valence) is set in [0.1, 0.2, 0.3, 0.4, 0.5]. The related experimental results are demonstrated in Figure 8, which shows the impact of parameters on our dataset. Experimental results show that our models achieve the highest prediction accuracy when the threshold values of strong and weak arousal (valence) are equal to 0.3 or 0.4. When using the GoogLeNetv1and GoogLeNetv4 models, the experiment parameter is equal to 0.3 performed significantly better than 0.2 and 0.4. Therefore, we choose 0.3 as the parameter of our model. Thus, threshold values of strong and weak arousal (valence) can be set at the abovementioned values in real scenarios.

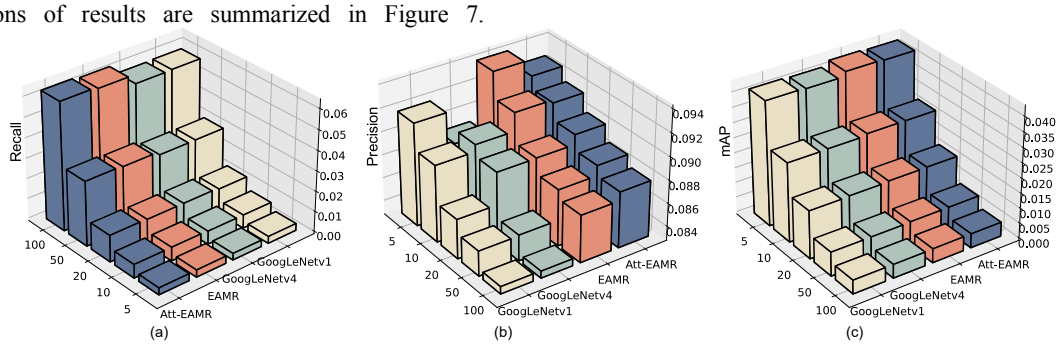

Figure 7. Comparison of music recommendation results by using the proposed methods and comparison methods.

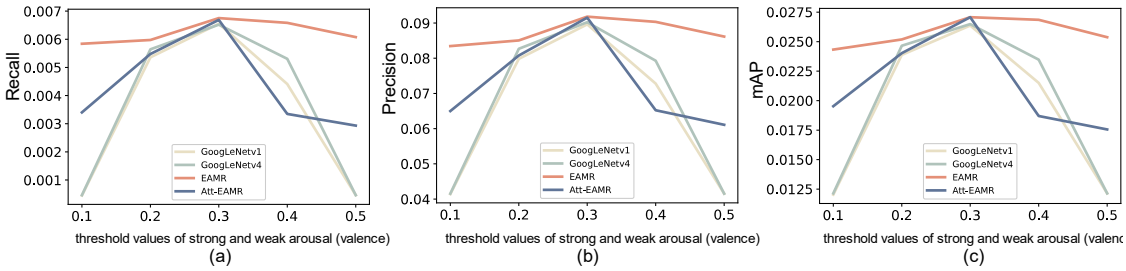

Figure 8. The comparison Results in Recall, Precision, and mAP on different threshold values of strong and weak arousal (valence).

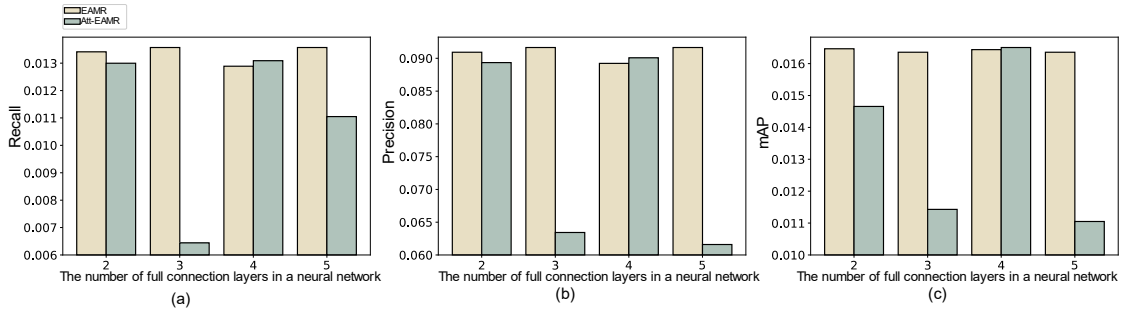

Figure 9. The comparison Results in Recall, Precision, and mAP on the attention mechanism.

C. *Attention Mechanism Analysis*

The attention mechanism plays an important role in enhancing the generalization ability of the models. To explore the effect of the attention mechanism on the model, we conduct experiments for different numbers of full connection layers in a neural network. The experimental results are shown in Figure 9. The experimental results show that the overall model performance is best when the full connection layer equals 4. Moreover, the experiments further show that the Att-EAMR model achieves better results in Recall, Precision, and mAP compared to EAMR.

This is because different convolutional kernels focus on different audio features. And users tend to focus on the types of music clips they are interested in. Therefore, the attention-based mechanism model Att-EAMR can further improve the model's accuracy compared to EAMR.

VI. CONCLUSION

This article presents an emotion-aware music recommendation system (EAMR) based on a log-Mel spectrogram and an arousal-valence model based on emotional signals hidden in music in this research. The music is first separated into many pieces and converted into a log-Mel spectrogram image. Then a convolutional neural network is used to downscale each image to a sixteen-dimensional one-hot vector, which is then multiplied by the total of the loudness of the associated segments. The sentiment vector for each piece of music is then calculated by adding all of the vectors and normalizing them. Finally, the similarity between the music is defined using the Euclidean distance of the sentiment feature vector to recommend the music with the highest similarity. Meanwhile, the improvement of the model performance by the attention mechanism is also considered, and the Att-EAMR model is proposed. The experimental result shows that our methods achieve promising results when facing the cold start problem. In the future, we will incorporate the emotion contained in the lyrics into the music emotion feature vector to improve recommendation accuracy.

ACKNOWLEDGMENT

This work was supported by the Wuhan Sports University Young Teachers' Research Fund Project Grant (No. 2022S10).

REFERENCES

[1] R. Liu and X. Hu, "A Multimodal Music Recommendation System with Listeners' Personality and Physiological Signals," Aug. 2020, pp. 357–360.

[2] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, "Current challenges and visions in music recommender systems research," *Int. J. Multimed. Inf. Retr.*, vol. 7, no. 2, pp. 95–116, Jun. 2018.

[3] Y. Guo and Z. Yan, "Recommended System: Attentive Neural Collaborative Filtering," *IEEE Access*, vol. 8, pp. 125953–125960, 2020.

[4] A. Ferraro, "Music cold-start and long-tail recommendation: bias in deep representations," *Proc. 13th ACM Conf. Recomm. Syst.*, 2019.

[5] A. Niyazov, E. Mikhailova, and O. Egorova, "Content-based Music Recommendation System," in *29th Conference of Open Innovations Association (FRUCT)*, 2021, pp. 274–279.

[6] P. Knees and M. Schedl, "A Survey of Music Similarity and Recommendation from Music Context Data," ACM Trans. Multimed. Comput. Commun. Appl. TOMCCAP, vol. 10, Dec. 2013.

[7] J. A. Russell, "Core affect and the psychological construction of emotion.," *Psychol. Rev.*, vol. 110, no. 1, pp. 145–172, Jan. 2003.

[8] J. R. Jennings and B. Allen, "Methodology," in *Handbook of Psychophysiology*, 4th ed., J. T. Cacioppo, L. G. Tassinary, and G. G. E. Berntson, Eds. Cambridge University Press, 2016, pp. 583–611.

[9] A. Paudel, B. R. Bajracharya, M. Ghimire, N. Bhattarai, and D. S. Baral, "Using Personality Traits Information from Social Media for Music Recommendation," in IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), 2018, pp. 116–121.

[10] Z. He, Y. Zhong, and J. Pan, "Joint Temporal Convolutional Networks and Adversarial Discriminative Domain Adaptation for EEG-Based Cross-Subject Emotion Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3214–3218.

[11] S. Sulun, M. E. P. Davies, and P. Viana, "Symbolic Music Generation Conditioned on Continuous-Valued Emotions," *IEEE Access*, vol. 10, pp. 44617–44626, 2022.

[12] W. Wenzhen, "Personalized Music Recommendation Algorithm Based on Hybrid Collaborative Filtering Technology," in *International Conference on Smart Grid and Electrical Automation (ICSGEA)*, 2019, pp. 280–283.

[13] Y. Wang, Y. Zhou, T. Chen, J. Zhang, W. Yang, and Z. Huang, "Hybrid Collaborative Filtering-Based API Recommendation," in *IEEE 21st International Conference on Software Quality, Reliability and Security (QRS)*, 2021, pp. 906–914.

[14] K. Okada, P. X. Tan, and E. Kamioka, "Five-Factor Musical Preference Prediction for Solving New User Cold-Start Problem in Content-Based Music Recommender System," in *International Conference on Information, Intelligence, Systems Applications (IISA)*, 2021, pp. 1–7.

[15] K. Chen, B. Liang, X. Ma, and M. Gu, "Learning Audio Embeddings with User Listening Data for Content-Based Music Recommendation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3015–3019.

[16] A. Gatzioura, J. Vinagre, A. M. Jorge, and M. Sànchez-Marrè, "A Hybrid Recommender System for Improving Automatic Playlist Continuation," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 5, pp. 1819–1830, 2021.

[17] C. Szegedy et al., "Going deeper with convolutions," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 1–9.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence,* 2017, pp. 4278–4284.

[19] A. Alajanki, Y.-H. Yang, and M. Soleymani, "Benchmarking music emotion recognition systems," PLOS ONE, 2016.

[20] C.-W. Chen, P. Lamere, M. Schedl, and H. Zamani, "Recsys Challenge 2018: Automatic Music Playlist Continuation," in Proceedings of the 12th ACM Conference on Recommender Systems, New York, NY, USA, 2018, pp. 527–528.

[21] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A Dataset for Music Analysis," in Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR, 2017, pp. 316-323.