Construction of multimodal music automatic annotation model based on neural network algorithm

Zhong Miao, Chaozhi Cheng*
College of Music and Dance, Huaihua University, Huaihua, China

ABSTRACT

Improving the effect of music annotation task through various advanced technologies has become a hot direction in the field of music information retrieval. The research of music automatic annotation has high value in both theoretical research and practical application. The deep CNN (Convolutional Neural Network) model in the field of deep learning has achieved good results in the fields of image and voice. In this paper, the construction of multimodal music automatic labeling model based on neural network algorithm is launched. In this paper, CNN combined with SAM(Self-attention mechanism) is used to learn the appropriate feature representation from the low-level Mel spectrum description of music and the original audio waveform data. Two-dimensional convolution is applied to the Mel spectrum input of music, and one-dimensional convolution is applied to the original audio waveform input, so as to better capture various structural features of music and annotate it. The results show that the accuracy of CNN combined with SAM method is 3.9% higher than that of linear weighted fusion method, and the AUC value is 8.5% higher than that of linear weighted fusion method. The comparison results show that the multi-modal music automatic annotation model framework proposed by CNN and SAM in this paper is effective for the automatic annotation task of music.

Keywords: Neural network; Music automatic labeling; Convolutional neural network; Self-attention mechanism

1. INTRODUCTION

Voice is one of the important ways people use to communicate information in their daily lives. From the frequency point of view, music and natural sounds are distributed in all frequency ranges, and the frequency range of speech is usually between 300 Hz and 4 kHz [1]. In recent years, with the continuous development of Internet technology and the rapid emergence of various multimedia applications, more composers, singers and bands can upload their music works to the online digital music library, thus having the opportunity to be listened to and loved by more listeners. Music subscription service has maintained a sharp upward trend in recent years and has become the main driving force for the growth of digital music, and its paying users have increased steadily in recent years. In this context, structured information organization such as music labeling is becoming more and more important, and the concept of music label has attracted more and more attention.

The problem of music automatic labeling is related to the general regression problem of music classification, such as genre classification and emotion prediction. Music is the media that people often contact, and it has many storage forms, such as MIDI, MP3, various compressed music products, real-time music broadcasting and so on [2-3]. Reference [4] proposes to train AdaBoost model by using MFCC (Mel frequency cepstrum coefficient) features, which can be self-labeled according to audio features and community tags. Literature [5] combines the word bag feature and audio feature of Chinese lyrics to automatically label Chinese songs. Literature [6] extracts various features of audio content as feature data, manually marks them, and uses KNN(K-Nearest Neighbor) algorithm to classify music emotion categories, and finally divides the song audio into 11 emotion categories. Literature [7] considers the sequence structure of music features, and its annotation network is composed of a series of one-dimensional convolution layers and one-dimensional pooling layers only along the time axis, which has achieved good annotation performance.

*Email: ccz175@163.com

Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023), edited by Srikanta Patnaik,
Tao Shen, Proc. of SPIE Vol. 12779, 127791Q · © 2023 SPIE · 0277-786X · Published under a
Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2689482

Under the background of music market transformation, as a structured way of music information organization, the concept of music label is prominent. Improving the effect of music labeling task through various advanced technologies has become a hot topic in the field of music information retrieval. The research on music automatic labeling has high value in both theoretical research and practical application. The deep CNN (Convolutional Neural Network) model in the field of deep learning has achieved good results in the fields of image and voice [8-9]. In this paper, the word vector of lyrics is used as input information, and a multimodal music automatic annotation model based on CNN and SAM(Self-attention mechanism) is proposed. The effects of different input representation methods, network structure and hyperparameters on the model performance are discussed through experiments, and the excellent performance of the music annotation model based on CNN is verified.

2. RESEARCH METHOD

2.1 Musical feature representation

Music automatic tagging refers to automatically adding a set of semantic text tags to music. This task has high research value for many music-related application scenarios such as music data management, indexing, storage and recommendation. Music tags are usually composed of several words, which are used to represent the attributes of the music segment and are the summary of the highest semantic level in audio content [10]. Nowadays, automatic music labeling is a content-based music classification, which predicts various music labels according to the content of music, so as to label them. The automatic music labeling method automatically processes the audio signal through an algorithm, and extracts a higher level of meaning from the waveform music signal. For the early music automatic annotation methods, the main tasks are melody extraction, harmony extraction, note recognition, rhythm recognition, musical instrument recognition and so on.

In order to realize automatic music annotation, we must first extract the audio feature representation for music samples, and then find the appropriate feature representation for music through feature combination, feature selection or unsupervised feature learning. The trained tagging model is then applied to the test set to predict the text tags related to each test music sample. The text tags predicted by the model will be compared with the real music tagging, and the corresponding evaluation index will be calculated to evaluate the tagging performance of the model. With the explosive growth of texts on the Internet, the importance of text classification has become increasingly prominent, which provides an important organizational form for massive texts and facilitates the rapid retrieval and efficient management of texts.

Most of the existing music annotation methods represent a piece of music as a whole, or simply divide the music into pieces of fixed length. In order to better model the relationship between music internal segments in the future, this paper uses a hierarchical adaptive music representation method. Given the audio signal sequence of music, the goal is to divide it into a group of segments with consistent audio characteristics. The length of these music segments is not fixed, but determined by their signal characteristics. Specifically, we should first calculate the similarity between the spectral representations of music signal sequences and construct the self-similarity matrix between audio frames. Compared with the fixed-length segmentation scheme, the adaptive music segmentation method proposed in this paper is helpful to extract the local audio features of music more accurately, thus reflecting the dynamic changes of music along the time axis more flexibly [11].

MFCC is the most commonly used feature parameter in the field of music emotion recognition, and the signal passes through a filter to enhance or compensate the high-frequency part that is suppressed in the process of speech generation. The relationship between the input and output signals is shown in Formula (1).

$$Y[n] = X[n] - \alpha X[n-1] \tag{1}$$

Where $Y[n]$ is the output pre-emphasis signal, $X[n]$ is the input signal, and α is in the range of [0.95,0.97], and the default value is 0.97.

Chinese word segmentation refers to the segmentation of Chinese character sequences into individual words. The purpose of word segmentation is to recombine word sequences according to certain rules and generate word sequences with certain significance. Filtering out stop words in natural text can not only effectively improve the efficiency of computer text processing, but also save a lot of storage space.In order to solve the problems in word frequency statistics, TF-IDF(Term Frequency Inverse Document Frequency) weighting method was introduced into text feature extraction.

Suppose there are D documents in this paper. The TF_{ij} of the word i for the document j is as follows:

$$TF_{ij} = \frac{f_{ij}}{\max\limits_{k} f_{kj}} \tag{2}$$

Where f_{ij} represents the frequency with which the word i appears in the document j. This frequency is normalized by the maximum word frequency in document j, which is the TF_{ij} of word i to document j. Therefore, the value of TF_{ij} is between [0,1].

In this paper, it is proposed that instead of word segmentation, words are directly used as units for text processing, and distributed text representation is generated as the input of deep neural network. The specific process of the lyrics text representation of a single song is as follows: unsupervised training is carried out by using a large-scale corpus to construct a dictionary of word embedding; For each word in the lyrics, find the corresponding word vector from the word embedding dictionary one by one. Then these word vectors are spliced according to the order of words in the lyrics to form a text representation matrix of the lyrics, which is used as the input of the deep neural network.

2.2 Automatic annotation of multimodal music

The main task of this method is to extract advanced information about music content from the original audio signal, so as to predict various music labels [12]. The main advantage of CNN is to extract the local features of music and extract more advanced features layer by layer; The main advantage of cyclic neural network is that it can process music sequences of different lengths and save the sequence information in the sequence. The model proposed in this paper assigns an adaptive weight to the feature vector corresponding to each moment in the feature sequence of music, which is obtained based on SAM calculation, and further calculates the weight of the feature and the feature description as the whole music to label the music appropriately.

CNN is a kind of neural network with some limitations in network topology and parameterization-at least one layer of the network uses convolution operation. CNN aims to provide a method to learn highly robust features, while learning highly robust features aims to achieve local invariance, translation invariance, distortion invariance and so on in response to visual objects. Although deep learning has advantages in feature learning, it is still necessary to carefully consider the characteristics (such as invariance) and the degree to which the characteristics need to be achieved when designing the structure of neural networks.

In this method, convolutional cyclic neural network is used to realize automatic music labeling. For the task of automatic music labeling, CNN combined with SAM is used in this paper to learn the appropriate feature representation from the low-level Mel spectrum description of music and the original audio waveform data. Two-dimensional convolution is applied to the Mel spectrum input of music, and one-dimensional convolution is applied to the original waveform input of audio, so as to better capture various structural features of music and annotate it. The structure of CNN combined with SAM proposed in this paper for automatic annotation of multimodal music is shown in Figure 1:

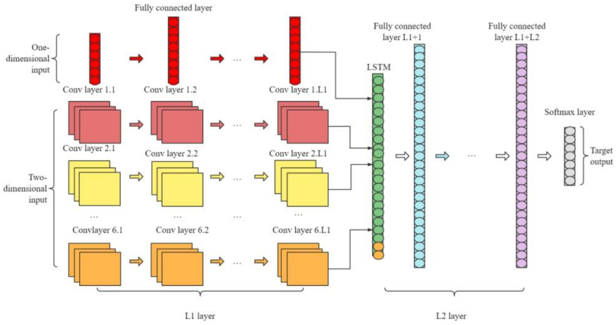

Figure 1. CNN combined with SAM structure

In this paper, it is assumed that the number of layers of a convolution layer is l, then the number of layers of the pool layer below it is $l+1$. In order to calculate the sensitivity of l layer, this paper needs to sample the sensitivity corresponding to the pool layer, so that the dimension of the sensitivity is consistent with the dimension of the convolution layer output.

Therefore, the sensitivity to the j channel of the l layer is:

$$\delta_j^l = \beta_j^{l+1}\left(f^{'}\left(u_j^l\right)\circ up\left(\delta_j^{l+1}\right)\right) \tag{3}$$

Where β_j^{l+1} represents the weight corresponding to the pool layer; $up(\cdot)$ stands for up-sampling operation. For example, if up-sampling is performed by taking n as a factor, it is equivalent to copying each node for n times in horizontal and vertical directions. This function can be realized by Kronecker product.

LSTM(Long Short-Term Memory) is a variant of RNN(Recurrent Neural Network), which is widely used to model sequence data. It introduces a cell state variable and three control gates. In the model proposed in this paper, a bidirectional LSTM model is used, that is, a forward LSTM is trained on the input feature sequence, and a backward LSTM is trained by inverting the input feature sequence, and then the outputs of the two LSTM are combined at corresponding times.

In order to keep more information related to semantic tags in the feature representation of music as a whole, similar to SAM used in work, this paper chooses to calculate multiple groups of attention weights simultaneously on the feature sequence. Figure 2 shows SAM.

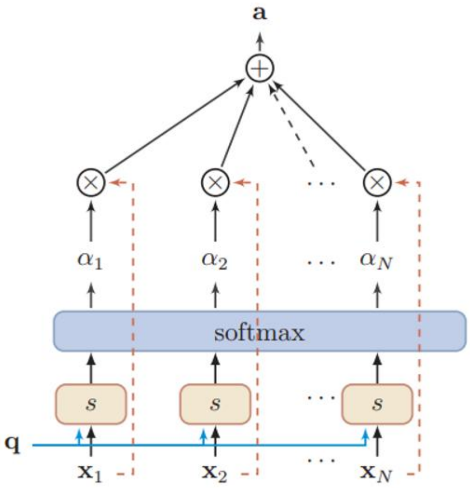

Figure 2. SAM schematic diagram

The parameter vector $w_2 \in R^{D_x}$ is extended to a parameter matrix $w_2 \in R^{r\times D_x}$, where r is a hyperparameter indicating the number of attention weight vectors. Thus, the attention weight matrix $A=[a_1,\cdots,a_r]^T$ can be obtained, and the calculation method is as follows:

$$A = soft\max\left(w_2\phi\left(w_1x^T\right)\right) \tag{4}$$

In the above formula, the calculation of $soft\max(\cdot)$ is carried out along the second dimension of the input matrix, thus ensuring that each group of weight vectors can satisfy the sum of 1. The superparameter r, which indicates the number of weight vectors, was set to 4 in the experiment.

On the basis of this attention weight matrix, this paper further calculates a two-dimensional embedding matrix M to aggregate the feature sequence of music, which is calculated as follows:

$$M = AX \tag{5}$$

The i -th line vector m_i in M is the weighted sum of the corresponding feature vectors at all times in the feature sequence. In the embedded representation matrix M of music, each individual vector actually pays attention to different parts of the whole feature sequence, so that it can retain many different musical features of the sequence features.

In order to highlight the importance of music segments being correctly labeled, we add the loss weight of judging as 0 on the basis of the original binary cross entropy, and increase the punishment of not labeling a certain music through this operation. The expression of the loss function with added penalty factor is shown in Formula (6).

$$L(y_n, \hat{y}_n) = -[\alpha y_n \log \hat{y}_n + (1-\hat{y}_n)\log(1-\hat{y}_n)] \tag{6}$$

Where α is the penalty factor, n is the sample number, and $L(y_n, \hat{y}_n)$ is the loss function of the n th sample about a certain music label. $\alpha = 1$ is the standard binary cross entropy, and $\alpha > 1$ is the binary cross entropy with the loss weight of judging 1 as 0, which is also the loss function finally adopted in this method.

3. EXPERIMENTAL ANALYSIS

This paper uses the widely used MTAT (Magna Tag ATune) data set to analyze and verify the performance of the proposed music annotation method. The MTAT dataset contains a total of 25863 music samples, each of which is 29.1 seconds long, and is stored in an mp3 format file at a 16kHz sampling rate. All audio files are stored in 16 subdirectories. In this experiment, the music samples in the first 12 subdirectories are used as the training set, the music samples in the 13th subdirectory are used as the verification set, and the music samples in the 14th to 16th subdirectories are used as the test set.

The audio format of the data set is MP3, and it is sampled or resampled according to the sampling frequency of 16kHz. Both the audio feature extraction network and the tag vector extraction module network use ELU as the activation function of the hidden layer. Dropout parameters are all set to 0.5 to prevent over-fitting, the activation functions of the output layer are all softmax functions, and the cross entropy function is used as the loss function. After each hidden layer of the network, a specification layer will be added to normalize the output results to make the learning converge quickly.

In the same experimental environment, using the same training set and test set, five groups of audio +VGG16 (two categories), audio +VGG16 (four categories), lyrics +TextCNN (two categories), linear weighted fusion and CNN combined with SAM were designed respectively, and comparative experiments were carried out. The classification of comparative experiments is shown in Figure 3:

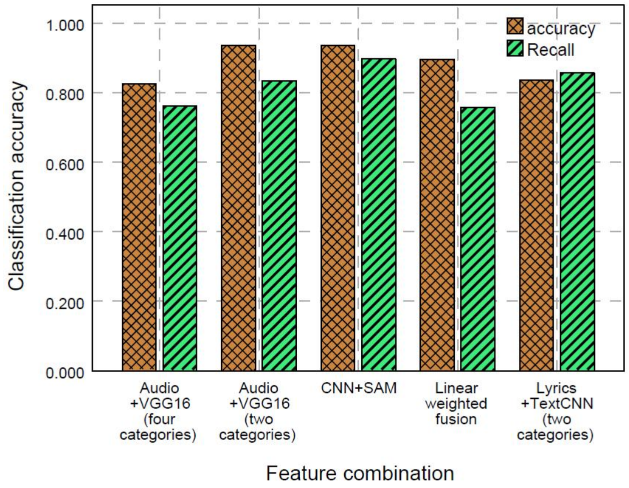

Figure 3. Compare the classification of the experimental group

It can be seen that the accuracy of song classification in single mode is lower than that in multi-mode fusion; In multimodal fusion, the accuracy of CNN combined with SAM method is 3.9% higher than that of linear weighted fusion method, because CNN combined with SAM method takes into account the different performances of audio and lyrics in energy and pressure respectively, and the classification effect of this fusion method in energy dimension is significantly improved than that of linear weighted fusion method.

In order to evaluate the effect of CNN combined with SAM method more comprehensively, we compare the AUC value of this method with that of other music automatic tagging methods on the same data set, and the results are shown in Figure 4.

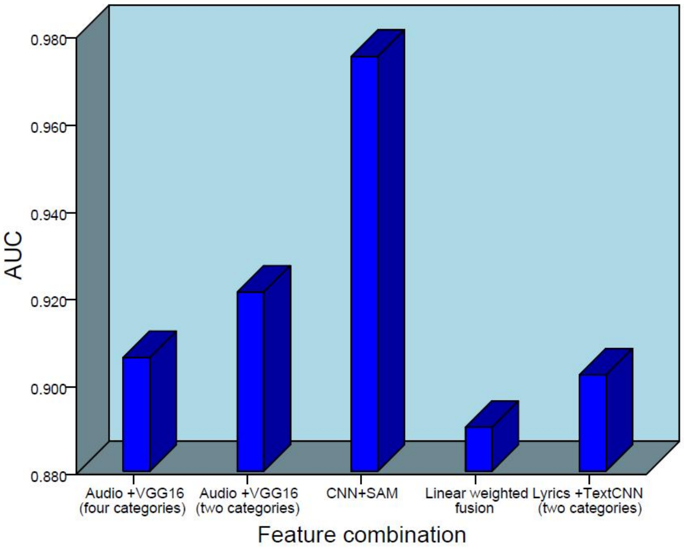

Figure 4. Comparison of experimental results of AUC value between this method and other methods

It can be seen that CNN combined with SAM method has the best music automatic annotation effect on MTAT data set at present. AUC value is 8.5% higher than that of linear weighted fusion method. The comparison results show that the multi-modal music automatic annotation model framework proposed by CNN and SAM in this paper is effective for the automatic annotation task of music.

4. CONCLUSION

The problem of music automatic labeling is related to the general regression problem of music classification, such as genre classification and emotion prediction. Under the background of the transformation of the music market, as a structured way of organizing music information, the concept of music label is of great importance. In this paper, the word vector of lyrics text is used as input information, and a multimodal music automatic labeling model based on CNN and SAM is proposed. The results show that the accuracy of CNN combined with SAM method is 3.9% higher than that of linear weighted fusion method, and the AUC value is 8.5% higher than that of linear weighted fusion method. The comparison results show that the multi-modal music automatic annotation model framework proposed by CNN and SAM in this paper is effective for the automatic annotation task of music.

REFERENCES

[1] Wang, Z. Y., Ray, Gao, Y. X., et al. Music automatic tagging algorithm based on tag depth analysis [J]. Journal of South China University of Technology: Natural Science Edition, 47(8):6 (2019).

[2] He, X. M., Music Common Semantic Annotation Based on Conditional Random Fields [J]. Electronic Measurement Technology, (8):5 (2016).

[3] He, L., Yuan, B., Classification of music genres using long-term and short-term memory networks [J]. Computer Technology and Development, 29(11):5 (2019).

[4] Li, Y. X., Li, Y. X., et al., Classification method of background music based on emotional characteristics [J]. Modern Electronic Technology, 40(15):4 (2017).

[5] Chen, S. J., Wang, C. Y., Xie, L., Music generation system based on the motion state of smart bracelet [J]. Journal of Zhengzhou University (Science Edition), 053(004):95-101 (2021).

[6] Dong, A. M., Liu, Z. Y., Yu, J. G., et al., Automatic classification of music genres based on visual transformation network [J]. Computer Application, 42(1):5 (2022).

[7] Morales-Kastresana, A., Telford, B., Musich, T. A., et al., Labeling Extracellular Vesicles for Nanoscale Flow Cytometry[J]. Scientific Reports, 7(1):1878 (2017).

[8] Yadati, K., Larson, M., Liem, C. C. S., et al., Detecting Socially Significant Music Events using Temporally Noisy Labels[J]. IEEE Transactions on Multimedia, 20(9):1-1 (2018).

[9] Chang, C. W., Christian, M., Chang, D. H., et al., Deep learning approach based on superpixel segmentation assisted labeling for automatic pressure ulcer diagnosis[J]. PLoS ONE, 17(2):e0264139 (2022).

[10] Ren, J. M., Wu, M. J., Jang, J. S. R., Automatic Music Mood Classification Based on Timbre and Modulation Features[J]. IEEE Transactions on Affective Computing, 6(3):236-246 (2015).

[11] Cazau, D., Wang, Y., Adam, O., et al., Calibration of a two-state pitch-wise HMM method for note segmentation in Automatic Music Transcription systems[J]. Chemistry - A European Journal, 18(28):8795-8799 (2017).

[12] Felouat, H., Oukid-Khouas, S., Graph matching approach and generalized median graph for automatic labeling of cortical sulci with parallel and distributed algorithms[J]. Cognitive Systems Research, 54(5):62-73 (2019).