MULTIMODAL MUSIC EMOTION RECOGNITION WITH HIERARCHICAL CROSS-MODAL ATTENTION NETWORK

Jiahao Zhao[1], *Ganghui Ru*[1], *Yi Yu*[2*], *Yulun Wu*[1], *Dichucheng Li*[1], *Wei Li*[1,3*]

[1] School of Computer Science and Technology, Fudan University, Shanghai, China
[2]Digital Content and Media Sciences Research Division, National Institute of Informatics (NII), Tokyo, Japan
[3] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9859812

ABSTRACT

Computational music emotion recognition is to recognize the emotional content in music tracks. In computational music emotion recognition studies, researchers have paid close attention to the audio content of the music tracks. Although lyrics content and music context contribute greatly to the perceived emotion, these kinds of emotional information are usually ignored. Based on this finding, we propose a multimodal music emotion recognition method jointly predicting the valence and arousal values by combining the audio, lyrics, track name, and artist of a given track. Audio features, lyrics features and context features are extracted separately and fused by a cross-modal attention mechanism, forming a hierarchical structure. Our proposed model outperforms two baselines by a large margin and achieves state-of-the-art performance on two public datasets.

Index Terms— Music Emotion Recognition, Multimodal Machine Learning, Deep Learning, Natural Language Processing

1. INTRODUCTION

Music Emotion Recognition (MER) is to recognize the emotional content in music tracks or the induced emotions of music listeners [1]. MER is a critical high-level task in Music Information Retrieval (MIR) and has attracted increasing attention in both the academic and the industrial communities due to its wide application prospects. MER can be used in many applications of information retrieval such as the categorization and recommendation of music tracks by emotion. However, MER is one of the most challenging tasks in MIR, due to its complicated nature and involvement of multiple disciplines such as music theory, music psychology, neuroscience, signal processing, and machine learning.

User-independent MER methods aim to compute perceived emotion with the music content (e.g., melody, rhythm) and music context (e.g., track name, artist, album photo) [1]. Early studies were mainly based on signal processing methods, and used some hand-crafted features (e.g., Mel-Frequency Cepstral Coefficients, Chroma). As the size of datasets grows rapidly, data-driven MER methods are becoming the mainstream of MER studies. Data-driven methods mainly use musical audio content, and have shown their great potential. Recently, Jacopo de Berardinis et al. proposed a MER model with promising performance by combining the Music Source Separation (MSS) task [2]. With the development of Natural Language Processing (NLP), many researchers also use lyrics content as the input. Y. Agrawal et al. proposed a lyric emotion recognition method based on Transformer [3].

Except for these uni-modal methods, multimodal MER methods have gradually shown their potential. In [4], Xiao Hu et al. designed a series of experiments proving that multimodal MER systems generally outperform both audio-only and lyrics-only systems. Further studies have also proven the importance of incorporating more different factors such as the lyrics content and the music context [5]. In those existing works we mentioned above, music context is available in most cases. Taking use of these metadata to improve MER performance is critical in practical scenarios. Nonetheless, to the best of our knowledge, no researchers have taken use of the music content as well as the music context in multimodal MER researches.

The main contributions of this paper are as follows: 1) We propose a method to improve MER task by using music context; 2) We design a hierarchical cross-modal attention network, which can extract and fuse features from low-level semantic information to high-level semantic information in different modalities; 3) Our proposed model shows state-of-the-art performance on two public datasets.

2. RELATED STUDIES

In this section, we introduce the taxonomies and the concept of induced&perceived emotions. Main idea of the multimodal methods and some significant multimodal MER studies are also introduced in detail.

* Corresponding authors. This work was supported by National Key R&D Program of China (2019YFC1711800) , NSFC (62171138) .

2.1. MER Backgrounds

The selection of taxonomy is first considered when designing an MER system. The overall taxonomies can be divided into two categories: 1) discrete label annotation (e.g., sad, happy, angry) ; 2) dimensional/continuous representation of valence and arousal. Both datasets in this paper use dimensional/continuous taxonomy, one is annotated with continuous valence and arousal values and the other is annotated with 4-Quadrant taxonomy based on Russel's emotional model [6].

As shown in Fig.1 **??**, the 4-Quadrant taxonomy actually classifies a music track according to its valence and arousal values. Particular thresholds (namely A_t and V_t) are used for the classification. Therefore, even though different taxonomies are adopted in both datasets we used, the two tasks are substantially the same. Note that in both datasets static labels are applied, which means there is only one annotation for a single music track. Some researchers also focus on dynamic label, Xinxing Li et al. proposed a DBLSTM-based multi-scale regression method to predict valence and arousal values dynamically [7].

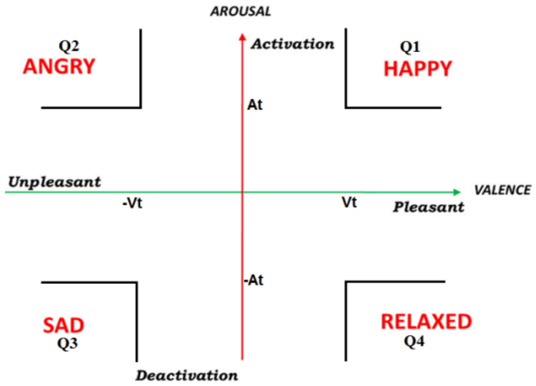

Fig. 1. 4-Quadrant classification adopted in [8], A_t and V_t are the arousal threshold and the valence threshold respectively.

Another conception we need to introduce is the induced and perceived emotion. Induced emotion refers to the actual emotion that a listener feels while listening to a music track. Calculations on induced emotion are uncertain and user-specific [9]. Perceived emotion refers to the emotional meaning in a music track decided by its content and context such as tempo, rhythm, lyric, track name, etc. As perceived emotion is more objective and independent of user context, it is the main issue of user-independent computational MER studies. Therefore, only perceived emotion is considered in this paper.

2.2. Multimodal MER Studies

Multimodal methods aim to process and relate information from multiple modalities (e.g., sound, image, language) [10]. Each modality holds its exclusive information, so complementarity exists in multiple modalities. Multiple modalities also show redundancy because similar or the same information exists in different modalities. Therefore, how to balance the complementarity and the redundancy is a critical issue in multimodal methods. This issue mainly relies on the feature fusion procedure, which can be divided into three categories: 1) Early fusion: Input features (e.g., spectrogram, Mel-spectrogram) are directly fused without further processing; 2) Middle fusion: Processed high-dimensional features are fused; 3) Late fusion: Output predictions (decisions) from each modality are fused. Middle fusion generally shows the best balance between complementarity and redundancy.

Multimodal methods in other multimedia areas are developing well and have shown many inspiring researches [11–14]. In MER studies, multimodal mechanism is applied in traditional machine learning methods in early researches. Yi-Hsuan Yang et al. proposed a bi-modal MER system with SVM as a classifier [15]. Recently, as more large datasets with lyrics become available, data-driven multimodal MER methods are more common. In [16], R. Delbouys et al. proposed a uni-modal audio model using Convolutional Neural Networks (CNN) and a uni-modal lyrics model using CNN and Long Short-Term Memory (LSTM) structure. The outputs of the two models were directly concatenated and then input into two fully connected layers for prediction. Their proposed model was inspiring for deep-learning-based multimodal MER methods. In [17], K. Pyrovolakis et al. proposed a bi-modal MER model using pre-trained BERT(Bidirectional Encoder Representations from Transformers) to process the lyrics and a CNN-based model to process the audio, the outputs were also directly concatenated. Their proposed model achieved state-of-the-art performance on the MoodyLyrics dataset [8].

Nonetheless, several improvements can be made in our opinion. We add a new modality of the music context (track name and artist) to improve the task. As the information of the track name and the artist is concise, we propose a hierarchical fusing strategy. Fusing the information from low-level semantics to high-level semantics with cross-modal attention mechanism is much more efficient than the direct concatenation adopted in the existing works such as [16] and [17].

3. METHODOLOGY

Our proposed model consists of three modules: the audio feature extraction module, the text feature extraction module, and the hierarchical cross-modal feature fusion module. The model predicts continuous valence and arousal values, where valence relates to pleasantness or positiveness and arousal refers to energy or activation. Note that when using 4-Quadrant taxonomy, an extra classifying layer is added.

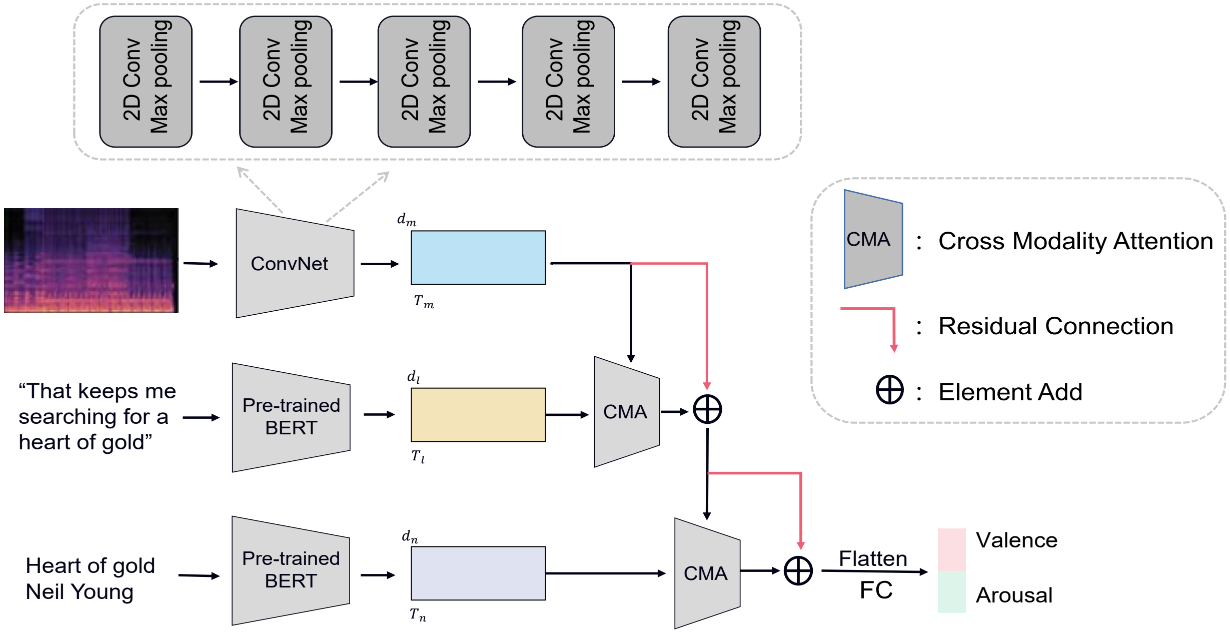

Fig. 2. The structure of our proposed model. The three branches extract features from the audio, text and context modality respectively. The CMA module fuses information from two modalities hierarchically. The fusion is applied from detailed low-level semantics to high-level semantics.

3.1. Audio Feature Extraction

The audio input is a Mel-spectrogram, calculated with 40 Mel-filters and 1024 sample-long Hanning window with no overlapping. This pre-processing procedure is adopted from [16]. A ConvNet module is used for feature extraction for its strong ability to extract and abstract audio features from spectrogram.

As shown on the top of Fig.2, the ConvNet module consists of five convolutional layers (Conv2D), each followed by a 2D max-pooling layer. The five convolutional layers have 32, 64, 128, 256, 128 filters respectively. Their convolution kernel size is set to 3x3, and the stride is set to 1. All the max-pooling layers share a window size of 2x2 to apply down-sampling. Batch Normalization (BN) is also applied in each convolutional layer in order to prevent over-fitting. Extracted features are flattened and spliced in the channel dimension as the output.

3.2. Text Feature Extraction

As shown in Fig.2, two branches processing lyrics and track name&artist share this text feature extraction structure. A pre-trained BERT (Bidirectional Encoder Representations from Transformers) [18] is used for extracting the text features. BERT has got large corpus and strong ability to extract semantic information from text. Therefore we choose it instead of LSTM with a limited corpus. The input is split into several segments and their average feature will be the output when its size is bigger than the input limitation (128). As BERT is a famous and well-developed model in NLP and we did not change or improve its structure, we do not explain more details here.

3.3. Cross-Modal Attention

The Cross-Modal Attention (CMA) module [19] is used for fusing the features from two modalities. In this module, an 8-headed attention mechanism is adopted. The specific calculation of a single head is shown in Eq.1 below:

$$\begin{aligned} Attention(F_Q, F_K, F_V) &= softmax(\frac{F_Q F_K^T}{\sqrt{d}})F_V \\ &= softmax(\frac{F_\alpha W_Q (F_\beta W_K)^T}{\sqrt{d}})F_\beta W_V \qquad (1) \end{aligned}$$

Where F_Q, F_K, F_V represent $Query$, Key and $Value$ in attention mechanism respectively, F_α and F_β are the extracted features of modality α and β, and d is the dimension size, W_Q, W_K, W_V are the learnable weight matrix of $Query$, Key and $Value$ respectively. The 8-headed attention mechanism can be defined as:

$$MHA(F_\alpha, F_\beta) = Concat(head_1, \ldots, head_8)W_O \qquad (2)$$

$$head_i = Attention(F_\alpha W_Q^i, F_\beta W_K^i, F_\beta W_V^i) \qquad (3)$$

Where W_O is a learnable weight matrix. This 8-headed attention mechanism emphasizes the important parts of each modality while direct concatenation fails to do so. As shown in Fig.2, a residual connection is also applied when calculating cross-modal attention, which helps the ConvNet converge by directly propagating the gradients to it. It can also avoid gradient vanishing or exploding. The overall structure of the CMA module is shown in Fig.3.

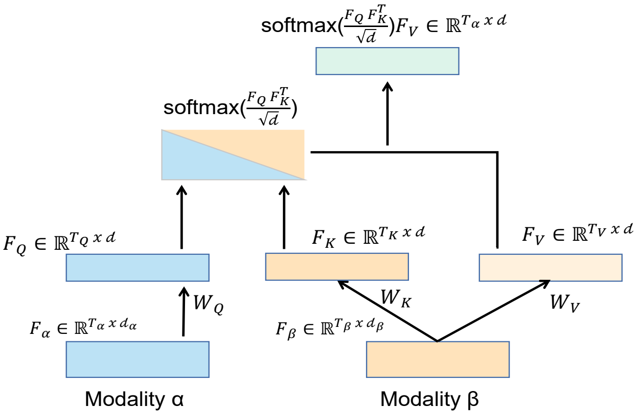
$\text{softmax}(\frac{F_Q F_K^T}{\sqrt{d}})F_V \in \mathbb{R}^{T_\alpha \times d}$
$\text{softmax}(\frac{F_Q F_K^T}{\sqrt{d}})$
$F_Q \in \mathbb{R}^{T_Q \times d}$
$F_K \in \mathbb{R}^{T_K \times d}$
$F_V \in \mathbb{R}^{T_V \times d}$
W_Q
W_K
W_V
$F_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$
$F_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$
Modality α
Modality β

Fig. 3. Structure of the cross-modal attention module, which fuses output features from modality α and β.

3.4. Hierarchical Structure

A common implementation of using music context is to concatenate the track name and artist to the lyrics input, and process the input by a single text model. However, track name is the concise summary of a music track given by the artist, and the name of the artist is also highly relevant to the genre or the perceived emotion of the music track [5]. When processed together with the lyrics, concise information in music context can not be emphasized and used properly, interference also occurs in this case. Therefore, we use track name and artist as a new modality and process them individually.

Except for the feature extracting procedure, the sequence that features fuse is also critical. So we design a hierarchical network structure to fuse the information from the three modalities in a proper sequence. Since audio modality and lyrics modality hold the most details and low-level semantics, fusing these two modalities firstly helps abstract high-level semantic information. Then the abstracted information is fused with features from track name and artist, which provide complementary information. As shown in Fig.2, the fusion is applied hierarchically from the top to the bottom, contributing to obtain a final representation with richer high-level semantics.

4. EXPERIMENTS

In this section, we carry out series of experiments on two public datasets. Discussion on the experimental results and the ablation studies shows the effect of the proposed improvements.

4.1. Dataset

Our proposed model is trained and tested on two public datasets: Million Song Dataset [20] annotated by Deezer (MSDD) [16] and the MoodyLyrics dataset [8]. As shown in Table 1, MSDD has 18,644 labeled songs, along with its Deezer song identifiers, MSD identifiers, artist and track titles, annotated with its valence and arousal values. MoodyLyrics has 2,595 labeled songs, along with its track titles and artists, annotated with a label from a set of four moods {happy, angry, sad, relaxed} based on Russel's model [6].

Table 1. Details of MSDD and MoodyLyrics Datasets

Dataset	Track Amount	Taxonomy
MSDD	18,644	valence & arousal values
MoodyLyrics	2,595	Russel's 4Q classification

In the MSDD dataset, 2000 songs are selected randomly in the provided metadata list and are spilt into 3 subsets with approximately 60%, 20%, 20% songs respectively for the training set, the validation set, and the test set. Note that the audio files and lyrics files are gathered from the web by provided metadata (track title, artist name, album title).

In the MoodyLyrics dataset, 2000 tracks balanced in 4 moods are selected from 2595 original tracks in [8], audio files and lyrics files are gathered in the same way as the MSDD dataset, and are split into 3 subsets with approximately 70%, 10%, 20% songs respectively for the training set, the validation set, and the test set. For fair comparison, we split the dataset in the same way as in [17].

In both datasets, data augmentation is applied by doing pitch shifting and downsampling.

4.2. Metrics

Different metrics is adopted on two datasets due to the difference of their taxonomies. In the MoodyLyrics dataset, F1-measure is used to evaluate the classification task. In the MSDD dataset, coefficient of determination (R^2) is used to evaluate the regression task, which is calculated as follows:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

Where n is the number of the evaluated samples, $\hat{y}_i$ is the i_{th} predicted value, y_i is the i_{th} ground truth value, $\bar{y}$ is the average value of all n ground truths. R^2 is a common metric for evaluating the fitting degree of two variables.

Table 2. Comparison between existing approaches with R^2 evaluated on the MSDD dataset, F1 measure evaluated on the MoodyLyrics dataset. Results in the second row are provided by [16], therefore are not evaluated on the MoodyLyrics dataset.

Method	$\mathrm{R}^2 - valence$	$\mathrm{R}^2 - arousal$	$\mathrm{R}^2 - mean$	F1 measure
Our Proposed Model	**0.306**	**0.311**	**0.309**	**96.14%**
R. Delbouys'model [16]	0.219	0.232	0.226	/
R. Delbouys' model_Reproduced	0.197	0.242	0.220	86.23%
K. Pyrovolakis's model [17]	0.258	0.270	0.264	94.32%

Table 3. Ablation Studies with R^2 evaluated on the MSDD dataset, F1 measure evaluated on the MoodyLyrics dataset.

Method	$\mathrm{R}^2 - valence$	$\mathrm{R}^2 - arousal$	$\mathrm{R}^2 - mean$	F1 measure
Our Proposed Model	**0.306**	**0.311**	**0.309**	**96.14%**
Non-Hierarchical	0.289	0.297	0.293	95.62%
Direct Fusion	0.263	0.279	0.271	93.87%
Bi-Modalities	0.278	0.291	0.285	95.28%

4.3. Results & Discussion

We trained and evaluated our proposed model on both datasets using the configuration mentioned above respectively. We also reproduced R. Delbouys' model [16] and K. Pyrovolakis' model [17] for comparison. The results are shown in Table 2, note that R^2 is evaluated on the MSDD dataset and the F1 measure is evaluated on the MoodyLyrics dataset.

Our proposed model outperformed R. Delbouys' model [16] and the reproduced model by a large margin in both valence&arousal metrics and F1 measure. The reproduced model showed slightly weaker performance, which was most likely caused by the reduced size of the dataset and the change of the data distribution. The improvement on valence prediction is more significant. This is mainly because our model takes better use of lyrics content and music context, which is highly interrelated to the valence element in music track.

We also compared our model with K. Pyrovolakis' model [17], which also used pre-trained BERT for lyrics analysis and showed state-of-the-art performance on the MoodyLyrics dataset. Although two models use same architecture in text feature extraction, their performances on valence prediction are quite different. The gap on valence prediction are even bigger than that on arousal prediction (0.048 and 0.041 respectively). This indicates that the hierarchical cross-modal attention mechanism can extract and emphasize the text information much better than direct concatenation. The results show that our model has better performance on both datasets than K. Pyrovolakis' model.

As there are only few researches on multimodal MER, we are not able to compare our proposed model with more baselines. Based on the results above, our model outperformed two baselines and achieved state-of-the-art performance on both public datasets.

4.4. Ablation Studies

In the ablation studies we evaluated three components: 1) Non-hierarchical: We fuse these modalities in a different sequence instead of fusing them hierarchically, in this experiment we fuse the audio modality and the context modality firstly; 2) Direct fusion: We directly concatenate the output of each modality after padding; 3) Bi-modalities: We only use the lyrics and the audio as the input and fuse the two modalities using cross-modal attention mechanism. We use these three experiments to investigate the effectiveness of the hierarchical fusing strategy, the CMA module and the additional context modality respectively.

The results are shown in Table 3. Compared with the non-hierarchical model, our model outperforms on all metrics, which indicates that the hierarchical fusing strategy are effective. The direct fusion model shows the worst performance because it is unable to balance the complementarity and redundancy of information from all three modalities, and the concise information from the context modality can not be extracted and emphasized without the attention mechanism. This proves that the CMA module contributes greatly to the total improvement. By comparing with the bi-modalities model, we can see that additional music context brings considerable improvement on all metrics.

By comparing the results of all three experiments, we can see that all three models show much worse performance on the valence prediction than on the arousal prediction. As the accuracy of the valence prediction is much lower than that of the arousal prediction in most recent researches [16, 21], our method can be used to solve this problem to some extent.

5. CONCLUSION

In this paper, we proposed a multimodal neural network with a hierarchical cross-modal attention mechanism for the MER task. Our proposed model uses music content (audio content, lyrics content) as well as music context (track name and artist) to predict the emotion of a given music track. Ablation studies have proved the effectiveness of the hierarchical fusing strategy, and our model achieved state-of-the-art performance on both MSDD and MoodyLyrics datasets.

Restricted by the large amount of work, we were unable to use the full MSDD dataset, experiments on full dataset will be carried out in our future works. For other future works in MER, we are currently trying to map the emotion labels (e.g. happy, sad) to a higher-dimensional semantic space, and therefore use a graph convolutional neural network for more efficient learning.

6. REFERENCES

[1] Juan Sebastián Gómez Cañón, Estefanía Cano, Tuomas Eerola, Perfecto Herrera, Xiao Hu, Yi-Hsuan Yang, and Emilia Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 106–114, 2021.

[2] Jacopo de Berardinis, Angelo Cangelosi, and Eduardo Coutinho, "The multiple voices of musical emotions: Source separation for improving music emotion recognition models and their interpretability," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*, 2020, pp. 310–317.

[3] Yudhik Agrawal, Ramaguru Guru Ravi Shanker, and Vinoo Alluri, "Transformer-based approach towards music emotion recognition from lyrics," in *European Conference on Information Retrieval*. Springer, 2021, pp. 167–175.

[4] X Hu, K. Choi, and J. S. Downie, "A framework for evaluating multimodal music mood classification," *Journal of the American Society for Information Science*, vol. 68, no. 2, pp. 273–285, 2017.

[5] M. Schedl, A. Flexer, and Julian Urbano, "The neglected user in music information retrieval research," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 523–539, 2013.

[6] James A Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161, 1980.

[7] Xinxing Li, Jiashen Tian, Mingxing Xu, Yishuang Ning, and Lianhong Cai, "Dblstm-based multi-scale fusion for dynamic emotion prediction in music," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[8] Erion Çano and Maurizio Morisio, "Moodylyrics: A sentiment annotated lyrics dataset," in *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, 2017, pp. 118–124.

[9] Wei-Hao Chang, Jeng-Lin Li, Yun-Shao Lin, and Chi-Chun Lee, "A genre-affect relationship network with task-specific uncertainty weighting for recognizing induced emotion in music," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[10] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.

[11] Andrew Owens and Alexei A Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[12] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba, "Music gesture for visual sound separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10478–10487.

[13] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba, "The sound of motions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1735–1744.

[14] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba, "The sound of pixels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 570–586.

[15] Yi-Hsuan Yang, Yu-Ching Lin, Heng-Tze Cheng, I-Bin Liao, Yeh-Chin Ho, and Homer H Chen, "Toward multimodal music emotion classification," in *Pacific-Rim Conference on Multimedia*. Springer, 2008, pp. 70–79.

[16] Rémi Delbouys, Romain Hennequin, Francesco Piccoli, Jimena Royo-Letelier, and Manuel Moussallam, "Music mood detection based on audio and lyrics with deep neural net," *arXiv preprint arXiv:1809.07276*, 2018.

[17] Konstantinos Pyrovolakis, Paraskevi Tzouveli, and George Stamou, "Mood detection analyzing lyrics and audio signal based on deep learning architectures," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 9363–9370.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[19] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, 2019, vol. 2019, p. 6558.

[20] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere, "The million song dataset," 2011.

[21] Yizhuo Dong, Xinyu Yang, Xi Zhao, and Juan Li, "Bidirectional convolutional recurrent sparse network (bcrsn): An efficient model for music emotion recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3150–3163, 2019.