Channel-wise Attention Mechanism in Convolutional Neural Networks for Music Emotion Recognition

Xi Chen[1], Lei Wang[1], Andi Pan[1], and Wei Li[1,2(✉)]

[1] School of Computer Science, Fudan University, Shanghai 201203, China
{19210240230,18210240192,18210240151,weili-fudan}@fudan.edu.cn
[2] Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 201203, China

Abstract. Music Emotion Recognition (MER), a subject of affective computing, aims to identify the emotion of a musical track. With the fast development of deep learning, neural networks, such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), have been applied recently. However, while using convolutional kernels, channels are treated equally, which means treating different aspects (such as tempo and vibrato related features) of a music clip equally. It's against the rule of human perception. Therefore, Channel-wise Attention Mechanism is introduced into the task of Music Emotion Recognition. The performance could be improved to a certain extent.

Keywords: Music Emotion Recognition · Channel-wise Attention Mechanism

1 Introduction

Music Emotion Recognition (MER), a subject of both Music Information Retrieval and Affective Computing, aims to identify the emotion conveyed by a musical clip [18]. Driven by large demands in the music industry such as providing a content-oriented categorization scheme, generating playlist automatically and music recommendation [6,10,22], MER has developed rapidly in recent years [1].

Traditional methods are feature-based ones. The most commonly used acoustic features (e.g. Mel Frequency Cepstrum Coefficient, spectral shape in timber, Chromagram and Rhythm strength) are summarized in [9]. On the one hand, since there are hundreds and thousands of features to be considered, feature selection methods [16] for removing redundant ones and principle component analysis (PCA) methods [14] for dimension reduction are introduced. On the other hand, manual features sophisticated designed to express the nature of music emotion have always been an interest in the field. Since most features are low level and related to tone color, in 2018, [15] designed musical texture related and expressive technique related features.

X. Shao et al. (Eds.): CSMT 2020, LNEE 761, pp. 43–54, 2021.
https://doi.org/10.1007/978-981-16-1649-5_4

In recent years, with the development of computer hardware and the large available online data, deep learning has demonstrated its great power in many fields including Music Emotion Recognition. Instead of designing specific features which is a challenging task [12] and demands large human labor, a Neural Network itself could extract the very pertinent representations.

Besides simply using deep learning models, such as CNN used in [12], different mechanisms are used to improve the performance. Similar to the multitask learning theory, hoping the first several layers to extract commonly acoustic features and the last several layers to extract target-oriented features, [13] stacked one CNN layer with two RNN branches for arousal and valence regression. Some hope to utilize auxiliary information. Inspired by speech emotion recognition tasks considering spoken content [19], [5] proposed a multimodal architecture based on audio and lyric. In [11], additional harmonic and percussive features are fed into the bi-directional LSTM model. Because of the lacking of training data, others also use the transfer learning method, aiming to make use of excellent features in related domains [3].

As we can see above, despite different mechanisms, the base architectures are usually CNN. It is known that a convolutional neural network learns hierarchical features from level to level [21] and that a higher-layer feature maps depend on lower-layer maps [2]. For instance, in the early layers, low-level information such as tempo, pitch, (local) harmony or envelop might be extracted [3], while high-level semantic patterns such as expressivity and musical texture features would be detected [15] in later layers. However, on the one hand, while doing convolutional operations, the low-resolution features which contain abundant low-frequency information are treated equally across channels [23] (i.e. tempo and pitch information may not contribute exactly the same), hence, the extracted features are not powerful enough. On the other hand, without processing the whole music clip, only understanding a few important aspects, one could recognize its emotion, whereas, a CNN would process all the feature maps which is in contrast to human perception [4].

Fortunately, the problems existed could just be solved by the Channel-wise Attention Mechanism, which has been successfully applied in Computer Vision [20], Natural Language Processing, and Speech Processing. The Channel-wise Attention Mechanism could re-weight feature maps in channels. Moreover, since a feature map is computed from earlier ones, it is natural to apply attention mechanism in multiple layers [2]. In this way, multiple semantic abstractions could be gained [2]. The applied Channel-wise attention mechanism is a sophisticatedly chose one, detailed in Sect. 2.

Music Emotion Recognition tasks could be categorized into a classification one and a regression one. The proposed method is tested on both tasks and the performance has been improved. It should be mentioned that, since public music emotion classification datasets are small, which will even limit the performance of the baseline network, a larger music emotion classification dataset is made.

In summary, the contribution of this paper could be put as follows:

(I). To solve the problem of treating each feature map equally while recognizing musical emotion patterns, Channel-wise Attention Mechanism is applied in multi-layers.
(II). The Channel-wise Attention Mechanism is a sophisticated chosen one.
(III). The procedure of how to make a large musical emotion classification dataset is introduced.
(IV). The proposed method could be proven useful to a certain extent.

2 Proposed Method

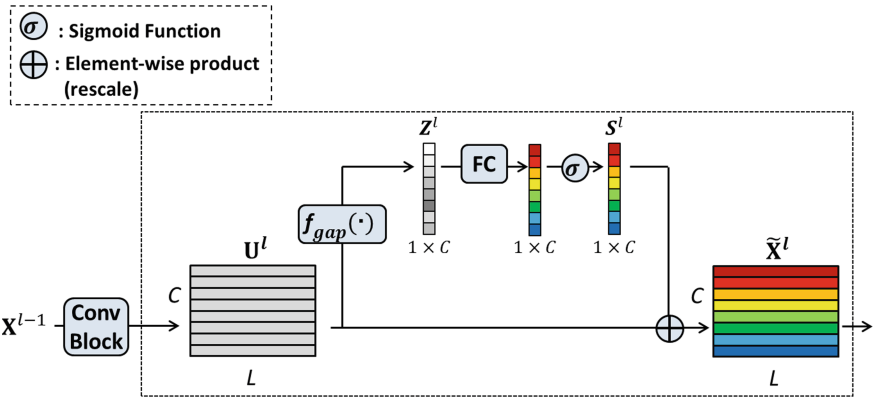

Fig. 1. The Channel-wise Attention Mechanism for Music Emotion Recognition

There are many sophisticated channel-wise attention mechanisms in literature, such as that in SENet [7], that in RCAB [23]. Obviously, we should not simply draw one of them to use, we should choose or design one on our needs. Firstly, to fully consider the interrelationships among all channels, the channel-wise attention mechanism is designed with the fully connected layer and the activation function, like that in SENet [7], rather than using convolutional ones whose receptive field is limited to only a few channels, such as that in RCAB [23]. Secondly, to learn a non-mutually-exclusive relationship, some channel-wise mechanism is under the mode of an encoder and a decoder scheme. However, after an encoder operation, whether it depends on doting with a weight matrix or convolutional operations, the rank is decreased after encoding, meaning some information from the feature map (though less important) will be lost. This is undesired. Thirdly, considering of computational efficiency, the designed channel-wise attention mechanism is lightweight.

Next, the proposed channel-wise attention mechanism and the backbone architecture will be introduced.

2.1 Channel-wise Attention Mechanism

The channel-wise attention mechanism block is a transformation block. On the above reasons, it is designed with a simple gating mechanism with an activation function. Figure 1 illustrates the mechanism along with the operations and the variables.

As a whole, it is a reweighting operation from $\mathbf{U}^l \in \mathbb{R}^{L \times C}$ to $\tilde{\mathbf{X}}^l \in \mathbb{R}^{L \times C}$, where $\mathbf{U}^l$ is the output feature map with the length of L and channel number of C after the l-th convolutional block with input $\mathbf{X}^{l-1}$. All of the superscripts in notation refer to the layer index.

First, each convolutional kernel with a fixed size receptive field serves as a local semantic information extractor. Therefore, each or some of the value in a feature map could not represent what the channel learns [7]. To mitigate this problem, the channel-wise statistic $\mathbf{Z}^l = \left[{z_1}^l,\ {z_2}^l,\ \ldots,\ {z_c}^l,\ \ldots,\ {z_C}^l\right]$ obtained by using global average pooling $\boldsymbol{f}_{gap}(\cdot)$ is used as the channel feature descriptor. For detail, ${z_c}^l$ is calculated by:

$$z_c^l = \boldsymbol{f}_{gap}\left({\boldsymbol{x_c}}^l\right) = \frac{1}{L}\sum_{i=1}^{L} {\boldsymbol{x_c}}^l(i) \tag{1}$$

More sophisticated channel descriptors could also be considered.

Next, inter-channel dependencies will be exploit by Eq. (2):

$$\mathbf{S}^l = \sigma\left(g\left(\mathbf{Z}^l\right)\right) = \sigma\left(\mathbf{W}^l \cdot \mathbf{Z}^l\right), \tag{2}$$

Where $\sigma(\cdot)$ denotes the sigmoid activation and $\mathbf{W}^l \in \mathbb{R}^{C \times C}$. Obviously, $g(\cdot)$ could also be interpreted as a fully connected layer with $\mathbf{W}^l$ as the corresponding parameter.

Finally, the attended feature map could be obtained by modulating $\mathbf{U}^l$ with $\mathbf{S}^l$, for each channel

$$\tilde{x}_c^l = f_{\mathbf{rs}}\left(u_c^l,\ s_c^l\right). \tag{3}$$

In Eq. (3) $f_{\mathbf{rs}}$ means rescaling u_c^l with scalar s_c^l.

2.2 Backbone Architecture

The backbone architecture we adopted is following the audio subnet of the Audio-Lyric Bimodal in [5]. It is originally used for the emotion value regression task. While in this paper, the backbone architecture would be applied to both regression and classification with different outputs.

It is composed of two convolutional blocks and two dense blocks. For one thing, as for the convolutional block, a convolutional layer, a max pooling layer and batch normalization are consecutive. The (the number of kernels, kernel size, stride) for convolutional layer are (32, 8, 1) and (16, 8, 1) separately, while the (kernel size, stride) for the max pooling layer are all in (4, 4). For another thing, the dense block includes a dropout and a fully connected layer. The intermedia neural number for the two dense blocks is 64 [5].

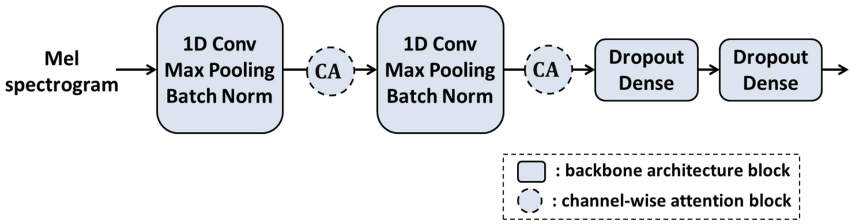

Fig. 2. Music Emotion Classification Model with dotted line representing the attention block and solid line representing the backbone architecture block

3 Evaluation

Datasets, metrics and experiment settings will be talked here. More, the method for making a large music emotion classification dataset is presented under the hope of helping other researchers to design much more powerful systems.

3.1 Dataset

Music Emotion Recognition tasks tend to use either categorical psychometrics or scalar/dimensional psychometrics for classification or regression [9]. Both music emotion representations are under the supporting of psychological theories [9]. Under categorical approaches, emotion tags are clustered into several classes. The well-known MIREX Audio Mood Classification Competition just uses this kind of psychometric [8]. While under continuous descriptors, a certain kind of emotion could be represented by a point in the Valence-Arousal (V-A) space [17]. Though there are two kinds of music descriptors, under the Circumflex Model of Affect [17], they could be transformed to each other.

Classification Task Dataset. For the reason that public music emotion classification datasets are small, containing only less than 1,000 clips, even the baseline deep learning neural network could not demonstrated its great power, not to mention the proposed channel-wise attention mechanism. Hence, under the guidance of the Circumflex Model of Affect by Russell [17], with the help of emotion related playlist (those which have been created intentionally and listened millions of times) on the mainstream music software, a large reliable dataset with thousands of music clips could be made with less human labor.

To begin with, a set of music emotion tags are chose according to human experience and psychology theory. Six tags, Stirring, Empowering, Angry, Somber, Peaceful and Upbeat are finally determined. Definitely, they are under the constrain of Circumflex Model of Affect, sparsely located on the model, which means that the gap between music emotion tags are large enough to make them separate well. Figure 3 will illustrate the relationship between the music tags and the Circumflex Model of Affect.

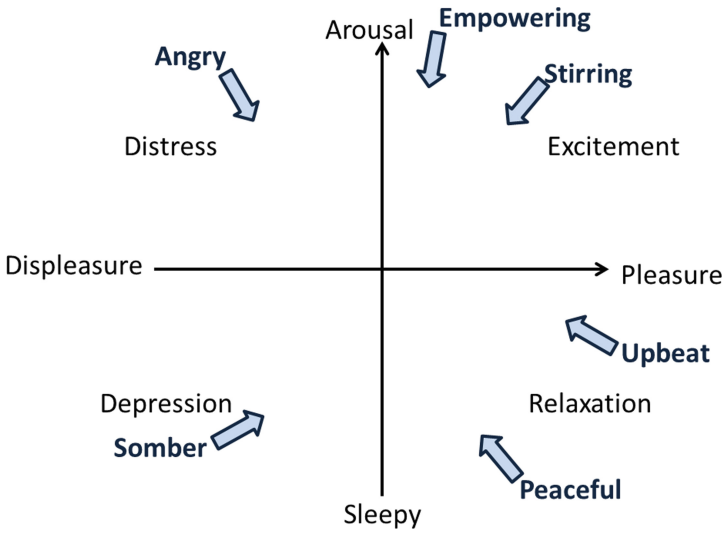

Fig. 3. Mapping the selected music emotion tags on the Circumflex Model of Affect

Next, searching the tag related playlists on popular music website such as NetEase cloud music and QQMusic, top played ones would be considered. By referencing to the comments of the playlists and humanly verifying each song, the final ones would be determined. After that, for each song, the first 5 s would be thrown away considering the emotion there might be different from the whole song, and then they would be cut into 30 s ones.

Finally, using this method, more than 4,000 music clips with sample rate of 44,100 are got.

Since annotated music excerpts are collected from website and are copyrighted, the dataset could not be made public. However, using the above mentioned method, researchers could make their own dataset easily.

Regression Task Dataset. As for the baseline architecture, the used continuous descriptor is song leveled, it uses an arousal value and a valence value to describe a 30 s music expert. Unfortunately, the dataset is not a public one. Mainstream public ones are all dynamically annotated, which means that they consider emotion variation in a music [1] and are annotated every once in a while. To mitigate this problem, we choose the averaged value of all annotations in a song to represent the song level descriptor.

For the dynamically annotated public dataset, we utilize a largest one, Emotional Analysis in Music (DEAM) [1]. It contains 1,802 songs including 1,744 45 s clips and 58 full length clips. The time resolution for annotation 2 Hz, meaning annotating per 500 ms. The annotated values are scaled in $[-1, +1]$.

In our experiment, since baseline architecture is designed for 30 s clips, only the middle 30 s (from the 7th to the 36th second of the clip) audio is preserved.

And, 58 full length clips are too long, using the song level averaged annotation to represent each segment is not a smart choice, therefore, they are thrown away. Finally, after processing, 1,744 clips are remained.

3.2 Metric

In notation, N, y^i, $\hat{y^i}$ corresponds to the number of samples, predicted label/value and real label/value separately, where $i \in [0, N-1]$.

Metric for Music Emotion Classification Task. Accuracy score, shorten as acc, represents the ratio of correctly classified samples to total number, could be written as:

$$acc\left(\hat{y^i}, y^i\right) = \frac{1}{N}\sum_{i=0}^{N-1} 1\left(\hat{y^i} == y^i\right). \tag{4}$$

Confusion matrix shows more detail information than acc. It is much easier to see how the system confusing among them. Let $\mathbf{C}$ to be matrix, $\mathbf{C}_{i,j}$ means the proportion of sample observed in class i but classified into class j. Specifically, $\mathbf{C}_{i,i}$ corresponds to the accuracy score for the i-th class.

Matric for Music Emotion Regression Task. Root Mean Square Error (RMSE) is a typical matric for regression tasks, meaning how far is the predicted value from the real one.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N-1}\left(\hat{y^i} - y^i\right)^2}. \tag{5}$$

3.3 Settings

As mentioned in Sect. 2, the backbone architecture followed from [5] will be used for both regression and classification tasks.

For each audio clip, after upsampling to 44,100 Hz, an Mel-spectrogram is extracted with 40 mel bands, 1024 sample long Hann window with no overlapping as input [5]. In baseline, it uses pitch shifting to argument the data. However, pitch is an emotion related feature [11,15,18]. Therefore, data argument method in baseline is not adapted.

While training, we use Cross Entrpy loss for classification task, Mean Square Error loss for regression task, Adam is the optimizer.

4 Experiments and Results

In this section, experiments conducted to validate the effectiveness of the proposed method will be presented. In notation, AudioNet represents the baseline, Layer1, Layer2, LayerALL means the location of the added channel-wise attention mechanism (i.e., Layer1 meaning adding the attention mechanism after the first layer).

4.1 Validating the Proposed Method

In the first set of experiments, we would like to validate the power of the proposed attention mechanism's power and that of the multi-layer attention scheme by using the classification dataset. Since the baseline has two convolutional blocks, three architectures'performances will be evaluated, AudioNet, AudioNet_Layer1, AudioNet_Layer2, AudioNet_LayerALL.

Table 1. Overall and class level accuracy score for the baseline and attention added ones in different locations.

Overall accuracy	
Network	On Whole Dataset
AudioNet	0.665
AudioNet_Layer1	0.695
AudioNet_Layer2	0.736
AudioNet_LayerALL	**0.741**

Class Level accuracy						
Network	Angry	Somber	Upbeat	Peaceful	Empower	Stirring
AudioNet	0.792	0.583	0.736	0.626	0.657	0.458
AudioNet_Layer1	**0.877**	**0.613**	0.725	0.712	0.614	0.615
AudioNet_Layer2	0.836	0.603	0.750	0.720	**0.754**	**0.789**
AudioNet_LayerALL	0.866	0.557	**0.895**	**0.818**	0.732	0.567

Experiment results will be seen in Table 1. As we can see, whether adding channel-wise attention after layer 1 or layer 2, the performance could be improved distinctly; this can verify channel-wise attention's ability. When adding attention mechanism after all layers, the performance could be improved further, this could demonstrate the rationality of adding attention to multi layers.

It is interesting to find that the architecture adding attention to the later layer will demonstrate more power than that adding to the earlier one. The reason behind might be that earlier layers extract low-level represents while later ones extract class-specific features [7]. Emphasizing more on class-specific features will help more than extracting better common music characteristics, for example, better musical texture features will help more than pitch features in music emotion recognition [15].

4.2 Performance on Classification Task

Performance in overall between baseline and the proposed one has been demonstrate above, Table 2 gives more detail by using confusion matrix.

Table 2. Confusion matrix for baseline and the proposed

AudioNet						
	Angry	Somber	Upbeat	Peaceful	Empowering	Stirring
Angry	**0.792**	0.028	0.028	0.000	0.083	0.069
Somber	0.000	**0.583**	0.042	0.188	0.021	0.167
Upbeat	0.014	0.028	**0.736**	0.000	0.167	0.056
Peaceful	0.009	0.261	0.017	**0.626**	0.000	0.087
Empowering	0.171	0.029	0.071	0.014	**0.657**	0.057
Stirring	0.125	0.083	0.042	0.125	0.167	**0.458**
AudioNet_LayerALL						
	Angry	Somber	Upbeat	Peaceful	Empowering	Stirring
Angry	**0.866**	0.015	0.015	0.000	0.015	0.090
Somber	0.011	**0.557**	0.034	0.250	0.000	0.148
Upbeat	0.000	0.000	**0.895**	0.000	0.070	0.035
Peaceful	0.000	0.143	0.000	**0.818**	0.000	0.039
Empowering	0.134	0.037	0.085	0.000	**0.732**	0.012
Stirring	0.133	0.067	0.100	0.000	0.133	**0.567**

Except for the overall accuracy, in the six classes, five classes' accuracy scores have been lift.

As seen in baseline's result, $\mathbf{C}_{Somber,Peaceful}$ and $\mathbf{C}_{Peaceful,Somber}$ are both not small. Since Somber and Peaceful are both less arousal, the network tend to be confused with each other. If more attention is put on valence related features, this phenomenon could be eased to some extent. After adding channel-wise attention mechanism, though accuracy for Somber has been reduced by 0.026, that for Peaceful has been improved by 0.192. This illustrates channel-wise attention mechanism's ability to re-weight and concentrate more on target-related feature maps.

As for Stirring, the baseline's accuracy score for which is the lowest. It is easily misclassified into Angry or Empowering because they are all more arousal. After adding the attention mechanism, the accuracy score for it has been improved.

4.3 Performance on Regression Task

Since the baseline has been proposed for song level emotion detection, rather than dynamic one, and there is no such type of dataset, we processed dynamic annotations in public dataset to generate the corresponding song level one, detailed in Sect. 3.2. In experiment, we conduct two set of experiments for arousal regression and valence regression.

Table 3. RMSE for the baseline and the proposed

Network	Arousal	Valence
AudioNet	0.301	**0.251**
AudioNet_LayerALL	**0.284**	0.253

As seen in Table 3, for arousal regression, the proposed performs better with a obviously smaller RMSE, while for valence regression, the baseline performs slightly better.

5 Conclusion

In this paper, channel-wise attention mechanism is introduced and designed to make the network focus more on the emotion related feature maps. Experiment results have verify the utility of the proposed method on both classification and regression tasks. In future work, we will concentrate more on the attention scheme, such as designing more sophisticate and accurate channel descriptors or introducing spatial attention mechanism.

Ackowlegement. This work was supported in part by National Key R&D Program of China (2019YFC1711800), NSFC (61671156).

References

1. Aljanaki, A., Yang, Y.H., Soleymani, M.: Developing a benchmark for emotional analysis of music. PloS One **12**(3), e0173392 (2017)
2. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.: SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 6298–6306. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.667

3. Choi, K., Fazekas, G., Sandler, M.B., Cho, K.: Transfer learning for music classification and regression tasks. In: Cunningham, S.J., Duan, Z., Hu, X., Turnbull, D. (eds.) Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, 23–27 October 2017, pp. 141–149 (2017). https://ismir2017.smcnus.org/wp-content/uploads/2017/10/12_Paper.pdf
4. Corbetta, M., Shulman, G.L.: Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci. **3**(3), 201–215 (2002)
5. Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J., Moussallam, M.: Music mood detection based on audio and lyrics with deep neural net. In: Gómez, E., Hu, X., Humphrey, E., Benetos, E. (eds.) Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, 23–27 September 2018, pp. 370–375 (2018). http://ismir2018.ircam.fr/doc/pdfs/99_Paper.pdf
6. Flexer, A., Schnitzer, D., Gasser, M., Widmer, G.: Playlist generation using start and end songs. In: Bello, J.P., Chew, E., Turnbull, D. (eds.) ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, 14–18 September 2008, pp. 173–178 (2008). http://ismir2008.ismir.net/papers/ISMIR2008_143.pdf
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. CoRR abs/1709.01507 (2017). http://arxiv.org/abs/1709.01507
8. Hu, X., Downie, J.S., Laurier, C., Bay, M., Ehmann, A.F.: The 2007 MIREX audio mood classification task: lessons learned. In: Bello, J.P., Chew, E., Turnbull, D. (eds.) ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, 14–18 September 2008, pp. 462–467 (2008). http://ismir2008.ismir.net/papers/ISMIR2008_263.pdf
9. Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G., Richardson, P., Scott, J.J., Speck, J.A., Turnbull, D.: State of the art report: music emotion recognition: a state of the art review. In: Downie, J.S., Veltkamp, R.C. (eds.) Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, 9–13 August 2010, pp. 255–266. International Society for Music Information Retrieval (2010). http://ismir2010.ismir.net/proceedings/ismir2010-45.pdf
10. Li, T., Ogihara, M.: Content-based music similarity search and emotion detection. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, 17–21 May 2004, pp. 705–708. IEEE (2004). https://doi.org/10.1109/ICASSP.2004.1327208
11. Liu, H., Fang, Y., Huang, Q.: Music emotion recognition using a variant of recurrent neural network. In: 2018 International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018). Atlantis Press (2019)
12. Liu, X., Chen, Q., Wu, X., Liu, Y., Liu, Y.: CNN based music emotion classification. CoRR abs/1704.05665 (2017). http://arxiv.org/abs/1704.05665
13. Malik, M., Adavanne, S., Drossos, K., Virtanen, T., Ticha, D., Jarina, R.: Stacked convolutional and recurrent neural networks for music emotion recognition. CoRR abs/1706.02292 (2017). http://arxiv.org/abs/1706.02292
14. Mion, L., Poli, G.D.: Score-independent audio features for description of music expression. IEEE Trans. Speech Audio Process. **16**(2), 458–466 (2008). https://doi.org/10.1109/TASL.2007.913743

15. Panda, R., Malheiro, R., Paiva, R.P.: Musical texture and expressivity features for music emotion recognition. In: Gómez, E., Hu, X., Humphrey, E., Benetos, E. (eds.) Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, 23–27 September 2018, pp. 383–391 (2018). http://ismir2018.ircam.fr/doc/pdfs/250_Paper.pdf
16. Robnik-Sikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn **53**(1–2), 23–69 (2003). https://doi.org/10.1023/A:1025667309714
17. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**(6), 1161 (1980)
18. Yang, X., Dong, Y., Li, J.: Review of data features-based music emotion recognition methods. Multimedia Syst. **24**(4), 365–389 (2018)
19. Yoon, S., Byun, S., Jung, K.: Multimodal speech emotion recognition using audio and text. In: 2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, 18–21 December 2018, pp. 112–118. IEEE (2018). https://doi.org/10.1109/SLT.2018.8639583
20. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 4651–4659. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.503
21. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, 6–12 September 2014, Proceedings, Part I, Lecture Notes in Computer Science, vol. 8689, pp. 818–833. Springer (2014). https://doi.org/10.1007/978-3-319-10590-1_53
22. Zhang, S., Tian, Q., Jiang, S., Huang, Q., Gao, W.: Affective MTV analysis based on arousal and valence features. In: Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, ICME 2008, 23–26 June 2008, Hannover, Germany, pp. 1369–1372. IEEE Computer Society (2008). https://doi.org/10.1109/ICME.2008.4607698
23. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, 8–14 September 2018, Proceedings, Part VII, Lecture Notes in Computer Science, vol. 11211, pp. 294–310. Springer (2018). https://doi.org/10.1007/978-3-030-01234-2_18