Multi-view Neural Networks for Raw Audio-based Music Emotion Recognition

NA HE
School of Computer Science, Faculty of Engineering & IT
University of Technology Sydney
Sydney, Australia
winterhn@gmail.com

Sam Ferguson
School of Computer Science, Faculty of Engineering & IT
University of Technology Sydney
Sydney, Australia
samuel.ferguson@uts.edu.au

***Abstract*—In Music Emotion Recognition (MER) research, most existing research uses human engineered audio features as learning model inputs, which require domain knowledge and much effort for feature extraction. We propose a novel end-to-end deep learning approach to address music emotion recognition as a regression problem, using the raw audio signal as input. We adopt multi-view convolutional neural networks as feature extractors to learn feature representations automatically. Then the extracted feature vectors are merged and fed into two layers of Bidirectional Long Short-Term Memory to capture temporal context sufficiently. In this way, our model is capable of recognizing dynamic music emotion without requiring too much workload on domain knowledge learning and audio feature processing. Combined with data augmentation strategies, the experimental results show that our model outperforms the state-of-the-art baseline with a significant margin in terms of R^2 score (approximately 16%) on the Emotion in Music Database.**

***Keywords*-deep learning; music emotion recognition; multi-view neural networks; raw audio**

I. Introduction

Facing enormous amounts of online music resources, Music Information Retrieval (MIR) plays an important role for music enthusiasts who wish to search and organize information. In particular, with increasing demand for retrieving music by emotion, Music Emotion Recognition (MER) is one fast-growing branch of MIR, benefiting emotion-related music applications as well as personalized experiences such as recommendation systems, music psychology, and artificial intelligence.

Generally, there are two kinds of emotion representation used in MER research: the categorical approach and the dimensional approach. The categorical approach maps emotion descriptions into some typical discrete emotion terms (such as happy, angry, sad and relaxed) [1], clusters [2] or multiple labels [3], which turns MER into a classification problem. By contrast, the dimensional approach allows us to label emotion within a continuous N-dimensional space, which turns MER into a regression problem. This approach is thought of as better suited to reduce ambiguity issues and to reflect time-series emotion variation [4]. One of the typical dimensional models is a 2-dimension circumplex plane articulated by Russell [5], with a horizontal axis of *valence* (positive-negative) and a vertical axis of *arousal* (active-inactive). Then music emotion could be measured as a pair of continuous values representing the degree of valence and arousal. Such emotion model is widely used in variety of emotion recognition systems [6], [7]. Since our research focuses on time-series emotion detection, this 2-dimension model is adopted.

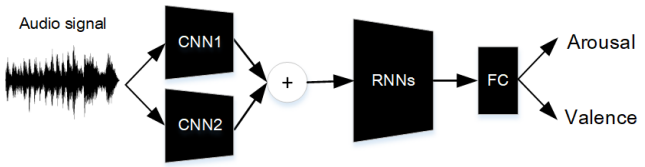

Figure 1. Overview of our multi-view neural networks solution. CNN represents convolutional neural network, RNNs represents recurrent neural networks, FC represents fully connected layer.

As for the input of MER training models, most previous research tends to use pre-processed audio features rather than raw audio [8], [9]. That usually require professional-level acoustic domain knowledge and heavy workload for feature extraction. Considering these issues, we propose using raw audio signal data as training model input directly. In this way, we can avoid expending too much effort on prior knowledge learning and hand-engineered feature processing.

To serve this purpose, we propose a novel architecture of deep neural networks model for emotion prediction illustrated in Fig. 1. We utilize multi-view Convolutional Neural Networks (CNNs) that we regard as multiple feature extractors to learn music features from raw audio automatically. Then these features are aggregated and fed into Recurrent Neural Networks (RNNs) to learn time-varying information for dynamic emotion variation. Finally, the result is fully connected and outputs 2 continuous values representing arousal and valence. Based on this structure, we term this stacked of multi-view convolutional recurrent neural networks as MCRNN. To the best of our knowledge, our model is the first multi-view neural networks for music emotion recognition using raw audio signals.

II. Related Work

In this section, we introduce related work about feature engineering and training models for MER.

A. Feature Engineering

In traditional machine learning models, human engineered audio features are usually used as training inputs. Researchers explored and summarised audio features in different categories (such as energy, timbre and rhythm), and compute novel high-level features to benefit recognition [9] or evaluate the usefulness of features for valence-arousal prediction [10], [11]. However, collecting hundreds of features gives rise to high time and labour cost regarding domain knowledge learning and feature extraction.

In recent years, deep neural networks have proved their powerful capability of feature learning so that some low-level features or raw audio signals could be taken as model inputs directly. Among them, the time-frequency representation of audio such as a mel-spectrogram has become an increasingly popular feature used for MER [12], [13]. Besides that, multimodal detection combining lyrics or music profile information with audio features trained in deep learning models are adopted widely to boost performance [14], [15]. Still, they need much effort for feature preparation.

B. Modelling

On the basis of dimensional Arousal-Valence (A-V) emotion representation, music emotion recognition is commonly defined as a regression problem. Before the emerging of deep learning, machine learning methods such as Multiple Linear Regression (MLR) and Support Vector Regression (SVR) were applied extensively [6], [8]. More recently, inspired by the success of deep learning in image detection, Convolutional Neural Networks (CNNs) have been explored as a regressor in music research [16], [17]. CNNs can learn representations from inputs automatically for each frame or clip of music, but fail to capture sequential information that is a crucial factor to improve the performance of music emotion variation detection. Due to this, RNNs have been proposed to analyse contextual music information over time to strengthen the fitness of regression. Especially, Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM) and Bidirectional Gated Reccurent Unit (BiGRU) show superiority in sequential data processing [18], [19]. Further, stacking these neural networks have been implemented and gained better performance [20], [21]. In addition, multi-view representation learning, that means learning features from multiple perspectives/views/modals, has shown its strong advantage of improving training model performance [22], [23]. However, there is no research applying multi-view concept to neural networks design for MER.

III. METHODOLOGY

The architecture of our MCRNN model is illustrated in Fig. 2. It consists of two stages: feature learning and sequence learning. Moreover, data augmentation method is utilized to promote performance.

A. Multi-view Feature Learning

In the first part of our model, two parallel CNN modules are employed as multi-view feature extractors. In detail, we define these two CNN modules as fine-view CNN and coarse-view CNN based on different kernel sizes. To adapt the model input, both of CNNs apply one-dimensional convolution (Conv1D) layers to learn the same sample sequence. For fine-view CNN, Conv1D is configured with 32×1 kernel size and 8 strides, where the rectified linear unit (ReLU) activation. The output of Conv1D is 8 feature maps, further handled by the BatchNormalization[24] layer and one-dimensional max pooling (MaxPooling1D) layer with 8×1 window size to avoid overfitting issue. Meanwhile, Conv1D in coarse-view CNN adopts 128×1 kernel size and 32 strides, as well as ReLU activation applied. The following processing is similar to fine-view CNN except changing window size to 2×1 in MaxPooling1D layer. Different window sizes serve another purpose of tuning into the same shape of the outputs from two CNN modules so that they could be merged into one single tensor for subsequent training layers.

These two views are similar to sample-level learning and frame-level learning mentioned in [25]. The sample-level learning uses relatively small kernel size to detect phase variations within a frame, while frame-level learning uses relatively long sample length to capture all possible audio pattern in periodic waveforms. Based on this point, two views are appropriate to learn feature representations for raw audio input.

B. Sequence Learning

The second part of our model is sequence learning part. Since emotion is associated with the context of music, Bidirectional Long Short-Term Memory (BiLSTM) is considered as a better choice because of its ability to capture both preceding and succeeding information. Besides this, increasing the depth of LSTM neural networks is taken into account. Additional hidden layers can recombine the learned representation from prior layers and create new representations at high levels of abstraction, and hence disentangle underlying relationship in temporal structure more easily [26]. However, we still need to balance learning efficiency and training difficulty when increasing the size and the depth of LSTM models. In this scenario, we adopt two bidirectional LSTM layers with 32 output units.

Finally, the dense layer connects all sequential learning results and output two regression values representing arousal and valence from a continuous range[-1,1]. In the procedure

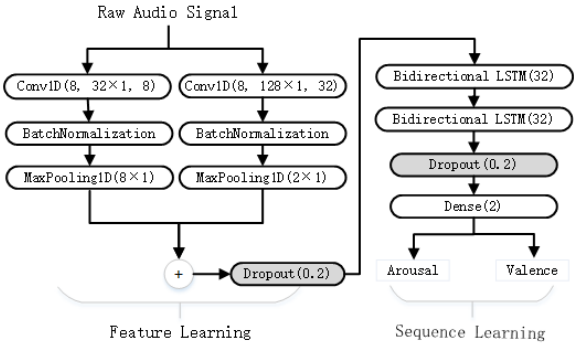

Figure 2. Architecture of MCRNN model

of training, we also add dropout function (labeled in grey color in Fig. 2) to further prevent overfitting.

C. Data Augmentation

Data augmentation (DA) is a method to generate synthetic data to increase the diversity of data for training models. In this way, the model could learn features from more relevant data and reduce overfitting effectively especially for small-scale dataset. Considering music audio characteristics and implementation cost, two approaches of data augmentation are adopted. One approach is pitch shifting that shift the pitch of audio clips. Here we lower the pitch of a waveform by a semitone. Such slightly perturbations would increase sample diversity but not impact the original music expression. Another approach is data flipping inspired by image processing [27] and time-series application[28]. In this research, we reverse the raw audio sequences in each annotation interval. This reversed data could enhance sequence learning through bidirectional LSTM in our model architecture. These two methods keep the same size of model input without changing duration and target labels. All synthetic data are generated from raw audio samples by using librosa API[1].

IV. Experiments

A. Data Description

The proposed model is applied to the Emotion in Music Database[2] proposed in [29]. This dataset is collected from the Free Music Archive (FMA)[3], including raw audio in mp3 format. Removing a set of duplicates in the initial 1000 songs, 744 songs are left actually. For the target labels, songs have been annotated via crowd-sourcing on Amazon Mechanical Turk (AMT) in the dimensions of arousal and valence independently, including static ratings given to the whole 45 seconds clips and dynamic annotation for each 0.5 second of last 30-second clips. The validation of annotations is documented in [29]. All music clips are sampled at 44100Hz. Corresponding to the dynamic annotations, 22050 sequential samples in each time step as an input are fed to our model to predict a pair of A-V values. Meanwhile, data augmentation methods generate 2 times more data to enrich our training dataset.

We take the DNN model proposed in [21] as the baseline. The dataset it used contains a subset of songs from Emotion in Music Database where the development sets have same source and data distribution. We argue that models applied to these two datasets are comparable. Moreover, we reproduce this baseline model on Emotion in Music Database. By using same dataset and evaluation method, we could keep two models in the same conditions so as to compare model architectures more convincingly.

B. Evaluation

We evaluate models with 10-fold cross validation. For each fold, we split training/validation/test sets with the ratio of 8:1:1. In accordance with previous research, Root Mean Square Error (RMSE) is used for model evaluation. RMSE is a measure of the average deviation of the estimates from the observed values, which is considered as an absolute measure of fit. Additionally, we add R^2 scores to compute coefficient of determination, which is considered a relative measure of fit. Through this approach, we can evaluate our model more comprehensively.

C. Implementation Details

We define batch size of 32, use Adam [30] as the optimizer with the learning rate of 0.001, and use Mean Square Error (MSE) as the loss function. We conduct training for each fold with early stopping strategy adopted where the patience is set to 10 on the validation dataset, and then we evaluate metrics on test set. The training and evaluation are implemented using the Keras library running on top of a tensorflow backend in python.

Apart from model definition mentioned in Section III, L2 regularization is added with setting the factor as 0.0001 to reduce overfitting in the fine-view CNN layer. Due to no pre-trained procedure in our MCRNN model, it is crucial to have a good initialization during training. Here we adopt the normal initializer proposed in [31] instead of Glorot uniform initializer, which produces the better performance.

V. Performance Discussion

In this section, we analyse input types and model structures based on experimental results. Then we conduct ablation study to demonstrate the effectiveness of our solution.

A. Performance Results Analysis

First, we compare models with different input types. The metrics in Table I show that DNN [21] model using raw audio as input outperforms DBLSTM[19] and CRNN-NB[20]

[1]https://librosa.org/doc/
[2]http://cvml.unige.ch/databases/emoMusic/
[3]https://freemusicarchive.org/

Table I
RMSE OF DIFFERENT NEURAL NETWORK MODELS IN AROUSAL AND VALENCE DIMENSION

Model	Model Input Type	Arousal	Valence	Average
DBLSTM[19]	engineered features	0.225	0.285	0.255
CRNN-NB[20]	engineered features	0.231	0.279	0.255
DNN[21]	raw audio	0.214	0.240	0.227
DNN[a]	raw audio	0.218	0.227	0.223
MCRNN	raw audio	**0.212**	**0.219**	**0.215**

[a]Reproduced DNN on Emotion in Music Database

Table II
R^2 SCORES COMPARISON WITH THE BASELINE IN AROUSAL AND VALENCE DIMENSION

Model	Arousal	Valence	Average
DNN[a]	0.405	0.08	0.243
MCRNN	**0.430**	**0.133**	**0.282**

[a]Reproduced DNN on Emotion in Music Database

models that use human engineered audio features as input. In the Emotion in Music Database, raw audio inputs contribute to good performance especially in valence recognition. So we argue that using raw audio signals with appropriate deep neural networks could model features well and gain better performance comparing with traditional engineering-feature-based models in this application.

Then, we compare our MCRNN model with DNN model in the same dataset based on RMSE and R^2 scores. Table I shows that our model gains lower RMSE scores than the baseline model in both of arousal and valence dimension with 4% improvement averagely. Regarding the R^2 scores as shown in Table II, the metric increases approximately 16% on the average. Especially in valence dimension, the result shows the great increment of 66%. Further, to prove statistical significance of model improvement, we carry out paired t-test on 10 folds of RMSE and R^2 scores for these two models, the p-value is less than 0.023 and 0.028 respectively. Compared with our model, DNN model only focuses on frame-level feature extraction but ignore phase variation in sample level. The results confirm that learning sample-level features could benefit valence recognition more.

B. Ablation Study

We conduct ablation study to evaluate the effect of multi-view structure and data augmentation. The results are illustrated in Fig. 3. Based on R^2 and RMSE scores on the aspects of arousal, valence and their average, it can be seen that multi-view CNNs model outperforms single-view models. By using multi-view architecture, the data detected from different perspectives could reveal more complementary information thereby enhancing modeling capability to learn more comprehensive features than those of single-view learning solution making. On the other hand, data augmentation takes effect on improving emotion prediction.

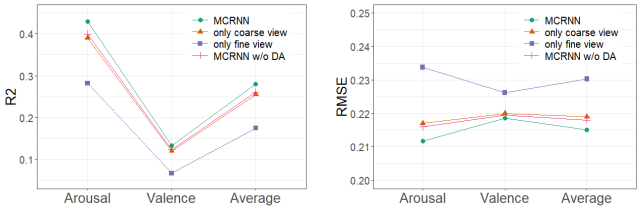

Figure 3. Ablation study based on MCRNN model measured by R^2 and RMSE of Arousal, Valence and their average. 4 situations are compared. That is, our MCRNN solution, single coarse-view CNN, single fine-view CNN, multi-view CNNs without data augmentation.

VI. CONCLUSION

This paper introduces novel multi-view neural networks trained in an end-to-end manner using raw audio signal directly to predict dynamic music emotion in dimensional arousal-valence space. In contrast with conventional music recognition methods, our solution does not use human engineered audio features, thus avoiding professional acoustic knowledge learning and intense feature engineering effort. Moreover, our model employs multi-view convolutional neural networks stacked by double bidirectional LSTM layers, which could capture more features from multiple perspectives combined with time-series analysis to improve the recognition performance. Also, we apply data augmentation methods (pitch shifting and data flipping) to increase the diversity of training data for this small-scale dataset to enhance model training performance. The experimental results demonstrate that our MCRNN model could achieve better performance than models using pre-processed audio features and models using single-view architecture.

REFERENCES

[1] C. Laurier, O. Meyers, J. Serrà, M. Blech, P. Herrera, and X. Serra, "Indexing music by mood: Design and integration of an automatic content-based annotator," in *Multimedia Tools and Applications*, vol. 48, no. 1, 2010, pp. 161–184.

[2] A. Bhattacharya and K. V. Kadambari, "A Multimodal Approach towards Emotion Recognition of Music using Audio and Lyrical Content," *arXiv preprint arXiv:1811.05760*, 2018. [Online]. Available: http://arxiv.org/abs/1811.05760

[3] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 2392–2396.

[4] A. Aljanaki, Y. H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PLoS ONE*, vol. 12, no. 3, 2017.

[5] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[6] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 448–457, 2008.

[7] K. W. Cheuk, Y.-J. Luo, B. B, T, G. Roig, and D. Herremans, "Regression-based music emotion prediction using triplet neural networks," *arXiv preprint arXiv:2001.09988*, 2020. [Online]. Available: http://arxiv.org/abs/2001.09988

[8] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *MIR 2010 - Proceedings of the 2010 ACM SIGMM International Conference on Multimedia Information Retrieval*, 2010, pp. 267–273.

[9] R. Panda, R. M. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, 3 2018.

[10] I. Vatolkin and A. Nagathil, "Evaluation of Audio Feature Groups for the Prediction of Arousal and Valence in Music," in *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, 2019, pp. 305–326.

[11] J. Grekow, "Audio features dedicated to the detection of arousal and valence in music recordings," in *Proceedings - 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications, INISTA 2017*, 2017, pp. 40–44.

[12] W. Bian, J. Wang, B. Zhuang, J. Yang, S. Wang, and J. Xiao, "Audio-Based Music Classification with DenseNet and Data Augmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11672 LNAI, 2019, pp. 56–65.

[13] J. S. Nayal, A. Joshi, and B. Kumar, "Emotion recognition in songs via Bayesian deep learning," in *ACM International Conference Proceeding Series*. Association for Computing Machinery, 9 2019, pp. 235–238.

[14] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018*, 2018, pp. 370–375.

[15] B. G. Patra, D. Das, and S. Bandyopadhyay, "Multimodal mood classification framework for Hindi songs," *Computacion y Sistemas*, vol. 20, no. 3, pp. 515–526, 2016.

[16] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016*, 2016, pp. 805–811.

[17] C. Senac, T. Pellegrini, F. Mouret, and J. Pinquier, "Music feature maps with convolutional neural networks for music genre classification," in *ACM International Conference Proceeding Series*, vol. Part F1301. ACM, 2017, p. 19.

[18] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2014.

[19] X. Li, J. Tian, M. Xu, Y. Ning, and L. Cai, "DBLSTM-based multi-scale fusion for dynamic emotion prediction in music," in *Proceedings - IEEE International Conference on Multimedia and Expo*, vol. 2016-Augus, 2016.

[20] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked convolutional and recurrent neural networks for music emotion recognition," in *Proceedings of the 14th Sound and Music Computing Conference 2017, SMC 2017*, 2017, pp. 208–213.

[21] R. Orjesek, R. Jarina, M. Chmulik, and M. Kuba, "DNN based music emotion recognition from raw audio signal," in *2019 29th International Conference Radioelektronika, RADIOELEKTRONIKA 2019 - Microwave and Radio Electronics Week, MAREW 2019*. IEEE, 2019, pp. 1–4.

[22] B. Wu, E. Zhong, A. Horner, and Q. Yang, "Music emotion recognition by multi-label multi-layer multi-instance multi-view learning," in *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, 2014, pp. 117–126.

[23] Y. Li, M. Yang, and Z. Zhang, "A Survey of Multi-View Representation Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1863–1883, 2019.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, 2015, pp. 448–456.

[25] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," in *Proceedings of the 14th Sound and Music Computing Conference 2017, SMC 2017*, 2017, pp. 220–226.

[26] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Communications of the ACM*, vol. 60, no. 6, 2017, pp. 84–90.

[28] Q. Wen, L. Sun, X. Song, J. Gao, X. Wang, and H. Xu, "Time Series Data Augmentation for Deep Learning: A Survey," *arXiv preprint arXiv:2002.12478*, 2020. [Online]. Available: http://arxiv.org/abs/2002.12478

[29] M. Soleymani, M. N. Caro, E. M. Schmidt, C. Y. Sha, and Y. H. Yang, "1000 Songs for Emotional Analysis of Music," in *CrowdMM 2013 - Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, 2013, pp. 1–6.

[30] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, 2015, pp. 1026–1034.