Multi-Representational Music Emotion Recognition using Deep Convolution Neural Networks

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

20-11-2021 / 29-11-2021

CITATION

Bathigama, Tharindu; Madushika, Sarasi (2021): Multi-Representational Music Emotion Recognition using Deep Convolution Neural Networks. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.17054549.v1>

DOI

[10.36227/techrxiv.17054549.v1](https://doi.org/10.36227/techrxiv.17054549.v1)

Multi-Representational Music Emotion Recognition using Deep Convolution Neural Networks

Tharindu Bathigama, Sarasi Munasinghe

Abstract—Emotion is an integral part of music and a complex aspect of music that is not easily understood by machines. Although the nature of this domain is complex, some progress has been made in this area suggesting that it might be feasible to develop models that can be used in real-world applications. But the truth is that recognizing music emotion is still unsaturated. On the other hand, models based on deep learning have been proposed to analyze emotions in the music due to its potential capability to exploit significant information from the music data. In this research we introduce a novel deep learning model that recognizes emotion in music using multiple representations of the audio, i.e. using the raw signal as well as computed features. DEAM dataset was used as the data for the model. The models consist primarily of convolution layers. The models predicted the mean and standard deviation of the arousal and valence dimensions. We were able to achieve a RMSE of 0.1501 and 0.1682 for dynamic prediction of arousal and valence respectively. For static arousal and valence prediction, we were able to achieve a RMSE of 0.2003 and 0.1938 for arousal and valence respectively. As these results are better than some of the state-of-the-art methods on the DEAM dataset, the proposed model has the potential value within the music emotion recognition community for effectively managing music emotions.

Index Terms—Music Emotion Recognition, Multiple Representations, Convolution Layers, Arousal and Valence Prediction.

1 INTRODUCTION

MUSIC is one of the oldest forms of art that we have as humans. All the cultures in the world have in some form or another an art-form that can be considered as music. It plays a major role in our lives such as in special events and celebrations. With the recent advances in technology music has become ever increasingly ubiquitous in our day-to-day lives. It is available through the radio, internet, and other forms of media outlets. With these advances, and abundance of music everywhere, the issue of effectively retrieving and cataloging music has become an attractive area of inquiry. To accomplish that task, Music Information Retrieval (MIR) has emerged as an area of study to effectively capture different aspects of music.

An important aspect of music is its relation to emotion, or the ability to invoke emotions. The ability to recognize emotion in music has many implications in several areas such as cognitive sciences and other areas which are concerned with how the brain works. These systems can also be used in several practical areas such as music recommendation, playlist generation, and in human computer interaction systems. Broadly we can categorize Music Emotion Recognition (MER) systems into two by considering the type of input it takes into the system. First, the methods based on the handcrafted features which are domain specific pre-processing techniques used on audio before it is given to the system. Second, the more recent methods, which are based on the automated feature extraction using deep neural networks.

Handcrafted feature extraction of audio is the extraction of domain specific information such as the power of the signal, zero-crossing-rate, etc. Before the introduction of deep learning, such feature extraction methods were necessary to provide the learning model such as Support Vector

Machines (SVM) to have the necessary input regarding the music. To extract these features, software such as Marsyas[1] have been developed.

With the development of deep learning techniques and more computing power, we can design models that automatically extract features from the audio. Automated feature extraction is the process of using the model itself to learn different feature representation of the input data. In the domain of image recognition, systems such as VGGNet[2] were able to out-perform the state-of-the-art by using convolution layers which were essentially learning the feature representations by itself. By adopting these ideas to the realm of audio processing and emotion recognition, a study used the raw audio signal with convolution layers for speech emotion recognition [3]. Likewise, a study which directly relates to this paper too has used raw audio as input for a music emotion recognition system input [4].

Automated feature-based models have shown to better results than the models that have used handcrafted features. But overall, models created for audio related tasks have not been able to achieve the level of performance as the models created for image related tasks. This is partly due to the large amount of data in raw audio and the loss of information when computing low-level features such as the spectrogram. In recent years, there have been studies that have studied the effects of multiple representations of the input data to deep learning models [5], [6]. They have shown that deep learning models do improve their performance on being able to take in different representations of the input. As such by exploring different audio representations for music emotion recognition, the performance of the models could be improved.

In this paper, we introduce a deep learning architecture

that combines different representations of the audio signal in determining music emotion. The aspects that we explore are the behavior of the system when raw audio or computed representations of the audio is input to the model. Specifically, the contributions of this paper are as follows,

- Designing a network architecture to predict music emotion using different audio representations such as, raw audio, spectrogram, mel-spectrogram and MFCCs
- Studying the learning ability when different representations are given to the system
- Attempt to address the subjectivity of emotion recognition by predicting the uncertainty of the prediction

The remainder of this article is structured as follows. In Section 2, we briefly discuss the related works on both handcrafted and automated feature-based MER models. Then, the method that we used, including the datasets, model architectures are discussed in Section 3. After that, in Section 4 the results obtained on the dataset are presented and discussed. Finally, in Section 5, we present our conclusions and promising future work.

2 RELATED WORKS

This section summarizes the related works on music emotion recognition using handcrafted and automated features.

2.1 Handcrafted Feature Extraction

In recent year handcrafted feature extracted systems for MER has been studied to a great extent [7]–[13]. There have been multiple studies that have been carried out to determine the effect of many of such features [8], [11], [13]. Some studies have extracted a set of fixed size feature vectors from the audio for the recognition of emotion [7], [14]. Mainly five sets of features which were extracted from the audio signal. These sets are energy, rhythm, temporal, spectrum and harmony [15]. For each feature set there exists several features and different software tools that are developed to extract them.

Some studies have extracted a fixed set of features from the audio and tested the limits of the models limits by considering different architectures [10], [11], [14]. These models were able to achieve good results in relation to the dataset that they have used. Combination of different extracted features have led to the most progress in these models. Another notable study is [12] where a personalized MER system was proposed. This study first creates a general MER system and then specializes to a certain group of annotations and produce the personalized MER system.

2.2 Automated Feature Extraction

The introduction of powerful deep learning algorithms and new hardware capabilities such as the GPU have enabled us to work on much more computationally challenging problems. One such challenging problem is extraction of features from audio. Therefore, it is logical that studies have been focused on automated feature extraction to tackle problems such as emotion recognition in music. The studies that follow are some such studies that have influenced the

field of MER and have given to insights that can be used in future research in this field.

A network architecture consisting of less parameters, but efficient, was proposed in [16]. Here, they have proposed the arousal and valence values in two separate network branches. A network which takes in raw audio was proposed in [4], 1d convolution layers were primarily used in order to extract the relevant features from the audio. A VGGNet[2] inspired model was introduced in [17] which was able to achieve good results. They also introduced a run-length algorithm to compute a final emotion result for the whole signal after having predicted smaller segments of the audio.

To conclude the related works section, we will present in brief the short comings, and areas of improvement that was found. First, it was clear that there are multiple methods that audio has been presented to deep learning models. Secondly, there seem to be no standard method to detect the emotion of music as the music is progressing over time. The subjectiveness of the emotion was also not addressed in those studies. To that end, we hypothesized that the results might be improved by combining different representations of the audio, i.e., using handcrafted and automated features. We also attempted to address the subjectiveness of the emotion by trying to predict the uncertainty of the prediction given by the model.

3 METHODOLOGY

In this section the methodology of the study is presented. We first describe how emotion is conceptualized for emotion recognition in music. Then the dataset used for this study is described. After that, a brief description is given on the different representations that we give the models. Finally, the model architectures we developed are presented.

3.1 Emotion Conceptualization

For the quantization or measuring of emotion we have used the arousal and valence emotion space introduced by Russell[18] and Thayer[19]. Here the arousal dimension describes the energy level of the emotion, and the valence dimension describes the pleasantness of the emotion. The figure 1 shows the emotion space as a 2d coordinate system.

A piece of music can contain a single feel/emotion throughout the track which is referred as static emotion annotation. In the meantime, emotion expressed in different sections of a track might be different which is expressed as dynamic emotion annotation. Therefore, MER systems have focused on both these views of the temporal aspect of emotion. Here, static emotion annotation implies that the track contains a single annotation of the emotion, either an emotion class or an arousal valence value. An example of how a system might predict the single emotion from a track is shown in figure 2. The dynamic emotion annotation implies that the track contains different annotations with respect to the time, an example of this is shown in figure 3.

3.2 Dataset

We have used the DEAM[20] dataset since multiple MER systems have used this dataset. This would allow us to

compare our method with other methods that have been proposed. This dataset consists of 1802 audio tracks which have been annotated using arousal and valence values. Since the annotation has been done by multiple annotators, the mean and the standard deviation of the annotations are also calculated. All the audio tracks are of 45 seconds with the latter 30 seconds being annotated. We have used both the static annotations and the dynamic annotations from the dataset to predict the static and dynamic emotion in music.

3.3 Audio Representations for Neural Networks

We have divided the audio representations into two sets, the raw audio set, and the computed feature set. In raw audio we only have the raw input audio as the input, while in the computed feature set, we have the spectrogram, mel-spectrogram and MFCCs (Mel-frequency cepstral coefficients). Both these sets individually and combined were used as input to the models.

3.4 Model Architectures

For the model architectures we have mainly used convolution layers since they have proven to be useful in MER [21]. For the preliminary studies, we tested the architectures using 1D convolution as well but 2D convolution-based models performed better than the 1D convolution ones. Therefore, in this paper the 2D convolution models are explained.

All the models described in this section share a common architecture for predicting the result. This is shown in figure 4. Each model takes in 5 seconds of audio, sampled at 22050Hz. This is passed to a feature extraction stage, which differentiate each of the models in the following sections. After feature extraction, the feature vector is passed through a *DenseBlockDP*, which is a serial connection of a linear layer, dropout, and a ReLU activation. *DenseBlock* is like the *DenseBlockDP*, without the dropout. After that the feature vector is passed to two branches of dense layers to predict the mean and the standard deviation of the arousal and valence values separately. The results are then concatenated into a single vector.

Fig. 1. Arousal Valence Emotion Space

Fig. 2. Static Emotion

Fig. 3. Dynamic Emotion

Fig. 4. Shared Architecture

3.4.1 Audio-Only 2D Convolution Model (A2DConv)

The *A2DConv* model takes the raw audio and uses convolution layers to extract features. *1DConvBlock* is a serial connection of a 1D convolution layer, batch normalize, dropout, and a ReLU activation. Similarly, the *2DConvBlockMP* is a serial connection of a 2D convolution layer, max pool, batch normalize, dropout, and a ReLU activation. After these layers, an adaptive average pooling layer is used to output a fixed size feature vector which is then flattened and passed to the regressor which was described previously. The number of channels, kernel size and other attributes of the architecture are shown in figure 5. Each of the channels of

Fig. 5. Audio-Only 2D Convolution Model

the convolution layers are shown in the *2DConvBlockMP*.

3.4.2 Computed-Only 2D Convolution Model (*C2DConv*)

The *C2DConv* model takes in only the computed features, MFCC, Mel-Spectrogram and the STFT is shown in figure 6. Like the previous model architecture, the attribute values of each of the components are shown in figure 6. The model takes the raw audio and compute the MFCC, Mel-Spectrogram and the STFT of the signal as the computed features. To extract these features the nnAudio[22] PyTorch[23] library is used, this allows for training or freezing these extraction layers as well. After that the features are flattened and concatenated before the regression is carried out.

3.4.3 Audio and Computed 2D Convolution Model (*AC2DConv*)

This model is essentially a combination of the previous two models. The four branches of the neural network capture and process different representations of the audio which are then combined in the concatenate layer. The architecture of this model is shown in figure 7.

4 EXPERIMENTAL IMPLEMENTATION

To implement the models, we have used the PyTorch[23] deep learning library. These models were trained to predict the static and dynamic emotion separately. First, to keep track of each model, let us introduce a set of models with each taking in a different kind of input and predicting either static or dynamic emotion. These models are shown in table 1.

TABLE 1
Model Names

Model Name	Input Features	Prediction Type
A2DConvStat	Audio	Static Emotion
C2DConvStat	Computed	Static Emotion
AC2DConvStat	Audio + Computed	Static Emotion
A2DConvD	Audio	Dynamic Emotion
C2DConvD	Computed	Dynamic Emotion
AC2DConvD	Audio + Computed	Dynamic Emotion

4.1 Evaluation Metrics

To evaluate the models, we have used the RMSE (Root Mean Squared Error) and the R^2 Score. These have been selected used by considering the best practices and past studies.

4.2 Hyperparameter Optimization

For each of these models, the hyperparameters were tuned by parameter searching using the “Weights and Biases” platform [24]. For each model around 50 model runs were executed to decide on optimal parameters for each of these, the optimal parameters obtained are listed in table 2 and table 3. Here, we have tuned the model with respect to the *dropout*, *learning rate*, *momentum*, *optimizer*, and *weight decay*. We have also tested the models with keeping the spectrograms (MFCC, Mel-Spectrogram and STFT) fixed or learnable. The adaptive layer dimensions have *adaptive_layer_units_0* and *adaptive_layer_units_1* have also been tuned for best performance.

5 RESULTS AND DISCUSSION

5.1 DEAM Dataset Results

After hyperparameter optimization, the models were validated using 5-fold cross validation on the dataset and the results shown in table 4 and table 5 were obtained. The results show that predicting the dynamic emotion using both audio and computed features performed the best from all the models. Overall, in both dynamic and static emotion prediction, the model which used both audio and computed features have performed better than using either audio or computed features.

The attempt to address the subjectiveness of emotion using uncertainty in the annotation was not successful. The R^2 score of the arousal (std) and valence (std) reflect that since, they have given negative results which suggest the model was not able to properly fit the data. The prediction of the mean of arousal and valence dimensions were, in comparison with the standard deviation, much more successful. Overall, the arousal dimension seems to be the one that the model is more confident in as evident from the R^2 scores for the arousal, this is in line with the past studies.

The R^2 score and RMSE was used to measure the performance of regression models. The best model performed on the DEAM dataset was *AC2DConvD* model, here we present some findings of the model to state its validity in recognizing the arousal and valence values. As previously stated, the attempt to recognize the variance of the annotation in the arousal and valence dimensions was unsuccessful and

Fig. 6. Computed-Only 2D Convolution Model

Fig. 7. Audio and Computed 2D Convolution Model

as evident from figure 8 the R^2 score is very poor for the standard deviation prediction of the arousal and valence dimensions.

Although it performed poorly on the standard deviation prediction it did perform well in predicting the mean of the dimensions as evident from figure 8. Apart from these findings we can see that all the 5-folds performed relatively consistent with about 0.01 difference in the RMSE and about 0.19 difference in the R^2 scores of the 5-folds. With these findings it can be safely assumed that the model did perform well on the dataset apart from the standard deviation prediction. The loss obtained for the 5-folds are also consistent suggesting that the model learned consistently.

5.2 Comparison with Related Works

We have compared our results with the benchmark studies that have used the DEAM dataset. table 6 represents the performance comparison of our proposed models with benchmark methods of the state-of-the-art music emotion recognition. As shown in table 6, results of our proposed models are outperformed existing state-of-the-art methods. Our model which predicted the static emotion was also able to outperform these past works.

6 CONCLUSION AND FUTURE WORKS

In this paper we addressed the problem of emotion recognition in music using multiple representations of the audio. We also tried to address the problem of subjectiveness of emotion by considering the variance of the annotation as

TABLE 2
Hyperparameters - Static Models

Parameters	A2DConvStat	C2DConvStat	AC2DConvStat
dropout	0.15	0.2	0.15
lr	0.01	0.001	0.0005
momentum	-	-	0.9
optimizer	adam	adam	sgd
spec_trainable	-	TRUE	TRUE
std_activation	relu	relu	softplus
weight_decay	1.00E-07	1.00E-07	1.00E-07
adaptive_layer_units_0	2	1	2
adaptive_layer_units_1	4	2	4

TABLE 3
Hyperparameters - Dynamic Models

Parameters	A2DConvD	C2DConvD	AC2DConvD
dropout	0.15	0.15	0.15
lr	0.001	0.0005	0.001
momentum	-	-	0.9
optimizer	adam	adam	sgd
spec_trainable	-	TRUE	TRUE
std_activation	relu	softplus	softplus
weight_decay	1.00E-07	0.000001	0.000001
adaptive_layer_units_0	2	1	8
adaptive_layer_units_1	4	16	4

Fig. 8. AC2DConvD Arousal (Std) R^2 Score

TABLE 4
Results - RMSE

Model	LOSS (RMSE)	Mean RMSE	
		Arousal	Valence
AC2DConvD	0.1303	0.1501	0.1682
C2DConvD	0.1378	0.1618	0.1782
AC2DConvStat	0.1541	0.2003	0.1938
C2DConvStat	0.1701	0.2301	0.2038
A2DConvStat	0.2063	0.309	0.2336
A2DConvD	0.2336	0.3574	0.254

TABLE 5
Results - R^2 Score

Model	R2 Score			
	Arousal	Valence	Arousal (Std)	Valence (Std)
AC2DConvD	0.4966	0.2515	0.1674	-0.0865
C2DConvD	0.4002	0.1637	0.1387	-0.2259
AC2DConvStat	0.5375	0.162	-0.1122	-0.1365
C2DConvStat	0.357	0.0847	-0.5717	-0.3669
A2DConvStat	-0.0778	-0.1092	-0.3404	-0.4155
A2DConvD	-1.5224	-0.5619	-0.0145	-0.8509

TABLE 6
Comparison

Model	LOSS (RMSE)	RMSE	
		Arousal	Valence
AC2DConvD	0.1303	0.1501	0.1682
C2DConvD	0.1378	0.1618	0.1782
AC2DConvStat	0.1541	0.2003	0.1938
C2DConvStat	0.1701	0.2301	0.2038
A2DConvStat	0.2063	0.309	0.2336
A2DConvD	0.2336	0.3574	0.254
R. Orjesek et al.[4]	-	0.214	0.24
M. Malik et al.[16]	-	0.202	0.268
X. Li et al.[14]	-	0.23	0.308

well, but it was not successful. But we were able to achieve better results on predicting the mean of arousal and valence values.

As for future work, the improvement of the model architectures to closely resemble the processes in our brain might be a good direction to pursue. The construction of larger datasets on music emotion recognition is also needed if we are to improve the ability of recognizing emotion in music. Development of personal music emotion recognition systems using music players might be able to address the issue of subjectiveness.

REFERENCES

- [1] G. Tzanetakis, "Marsyas." [Online]. Available: <http://marsyas.info/>
- [2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.
- [3] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1746809418302337>
- [4] R. Orjesek, R. Jarina, M. Chmulik, and M. Kuba, "DNN based music emotion recognition from raw audio signal," 2019 29th International Conference Radioelektronika, RADIOELEKTRONIKA 2019 - Microwave and Radio Electronics Week, MAREW 2019, pp. 1–4, 2019.
- [5] D. Williams, R. Hamon, and I. Gerg, "On the Benefit of Multiple Representations With Convolutional Neural Networks for Improved Target Classification Using Sonar Data," 2019.
- [6] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [7] 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.
- [8] J. Grekow, "Audio features dedicated to the detection of arousal and valence in music recordings," *Proceedings - 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications, INISTA 2017*, pp. 40–44, 2017.
- [9] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.
- [10] R. Panda, B. Rocha, and R. P. Paiva, "Dimensional music emotion recognition: Combining standard and melodic audio features," *10th International Symposium on Computer Music Multidisciplinary Research - CMMR 2013*, no. June 2014, pp. 583–593, 2013.
- [11] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," *MIR 2010 - Proceedings of the 2010 ACM SIGMM International Conference on Multimedia Information Retrieval*, pp. 267–273, 2010.
- [12] Y. H. Yang, Y. C. Lin, and H. Chen, "Personalized music emotion recognition," *Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, pp. 748–749, 2009.
- [13] X. Yang, Y. Dong, and J. Li, "Review of data features-based music emotion recognition methods," *Multimedia Systems*, vol. 24, no. 4, pp. 365–389, 2018.
- [14] X. Li, H. Xianyu, J. Tian, W. Chen, F. Meng, M. Xu, and L. Cai, "A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May, pp. 544–548, 2016.
- [15] Y.-H. Yang and H. H. Chen, *Music Emotion Recognition*, 2011.
- [16] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked convolutional and recurrent neural networks for music emotion recognition," *Proceedings of the 14th Sound and Music Computing Conference 2017, SMC 2017*, pp. 208–213, 2019.
- [17] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, and S. K. Saha, "Recognition of emotion in music based on deep convolutional neural network," *Multimedia Tools and Applications*, vol. 79, no. 1-2, pp. 765–783, 2020.
- [18] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [19] R. E. Thayer, *The biopsychology of mood and arousal*. New York, NY, US: Oxford University Press, 1989.
- [20] M. Soleymani, A. Aljanaki, and Y.-H. Yang, "DEAM: MediaEval Database for Emotional Analysis in Music," pp. 3–5, 2016. [Online]. Available: <http://cvml.unige.ch/databases/DEAM/manual.pdf>
- [21] X. Liu, Q. Chen, X. Wu, Y. Liu, and Y. Liu, "CNN based music emotion classification," 2017. [Online]. Available: <http://arxiv.org/abs/1704.05665>
- [22] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans, "NnAudio: An on-the-Fly GPU Audio to Spectrogram Conversion Toolbox Using 1D Convolutional Neural Networks," *Tech. Rep.*, 2020. [Online]. Available: <https://github.com/pseeth/torch-stft>
- [23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett,

- Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [24] L. Biewald, “Experiment Tracking with Weights and Biases,” 2020. [Online]. Available: <https://www.wandb.com/>