Improving Deep Learning Methodologies for Music Emotion Recognition

Pedro Lima Louro[1,]
pedrolouro@dei.uc.pt
Hugo Redinho[1,]
redinho@student.dei.uc.pt
Ricardo Malheiro[1, 2]
rsmal@dei.uc.pt
Rui Pedro Paiva[1,]
ruipedro@dei.uc.pt
Renato Panda[1, 3]
panda@dei.uc.pt

[1] Centre for Informatics and Systems of the University of Coimbra (CISUC), Department of Informatics Engineering, LASI
University of Coimbra
Coimbra, PT

[2] Polytechnic Institute of Leiria School of Technology and Management
Leiria, PT

[3] Ci2 — Smart Cities Research Center
Polytechnic Institute of Tomar
Tomar, PT

Abstract

Music Emotion Recognition (MER) has traditionally relied on classical machine learning techniques. Progress on these techniques has plateaued due to the demanding process of crafting new, emotionally-relevant audio features. Recently, deep learning (DL) methods have surged in popularity within MER, due to their ability of automatically learning features from the input data. Nonetheless, these methods need large, high-quality labeled datasets, a well-known hurdle in MER studies. We present a comparative study of various classical and DL techniques carried out to evaluate these approaches. Most of the presented methodologies were developed by our team, if not stated otherwise. It was found that a combination of Dense Neural Networks (DNN) and Convolutional Neural Networks (CNN) achieved an 80.20% F1-score, marking an improvement of approximately 5% over the best previous results. This indicates that future research should blend both manual feature engineering and automated feature learning to enhance results.

1 Introduction

In the development of MER systems, initial efforts focused on classical machine learning (ML) techniques, requiring extensive feature engineering to capture musical dimensions like melody, rhythm, and timbre [6]. However, the approach by Panda *et al*. [7], which presented novel features that yielded a 76% F1-score, faced challenges due to the intricate design process. Recently, DL has emerged as a promising method, automatically learning features through architectures like convolutional and recurrent neural networks, despite limitations like the need for large datasets [1]. Studies have addressed issues like neural networks' lack of interpretability and the challenge of ensuring these learn relevant features, showing that neural networks can indeed capture significant musical properties [10].

We present a comparative study on proposed improvements to DL-based methodologies using the 4QAED dataset and its extensions, highlighting DL's potential despite challenges like data scarcity and class imbalance. The main contribution of this study is an ensemble of a DNN and CNN achieving a state-of-the-art F1 Score of 80.20%.

2 Background

The relationship between music and emotions is a key area of interest in music psychology, focusing on how emotional content in music can be understood from expressed, perceived, and induced viewpoints. Perceived emotion, which provides the most objective perspective, is commonly the main subject of study within the existing literature, as highlighted by Pannese [8]. This research area explores different frameworks for categorizing human emotions derived from musical experiences. These frameworks are broadly classified into categorical models, like Hevner's Adjective Circle [3], which groups similar emotional adjectives together, and dimensional models, such as Russell's Circumplex Model [9], which organizes emotions on a plane based on different biological systems' roles in emotion processing. However, concerns have been raised about the categorical models' inability to capture emotions' continuous nature and complexity, while dimensional models require more cognitive effort from annotators and prior knowledge of the axes' meanings [2].

Our team introduced the 4QAED dataset alongside a new state-of-the-art methodology [7]. This dataset employs expert-curated labels from the AllMusic API, translating these into arousal and valence (AV) values. Instead of using these continuous values, the 4QAED dataset categorizes emotions into one of four quadrants corresponding to those values, offering a more simplified categorization that still leverages the insights of dimensional emotion models. The methodology proposed consists on an set of standard and novel emotionally-relevant audio features, achieving better results than using standard features only. This method paves the way for further detailed exploration of the dataset and its potential expansions in understanding music's emotional dimensions.

3 Conducted Experiments

The experimented methodologies are discussed in this section. It begins by describing the established classical and DL baselines, followed by the various improvements applied to our base neural network model.

3.1 Baseline Methodologies

A classical baseline was developed based on Panda *et al*.'s work, considering the 100 top-ranked standard and novel features. An SVM was trained using these features, and the hyperparameters were found using a Bayesian optimization method. Concurrently, we developed a novel CNN architecture, which accepts Mel-spectrograms as input, inspired by the work of Choi *et al*., modifying it to output one of the four quadrants of Russell's model. We used an early stopping mechanism to counteract overfitting once training accuracy hit or surpassed 90%.

3.2 Architectural Improvements

We enhanced our baseline DL model's architecture by integrating two Gated Recurrent Units (GRU) to focus on learning time-domain-specific features while also testing an adapted CRNN architecture with the same end. Furthermore, our experiments included a simple ensemble method, named Hybrid Augmented from herein, combining a pre-trained CNN and DNN. The DNN processed 1714 relevant features for improved classification outcomes, and synthetic samples were used for pre-training the CNN portion. This integrated approach, which leveraged data from both models, significantly boosted the overall classification performance.

Most of the proposed methodologies use the entirety of the 30-second audio samples from 4QAED as input for models. However, considering that humans can identify emotions in shorter clips, we experimented with smaller, 3.5-second segments. We adapted from the ShortChunk CNN approach introduced in [11], treating each segment as an individual training sample while consolidating segment predictions for the final evaluation. This also indirectly increased the amount of training data for these models. Additionally, in contrast to previous DL efforts that relied on convolutional layers, we adopted Lee et al.'s [5] end-to-end Sample CNN approach.

3.3 Methodological Improvements

Embeddings offer a promising approach when exploring alternatives to feeding Mel-spectrograms directly to models. We experimented with the approach by Koh *et al*. [4], who utilized the OpenL3 library to extract embeddings from spectral representations and feeding these to a Random

Forest classifier. We also experimented with training the same classifier with the embeddings obtained from a pre-trained autoencoder.

Data augmentation was also explored, applying various audio augmentation techniques, like Time Shifting, Time Stretching, Pitch Shifting, and Power Shifting, directly to the audio signal. Time-Frequency Masking, Seven-Band Parametric Equalization, and Random Gain were also considered. For DL, we turned to SMOTE for sample generation, eventually employing an autoencoder to reduce audio signal dimensionality.

Transfer learning was also explored, specifically using architectures trained on artist recognition to leverage possible learned emotionally-relevant features. In these efforts, we also explored the benefits of using larger music datasets, specifically MagnaTagATune[1], for weight initialization, adopting a CRNN model optimized for multi-label classification, as per the implementation in Won *et al*. [11].

4 Evaluation Approach

The 4QAED dataset[2], comprising 900 samples evenly distributed across the four quadrants of Russell's emotion model, was used for evaluation. These quadrants represent different emotions: happiness and excitement (Q1), anger and frustration (Q2), sadness and melancholy (Q3), and serenity and contentment (Q4). The dataset includes 30-second song excerpts and provides both a broad set of 1714 emotion-relevant features and the 100 top-ranked features already mentioned, along with categorical labels for quadrant classification. A proposed expansion, referred to as New-4QAED, increases the sample count to 1629, including a balanced subset of 1372 samples, ensuring equitable genre representation across quadrants.

To assess the proposed methodologies, F1-score, the harmonic mean of Precision and Recall, was used. The model's hyperparameters were first optimized on Original-4QAED using a grid search strategy, which was then applied with New-4QAED to ensure a fair comparison. A 10-fold and 10-repetition stratified cross-validation strategy was followed.

5 Results and Discussion

A summary of the results is shown in Table 1 Using larger datasets improved the performance in relation with the baseline CNN with the addition of GRU units and the CRNN architecture. The improvement was evident in the F1-scores for both methodologies when comparing the Original- and New-4QAED C datasets, indicating better outcomes than the DL Baseline. Moreover, the Hybrid Augmented methodology performed the best, reaching an F1-score of 80.20% on New-4QAED B, underscoring the impact of dataset size and quadrant distribution. Furthermore, applying Time-Frequency Masking, Seven-Band Parametric Equalization, and Random Gain techniques introduced a significant improvement.

The remaining methodologies underperformed in comparison with the DL baseline. Segment-level methodologies did not show improvements, likely due to the limited size of datasets used for training and the challenges of learning from smaller data segments. Similarly, approaches involving knowledge transfer and data representation may suffer from the same problem and that they might not be well-suited for emotion recognition tasks, given the characteristics of those tasks' datases. Embeddings also underperformed, indicating the complexity of finding effective data enhancement strategies for MER.

6 Conclusion and Future Work

The study comprehensively compared ML and DL methodologies for static emotion recognition in music, focusing primarily on DL to overcome the limitations of traditional ML techniques. The Hybrid Augmented methodology showcased superior performance on the balanced New-4QAED dataset, achieving an 80.20% F1 Score. This ensemble method, leveraging Mel-spectrogram representations and synthesized samples for training, outperformed other techniques, including those utilizing classical data augmentation. The analysis also highlighted the significant impact of dataset size over class balance on the classification outcomes.

Future research on developing novel classical features and refining DL architectures is clearly needed. A promising direction involves tailoring data augmentation techniques specifically for MER to maximize the inherent feature-extraction capabilities of DL models. Additionally, with the availability of larger training datasets, integrating an RNN component into future DL models could be advantageous for capturing time-domain-specific features. Exploring various spectral representations as inputs also presents a valuable area for future work, although it necessitates addressing the stability issues of these approaches before they can be more widely adopted.

Table 1: Results for the experimented methodologies.

Methods	Original-4QAED	New-4QAED C	New-4QAED B
SVM Baseline	75.59%	69.79%	69.82%
DL Baseline	60.62%	61.66%	60.28%
Baseline with GRU	60.07%	61.99%	58.85%
CRNN	**64.63%**	**64.09%**	**62.54%**
Hybrid Augmented	**68.04%**	**67.85%**	***80.24%***
ShortChunk CNN	60.61%	61.84%	57.07%
Sample CNN	60.92%	60.78%	54.46%
OpenL3	55.70%	53.62%	52.85%
Autoencoder	50.18%	53.56%	53.69%
Baseline + TFM	**62.03%**	61.82%	61.39%
Baseline + SB	**62.12%**	61.73%	61.01%
Baseline + RG	**62.24%**	62.08%	61.36%
Baseline + DeepSMOTE	60.70%	61.47%	60.48%
Artists CNN	50.85%	50.27%	50.22%
CRNN Pre-trained MTAT	50.21%	51.70%	51.44%

References

[1] K. Choi, G. Fazekas, M. Sandler, and K. Cho. Convolutional recurrent neural networks for music classification. In *Proceedings of the 2017 International Conference on Acoustics, Speech and Signal Processing*, pages 2392–2396, 2017.

[2] T. Eerola and J. K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39: 18–49, 2011.

[3] K. Hevner. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48:246–268, 1936.

[4] E. Koh and S. Dubnov. Comparison and analysis of deep audio embeddings for music emotion recognition, 2021. URL `https://arxiv.org/abs/2104.06517`.

[5] 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*, volume 14, pages 220–226, 2017.

[6] 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*, 14:5–18, 2006.

[7] R. Panda, R. Malheiro, and R. P. Paiva. Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, 11:614–626, 2020.

[8] A. Pannese, M.-A. Rappaz, and D. Grandjean. Metaphor and music emotion: Ancient views and future directions. *Consciousness and Cognition*, 44:61–71, 2016.

[9] J. Posner, J. A. Russell, and B. S. Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17:715–734, 2005.

[10] M. Won, S. Chun, and X. Serra. Toward interpretable music tagging with self-attention, 2019. URL `https://arxiv.org/abs/1906.04972`.

[11] M. Won, A. Ferraro, D. Bogdanov, and X. Serra. Evaluation of cnn-based automatic music tagging models. In *Proceedings of the 17th Sound and Music Computing Conference*, pages 331–337, 2020.

[1] Available at: https://mirg.city.ac.uk/codeapps/the-magnatagatune-dataset

[2] Available at: http://mir.dei.uc.pt/resources/MER_audio_taffc_dataset.zip