Emotion and Themes Recognition in Music Utilising Convolutional and Recurrent Neural Networks

Shahin Amiriparian[1], Maurice Gerczuk[1], Eduardo Coutinho[2], Alice Baird[1], Sandra Ottl[1], Manuel Milling[1], Björn Schuller[1,3]
[1]ZD.B. Chair of Embedded Intelligence for Health Care & Wellbeing, Univeristy of Augsburg, Germany
[2]Applied Music Research Lab, Department of Music, University of Liverpool, U. K.
[3]GLAM – Group on Language, Audio & Music, Imperial College London, U. K.
amiriparian@ieee.org

ABSTRACT

Emotion is an inherent aspect of music, and associations to music can be made via both life experience and specific musical techniques applied by the composer. Computational approaches for music recognition have been well-established in the research community; however, deep approaches have been limited and not yet comparable to conventional approaches. In this study, we present our fusion system of end-to-end convolutional recurrent neural networks (CRNN) and pre-trained convolutional feature extractors for music emotion and theme recognition. We train 9 models and conduct various late fusion experiments. Our best performing model achieves 74.2 % ROC-AUC on the test partition which is 1.6 percentage points over the baseline system of the MediaEval 2019 Emotion & Themes in Music task.

1 INTRODUCTION

The ability of music to express and induce emotions is a well-known and demonstrable fact [21]. It communicates and induces similar emotional states in all listeners because musical parameters (e. g., rhythm, melody, timbre, dynamics) encode affective information that is implicitly decoded by listeners [14, 18]. Furthermore, both music psychologists and computer scientists have provided plenty of evidence that listeners construe emotional meaning by attending to structural aspects of the acoustic signal at various levels [10, 13, 22]. Recent deep learning solutions demonstrate the suitability of recurrent neural networks (RNNs), autoencoders, and convolutional neural networks (CNNs) for the task of audio-based music emotion recognition (MER) [17, 23, 25]. In [12], we have utilised denoising autoencoders and a transfer learning approach for time-continuous predictions of emotion in music and speech. Furthermore, we have conducted both psychological and computational experiments that aimed at clarifying the role of music structure in the expression and induction of musical emotions [11, 15]. In this paper, we introduce our end-to-end architecture for the task of emotion and theme recognition in music at MediaEval 2019 [7].

2 APPROACH

Our framework – which is motivated by our previous works with CRNNs [1, 5] – is depicted in Figure 1. It consists of two models whose predictions are fused to obtain the final predictions. These models capture both shift-invariant, high-level features (convolutional block), and long-term temporal context (recurrent block) from the musical inputs [7, 8]. The MTG-Jamendo dataset [8] includes 18 486 audio tracks with 56 distinct mood and theme annotations/tags. All audio files have at least one tag. The dataset provides 60-20-20 % splits for training, validation, and testing. For the full description of the challenge data, please refer to [8].

2.1 Convolutional Recurrent Neural Network

The CRNN system (upper part of Figure 1) consists of a vgg-ish model (which is trained on the Audioset dataset [19]) with the final global average pooling layer replaced by an RNN. Specifically, we add 2 recurrent layers with 256 units (we tried 128, 256, and 512 units) and a dropout [27] of 0.3 (out of [0.2, 0.3, 0.4]) for each layer, followed by a 1 024 unit dense layer, batch normalisation [20], ReLU activation [24] and a dropout of 0.3. Tagging is performed by a 52 unit dense layer with sigmoid activation. We initialise the convolutional feature extractor with the official SoundNet trained weights [6]. Subsequently, sequences of log Mel spectrograms are generated using the kapre keras library [9]. Afterwards, the input is resampled to 16k Hz, and 64 Mel filters and an FFT window of 512 samples with a hop size of 256 are used. During training, we sample a random 20 s chunk of every song and apply random Gaussian noise with a maximum power of 0.2. For evaluation, we use the centre 20 s chunk of each song. We apply the RMSprop optimiser [28] and train the network with a batch size of 32. We first train only the top RNN and tagging layers for 20 epochs with a learning rate of 0.001, keeping the weights of the pre-trained vgg-ish frozen. We then unfreeze the feature extraction layers and resume training from the best checkpoint – measured in validation Receiver OperatingCharacteristic Curve (ROC-AUC) – with a reduced learning rate of 0.0001 for another 80 epochs. Finally, the best overall model is restored and evaluated on the test partition.

2.2 Utilising pre-trained CNNs

The second model (see bottom part of Figure 1) uses our Deep Spectrum system[1][3] to extract pre-trained CNN features from Mel spectrograms (128 Mel filters) of the songs, which have been shown to outperform engineered feature sets on a variety of acoustic tasks [2–4]. We use an ImageNet [16] pre-trained VGG16 architecture and forward plots of 1 and 5 second audio chunks through the network [26]. The activations of the penultimate layer then form our feature vectors. We extract these features for the first 30 seconds (the minimum song duration in the dataset [8]) of each song and use them as sequenced input for training RNNs. For both

MediaEval'19, 27-29 October 2019, Sophia Antipolis, France

[1] https://github.com/DeepSpectrum/DeepSpectrum

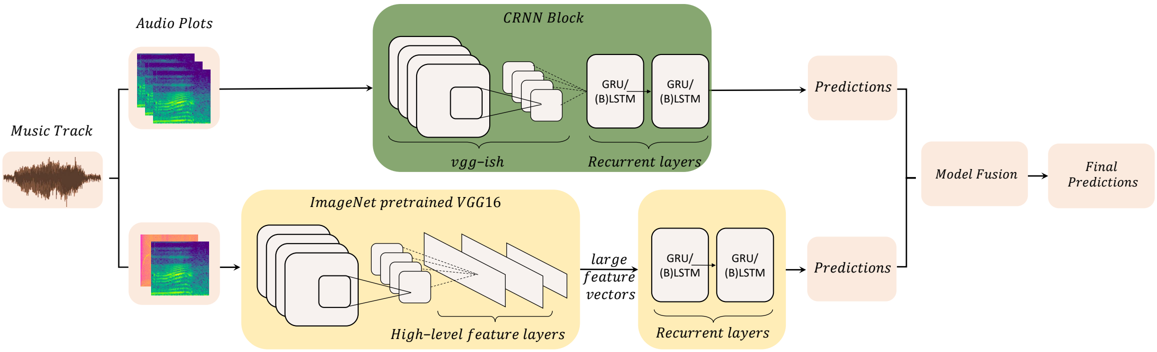

Figure 1: An overview of our system composed of two CRNN blocks. For a detailed account on the framework refer to Section 2.

feature types, three RNN architectures are trained which differ in the choice of recurrent cells, as with the CRNN. We chose an architecture with 2 recurrent layers of size 1 024 units each, followed by a dense layer with the same number of units before the final densely connected prediction layer. Afterwards, batch normalisation is used after each of the recurrent layers and the penultimate dense layer. Finally, a dropout of 0.4 is applied to the activations of the hidden layers. We train the model using RMSprop with a learning rate of 0.001 and batch size 32 for a maximum of 1 000 epochs, but perform early stopping if the validation ROC-AUC does not increase for over 50 epochs. Thus, none of our models was trained for more than 200 epochs. As for the CRNN, we restore the best model checkpoint before evaluating on the test partition.

2.3 Fusion Experiments

To explore further potential performance improvements, we apply model fusion experiments by averaging the prediction scores returned by our networks for the test partition. From these scores, we generate corresponding tag decisions with the official challenge script. In total, we evaluate five different fusion scenarios: fusion of all systems, fusion of all Deep Spectrum , fusion of all CRNN systems and fusion of Deep Spectrum systems trained on 1 s and 5 s feature windows, respectively.

3 RESULTS AND ANALYSIS

The results of our experiments are shown in Table 1. Our best CRNN model with GRU layers reaches 69.5 % ROC-AUC on the test set, while a bi-directional LSTM trained on 1 s Deep Spectrum features achieves 71.0 % ROC-AUC. These results can be explained by the fact that we use a fixed size chunk of each song (20 s for CRNN and 30 s for Deep Spectrum + RNN) instead of the whole song. We made this choice because training of the RNN models on longer sequences quickly becomes computationally infeasible. Nonetheless, we can see that fusion leads to an increase in performance. For each type of system, in-group fusion only leads to marginal performance boosts. We notice a larger positive effect by combining various system types hinting at complimentary information found on different scales. Finally, fusing all 9 systems increases the performance to 74.2 % ROC-AUC on the test set. This shows that the features extracted from spectrograms with an ImageNet pre-trained CNN provide further information not found by training on audio data alone. Our fusion configuration further achieves a macro average F1 of 17.5 % and a macro PR-AUC of 11.7 %.

Table 1: Performance of our proposed approaches. All results are given in macro ROC-AUC. Baseline accuracy on the test set is 72.5 % ROC-AUC [7].

	CRNN		
	RNN type	validation	test
	LSTM	71.4	69.4
	GRU	72.6	**69.5**
	BLSTM	71.9	68.2
	Deep Spectrum [3] + RNN		
spectrogram width (s)	RNN type	validation	test
1	LSTM	70.1	70.0
1	GRU	68.4	69.8
1	BLSTM	69.2	**71.0**
5	LSTM	69.0	70.3
5	GRU	68.8	69.9
5	BLSTM	68.4	70.8
	Fusion		
fused models			test
All CRNN (3 models)		–	70.7
All 1s Deep Spectrum (3 models)		–	71.5
All 5s Deep Spectrum (3 models)		–	71.6
All Deep Spectrum (6 models)		–	72.6
All systems (9 models)		–	**74.2**

4 DISCUSSION AND OUTLOOK

We outperformed the competitive challenge baseline of MediaEval 2019 Emotion & Themes in Music task after fusing the outputs of our two systems (cf. Table 1) . We also demonstrated that the Deep Spectrum + RNN approach (which makes use of CNNs pre-trained on ImageNet) yields better results than the CRNN with the vgg-ish model. For the future work, a systematic comparison between engineered and data-driven feature sets will be done by using the same machine learning models. Its aim will be to determine the usefulness of data-driven features for emotions and theme predictions in music. We believe that this research direction can lead to a better understanding of the relevant cues for emotion communications in music and improvements in automated emotion recognition systems.

REFERENCES

[1] Shahin Amiriparian, Alice Baird, Sahib Julka, Alyssa Alcorn, Sandra Ottl, Suncica Petrović, Eloise Ainger, Nicholas Cummins, and Björn Schuller. 2018. Recognition of Echolalic Autistic Child Vocalisations Utilising Convolutional Recurrent Neural Networks. In *Proceedings of INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association*. ISCA, Hyderabad, India, 2334–2338.

[2] Shahin Amiriparian, Nicholas Cummins, Maurice Gerczuk, Sergey Pugachevskiy, Sandra Ottl, and Björn Schuller. 2018. "Are You Playing a Shooter Again?!" Deep Representation Learning for Audio-based Video Game Genre Recognition. *IEEE Transactions on Games* 11 (2018).

[3] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, and Björn Schuller. 2017. Snore Sound Classification Using Image-based Deep Spectrum Features. In *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. ISCA, Stockholm, Sweden, 3512–3516.

[4] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Sergey Pugachevskiy, and Björn Schuller. 2018. Bag-of-Deep-Features: Noise-Robust Deep Feature Representations for Audio Analysis. In *Proceedings of the 31st International Joint Conference on Neural Networks (IJCNN)*. IEEE, Rio de Janeiro, Brazil, 2419–2425.

[5] Shahin Amiriparian, Sahib Julka, Nicholas Cummins, and Björn Schuller. 2018. Deep Convolutional Recurrent Neural Networks for Rare Sound Event Detection. In *Proceedings 44. Jahrestagung für Akustik, DAGA 2018*. DEGA, Deutsche Gesellschaft fÃijr Akustik e.V. (DEGA), Munich, Germany.

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., Barcelona, Spain, 892–900.

[7] Dmitry Bogdanov, Alastair Porter, Philip Tovstogan, and Minz Won. 2019. MediaEval 2019: Emotion and Theme Recognition in Music Using Jamendo. In *MediaEval Benchmarking Initiative for Multimedia Evaluation*. Sophia Antipolis, France.

[8] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019)*. ICML, Long Beach, CA, United States.

[9] Keunwoo Choi, Deokjin Joo, and Juho Kim. 2017. Kapre: On-GPU Audio Preprocessing Layers for a Quick Implementation of Deep Neural Network Models with Keras. In *Machine Learning for Music Discovery Workshop at 34th International Conference on Machine Learning*. ICML, International Conference on Machine Learning (ICML), Sydney, Australia.

[10] Eduardo Coutinho and Angelo Cangelosi. 2009. The Use of Spatio-Temporal Connectionist Models in Psychological Studies of Musical Emotions. *Music Perception: An Interdisciplinary Journal* 27, 1 (sep 2009), 1–15.

[11] Eduardo Coutinho and Angelo Cangelosi. 2011. Musical Emotions : Predicting Second-by-Second Subjective Feelings of Emotion From Low-Level Psychoacoustic Features and Physiological Measurements. *Emotion* 11, 4 (aug 2011), 921–937.

[12] Eduardo Coutinho, Jun Deng, and Björn Schuller. 2014. Transfer learning emotion manifestation across music and speech. In *2014 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 3592–3598.

[13] Eduardo Coutinho and Nicola Dibben. 2013. Psychoacoustic cues to emotion in speech prosody and music. *Cognition & Emotion* 27, 4 (jun 2013), 658–684.

[14] Eduardo Coutinho and Björn Schuller. 2017. Shared acoustic codes underlie emotional communication in music and speechâĂŤEvidence from deep transfer learning. *PloS one* 12, 6 (2017), e0179289.

[15] Eduardo Coutinho, Felix Weninger, Björn Schuller, and Klaus R. Scherer. 2014. The munich LSTM-RNN approach to the MediaEval 2014 "Emotion in Music" Task. In *CEUR Workshop Proceedings*, Martha Larson, Bogdan Ionescu, Xavier Anguera, Maria Eskevich, Pavel Korshunov, Markus Schedl, Mohammad Soleymani, Georgios Petkos, Richard Sutcliffe, Jaeyoung Choi, and Gareth J.F. Jones (Eds.), Vol. 1263. CEUR, Barcelona, Spain.

[16] J. Deng, W. Dong, R. Socher, L. J. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, FL, 248–255.

[17] Yizhuo Dong, Xinyu Yang, Xi Zhao, and Juan Li. 2019. Bidirectional Convolutional Recurrent Sparse Network (BCRSN): An Efficient Model for Music Emotion Recognition. *IEEE Transactions on Multimedia* (2019).

[18] Alf Gabrielsson and Erik Lindström. 2010. The role of structure in the musical expression of emotions. In *Handbook of music and emotion: Theory, research, applications*, Patrik N. Juslin and John Sloboda (Eds.). Oxford University Press, Oxford, 367–400.

[19] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, and others. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[20] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

[21] Patrik N. Juslin and John Sloboda (Eds.). 2011. *Handbook of music and emotion: Theory, research, applications*. Oxford University Press.

[22] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. 2010. Music emotion recognition: A state of the art review. In *Proceedings of ISMIR*, Vol. 86. Utrecht, Holland, 937–952.

[23] Huaping Liu, Yong Fang, and Qinghua Huang. 2019. 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.

[24] Vinod Nair and Geoffrey Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. Haifa, Israel, 807–814.

[25] Richard Orjesek, Roman Jarina, Michal Chmulik, and Michal Kuba. 2019. DNN Based Music Emotion Recognition from Raw Audio Signal. In *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 1–4.

[26] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[28] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4, 2 (2012), 26–31.