Emotion and Themes Recognition in Music with Convolutional and Recurrent Attention-Blocks

Maurice Gerczuk[1], Shahin Amiriparian[1], Sandra Ottl[1], Srividya Tirunellai Rajamani[1], Björn Schuller[1,2]
[1]Chair of Embedded Intelligence for Health Care & Wellbeing, Univeristy of Augsburg, Germany
[2]GLAM – Group on Language, Audio, & Music, Imperial College London, U. K.
maurice.gerczuk@uni-a.de

ABSTRACT

Emotion is an essential aspect of music, and its recognition is a prevalent research topic in the field of computer audition. Machine learning-based Music Emotion Recognition (MER) systems could boost the accessibility of music collections by providing standardised methodologies of music categorisation. In this paper, we introduce our (team name: *AugsBurger*) machine learning architecture sequentially composed of a convolutional feature extractor with block attention modules and a recurrent stack with self-attention for automatic MER. We train 5 models and conduct various late fusion experiments. Utilising a Convolutional Recurrent Neural Network (CRNN) with convolutional block attention applied throughout a 18-layer ResNet and a single recurrent layer with a Gated Recurrent Unit cell, a ROC-AUC of 73.9 % can be achieved on the test partition of the MediaEval 2020 Emotion & Themes in Music task. Applying late fusion on the individual model predictions and another challenge submission, this result is further increased to 75.3 % ROC-AUC.

1 INTRODUCTION

The ability of music to express emotions is a demonstrable and eminent fact [18]. Emotional experiences of music are complex and dependent on factors related to the states and traits of the listener, the performer, and the listening context, with research suggesting that musical structure alone is a key determinant of the emotional indication of music [25]. Different music emotion categories can induce emotional states, such as happiness, sadness, hope, excitement, and joy in listeners [16, 19]. This is primarily due to the affective information encoded in musical parameters, including melody, timbre, rhythm, and dynamics which are implicitly decoded by listeners [13, 17]. Conventional feature extraction methods (e. g., openSMILE [15]) have shown their suitability to extract such features from music recordings [28, 32]. However, the state-of-the-art for MER is defined by contemporary machine learning approaches which utilise convolutional and recurrent neural networks and learn data representations directly from the audio signals (or spectrograms) instead of extraction of pre-defined handcrafted features [3, 20, 21, 29]. Moreover, the integration of an attention mechanism in such systems has shown promise for various audio recognition tasks [6, 26]. Motivated by our previous works with CRNNs [2, 3, 5] and the success of attention mechanisms [4, 6, 26, 31], in this paper, we introduce an end-to-end framework composed of two attention blocks: a Convolutional Neural Network (CNN) with Convolutional Block Attention Modules (CBAMs), and a recurrent block with self-attention for the task of emotion and theme recognition in music [8–10].

MediaEval'20, December 14-15 2020, Online

2 APPROACH

A high-level overview of our approach is depicted in Figure 1. The framework consists of a CNN feature extractor enhanced by CBAMs and an RNN with self-attention. The convolutional block aims to learn high-level shift-invariant features, whilst long(er)-term temporal dependencies of music data [8–10] are mainly extracted by the recurrent block [1]. For all experiments in this paper, the MTG-Jamendo dataset [10] is solely used. The 18 486 audio tracks of the MTG-Jamendo dataset [10] are annotated in 56 distinct mood and theme categories, with every track having at least one tag. The dataset provides 60-20-20 % splits for training, validation, and testing. A full description of the dataset can be found in [10].

2.1 Pre-Processing and Augmentation

Our model uses the pre-computed mel-spectrograms that are part of the challenge dataset. Furthermore, we only use random windows of 8 seconds (500 timesteps) during training, reducing the memory footprint of our models and also serving as a form of data augmentation. Additionally, we apply SpecAugment [22], randomly applying masks of a maximum width of 10 (timesteps or frequency bands) to both the frequency and time domains of the spectrograms. We do not, however, use warping. During validation and testing, we take an 8 second chunk from the middle of each spectrogram and do not apply SpecAugment.

2.2 Convolutional Block Attention Module

In the CNN part of our modes, we use CBAMs [30] to refine the learn feature maps. CBAMs sequentially apply channel and spatial attention to the max and average pooled outputs of a convolutional layer. The attention maps are applied by element-wise multiplication. As these modules are a very lightweight extension, and show consistent performance increases for a wide range of popular image-recognition benchmarks [30], we evaluate their efficacy when added to a CRNN for music emotion and theme recognition.

2.3 Recurrent Self-Attention

In the RNN head of our CRNN framework, we use an attention mechanism to help the model focus on important parts of the feature sequences extracted by the CNN. This is done by applying self-attention [12] to the RNN outputs and states at each time step,

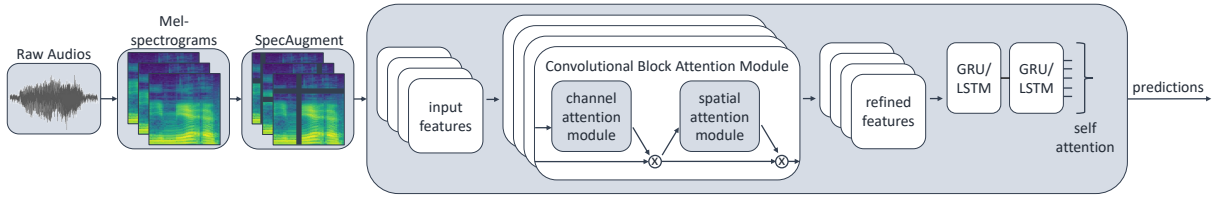

Figure 1: Overview of our system composed of a CBAM enhanced convolutional feature extractor followed by a Recurrent Neural Network (RNN) stack with self-attention. A detailed account of the framework is given in Section 2.

finally forming a compact representation for the whole sequence. We use the scaled dot product attention of Vaswani et al. [27].

2.4 Attention CRNN

Combining a CNN feature extractor with CBAMs and an RNN head with self-attention leads to our final attention CRNN. Specifically, we use an 18-layer ResNet architecture and replace the global pooling layer with an RNN stack. In the ResNet, we apply CBAMs in every convolutional block right before adding the residual. For the RNN, we evaluate using Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) cells. We started with a single layer with 256 units for both cell types. As we found GRUs to perform better, we additionally trained a model containing two recurrent layers of this type with 128 units each. Finally, the step-wise outputs and the final hidden state are used in the self-attention mechanism as keys and values, and query, respectively. Finally, a fully connected layer with sigmoid activation is used to perform the theme and mood tagging of the input audio samples. We train all of our models for a maximum of 100 epochs with an Adam optimiser with the learning rate set to 0.0003 but stop the training early if the validation ROC-AUC does not improve for 20 epochs. We use the weights from the best epoch (measured in validation ROC-AUC) for evaluation on the held-out test set.

2.5 Fusion Experiments

We apply late fusion to the results achieved by our attention CRNNs (models with CBAMs) through averaging prediction scores. Furthermore, we fuse our predictions with another system submitted to the challenge [23] which uses standalone self-attention an attention-based Rectified Linear Units (ReLUs) [11] added to the challenge baseline's vggish model [9].

3 RESULTS AND ANALYSIS

The results of our experiments are shown in Table 1. The best individual model can be found with a CBAM enhanced CRNN with a single GRU layer. This model reaches 73.9 % ROC-AUC on the test partition, compared to the challenge baseline of 72.5 %. Using the same architecture but without applying CBAMs, only 69.4 % are achieved on test. Furthermore, we observe that models with a single GRUs layer outperform their LSTM counterparts. Fusing the two best CRNNs – CBAMs in the CNN and GRU cells in the RNN – further leads to a slight performance boost to 74.1 %. More effective is fusing with the attention enhanced CNN from [23], achieving our best result of 75.3 % ROC-AUC and 13.1 % PR-AUC on test. This hints at complementarity of the two systems.

Table 1: Performance of our proposed approaches measured in macro ROC-AUC. Baseline ROC-AUC on test is 72.5 % [9].

CRNN				
conv. attention	RNN cell	RNN units	validation	test
no attention	GRU	256	69.9	69.4
no attention	LSTM	256	70.0	69.3
block attention	GRU	2×128	67.9	71.6
block attention	GRU	256	71.8	**73.9**
block attention	LSTM	256	69.4	69.4
attention CNN [23]	-	-	72.8	72.8
Fusion				
block attention with GRU				74.1
block attention with GRU+ [23]				**75.3**

A noteworthy characteristic of all the systems used in our submission to the challenge is that they do not make use of any external data and only train on short extracts of the songs (about 10 seconds long). Furthermore, none of the models were trained for more than 50 epochs, against the challenge baseline's 1 000 epochs. In this way, our attention models reduce data, memory and time requirements while achieving stronger performance than the baseline. Compared to the other challenge submissions, only one submission that does not rely on external data outperforms our best fusion model[1].

4 DISCUSSION AND OUTLOOK

We have introduced a CRNN architecture with attention modules for both convolutional (cf. Section 2.2) and recurrent blocks (cf. Section 2.3) for emotions and themes recognition in music. In the pre-processing step, in order to achieve a better model generalisation, we have augmented the spectrograms from the music recordings with SpecAugment [22] and trained our models with both challenge and augmented data (cf. Section 2.1). Furthermore, as a post-processing step, we have conducted a set of late (decision-level) fusion experiments to check the complementary of the predictions from each trained model (cf. Section 2.5). The results indicate the efficacy of our applied methodologies for this challenge(cf. Section 3). Considering the performance increase achieved by utilising attention-based ReLUs with the baseline's vggish architecture in [23], it is worth investigating this activation mechanism in combination with the CBAM enhanced CRNNs presented herein. As our models only make use of the challenge data itself, one should also consider exploiting external data, such as the Million Song Dataset [7], Music4All [24] or NSynth [14] for possible improvements in model accuracy.

[1] https://multimediaeval.github.io/2020-Emotion-and-Theme-Recognition-in-Music-Task/results

REFERENCES

[1] Shahin Amiriparian. 2019. *Deep representation learning techniques for audio signal processing*. Ph.D. Dissertation. Technische Universität München.

[2] 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.

[3] Shahin Amiriparian, Maurice Gerczuk, Eduardo Coutinho, Alice Baird, Sandra Ottl, Manuel Milling, and Björn Schuller. 2019. Emotion and themes recognition in music utilising convolutional and recurrent neural networks. In *MediaEval Benchmarking Initiative for Multimedia Evaluation*. Sophia Antipolis, France.

[4] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Alice Baird, Lukas Stappen, Lukas Koebe, and Björn Schuller. 2020. Towards Cross-Modal Pre-Training and Learning Tempo-Spatial Characteristics for Audio Recognition with Convolutional and Recurrent Neural Networks. *EURASIP Journal on Audio, Speech, and Music Processing* 2020 (2020), to appear.

[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 für Akustik e.V. (DEGA), Munich, Germany.

[6] Shahin Amiriparian, Pawel Winokurow, Vincent Karas, Sandra Ottl, Maurice Gerczuk, and Björn Schuller. 2020. Unsupervised Representation Learning with Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*. 11–17.

[7] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. (2011).

[8] 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.

[9] Dmitry Bogdanov, Alastair Porter, Philip Tovstogan, and Minz Won. 2020. MediaEval 2020: Emotion and Theme Recognition in Music Using Jamendo. In *MediaEval Benchmarking Initiative for Multimedia Evaluation*. Online.

[10] 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.

[11] Dengsheng Chen and Kai Xu. 2020. AReLU: Attention-based Rectified Linear Unit. (2020). arXiv:arXiv:2006.13858

[12] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 551–561. https://doi.org/10.18653/v1/D16-1053

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

[14] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*. PMLR, 1068–1077.

[15] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[16] Bruce Ferwerda and Markus Schedl. 2014. Enhancing Music Recommender Systems with Personality Information and Emotional States: A Proposal.. In *Umap workshops*.

[17] 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.

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

[19] Ai Kawakami, Kiyoshi Furukawa, Kentaro Katahira, and Kazuo Okanoya. 2013. Sad music induces pleasant emotion. *Frontiers in psychology* 4 (2013), 311.

[20] Khaled Koutini, Shreyan Chowdhury, Verena Haunschmid, Hamid Eghbal-zadeh, and Gerhard Widmer. 2019. Emotion and Theme Recognition in Music with Frequency-Aware RF-Regularized CNNs. *arXiv preprint arXiv:1911.05833* (2019).

[21] Maximilian Mayerl, Michael Vötter, Hsiao-Tzu Hung, Bo-Yu Chen, Yi-Hsuan Yang, and Eva Zangerle. 2019. Recognizing Song Mood and Theme Using Convolutional Recurrent Neural Networks. In *MediaEval Benchmarking Initiative for Multimedia Evaluation*. Sophia Antipolis, France.

[22] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779* (2019).

[23] Srividya Tirunellai Rajamani, Kumar Rajamani, and Björn Schuller. 2020. Emotion and Theme Recognition in Music using Attention-based Methods. In *MediaEval Benchmarking Initiative for Multimedia Evaluation*. Online.

[24] Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Valéria Delisandra Feltrim, Marcos Aurélio Domingues, and others. 2020. Music4All: A New Music Database and Its Applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. IEEE, 399–404.

[25] Klaus R. Scherer and Eduardo Coutinho. 2013. How music creates emotion: a multifactorial process approach. In *The Emotional Power of Music: Multidisciplinary Perspectives on Musical Arousal, Expression, and Social Control*, T Cochrane, B Fantini, and K R Scherer (Eds.). Number 10. Oxford University Press, 121–145.

[26] Lorenzo Tarantino, Philip N Garner, and Alexandros Lazaridis. 2019. Self-Attention for Speech Emotion Recognition.. In *INTERSPEECH*. 2578–2582.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[28] Felix Weninger, Florian Eyben, and Björn Schuller. 2013. The TUM approach to the MediaEval music emotion task using generic affective audio features. In *MediaEval Benchmarking Initiative for Multimedia Evaluation*. Barcelona, Spain.

[29] Minz Won, Andres Ferraro, Dmitry Bogdanov, and Xavier Serra. 2020. Evaluation of CNN-based Automatic Music Tagging Models. *arXiv preprint arXiv:2006.00751* (2020).

[30] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[31] Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. 2019. Speech emotion recognition using multi-hop attention mechanism. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2822–2826.

[32] Fan Zhang, Hongying Meng, and Maozhen Li. 2016. Emotion extraction and recognition from music. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. IEEE, 1728–1733.