Recognizing Music Mood and Theme Using Convolutional Neural Networks and Attention

Alish Dipani[1, 2, †], Gaurav Iyer[2, †], Veeky Baths[2]
[1]Upload AI LLC, USA
[2]Cognitive Neuroscience Lab, BITS Pilani, K.K.Birla Goa Campus, India
alish.dipani@uploadai.com,f20170544@goa.bits-pilani.ac.in,veeky@goa.bits-pilani.ac.in

ABSTRACT

We present the UAI-CNRL submission to MediaEval 2020 task on Emotion and Theme Recognition in Music. We make use of the ResNet34 architecture, coupled with a self-attention module to detect moods/themes in music tracks. The autotagging-moodtheme subset of the MTG-Jamendo dataset was used to train the model. We show that the proposed model outperforms the provided VGG-ish and popularity baselines.

1 INTRODUCTION

Music has been shown to induce a variety of emotions such as happiness, sadness, and anger [7, 8, 27]. This induction of emotions can be attributed to different intrinsic properties such as tempo, rhythm variations, intensity, mode and extrinsic properties such as the association of music with personal events and previous experiences [12, 23]. These emotional responses could also be one of the important motivators for humans to listen to music [20–22].

Automatic tagging and detection of emotions of music is a difficult task considering the subjectivity of human emotions. The MTG-Jamendo dataset [4] aims at tackling several such autotagging tasks by providing royalty-free audios of consistent quality with several tags for genre, instruments and mood/theme. The Emotion and Theme Recognition Task of MediaEval 2020 uses the mood/theme subset of the MTG-Jamendo dataset. The task is as follows - given audio, automatically detect one or multiple moods/themes out of 56 given tags, for example, fun, sad, romantic, happy [3].

In this paper, we describe our approach (team name: UAI-CNRL) for this task by using convolutional neural networks to extract features from the mel-spectrograms of the audios and multi-head self-attention to predict the mood/theme by processing the extracted features. Our approach achieves better performance than the baselines.

2 RELATED WORK

Convolutional neural networks (CNNs) have been successful in extracting meaningful features for tasks such as image recognition [10, 14] and object detection [10]. In the field of audio processing, CNNs have been used for a variety of tasks, such as automatic tagging [6], source separation [30], music emotion classification [16] and speaker identification [18].

Transformer networks which use self-attention layers [28] have been successful in tackling language tasks involving long-range dependencies. They have also been used in the field of audio processing for many tasks, such as automatic tagging [29], source separation [5], and speech recognition [2].

A combination of these methods have been demonstrated to achieve state-of-the-art performance [2, 9, 32]. Inspired by these, we use convolution layers to extract features from mel-spectrograms and self-attention layers to process those features to predict the moods/themes.

3 APPROACH

We make use of a popular convolutional neural network architecture, the ResNet [10] as a feature extractor to extract compact representations of our data. We pair this with self-attention [28] in order to capture long-term temporal attributes of the given data. We also make use of batch normalization [11] and dropout [24] in order to further regularize the model. We describe the model architecture in this section. Our code and trained model are available at this URL[§].

3.1 ResNet34

Residual connections make training deep neural networks easier, since they address the problem of vanishing gradients. We make use of a standard ResNet34 architecture to take advantage of this property. This is preceded by two convolutional layers in order to reshape the data into a form that can be fed into the ResNet. Another convolutional layer is used after the ResNet feature extractor to reduce the number of channels.

3.2 Self-Attention

The MTG-Jamendo dataset consists of tracks of varying lengths, a majority of which are over 200 seconds. Using self-attention, we attempt to capture long-range temporal attributes and summarize the sequence of music representation.
Our model architecture is inspired by the works done in [25], which uses multi-head attention along with positional encoding. 2 layers, each consisting of 4 attention heads were used. The input sequence length and embedding size used were unchanged.

3.3 Data Augmentation

3.3.1 Mixup. Previous submissions to MediaEval 2019 [25] for this task have shown that Mixup [31] greatly improves the performance of the model being used. Mixup creates a new training

[†]Authors Contributed Equally
[§]https://github.com/alishdipani/Multimediaeval2020-emotions-and-themes-in-music

MediaEval'20, December 14-15 2020, Online

example by linearly combining two random, existing training samples - in the feature space as well as in the label space. More formally, Mixup trains a neural network on convex combinations of pairs of examples and their labels. This helps the model alleviate unwanted behaviours, such as memorization, especially since the dataset size is relatively small.

3.3.2 SpecAugment. SpecAugment [19] is an augmentation technique used for speech recognition, which involves augmenting the spectrogram itself, instead of the waveform data. SpecAugment modifies the spectrogram by warping it in the time axis, masking blocks of frequency channels, and masking blocks of time steps. This makes the model more robust to missing information in terms of the input speech data as well as frequency information.

3.3.3 Other Augmentations. Other transformation techniques, such as random cropping and random scaling were used to further augment the given data.

4 TRAINING DETAILS

This section describes the details of data pre-processing, architecture and other training details.

4.1 Data Preparation

We use the mel-spectrograms provided in the MTG-Jamendo dataset for the purpose of training. Random cropping and scaling are used to augment and transform the data into a tensor of length 4096 (approximately 87.4 seconds). Additionally, SpecAugment is used to augment the dataset.

4.2 Architecture and Control Flow

- The input tensor of shape (1, 96, 4096) is divided into 16 segments length-wise, each new segment being of length 256.
- Each segment is then processed through 2 convolutional layers, in order to obtain a representation with 3 channels.
- The obtained representation is then passed into the ResNet34 feature extractor, followed by a convolutional layer to obtain an intermediate representation.
- The feature maps are then passed through the self-attention module, followed by a series of linear layers to obtain the final class scores. Dropout is used to regularise the training process.
- The model returns the outputs of the self-attention module and the feature maps (after passing them through the linear layers). Both outputs are used to compute the loss and perform backpropagation, but only the outputs of the self-attention module are used to make predictions.

4.3 Hyperparameters and Other Details

The model was trained with the Adam [13] optimizer, at a learning rate of 1e-4, for 35 epochs. The values of β_1 and β_2 were set to 0.9 and 0.999 respectively. Binary cross entropy loss was used as the loss function.

Table 1: Results

Metric	Ours	VGG-ish[3]	popularity[3]
ROC-AUC-macro	**0.7360**	0.7258	0.5000
PR-AUC-macro	**0.1275**	0.1077	0.03192
precision-macro	**0.1639**	0.1382	0.0014
recall-macro	**0.3487**	0.3086	0.0179
F-score-macro	**0.1884**	0.1657	0.0026
ROC-AUC-micro	**0.7865**	0.7750	0.5139
PR-AUC-micro	0.1369	**0.1409**	0.0341
precision-micro	0.1105	**0.1161**	0.0799
recall-micro	**0.4032**	0.3735	0.0447
F-score-micro	0.1735	**0.1771**	0.0573

5 RESULTS

The proposed model produces results that improve on those of the given VGG-ish and popularity baselines. We obtain an ROC-AUC-macro metric of 0.7360 and a PR-AUC-macro metric of 0.1275. For comparison, the baseline VGG-ish model produces an ROC-AUC macro of 0.7258 and a PR-AUC macro of 0.1077. Detailed results can be found in Table 1.

6 FUTURE WORK

In this section, we discuss other approaches that we considered towards the problem statement. These may be used as pointers towards future work on tasks involving this dataset.

Our approach can be broken down into two parts - first, the extraction of features from the audio data and second, processing the extracted features to predict the moods/themes. Both these parts could be potentially improved upon, and we mention a few ways to do so below.

With respect to feature extraction:

- Using a wider range of features to aid the classification task instead of using mel-spectrograms. For example, the LEAF frontend proposed by [1] can be used for this approach.
- Using self-supervised approach to extract features, such as wav2vec 2.0 [2]. This would also reduce reliance on labelled data.
- Using temporal convolutional networks [15] to extract features directly from audio instead of using mel-spectrograms.

With respect to the processing of extracted features:

- Using dual path processing inspired by [17] in order to capture long-term dependencies while also reducing computational load.
- Exploring ways of processing the raw audio data with more powerful models, such as WaveNet [26] in order to obtain better insights into the dataset, and theme recognition in general.

ACKNOWLEDGMENTS

We thank Shell Xu Hu for helpful discussions.

REFERENCES

[1] Anonymous. 2021. A Universal Learnable Audio Frontend. In *Submitted to International Conference on Learning Representations*. https://openreview.net/forum?id=jM76BCb6F9m under review.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. (2020). arXiv:cs.CL/2006.11477

[3] Dmitry Bogdanov, Alastair Porter, Philip Tovstogan, and Minz Won. 2020. Emotion and Theme Recognition in Music Using Jamendo. In *Working Notes Proceedings of the MediaEval 2020 Workshop*.

[4] 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)*. Long Beach, CA, United States. http://hdl.handle.net/10230/42015

[5] Jingjing Chen, Qirong Mao, and Dong Liu. 2020. Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. *arXiv preprint arXiv:2007.13975* (2020).

[6] Keunwoo Choi, George Fazekas, and Mark Sandler. 2016. Automatic tagging using deep convolutional neural networks. (2016). arXiv:cs.SD/1606.00298

[7] Hauke Egermann, Nathalie Fernando, Lorraine Chuen, and Stephen McAdams. 2015. Music induces universal emotion-related psychophysiological responses: comparing Canadian listeners to Congolese Pygmies. *Frontiers in psychology* 5 (2015), 1341.

[8] Thomas Fritz, Sebastian Jentschke, Nathalie Gosselin, Daniela Sammler, Isabelle Peretz, Robert Turner, Angela D Friederici, and Stefan Koelsch. 2009. Universal recognition of three basic emotions in music. *Current biology* 19, 7 (2009), 573–576.

[9] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. (2020). arXiv:eess.AS/2005.08100

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. (2015). arXiv:cs.CV/1512.03385

[11] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. (2015). arXiv:cs.LG/1502.03167

[12] Stéphanie Khalfa, Mathieu Roy, Pierre Rainville, Simone Dalla Bella, and Isabelle Peretz. 2008. Role of tempo entrainment in psychophysiological differentiation of happy and sad music? *International Journal of Psychophysiology* 68, 1 (2008), 17–26.

[13] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. (2017). arXiv:cs.LG/1412.6980

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[15] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager. 2016. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*. Springer, 47–54.

[16] Xin Liu, Qingcai Chen, Xiangping Wu, Yan Liu, and Yang Liu. 2017. CNN based music emotion classification. (2017). arXiv:cs.MM/1704.05665

[17] Yi Luo, Zhuo Chen, and Takuya Yoshioka. 2020. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. (2020). arXiv:eess.AS/1910.06379

[18] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. *Interspeech 2017* (Aug 2017). https://doi.org/10.21437/interspeech.2017-950

[19] 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. *Interspeech 2019* (Sep 2019). https://doi.org/10.21437/interspeech.2019-2680

[20] Mark Reybrouck and Tuomas Eerola. 2017. Music and its inductive power: a psychobiological and evolutionary approach to musical emotions. *Frontiers in Psychology* 8 (2017), 494.

[21] Thomas Schäfer, Peter Sedlmeier, Christine Städtler, and David Huron. 2013. The psychological functions of music listening. *Frontiers in psychology* 4 (2013), 511.

[22] Roni Shifriss, Ehud Bodner, and Yuval Palgi. 2015. When you're down and troubled: Views on the regulatory power of music. *Psychology of Music* 43, 6 (2015), 793–807.

[23] John A Sloboda and Patrik N Juslin. 2001. Psychological perspectives on music and emotion. *Music and emotion: Theory and research* (2001), 71–104.

[24] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15, 56 (2014), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

[25] Manoj Sukhavasi and Sainath Adapa. 2019. Music theme recognition using CNN and self-attention. (2019). arXiv:cs.SD/1911.07041

[26] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. (2016). arXiv:cs.SD/1609.03499

[27] Daniel Västfjäll. 2001. Emotion induction through music: A review of the musical mood induction procedure. *Musicae Scientiae* 5, 1_suppl (2001), 173–211.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. (2017). arXiv:cs.CL/1706.03762

[29] Minz Won, Sanghyuk Chun, and Xavier Serra. 2019. Toward interpretable music tagging with self-attention. *arXiv preprint arXiv:1906.04972* (2019).

[30] Jeroen Zegers and Hugo Van hamme. 2019. CNN-LSTM models for Multi-Speaker Source Separation using Bayesian Hyper Parameter Optimization. (2019). arXiv:cs.LG/1912.09254

[31] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. (2018). arXiv:cs.LG/1710.09412

[32] Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu. 2020. Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition. (2020). arXiv:eess.AS/2010.10504