Emotion and Theme Recognition in Music Using Attention-Based Methods

Srividya Tirunellai Rajamani[1], Kumar Rajamani[2], Björn Schuller[1,3]
[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Institute of Medical Informatics, University of Lübeck, Germany
[3]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
srividya.tirunellai@informatik.uni-augsburg.de

ABSTRACT

Emotion and theme recognition in music plays a vital role in music information retrieval and recommendation systems. Deep learning based techniques have shown great promise in this regard. Realising optimal network configurations with least number of FLOPS and model parameters is of paramount importance to obtain efficient deployable models, especially for resource constrained hardware. Yet, not much research has happened in this direction especially in the context of music emotion recognition. As part of the *MediaEval 2020: Emotions and Themes in Music* challenge, we (team name: *AUGment*), propose novel integration of attention based techniques for the task of emotion/mood recognition in music. We demonstrate that using stand-alone self-attention in the later layers of a VGG-ish network, matches the baseline PR-AUC with 11 % fewer FLOPS and 22 % fewer parameters. Further, utilising the learnable Attention-based Rectified Linear Unit (AReLU) activation helps to achieve better performance than the baseline. As an additional gain, a late fusion of these two models with the baseline also improved the PR-AUC and ROC-AUC by 1 %.

1 INTRODUCTION

Automatic detection of mood/theme of music is a challenging and widely researched topic that aids in music tagging and recommendation systems. This involves acoustic feature extraction followed by single or multi-label classification. Conventional approaches used hand-crafted audio features representing physical or perceived aspects of sound as input to machine learning algorithms [14, 18, 21]. Contemporary methods make use of Deep Neural Networks (DNNs) with hand-crafted or automatically learnt features from audio [1, 10, 12, 13, 24].

Attention based mechanisms have shown great promise and achieved state-of-the-art results in several tasks such as Natural language processing (NLP) [23], image classification and segmentation [15], computer vision [22], as well as speech analysis [5, 9, 17, 26, 28]. The effectiveness of these mechanisms in the task of music mood/emotion recognition, however, is less explored. We perform an investigation of the effectiveness of different attention based techniques for multi-label music mood classification.

2 EXPERIMENTAL SETUP

The data used in the MediaEval 2020 task is a subset of the MTG-Jamendo dataset [4]. The subset used in the MediaEval 2020 task [3] includes 18 486 full-length audio tracks of varying length with mood and theme annotations.The dataset comprises of 56 distinct mood/themes tags. All tracks have at least one tag, but many have more than one making it a multi-label classification task.

The Mel-spectrogram is a widely used feature for audio related tasks such as boundary detection, tagging [11], and latent feature learning. It is also shown to be an effective time-frequency representation of audio for the task of automatic music tagging [8]. Using Mel-spectrogram as the input enables the use of image classification networks like Convolution Neural Networks (CNN) or Residual Neural Networks (ResNets). CNNs, including their variants like Visual Geometry Group (VGG) networks, have been successfully used for image recognition [25, 29], object detection [16, 20], and image segmentation [7]. VGG-like architectures that comprise of a stack of convolution layer followed by a fully connected layer are further shown to be well-suited for the task of music tagging [3].

We consider the first 1 400 time bins of the Mel-spectrogram of each track as input, since the central theme or mood is usually established in the opening of a track. This approach, as opposed to taking time bins from the center of the track or using random chunks, additionally ensures that the input is guaranteed to have non-silent segments. Optionally, trimming silence from the start would make it even more robust on tracks that potentially could have delayed onset. A VGG-ish architecture [8] with five 2D convolutional layers followed by a dense connection is used as the baseline for our experiments. We determine the effectiveness of various attention mechanisms on this baseline for the task of music mood/theme detection[1]. Training is done for a maximum of 100 epochs with early stopping if the validation ROC-AUC does not increase for over 35 epochs.

3 METHODS

3.1 Stand-alone self-attention

Self-attention is attention applied to a single context instead of across multiple contexts (i. e., the query, keys, and values are extracted from the same context). Stand-alone self-attention replaces spatial convolutions with a form of self-attention rather than using attention as an augmentation on top of convolutions. Stand-alone self-attention especially in later layers of a network is shown to outperform the baseline on image classification with far fewer floating point operations per second (FLOPS) and parameters [19]. We experiment using stand-alone self-attention in later layers of the baseline VGG-ish network.

[1]The source code is published at https://github.com/SrividyaTR/MediaEval2020-EmotionAndThemeInMusic

MediaEval'20, December 14-15 2020, Online

No.	Model	GFLOPS	# Parameters	ROC-AUC	PR-AUC
1	VGG-ish baseline	3.32	448 122	.725	.107
2	Self-attention in Layer3	**2.94**	**350 074**	.716	**.108**
3	Self-attention in Layer4	3.28	350 074	.723	.108
4	Self-attention in Layer5	3.32	399 098	.716	.101
5	AReLU activation in baseline	3.32	448 132	**.728**	.107
6	Late fusion of models 2 and 5	–	–	.732	.114
7	Late fusion of models 1, 2 and 5	–	–	**.735**	**.118**

Table 1: Results on MediaEval2020 test set. Using self-attention in Layer3 matches the baseline PR-AUC with 11 % fewer FLOPS and 22 % fewer parameters

3.2 Attention-based Rectified Linear Unit

The *Attention-based Rectified Linear Unit* (AReLU) is a learnable activation function [6]. It exploits an element-wise attention mechanism and amplifies positive elements and suppresses negative ones through learnt, data-adaptive parameters. The network training is more resistant to gradient vanishing as the attention module within AReLU learns element-wise residues of the activated part of the input. With only two extra learnable parameters (alpha and beta) per layer, AReLU enables fast network training under small learning rates. We experiment using AReLU activation in all of the 5 layers of the baseline VGG-ish network and observe improved performance.

3.3 Fusion Experiments

We perform late fusion experiments by averaging the prediction scores of our different models for the test partition. By a fusion of the prediction scores from the stand-alone self-attention based model, AReLU-activation based model, and the baseline, we further improve the performance as compared to the baseline.

4 SUBMISSIONS AND RESULTS

Figure 1 provides an overview of our approach and the different attention mechanisms that we utilise for the task of emotion and theme recognition in music. Overall, we submitted 4 models to the challenge. The first model is based on self-attention in Layer3 of the VGG-ish baseline and the second is based on using AReLU activation in all the 5 convolution layers of the baseline. The next 2 submissions are a late-fusion of these 2 models and with the baseline.

Table 1 summarises the results of our experiments. Using stand-alone self-attention instead of 2D convolution in Layer3 of the VGG-ish network resulted in a PR-AUC comparable to the baseline with **11 %** fewer FLOPS and **22 %** fewer parameters. Using AReLU activation in all of the 5 layers of the VGG-ish network improved the ROC-AUC as compared to the baseline. A late fusion of these 2 model's prediction resulted in about 1 % increase in both PR-AUC and ROC-AUC . A fusion of our model with the baseline model helped in further improving the performance.

We experimented using self-attention in other convolution layers of the baseline VGG-ish network, but the best performance with least trainable parameters was noted when using self-attention in Layer3. Using self-attention in Layer4 also gave comparable performance though with 1.2 % fewer FLOPS and **22 %** fewer parameters. Further, when using self-attention in initial layers (Layer1 or Layer2), the amount of memory required to hold the activations was significantly large, leading to the observation that it works best on down-sampled input. We also observed that using a batch-size of 16 and learning rate of 0.0001 helped in faster convergence to the best model. The best model was learnt within 25 epochs in all our experiments.

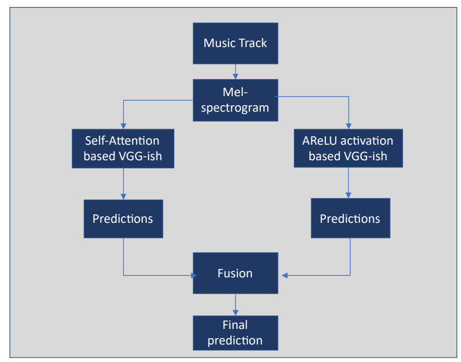

Figure 1: Overview of our different Attention-based approaches for Emotion and Theme Recognition in Music

5 DISCUSSION AND OUTLOOK

We demonstrated the effectiveness of a self-attention-based VGG-like network for multi-label emotion and theme recognition in music. This network's computational efficiency is particularly relevant when executing the model inference on a mobile device or other resource constrained computing hardware. We also established the performance benefits of using AReLU activation for this task. A potential future work is to evaluate the effectiveness of incorporating AReLU activation within a self-attention based VGG-like network instead of performing a late fusion. One should evaluate the effectiveness of other attention-based techniques like attention augmented convolution [2] for this task. Data Augmentation using *mix-up* [27] could also be evaluated to analyse the impact on performance.

REFERENCES

[1] 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 *Proceedings of the MediaEval 2019 Workshop*. Sophia Antipolis, France.

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc Le. 2019. Attention Augmented Convolutional Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea, 3285–3294.

[3] Dmitry Bogdanov, Alastair Porter, Philip Tovstogan, and Minz Won. 2020. MediaEval 2020: Emotion and Theme Recognition in Music Using Jamendo. In *Proceedings of the MediaEval 2020 Workshop*. Online, 14-15 December 2020.

[4] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Proceedings of the Machine Learning for Music Discovery Workshop, 36th International Conference on Machine Learning (ICML)*. California, United States.

[5] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China, 4960–4964.

[6] Dengsheng Chen, Jun Li, and Kai Xu. 2020. AReLU: Attention-based Recti□ed Linear Unit. (2020). arXiv:cs.LG/2006.13858

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan Yuille. 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (2016).

[8] Keunwoo Choi, György Fazekas, and Mark B. Sandler. 2016. Automatic Tagging Using Deep Convolutional Neural Networks. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*. New York City, United States, 805–811.

[9] Chorowski, Jan K and Bahdanau, Dzmitry and Serdyuk, Dmitriy and Cho, Kyunghyun and Bengio, Yoshua. 2015. Attention-Based Models for Speech Recognition. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*, Vol. 28. Montreal, Canada, 577–585.

[10] Eduardo Coutinho, Felix Weninger, Björn Schuller, and Klaus Scherer. 2014. The Munich LSTM-RNN Approach to the MediaEval 2014 Emotion in Music Task. *Proceedings of the MediaEval 2014 Workshop* (2014).

[11] Sander Dieleman and Benjamin Schrauwen. 2013. Multiscale Approaches To Music Audio Feature Learning.. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*. Curitiba, Brazil, 3–8.

[12] Sander Dieleman and Benjamin Schrauwen. 2014. End-to-end learning for music audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy, 6964–6968.

[13] Matthias Dorfer and Gerhard Widmer. 2018. Training general-purpose audio tagging networks with noisy labels and iterative self-veri□cation. In *"Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE)"*. Surrey, UK, 178–182.

[14] Florian Eyben, Klaus R. Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and A□ective Computing. *IEEE Transactions on Affective Computing* 7, 2 (2016), 190–202.

[15] Chaitanya Kaul, Suresh Manandhar, and Nick Pears. 2019. Focusnet: An Attention-Based Fully Convolutional Network for Medical Image Segmentation. In *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*. Venice, Italy, 455–458.

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, Hawaii, United States, 936–944.

[17] Shuo Liu, Jinlong Jiao, Ziping Zhao, Judith Dineley, Nicholas Cummins, and Björn Schuller. 2020. Hierarchical Component-attention Based Speaker Turn Embedding for Emotion Recognition. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*. Glasgow, Scotland, UK, 1–7.

[18] Lie Lu and Dan Liu. 2006. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing* 14 (2006), 5–18.

[19] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. 2019. Stand-Alone Self-Attention in Vision Models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vol. 32. Vancouver, Canada, 68–80.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2015).

[21] Björn Schuller, Gerhard Rigoll, and Manfred Lang. 2003. Hidden Markov model-based speech emotion recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Hong Kong, China, II–1 – II–4.

[22] Lukas Stappen, Georgios Rizos, and Björn Schuller. 2020. X-AWARE: ConteXt-AWARE Human-Environment Attention Fusion for Driver Gaze Prediction in the Wild. In *Proceedings of the International Conference on Multimodal Interaction (ICMI)*. 858–867.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Vol. 30. California, United States, 5998–6008.

[24] Felix Weninger, Florian Eyben, and Björn Schuller. 2013. The TUM Approach to the MediaEval Music Emotion Task Using Generic Affective Audio Features. *Proceedings of the MediaEval 2013 Workshop* (2013).

[25] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, Hawaii, United States, 5987–5995.

[26] Yeonguk Yu and Yoon-Joong Kim. 2020. Attention-LSTM-Attention Model for Speech Emotion Recognition and Analysis of IEMOCAP Database. *Electronics* 9 (2020), 713.

[27] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *Proceedings of the 6th International Conference on Learning Representations(ICLR)*. Vancouver, BC, Canada.

[28] Ziping Zhao, Zhongtian Bao, Zixing Zhang, Nicholas Cummins, Haishuai Wang, and Björn Schuller. 2020. Hierarchical Attention Transfer Networks for Depression Assessment from Speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7159–7163.

[29] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc Le. 2018. Learning Transferable Architectures for Scalable Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, Utah, United States, 8697–8710.