EXPLAINING PERCEIVED EMOTION PREDICTIONS IN MUSIC: AN ATTENTIVE APPROACH

Sanga Chaki[1] **Pranjal Doshi**[2] **Sourangshu Bhattacharya**[2]
Priyadarshi Patnaik[3]
[1] Advanced Technology Development Centre, IIT Kharagpur, India
[2] Department of Computer Science and Engineering, IIT Kharagpur, India
[3] Department of Humanities and Social Sciences, IIT Kharagpur, India
s.chaki27@gmail.com, sourangshu@gmail.com

ABSTRACT

Dynamic prediction of perceived emotions of music is a challenging problem with interesting applications. Utilization of relevant context in audio sequence is essential for effective prediction. Existing methods have used LSTMs with modest success. In this work we describe three attentive LSTM based approaches for dynamic emotion prediction from music clips. We validate our models through extensive experimentation on standard dataset annotated with arousal-valence values in continuous time, and choose the best performer. We find that the LSTM based attention models perform better than the state of the art transformers for the dynamic emotion prediction task, both in terms of R^2 and Kendall-τ metrics. We explore individual smaller feature sets in search of a more effective one and to understand how different features contribute to perceived emotion. The spectral features are found to perform at par with the generic ComPare feature set [1]. Through attention map analysis we visualize how attention is distributed over music clips' frames for emotion prediction. It is observed that the models attend to frames which contribute to changes in reported arousal-valence values and chroma to produce better emotion predictions, effectively capturing long-term dependencies.

1. INTRODUCTION

Automatic determination of perceived emotion in music is an active and major area of focus for the music information retrieval (MIR) community. The aim of dynamic perceived emotion prediction task is to output a sequence of time-synchronized arousal-valence labels when a music clip is given as input. It finds varied applications in the domains of personalized and/or generalized music recommendations, organizing music databases, automatic music creation, mood based music search etc. This task is challenging because: 1) perceived emotion might depend on the inherent relationship between different frames of music, distributed over time, and 2) emotion perception is inherently subjective in nature, highly contextual and personal. Thus, it is understandable that the emotions related to music are a time-continuous process, where the context of the sequential music frames play an immense role on the associated emotion. Relating this to the machine learning perspective, one can discern the need of context sensitive models like recurrent neural networks (RNNs) for the task at hand. In this study, we use attention mechanism with a deep RNN-LSTMs (Long Short Term Memory) and the Transformer [2], to predict the perceived emotion in each defined time frame of music continuously. We compare our approach with recent works [3] using only LSTM. We also attempt to understand the importance of types of features contributing to dynamic perceived emotion. Lastly, attention is visualized with the help of attention map analysis. The following are the major contributions of this work: 1) The LSTM based attention models are found to perform better than the state of the art Transformers for the dynamic emotion prediction task. 2) Spectral features are found to perform at par with the generic ComPare feature set [1]. 3) Attention maps are interpreted to observe that the attention models are able to focus on relevant music frames for dynamic emotion prediction task.

This paper is organized as follows. In section 2, relevant literature regarding music emotion recognition and attention is reviewed. Section 3 provides details of the attention based models and Transformer used in this work. All the experiments carried out and the observations are reported in section 4. Finally, the conclusions drawn from the present study are detailed in section 5.

2. RELATED WORK

2.1 Music Emotion Recognition

In the past, most music emotion prediction systems used features of timbre, pitch, MFCCs and/or lyrics and applied to classifiers like SVMs [4]. Current state-of-the-art methods for music emotion prediction are mostly based on deep neural networks like RNN-LSTMs. Coutinho et al. [5] proposed the use of this model for this task. Weninger et al. [3, 6, 7] used RNN-LSTM networks successfully to perform continuous time music emotion regression, using

Attribution: S. Chaki, P. Doshi, S. Bhattacharya and P. Patnaik, "Explaining Perceived Emotion Predictions in Music: An Attentive Approach", in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.,* Montréal, Canada, 2020.

a modified cost function, on the *1000 Songs for Emotional Analysis of Music* dataset [8]. Giamusso et al. [9] used neural networks to predict playlist emotions based on lyrics. Fan et al. [10] performed ranking based emotion recognition from experimental music. Delbouys et al. [11] used LSTM and ConvNet models on the Million Song Dataset [12] for audio and lyrics based bimodal music emotion detection.

2.2 Emotion Representation

Over the years, Discrete and Dimensional models of emotion representation have been used in MIR. Studies using discrete model either tag their musical data with single [13] or cluster [14] of simple tags. In dimensional models like Russel's Circumplex model [15], emotion is mapped into a 2-D plane, spanned by two axes denoting *arousal* and *valence*. Using this well known and satisfyingly exhaustive emotion representation, the problem of emotion recognition/prediction is turned into a two dimensional regression problem [16].

2.3 Attention in MIR tasks

Recently, attention mechanism and Transformer models have found application in a wide range of MIR tasks, with success. Balke et al. [17] used a soft-attention mechanism on input of synthesized piano data for audio sheet music retrieval. Their results indicate that attention increases the robustness of the retrieval system by focusing on different parts of the input representation based on the tempo of the audio. The improved results led them to argue for the potential of attention models as a very general tool for many MIR tasks. Gururani et al. [18] explored an attention mechanism for handling weakly labeled data for multi-label instrument recognition. Their results show that incorporating attention leads to overall improvement in classification accuracy metrics and enables models to *attend to* specific time segments in the audio relevant to each instrument label leading to interpretable results. Donahue et al. [19] used the Transformer architecture to improve performance for the task of generating multi-instrumental music scores. Chen et al. [20] proposed the Harmony Transformer, a multi-task music harmony analysis model aiming to improve chord recognition. Park et al. [21] utilized a bi-directional Transformer for chord recognition (BTC) which showed competitive performance. Through attention map, they visualized how attention was performed, and it was observed that the model was able to divide segments of chords by utilizing adaptive receptive field of the attention mechanism and capture long-term dependencies. These and other works have explored various feature sets like CQT (in [21]), Chroma (in [20]), along with other standard feature sets [1] (in [3]). These recent successes in varied MIR tasks in terms of model accuracy and interpretability, motivated us to apply the same in the music emotion regression task. To the best of our knowledge, neither attention models nor Transformers have been applied before to the task under examination.

3. ATTENTION BASED MODELS FOR EMOTION PREDICTION IN MUSIC

3.1 Attention Model (AT)

In the past, traditional LSTM-RNN approach has provided good results in music emotion regression [3]. In this work we propose the use of attention mechanism for dynamic emotion prediction in music. According to the *attention* model [22], to compute each output of a encoder-decoder architecture, a distinct *context vector* is used, which is a function of all the hidden states at the encoder side and not just the last one. The encoder encodes the input into a set of hidden states and attention is applied on them to produce target arousal and valence values over fixed length segments or time frames of the music audio signal. The encoder reads the input sequence $\mathbf{x} = (x_1, x_2, \ldots, x_T)$, which is a sequence of vectors, and produces the hidden states $(h_1, h_2, \ldots, h_T)$, using some RNN approach. In this work LSTM is used. In traditional attention mechanism [22], the whole set of hidden states $(h_1, h_2, \ldots, h_T)$ are available to compute the context vectors. Each time, the context vector c_i is calculated as a weighted sum of all the hidden states. Let the output be $\mathbf{y} = (y_1, y_2, \ldots, y_T)$. For the current problem, $\mathbf{y}$ can be defined as set of arousal or valence values associated with each music time frame. The t^{th} output, y_t, will be a function $\mathbf{g}()$ of a) the present hidden state h_t, b) the previous output y_{t-1}, c) the unique context vector c_t, as given by equation 1.

$$p(y_t | y_1, y_2, \ldots, y_{t-1}, \mathbf{x}) = g(h_t, y_{t-1}, c_t) \tag{1}$$

The unique context vector c_t depends on the sequence of annotations $(h_1, h_2, \ldots, h_T)$, and is computed as a weighted sum of these annotations h_j, as given in equation 2.

$$c_t = \sum_{j=1}^{T} \alpha_{tj} h_j \tag{2}$$

So, the model at time t, *attends* to each h_j corresponding to each of the inputs, with a weight of α_{tj}. To obtain each weight α_{tj} for each output y_t, the alignment between the corresponding h_t and each of h_j need to be calculated, where $1 \leq j \leq T$. So, the alignment model, when attending to h_j, is given by equation 3.

$$e_{tj} = a(h_{t-1}, h_j), 1 \leq j \leq (t-1) \tag{3}$$

This alignment is the measure of how well the inputs around position j and the output at position t match. Then, each of these scores e_{tj} are used to calculate the attention weights for each h_j as given in equation 4.

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{k=1}^{T} exp(e_{tk})} \tag{4}$$

So, for each output, the context vector will *attend* or focus on those parts of the entire input sequence, which are more relevant for that particular output, by assigning higher weights to the associated encoder-side hidden states, using an *alignment* model. These models are referred to as the *AT* models from hereon. The naming convention of the models is the acronym *AT* for attention, followed by the hidden layer dimensions.

Table 1: Model Selection for Dynamic Arousal Prediction

Model	**Parameter Search (Layer Size)**	**Best Model**	R^2_A	$\overline{\tau}_A$	MAE_A
Baseline [3]	400	-	0.60	0.14	0.11
LSTM (Single Layer)	128, 300, 400, 512, 700, 1024, 2048	1024	0.73	0.12	0.12
LSTM (Multi Layer)	(700_128), (700_400), (2048_1024), (2048_1024_700)	(700_128)	0.69	0.20	0.12
AT (Single Layer)	32, 64, 128, 300, 400, 512, 700, 1024, 2048	300	0.75	0.15	0.13
AT (Multi Layer)	(300_128), (400_128), (1024_400), (2048_1024), (2048_1024_512)	(2048_1024)	**0.78**	0.24	0.11
BAT (Single Layer)	400, 1024, 2048	1024	0.55	0.04	0.12
BAT (Multi Layer)	(300_128), (400_128), (1024_400), (1024_512), (2048_1024), (2048_1024_512)	(2048_1024)	0.58	0.06	0.12
Transformer	1-Layer, 2-Layer, 4-Layer	2-Layer	0.64	0.61	0.27

Table 2: Model Selection for Dynamic Valence Prediction

Model	**Parameter Search (Layer Size)**	**Best Model**	R^2_V	$\overline{\tau}_V$	MAE_V
Baseline [3]	400	-	0.29	0.08	0.16
LSTM (Single Layer)	128, 300, 400, 512, 700, 1024, 2048	700	0.39	0.10	0.15
LSTM (Multi Layer)	(700_128), (700_400), (2048_1024), (2048_1024_700)	(2048_1024)	0.29	0.17	0.15
AT (Single Layer)	32, 64, 128, 300, 400, 2048 512, 700, 1024, 2048	400	**0.53**	0.08	0.16
AT (Multi Layer)	(300_128), (400_128), (1024_400), (2048_1024), (2048_1024_512)	(300_128)	0.51	0.04	0.16
BAT (Single Layer)	400, 1024, 2048	2048	0.16	0.13	0.15
BAT (Multi Layer)	(300_128), (400_128), (1024_400), (1024_512), (2048_1024), (2048_1024_512)	(400_128)	0.21	0.16	0.14
Transformer	1-Layer, 2-Layer, 4-Layer	1-Layer	0.12	0.11	0.10

3.2 Backward Attention Model (BAT)

A modified form of the traditional attention mechanism [22] is also used in the current work, called Backward Attention (BAT) models. In these models, for emotion prediction at each t^{th} time frame, attention is distributed only among h_k hidden states, where, $1 \leq k \leq (t-1)$.

3.3 Transformers

The transformer architecture as proposed in Vaswani et. al. [2] is used in this work, with changes in the number of encoder side layers, as appropriate for the experiments. Attention is calculated as in equation 5.

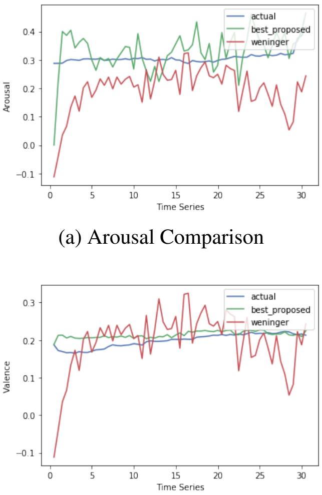
$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \qquad (5)$$

where, Q, K and V are matrices representing the set of queries, keys and values respectively and d_k is the key dimension.

4. EXPERIMENTS

4.1 Data Description and Experimental Setup

We use the *1000 Songs for Emotional Analysis of Music* dataset [8] for all experiments. Of the thousand clips,the dataset provides arousal and valence annotations for only 744 clips, which are used as *ground truth* values. According to the dataset manual [8], arousal-valence continuous annotations for each song (second 15-45), with 2Hz sampling frequency are available in the dataset. We define each non-overlapping 500ms of the clips as *one music frame*. Thus, the last 30s or the last 61 frames of each clip are used for this work, since only those 61 emotion (arousal-valence) tags are available. 10-fold cross validation was used on the training and test sets. We used the Mean squared error (MSE) as the loss function. RMSProp, with the default learning rate of 0.001 was used for optimizing the loss with a batch size 20, and maximum 50 epochs. An early stopping strategy is also used, if validation error shows no improvement over 10^{-4} after 5 epochs, processing is stopped. Sequences are presented in random order during training. All hyper-parameters not explicitly mentioned here are left to their default values as in Tensorflow 1.14. The feature sets used for different experiments are described below.



(a) Arousal Comparison

(b) Valence Comparison

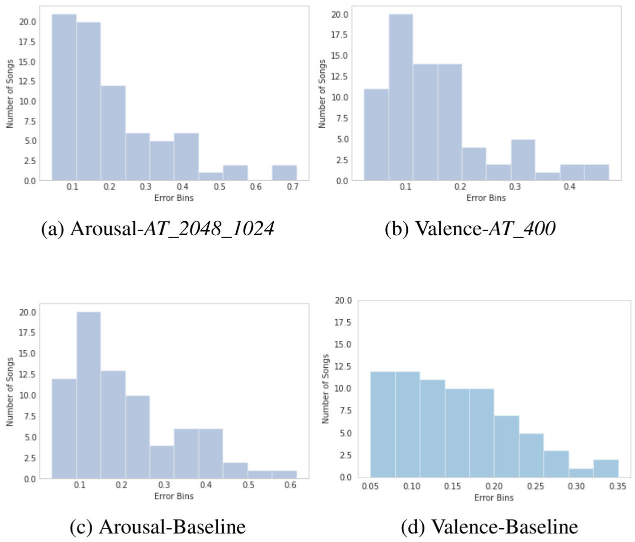
Figure 1: Dynamic Emotion Predictions for Clip 584

4.1.1 ComPare Feature Set

The 2013 Computational Paralinguistics Evaluation (ComParE) tasks featureset [1], containing 6670 features is used for all experiments in sections 4.2 and 4.4. TUM's open-source *openSMILE* feature extractor [23] is used to extract the ComParE featureset for each frame of each clip. Standard normalization was performed on the extracted feature values before the experiments. So, each clip is characterised by 61 feature vectors, each of size 6670.

4.1.2 Other Feature Sets

In experiments reported in section 4.3, subsets of the Compare feature set [1] and some other features are explored. These features extracted using Librosa [24] are detailed here. The *Chroma(STFT+CQT)* features [24] consist of chroma values derived using both STFT analysis and constant-Q transform (CQT) analysis implementations. The *CQT on Audio clip* features [24] are derived from the core Spectrogram operations of Librosa [24] suitable for pitch-based signal analysis. The *Spectral Features* [24] denote the distributions of energy over a set of frequencies and are very important in many MIR analysis techniques. These consist of: Chroma(24), CENs (12) MFCC (20), RMS (1), Mel-scaled spectrogram (128), spectral centroid (1), spectral bandwidth (1), spectral contrast (7), spectral flatness (1), spectral roll-off (1), zero crossing rate (1). All clips were re-sampled to 44100 Hz before feature extraction. All features were extracted for non-overlapping frames of 500 ms each, corresponding to the available arousal-valence labels of the dataset.



(a) Arousal-*AT_2048_1024*

(b) Valence-*AT_400*

(c) Arousal-Baseline

(d) Valence-Baseline

Figure 2: Emotion Error Histograms over Validation Set

4.1.3 Metrics

The metrics used for reporting the results are Coefficient of determination (R^2), average Kendall's τ per song ($\overline{\tau}$)and mean absolute error (MAE). The determination coefficient (R^2) is a key output of regression analysis, which provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. It can vary between 0 and 1. If a data set has n values marked $(y_1 \ldots y_n)$, and each associated with a predicted value $(f_1 \ldots f_n)$. So, R^2 is defined as $R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}}$ where, $SS_{res} = \sum_i (y_i - f_i)^2$ and $SS_{tot} = \sum_i (y_i - \overline{y})^2$, given $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Kendall's τ per song ($\overline{\tau}$) is a measure of how well the emotional profile of each song is captured by the regressor, as opposed to overall correlation. It measures the correspondence between two rankings. Values close to 1 indicate strong agreement, values close to -1 indicate strong disagreement. It is defined $\overline{\tau} = \frac{P-Q}{\sqrt{(P+Q+T)*(P+Q+U)}}$ where, P is the number of concordant pairs, Q the number of discordant pairs, T the number of ties only in target set $(y_1 \ldots y_n)$, and U the number of ties only in predicted set $(f_1 \ldots f_n)$. The mean absolute error (MAE) is given for reference. In the next section, we report the results of applying the proposed model for dynamic music emotion regression.

Baseline: It has been shown by Weninger et. al. [3, 6] that LSTMs can be used to produce good performance in emotion prediction, using the ComParE featureset. We try to reproduce their results using single layer LSTM-RNNs with hidden layer size of 400 units. These results are considered as *Baseline* in this work and are reported in the "Baseline" annotated rows of Table 1 and Table 2 for arousal and valence respectively.

4.2 Experiment 1: Model Selection

In the first set of experiments, we aim to find the best model for dynamic arousal and valence prediction, among the ones proposed in section 3. Accordingly, the models with attention (*AT*, *BAT*, *Transformers*) and without attention (*LSTM*) are executed with varying layer sizes and layer numbers. The findings for arousal and valence are reported in Table 1 and Table 2 respectively. For dynamic arousal prediction (Table 1) using the ComPare feature set [1] (sec 4.1.1), the best result is obtained with the multi-layer attention model *AT_2048_1024*. Comparable result is also obtained with single-layer attention model *AT_300*. The best model for dynamic valence prediction (Table 2) is found to be the single-layer attention model *AT_400*. Comparable result is also obtained with multi-layer attention model *AT_300_128*.

Table 3: Feature Sets for Arousal Prediction

Features Used	# Features	Best Model	R^2_A	$\overline{\tau}_A$	MAE_A
Chroma(STFT+CQT)	24	AT_64	0.15	0.04	0.19
CQT on Audio clip	252	AT_64	0.45	0.06	0.17
Chroma+CQT	276	AT_64	0.57	0.07	0.14
Spectral Features	197	AT_64	**0.70**	0.03	0.12

Table 4: Feature Sets for Valence Prediction

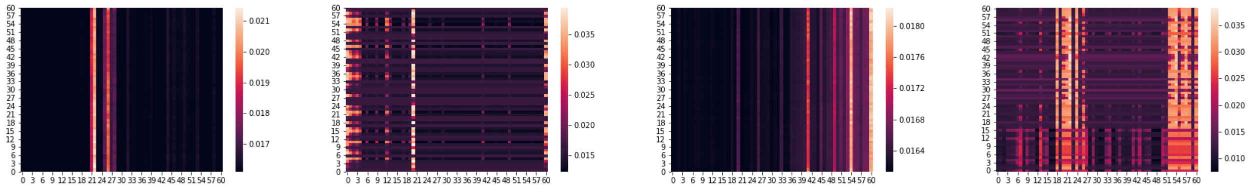
Features Used	# Features	Best Model	R^2_V	$\overline{\tau}_V$	MAE_V
Chroma(STFT+CQT)	24	AT_64	0.01	0.002	0.09
CQT on Audio clip	252	AT_64	0.07	0.01	0.17
Chroma+CQT	276	AT_64	0.17	0.06	0.14
Spectral Features	197	AT_128	**0.35**	0.07	0.16

The following are observed from this experiment: a) The best prediction performances reported in this section are better than that reported by the baseline methods (sec 4.1.3). b) Among all the experiments conducted, *AT* models fare best in dynamic arousal-valence prediction using the full ComPare feature set [1]. c) The best single and multi layer *AT* models' performances are comparable. d) Performance for arousal prediction (R^2_A and $\overline{\tau}_A$) in general is much better than valence (R^2_V and $\overline{\tau}_V$) - across all models tested. Though performance with respect to MAE are comparable.

In the following subsections, we demonstrate an illustrative example of dynamic emotion prediction using a clip chosen at random, followed by an error analysis of the predictions by the best proposed models, over the validation set clips.

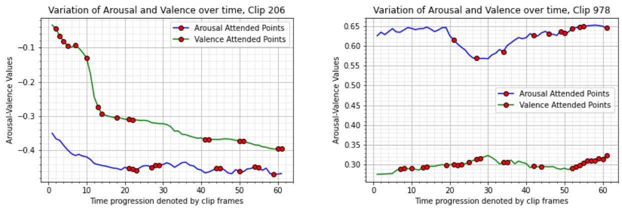
4.2.1 Illustrative examples

In this section, we demonstrate an illustrative example of dynamic emotion prediction pattern, with respect to ground truth (sec 4.1) and baseline (sec 4.1.3), using a clip chosen at random from the dataset [8]. The best models, *AT_2048_1024* for arousal and *AT_400* for valence, obtained in section 4.2 are used for dynamic (per 500 ms) arousal and valence prediction of music clip *584.mp3*. This is presented in Figure 1. Figure 1a and Figure 1b denote the time varying arousal and valence predictions respectively. In the figures, X-axis denote the time (in seconds), and the Y-axis denote arousal and valence values respectively. It is seen that the proposed best models follow the pattern of reported emotions more closely than baseline model.



(a) Clip 206 - Arousal (b) Clip 206 - Valence (c) Clip 978 - Arousal (d) Clip 978 - Valence

Figure 3: Attention Maps using AT models. X-axis = attention points (500ms clip frames), Y-axis = prediction points (clip's progression through time)



(a) Clip 206 (b) Clip 978

Figure 4: Comparing attended frames with ground truth Emotion ratings of dataset [8]

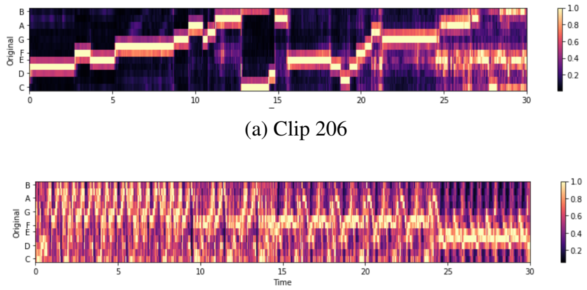
4.2.2 Errors Analysis

In this section we aim to observe patterns and biases in the best proposed models' (sec 4.1) emotion predictions, with respect to the baseline (sec 4.1.3). The respective predictions are utilized to group the validation set clips into error bins for this study. These are shown as histograms in figure 2. The X-axis denote the error bins of the models over the validation set clips. The Y-axis denote the number of clips of the validation set, which fall into each error bin. Comparing Figure 2a and Figure 2c, it can be seen that, for the proposed model, the number of clips with higher values of errors are less, in case of arousal. In case of valence, for the proposed model, almost all the clips are grouped into the error bins $\leq$ 0.05 (Figure 2b). Whereas for the baseline model ((Figure 2d)), a significant number of clips across bins are present.

4.3 Experiment 2: Exploring Other Feature Sets

In section 4.2, all the experiments use the full ComPare feature set [1]. Though it performs well in dynamic emotion prediction in music, it might be noted that it is generic, not music specific. It is large, which causes models to have large number of parameters. Also, there might be other relevant features, which might be used for this task, eg. Constant Q Transform features. In this section, we explore some smaller feature sets detailed in section 4.1.2, which might possibly produce similar or better results, over the same dataset [8], with the additional benefit of being smaller in size.

Single layer *AT* models were used to train on these new feature sets, since, it was observed in section 4.2 that they perform best and at par with multi layer models for emotion prediction. The results are presented in Table 3 and Table 4 for arousal and valence respectively. For arousal (Table 3), it is observed that *AT_64* performs well, when using the *Spectral Features* set, with a R_A^2 comparable to the best model *AT_2048_1024* using full ComPare [1] feature set. It is evident that Chroma features alone have negligible contribution in arousal prediction. *CQT* set performs moderately. For valence prediction (Table 4) also, *Spectral features* set performs best among all. *CQT* set does not contribute much to valence prediction. Thus, we conclude that there might be a possibility of a smaller featureset for emotion prediction.



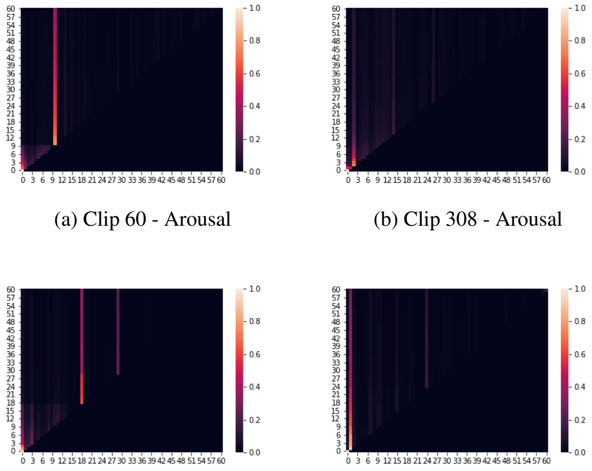
(a) Clip 206

(b) Clip 978

Figure 5: Chromagrams for Attention Map Analysis. X-axis = time (in seconds), Y-axis = Chroma. Vertical bars=Chroma intensities

4.4 Attention Maps for Emotion Prediction

Attention maps demonstrate the relative importance of layer activations at different 2D spatial locations with respect to arousal and valence predictions. In this section, the best *AT* and *BAT* models are used to generate the attention maps for both arousal and valence, for some clips chosen at random from the dataset [8], presented in Figure 3 and Figure 6. These maps provide information about those frames of the clip, which are attended to during emotion prediction. This in turn can yield valuable insights into specific audio features of those frames, conducive to certain emotion perception. For all the maps, X-axis signifies the attention points, which are the 500 ms frames of the clip the model attends to. The Y-axis signifies the prediction points, the clip's progression through time. It is to be noted that these 61 frames in the maps, correspond to the



(a) Clip 60 - Arousal

(b) Clip 308 - Arousal

(c) Clip 60 - Valence

(d) Clip 308 - Valence

Figure 6: Attention Maps using BAT models. X-axis = attention points (500ms clip frames), Y-axis = prediction points (clip's progression through time)

last 30 seconds of each clip, as per the dataset [8]. So, the s^{th} frame of a clip, is actually the $(15 + \frac{x-1}{2})^{th}$ second of the entire 45 second clip. The vertical bars on the right of each attention map give the attention weight values present in each map. The observations are discussed in the following subsections.

4.4.1 Attention Maps Using AT models

The attention maps for arousal and valence prediction, generated using *AT_2048_1024* and *AT_400*, for clips 128.mp3, 171.mp3, 206.mp3, and 978.mp3 from the dataset [8] are presented in Figure 3. Figure 3a and Figure 3b demonstrates the attention maps for arousal and valence prediction in clip 206.mp3 . As evident from the figure 3a, the model attends mostly to the clip frames 20-22, 26-28, and then again frames between 43-44, 49, 53-54 and 58 to predict arousal. From figure 3b, it is observed that the model attends to the frames 1-4, 6, 9, 12-13, 17, 20-21, 40-41, 49-50 and 59-61 to predict valence. Similar observations can be made about the other clips as well from Figure 3.

Observations: For arousal prediction, the model attends to comparatively fewer frames of the clip. These attended frames are observed to occur around 10 seconds (20 frames) after the clip has started. It can be concluded that the arousal generated in the later part of the music clip plays a significant role in determining the arousal perception of the entire clip. The attended frames have arousal ratings which are approximately average of all the arousal ratings for a particular clip. On the other hand, for valence prediction, attention is distributed across the clip, whenever there is perceptible change in valence ratings. Thus it can be concluded that reports of valence depends on momentary perception. Even small changes are registered. The attended frames have quite varied valence rating values within a particular clip.

For further investigation, we juxtapose our findings with a) The dynamic arousal and valence ratings provided by the dataset [8] - ground truth, given in Figure 4, and b) Chromagrams of the clips obtained using Librosa [24], presented in Figure 5. In each line graph of Figure 4, the X-axis denotes time frames, and Y-axis denotes the arousal and valence values.

It is to be noted here that the clips 206 and 978 are so chosen that they have significantly different arousal and valence ground truth values. In clip 206, the arousal values are lesser than the valence values. In clip 978, the reported arousal values are greater than the valence values. The blue and green lines denote arousal and valence respectively, the red dots highlight the time frames attended to by the *AT* models, as evident from Figure 3. In each subplot of Figure 5, the X-axis denotes time (in seconds), and the Y-axis denotes the Chroma. The vertical bars indicate the intensities of the Chroma. Figure 5a demonstrates the chromagram for clip 206.mp3.

Observations: For arousal prediction, the model attends on those frames with stable presence of higher notes (eg. A, B). For valence prediction, model attends all over the chroma bins, specially when there is a change in notes in the chroma sequence of the clip. Similar observations might be made from the other chromagrams as well.

4.4.2 Attention Maps using BAT models

The attention maps generated using the BAT models, *BAT_2048_1024* for arousal are presented in Figure 6. Figure 6a gives the attention map for arousal prediction in clip 60.mp3 of the dataset [8]. As evident from the figure, the attention of the model shifts continuously throughout the clip, as it progresses in time, though Segments 11-12 receive maximum attention overall. Similar trends are observed in Figure 6c as well, which represents the map for valence prediction for the same clip. Initially, the first few segments are attended to. As the clip progresses in time, the attention is shifted to later segments, with segments 18-19 and 29-30 being more prominent. As the clip progresses, the attention to initial segments reduces, rendering the lower right triangular region of the maps devoid of any attention traces.

5. CONCLUSION

We demonstrate that the state of the art models for continuous-time emotion prediction perform modestly, thus emphasizing the need for further research in this area. We have proposed an attentive LSTM based model which improves the state of the art performance significantly, on standard benchmark dataset with standard metrics. Further, we observe that a reduced, music-specific feature set achieves similar performance to the new state of the art model on arousal prediction, leading to much smaller models. Finally, we analyse attention maps for the full attention model to conclude that the model indeed attends to critical portions of the music in order to predict the dynamic emotions. We also observe that the nature of attention is different in case of arousal and valence prediction tasks.

6. REFERENCES

[1] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[3] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *ICASSP*. IEEE, 2014, pp. 5412–5416.

[4] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *ISMIR*, vol. 86, 2010, pp. 937–952.

[5] E. Coutinho, F. Weninger, B. W. Schuller, and K. R. Scherer, "The munich lstm-rnn approach to the mediaeval 2014 "emotion in music" task." in *MediaEval*, 2014.

[6] F. Weninger, F. Ringeval, E. Marchi, and B. W. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio." in *IJCAI*, vol. 2016, 2016, pp. 2196–2202.

[7] F. Weninger, F. Eyben, and B. Schuller, "The tum approach to the mediaeval music emotion task using generic affective audio features," in *Proceedings MediaEval 2013 Workshop*, 2013.

[8] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*. ACM, 2013, pp. 1–6.

[9] S. Giammusso, M. Guerriero, P. Lisena, E. Palumbo, and R. Troncy, "Predicting the emotion of playlists using track lyrics," *ISMIR, Late Breaking Session*, 2017.

[10] J. Fan, K. Tatar, M. Thorogood, and P. Pasquier, "Ranking-based emotion recognition for experimental music." in *ISMIR*, 2017, pp. 368–375.

[11] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," *ISMIR*, pp. 370–375, 2018.

[12] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," *ISMIR*, 2011.

[13] Y. Song, S. Dixon, and M. T. Pearce, "Evaluation of musical features for emotion classification." in *ISMIR*. Citeseer, 2012, pp. 523–528.

[14] R. Panda, R. Malheiro, and R. P. Paiva, "Musical texture and expressivity features for music emotion recognition." in *ISMIR*, 2018, pp. 383–391.

[15] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[16] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 2, pp. 448–457, 2008.

[17] S. Balke, M. Dorfer, L. Carvalho, A. Arzt, and G. Widmer, "Learning soft-attention models for tempo-invariant audio-sheet music retrieval," *ISMIR*, pp. 216–222, 2019.

[18] S. Gururani, M. Sharma, and A. Lerch, "An attention mechanism for musical instrument recognition," *ISMIR*, pp. 83–90, 2019.

[19] C. Donahue, H. H. Mao, Y. E. Li, G. W. Cottrell, and J. McAuley, "Lakhnes: Improving multi-instrumental music generation with cross-domain pre-training," *ISMIR*, pp. 685–692, 2019.

[20] T.-P. Chen and L. Su, "Harmony transformer: Incorporating chord segmentation into harmony recognition," *ISMIR*, pp. 259–267, 2019.

[21] J. Park, K. Choi, S. Jeon, D. Kim, and J. Park, "A bi-directional transformer for musical chord recognition," *ISMIR*, pp. 620–627, 2019.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR*, 2015.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[24] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.