Deep Learning for Emotion Recognition in Affective Virtual Reality and Music Applications

Jason Teo, Jia Tian Chia, Jie Yu Lee

Abstract: This paper presents a deep learning approach to emotion recognition as applied to virtual reality and music predictive analytics. Firstly, it investigates the deep parameter tuning of the multi-hidden layer neural networks, which are also commonly referred to simply as deep networks that are used to conduct emotion detection in virtual reality (VR)-electroencephalography (EEG) predictive analytics. Deep networks have been studied extensively over the last decade and have shown to be among the most accurate methods for predictive analytics in image recognition and speech processing domains. However, most predictive analytics deep network studies focus on the shallow parameter tuning when attempting to boost prediction accuracies, which includes deep network tuning parameters such as number of hidden layers, number of hidden nodes per hidden layer and the types of activation functions used in the hidden nodes. Much less effort has been put into investigating the tuning of deep parameters such as input dropout ratios, L1 (lasso) regularization and L2 (ridge regularization) parameters of the deep networks. As such, the goal of this study is to perform a parameter tuning investigation on these deep parameters of the deep networks for predicting emotions in a virtual reality environment using electroencephalography (EEG) signal obtained when the user is exposed to immersive content. The results show that deep tuning of deep networks in VR-EEG can improve the accuracies of predicting emotions. The best emotion prediction accuracy was improved to over 96% after deep tuning was conducted on the deep network parameters of input dropout ratio, L1 and L2 regularization parameters. Secondly, it investigates a similar possible approach when applied to 4-quadrant music emotion recognition. Recent studies have been characterizing music based on music genres and various classification techniques have been used to achieve the best accuracy rate. Several researches on deep learning have shown outstanding results in relation to dimensional music emotion recognition. Yet, there is no concrete and concise description to express music. In regards to this research gap, a research using more detailed metadata on two-dimensional emotion annotations based on the Russell's model is conducted. Rather than applying music genres or lyrics into machine learning algorithm to MER, higher representation of music information, acoustic features are used. In conjunction with the four classes classification problem, an available dataset named AMG1608 is feed into a training model built from deep neural network. The dataset is first preprocessed to get full access of variables before any machine learning is done. The classification rate is then collected by running the scripts in R environment. The preliminary result showed a classification rate of 46.0%. Experiments on architecture and hyper-parameter tuning as well as instance reduction were designed and conducted. The tuned parameters that increased the accuracy for deep learners were hidden layer architecture, number of epochs, instance reduction, input dropout ratio and ℓ1 and ℓ2 regularization. The final best prediction accuracy obtained was 61.7%, giving an overall improvement of more than 15% for music emotion recognition which are based purely on the music's acoustical features.

Revised Manuscript Received on June 22, 2019.

Jason Teo, Faculty of Computing & Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia

Jia Tian Chia, Faculty of Computing & Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia

Jie Yu Lee, Faculty of Computing & Informatics, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia

Index Terms: neuroinformatics, virtual reality, deep learning, electroencephalography, emotion classification, music emotion recognition, acoustic features.

I. INTRODUCTION

The ability to predict emotions [1] accurately from using only commercially-available wearable electroencephalography (EEG) devices [2] has important ramifications particularly in virtual reality applications. For example, being able to detect a computer game player's emotions as the player plays the game in virtual reality will enable unique custom-created immersive content that is procedurally generated in real-time based on the player's response to how he or she is currently being stimulated emotionally by the game; being able to detect the level of pleasure or frustration of a physically impaired patient undergoing physical therapy via virtual rehabilitation and adjusting the content of the virtual therapy in real-time based on the detection of the patient's current emotions while undergoing the virtual therapy; or being able to detect the arousal or boredom of learners in a virtual learning environment as he or she is being exposed to while being exposed to learning materials immersively and being able to tune, modify and optimize the presentation of subsequent learning content to the learner in real-time based on this learner's emotional condition while absorbing the learning material.

Our previous attempt at predicting emotions in a virtual reality environment which exposed the user to a high-adrenaline stimulus in the form of a roller-coaster experience as captured in a YouTube 360 video investigated a number of traditional classification algorithms including support vector machines (SVMs), k-Nearest Neighbor (kNN) classifiers, random forests (RFs), conventional single hidden layer artificial neural networks (ANNs), and decision trees (DTs), where the best prediction accuracy obtained was between 65-89%. In this current study, we investigate the use of deep learning classifiers and more importantly, to perform deep tuning of the deep networks parameters for input dropout ratio, L1 (lasso) regularization and L2 (ridge) regularization [3], in an attempt to further boost the prediction accuracy of our VR-EEG predictive analytics problem in emotion detection using commercially-available wearable headsets for acquiring the brainwave signals.

Moreover, the music industry has shifted to the digital distribution era where online stores and streaming services are in the lead. Few of the examples are Spotify, iTunes and Grooveshark. The shift had caused automatic music recommendation to become a relevant problem whereby it is essential to have an accurate matching between users' taste and their choice. These online stores are then able to gain more customer by aiming right targets to the right audience. With the enormous growing of data, there are many effective searching methods in indexing, classification or clustering.

Artificial Intelligence (AI) is fast becoming one of the key disruptive technologies across multitudes of industry. Industry heavyweights such as Google, Spotify, Microsoft, Uber, Facebook and Apple are all prime adopters of this new wave of technology with many investing significantly in the field of AI [4]. As an example of the music industry's customer base and data processing requirements, Spotify possesses about 24 million users, where users are able to access into a music database with more than 20 million songs, either with free access with advertisements or paid access without advertisements.

As Spotify adds approximately 20,000 songs per day, the business needs to handle this amount of processing of audio files has given rise to the potential use of AI in its daily operations. The company is acquiring a number of AI-startup companies to improve song recommendations and targeted advertisements. Automatic systems which can predict human emotions from speech, as well as music,are essential in developing new AI-driven applications in the entertainment industry.

Another example of an entertainment company that combines machine learning with its business needs is Netflix. The goal of its application of AI technology is to make recommendations based on its users' favorite shows and movies. It initially makes suggestions based on actors, genre and filming location of the movies, yet, issues such as incorrect recommendations have occurred and this has lead the company to investigate deep learning approaches in improving its recommendations [5]. The approach is to train its software by feeding large amounts of information to a learning machine called artificial neural networks, which mimic the human brain in terms of pattern identification to produce better recommendations. Emotion recognition is not only subjective but is also hard to analyze and quantify. Thus, generating a flexible model for music emotion recommendation is a prime research question.

In this research, a deep neural network (DNN) is used to train the data. The objective is to build a music emotion recommender system using deep architectures of artificial neural networks to make song recommendations to humans. Human emotions can be of different types, such as happiness, fear, anger and boredom, it is thus a challenging task to classify and make a prediction based on the current emotionsof the user while listening to music. The outcomes of the research would provide a better understanding on the source of inspired emotions for a given music.

Since this second part of the study deals with a very large amount of data, a deep architecture is employed to train and produce better prediction models. Deep networks typically utilize a number of hidden layers and due to the multiple layers, are better at learning certain hierarchical features beginning from lower-level features and moving to higher-level features in its learning process. Deep learning is an emerging area in both machine learning communities and data mining. The models can be trained either in a supervised or unsupervised mode. It is has been shown to have been applied successfully to computer vision, audio, speech as well as language processing domains. The deep learning approach is regularly reported to have outperformed many other state-of- the-art machine learning approaches.

The paper is divided into five sections. We introduce the study in the first section, followed by literature to explain the background of this study in the second section, then the methods adopted in this study are presented in the third section, followed by the presentation of the results and discussion in the fourth section, and finally the conclusion is given in the final section.

II. RELATED MATERIAL

A. Affective Virtual Reality

The use of electroencephalography as a non-invasive method of detecting emotion has gained significant traction over the last five years. Significant advancements have been made with diverse successful implementations ranging from diagnostic uses such as in detection of autism in young children [6], safety mechanisms in AI-assisted driving solutions [7], to personalized content in affective entertainment [8]. One crucial aspect necessary for the general adoption of EEG-based emotion detection among consumers is the availability of affordable, commercial-off-the-shelf EEG headsets commonly referred to as wearable EEG devices. Such devices allow for end-users such as consumers and hobbyists to acquire a number of widely available wearable EEG devices at sub-$200 prices. The widespread availability of such affordable wearable EEG devices has allowed for the field of emotion-based applications based on EEG to generate a great amount of interest from lay users and subsequently provides a very strong motivation for further research into EEG-based emotion detection using such consumer-grade devices.

Nonetheless, with the advent of such consumer-grade EEG devices comes another challenge, which is to be able to accurately perform emotion detection using much fewer sensors coupled with the fact that such sensors have lower signal quality compared to their medical-grade, laboratory-based EEG counterparts that cost anywhere from ten to a thousand times more than consumer grade EEG devices. Specifically, the challenge here is that the number of electrodes available to capture EEG signals from the user's scalp using consumer-grade devices is typically in the region of only 1-4 sensors as compared to medical grade

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

devices that typically have 32-128 sensors. Moreover, the signal acquired from using consumer-grade EEG devices tends to contain more noise than medical-grade EEG devices. As such, the ability to detect emotions is a much more challenging endeavor when using consumer-grade EEG devices compared to laboratory settings using medical-grade EEG devices.

Another aspect of emotion detection that has not received as much attention is in virtual reality (VR) environments. The value of the VR industry worldwide is projected to exceed USD200 billion by 2020 and one of the main areas of revenue generation in this industry comes from immersive entertainment such as VR computer gaming. The ability to create personalized VR content based on the user/gamer's emotion in real-time will allow for the step in the evolution of affective gaming. As such, being able to accurately detect emotions while immersed in VR environments using only affordable wearable EEG devices paired with similarly affordable, high-resolution VR headsets holds great promise in moving true affective entertainment into the consumer/gamer's home. Consequently, this forms the major motivation for conducting this study which attempts to perform deep tuning in order to improve the emotion detection accuracy based on the brainwave readings obtained from affordable wearable EEG devices while the user is immersed in virtual reality.

B. Affective Music Applications

Recently, deep neural architectures have been investigated in implementing entertainment recommender systems. As more researchers are working on this issue, the success in the field has encouraged researchers to propose more on the learning of latent factors from different sources. Multi-view deep models can be generated to learn a rich feature set for users from web browsing history and search queries. Emotions are a psychological and physiological state, related to a variety of thoughts, feelings, and behaviors. Research shows that there is a delicate relationship between music and emotion. Organized sound or music, can resonate with our nerve tissue. Emotion is always related to mood, temperament and personality. They are experienced from an individual point of view. It can be said that emotions are short-term, moods have longer term whereas personalities are in very long-term. Emotion recognition from musical stimuli represents a highly challenging task since the extraction and identification of effective musical features for emotion classification remains an open question. A community-based framework for evaluating MIR systems, Music Information Research Evaluation eXchange (MIREX), included audio music mood classification as a task in 2007 for the first time.

Yet, emotion has not been incorporated within the music metadata for music information retrieval (MIR) purpose. The reason is due to some unavoidable key problems found in music recommenders, which are cold starts, popularity bias and human effort. Cold starts, known as sparsity problems, mean that there is a lack of ratings, which results in poor prediction results. Popularity bias causes unpopular and new songs to get less recommendation. Lastly, a good recommender system should involve minimum human effort, since we are seeking more accurate results.

Machine learning excels at deciphering patterns from complex data. The results showed that the most significant features are danceability, energy, loudness, tempo and time signature. Moreover, there are no precise definitions for each genre. Musical genres are usually determined based on the name of the artist or the name of the album rather than on the individual musical recordings. This typically results in MP3 metadata tags that tend to have less than reliable annotations. Moreover, new musical genres are introduced from time to time, and as such the knowledge base of musical genres would naturally evolve and grow over time.

As for dataset preparation, the larger the input layer to neural network, the better the system performs, and the use of limited song features indicates a more robust system. Furthermore, input data should be reduced to reduce noise. Therefore, before performing any training, it is essential to understand and design the inputs so that the dataset is suitable for the deep architecture usage. Findings and results from prior studies have shown that the selection for dataset and feature are important in music classification, which will also have a direct impact on music recommendation.

When it comes to the context of music, when related to automated systems, very often it is associated with music genres. Yet, choosing songs is arguably not based on genres but rather as a result of emotions. A lot of work is being done in music emotion recognition. However, there is no standard benchmark for music recommendation, especially when it deals with human emotion sincehuman emotion varies from time-to-time. Deep learning methods mimics the architecture of human brains, which is therefore worthwhile to investigate its use for the analysis and classification of music data.

One of the first studies to apply deep learning in unsupervised music content analysis was reported by Lee et al. [9]. Using two hidden layers within a convolutional deep belief network, the deep learning architecture was trained using unsupervised learning to produce meaningful musical features. Subsequently, a conventional neural network was used to classify songs into four distinct types of musical genres which were techno, rock, classical, techno and pop. Using a Hidden Markov Model (HMM) with sequences of features, Shao et al. [10] reported on musical clustering of pop, country, jazz and classical music in an unsupervised machine learning approach to music content analysis.

Hamel & Eck [11]were considered the earliest to apply deep learning in supervised music content classification. The accuracy for genre classification was 84.3%. Li et al. [12] used Convolutional Neural Networks(CNNs) for MIR and was also one of the pioneer studies that adopted a deep learning approach in analysing musical content. Dahl et al. [13]showed that using optimal dropout ratios and Rectified Linear Units (ReLUs) can result in significant classification improvement with only minimal hyper-parameter tuning in a Large Vocabulary Continuous Speech Recognition (LVCSR) problem.

Sigtia & Dixon [14] improved the architecture of the deep learner by tuning more hyper-parameters in order to increase the music genre classification rate. In summary, neural networks typically need large amounts of training data, yet the understanding on the importance of feature reduction and hyper-parameter tuninglargely remainsunexplored particularly in the application of deep learning approaches in MIR. The number of features in the MIR datasetsas well as the deep learner's hyper-parameters is large and has very significant impact on the classification outcomes. It is therefore beneficial to find out the most optimal input features through feature reduction as well as tuning of deep learning hyper-parameter settings.

III. METHODS

A. Affective Virtual Reality

As previously highlighted, the underlying goal of this series of studies is to enable the deployment of consumer-grade EEG headsets for classifying emotions for VR users. To this end, we have chosen to use the Muse headset from Interaxon [15] as the EEG data acquisition device as shown in Fig. 1. The se-up is trivial since it is a dry-electrode device and just requires the user to put the headset on like putting on a headband. It also comes at a very affordable cost of below $200 per set. Another advantage of the Muse headset is that the user's head will remain fully mobile. This is due to the fact that this particular headband uses Bluetooth technology for connection and thus enables wireless data transmission, an aspect which is crucial for VR applications. Using the international 10-20 coordinates standard [16] for referencing the electrode placements on the skull, the Muse headband has four channels at TP9, AF7, AF8 and TP10 as shown in Fig. 2. A reference channel with 3 sensors are located at Fpz.

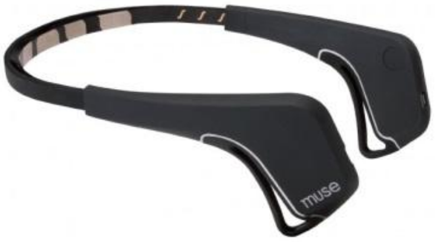


Fig. 1 The Muse wearable EEG headset from Interaxon (source: Interaxon)

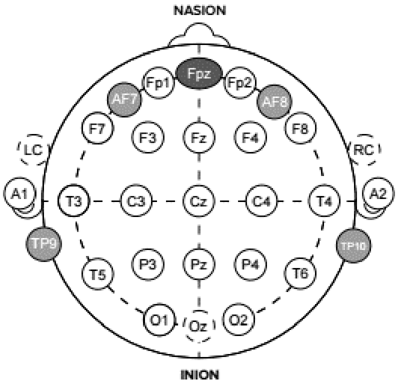


Fig. 2 Muse electrode locations according to the 10-20 international standard notation

For the immersive VR stimuli, we chose a roller-coaster video from YouTube 360 [17] and displayed using Google's Cardboard VR technology [18]. The reason we chose a roller coaster video is to elicit a strong excitement reaction to the VR stimuli, in which the video has two specific exciting sections, one which is a drop from a high peak and another comprising a series of high speed 360-degrees turns. Screengrabs of these two specific stimuli segments are as shown in Fig. 3 and Fig. 4.

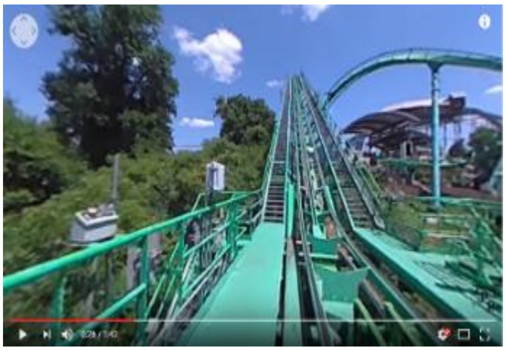


Fig. 3 Screengrab of the stimulus PRIOR to the segment which elicits the excited response

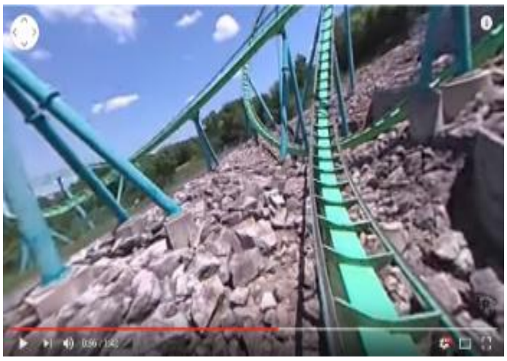


Fig. 4 Screengrab of the stimulus DURING one of the segments which elicits the excited response

The group of participants in this study comprised 24 VR users (12 females, 12 males) who had normal or corrected-to-normal vision with no history of psychiatric illnesses with the ages ranging between 20 to 28 years old. The VR users were asked to sit on a rotatable chair without any restriction to head movements while he or she was immersed within the VR stimuli environment. A photograph of a participant experiencing the VR session wearing the Muse headset is as shown in Fig. 5.

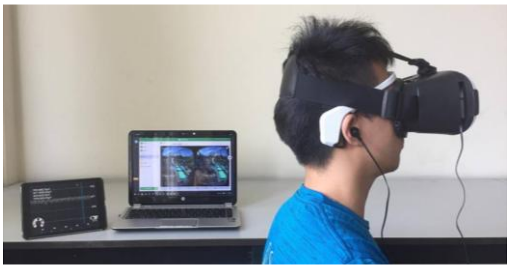


Fig. 5 View of experimental setup for Muse EEG

headset and VR headset on a human participant

B. Affective Music Applications

AMG1608 [19] is a dataset with 1608 30-second music clips that were manually annotated by 665 human listeners. The dataset is accessible online to the public. The songs contain Western music from AMG with 34 distinct mood categories. Songs are said to be distributed evenly in the emotion space. The valence arousal values are created from tag2VA algorithm. The quadrant column plays a significant role in this study's 4-class classification problem. Based on Russell'semotion model, each song will be annotated by the subject based on valence and arousal values. The song will be plotted into the emotion plane to indicate the emotional quadrant of that particular song. However, in this dataset, each song has at least 15 annotated emotions, and each instance cannot be removed and need to be treated as one unique transaction. The reason is that perceived emotion is very subjective, thus, all samples should be taken account.

RStudiois used to separate the dataset into valence and arousal space. From the original dataset downloaded online, it has a structure of 1608 x 665, where 1608 refers to the number of songs and 665 depicts the number of subjects who annotated the AMG1608 dataset. All NA values are then removed. One issue that was needed to be considered during processing is that each song will have a different number of annotations, where the minimum number of annotation for a song was 15, and the maximum was 32.

After processing all the NA values, the total number of instances, or annotated emotions for 1608 songs has a total of 26,914 instances. Since the original file of "song_label" are packaged into two layers, a validation on both valence and arousal values for each song is essential. This means that if there is an annotation found in one column, then the arousal layer should have the annotation as well, otherwise it is considered as an erroneous entry.

Lastly, a new column is added into the valence-arousal file, where the new column represents the quadrant of that particular annotation. For the MIR quadrants, the column consists of four types of classes, namely Q1 (happy), Q2 (upset), Q3 (bored) and Q4 (calm). The classes are based on valence and arousal values.

IV. EXPERIMENTAL RESULTS & DISCUSSION

A. Affective Virtual Reality

In order to conduct deep tuning of the deep learning neural networks, three parameters were investigated, namely the parameter settings for input dropout ratio, L1 (lasso) regularization and L2 (ridge) regularization [20,21]. The input dropout ratio determines the probability that an input feature will be suppressed during training in order to improve generalization of the deep network. L1 and L2 regularization is achieved by adding penalty values to the existing loss function via Equation (1) as follows:

$$L^{'}(W,B|j) = L(W,B|j) + \lambda_1 R_1(W,B|j) + \lambda_2 R_2(W,B|j)$$

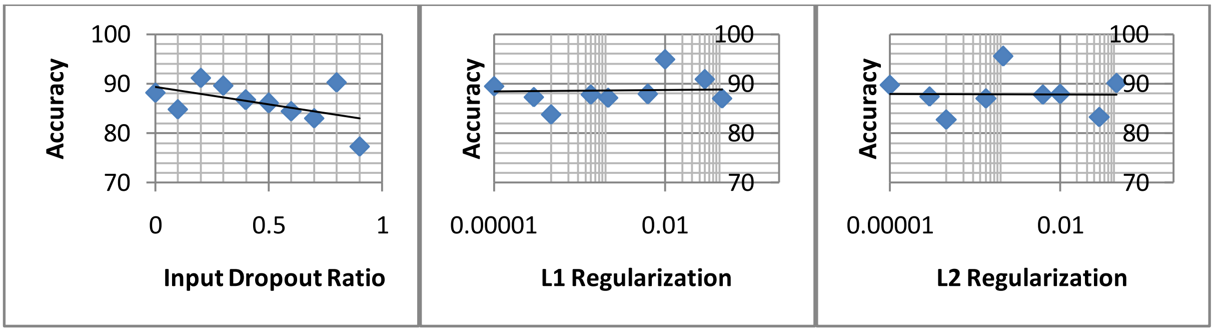
where L represents the loss function, W and B the weights and biases of the network; j the training instance; L1 regularization is achieved via R1(W,B|j) which represents of the sum of all absolute values of the weights and biases in the network; L2 regularization is achieved via R2(W;B|j) represents the sum of squares of all the weights and biases in the network; and the constants parameters λ1 and λ2 are usually set at a very small value such as 10-5. In essence, L1 reduces the complexity of the network in order to avoid overfitting and improve overall generalization whereas L2 regularization improves the overall learned model by reducing the estimate variance of the deep network classifier.

Preliminary testing using shallow tuning of the deep networks' parameters yielded neural network architectures that performed best when utilizing three hidden layers with 200 hidden nodes within each layer and where the initial weights were set using the uniform adaptive method [22] and using cross-entropy as the error function [23]. The hidden layer nodes utilized a rectified linear unit (ReLU) transfer function [24] with 50% dropout and adaptive learning rate while the output layer used a softmax transfer function. The deep neural learning architectures were tested using 10-fold cross-validation and run for 10 epochs in each experiment. Short-Time Fourier Transform [25] was used to decompose the raw signal from each electrode into 5 bands (delta, theta, alpha, beta, gamma) [26]. Each participant contributed one set of non-excited state data and two sets of excited state data giving a total of 72 observations (3 observations x 24 participants). For each class, 16 timepoints were obtained per class giving a total of 80 features (5 bands x 16 timepoints). As can be seen, this dataset is extremely challenging in that there are much less observations than there are input features. Thus, this challenging dataset presents a good test case for discerning the actual capabilities of the deep networks through deep tuning. The results obtained are tabulated below in Table 1.

The first test consisted of changing the input dropout ratio from 0 to 0.9 in increments of 0.1. As can be seen from Table 1, the best prediction result was obtained from an input dropout ratio of 0.2 at 91.15% whereas the worst was obtained with a setting of 0.9 at 77.30%. From Fig. 6(a), it appears that the trend points towards a smaller value being more suitable for the input dropout ratio as compared to larger values that tended to perform increasingly worse. The second and third test consisted of changing the regularization factor from 0 to 0.0 as shown in Table 1 column 3. Using the best input dropout ratio of 0.2 found from the first test, in the second test for L1 (lasso) regularization, the best prediction result obtained was using a setting of 0.01 at 94.92% whereas the worst was obtained with a setting of 0.0001 at 83.78%. Using the best input dropout ratio of 0.2 found from the first test and the best L1 regularization factor of 0.01 found from the second test, in the third test for L2 (ridge) regularization, the best prediction result obtained was using a setting of 0.001 at 95.50% whereas the worst was also obtained with a setting of 0.0001 at 82.80%. There did not appear to be any trends in terms of the best settings for L1 and L2 regularization.

Table. 1 Results of Prediction Accuracy after Deep Tuning of the Deep Networks for Emotion Detection

Parameter Setting for Input Dropout Ratio	Prediction Accuracy for Input Dropout Ratio Tuning	Parameter Setting for L1 & L2 Regularization	Prediction Accuracy for L1 Regularization Tuning	Prediction Accuracy for L2 Regularization Tuning
0	88.21	0	91.15	94.92
0.1	84.84	0.00001	89.53	89.78
0.2	**91.15**	0.00005	87.32	87.42
0.3	89.57	0.0001	83.78	82.80
0.4	86.70	0.0005	87.83	86.99
0.5	86.05	0.001	87.14	**95.50**
0.6	84.45	0.005	87.87	87.75
0.7	82.96	0.01	**94.92**	87.98
0.8	90.27	0.05	90.92	83.28
0.9	77.30	0.1	87.02	90.10



(a) (b) (c)

Fig. 6 Plots of prediction accuracy against deep parameter tuning settings. (a) Input dropout ratio tuning results (b) L1 regularization tuning results (c) L2 regularization tuning result.

B. Affective Music Applications

1) Preliminary Experiments

The results from the trained model used standard parameters with 10 epochs, 10-fold cross-validation, 500-50 hidden layers, Rectifier Linear Units (ReLU) as the activation function and 0.5 as dropout ratio. As a start, results of running 10 epochs and 50 epochs are compared as shown In Table 2. From the results, more epochs will result in higher accuracy, but the difference is not significant. Based on the preliminary result, an accuracy of 46.0% and 47.2% were obtained

Table. 2 Preliminary Results

Number of epochs	Accuracy
10	0.460
50	0.472

2) Experiment 1: The architecture of hidden layers

A deep learning model has a deep architecture due to the presence of multiple hidden layers. The complex network is important as it connects neuronal layers with the inputs and the outputs. When dealing with dense layers with complicated datasets, additional layers can be beneficial. Thus, questions on the most ideal number of hidden layers and hidden nodes arise. The accuracy is obtained through the standard experiment setting as mentioned previously, except that there is change on number of hidden layers. The experiments all have 500 nodes. For example, for number of hidden layers is two, the architecture of the deep learner is 500-500-500. The accuracies are obtained through running the model for 10 iterations, and the mean for the average accuracy.

a) Experiment 1a: Number of hidden layers

The first question that arises is how many hidden layers will result in higher accuracy? To answer the question, a few architectures of hidden layers were tested. Networks with 2, 3, 4, and 5 hidden layers were tested. Each hidden layer was set to the 500 hidden nodes each. From the experiment, the highest accuracy of 49.3% was obtained when 3 hidden layers were used.

Table. 3 Results of Accuracy based on number of hidden layers

Number of hidden layers	Accuracy
2	0.460
3	**0.490**
4	0.436
5	0.420

b) Experiment 1b: Number of hidden nodes

The overall deep learning architecture cannot be complete without defining the number of neurons in the hidden layers. Referring to the experiment on the number of hidden layers mentioned previously, the best setting was 3 hidden layers and it is used here in this experiment.

Four experiments with the above network architecture but with different settings for the number of hidden nodes are performed, in order to answer the inquiry on number of hidden nodes. The results showed that the 1000-1000-1000 architecture provided the highest accuracy of 49.3%. However, compared to the 500-500-500 hidden nodes design, the model took a longer time to process due to the larger network slide.

Table. 4 Results of Accuracy based on number of hidden nodes

Architecture of hidden nodes	Accuracy
50-50-50	0.475
100-100-100	0.489
500-500-500	0.490
1000-1000-1000	**0.491**

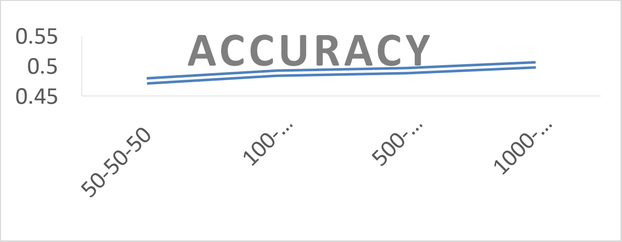


Fig. 7 Graph showing accuracies based on results on various hidden nodes number

c) Experiment 1c: Shrinking of number of nodes

Observing that three layers with 1000-1000-100 hidden layer setup performed the best, this experiment is carried out to check on the effect of shrinking the number of hidden nodes. The research question emerged where what is the most optimum number of hidden nodes to produce the highest accuracy. Table 5 shows the results of the experiments where the 500-100-50 architecture resulted in the highest accuracy.

Table. 5 Results based on various number of hidden nodes

Architecture of hidden nodes	Accuracy
2000-1000-500	0.475
1000-500-100	0.489
500-100-50	0.493

3) Experiment 2: Activation Function

Neurons are computational unit which take input and output, through activation function. This next experiment tested three of the commonly used activiation function, namely "Tanh", "Rectifier" and "Maxout" functions.

The assumption is that the activation function, Rectifier Linear Units (ReLU) will result in the highest accuracy, and this assumption was made since the preliminary experiment.

Table. 6 Accuracies based on different activation functions

Activation Functions	Accuracy
Rectifier	0.493
Tanh	0.477
Maxout	0.456

The results can be explained that ReLU caused the deep learning model fasten the forward and backward passes and meanwhile being able to maintain the non-linear nature of activation function. The hypothesis made is accepted based on the results showed above.

4) Experiment 3: Number of epochs

One of the arguments in R's h2o deep learning library is the number of epochs, which is the number of times the dataset should be passed through, or iterated. With the increase number of epochs, it is assumed that the network will remember the pairs of pattern and categories of the instances. Six experiments on different number of epochs along with the standard setting mentioned previously are performed. The design of the hidden layer is based on the result as per deduced from the earlier experiment, which is the 500-100-50 hidden layers. The results are displayed in Table 6.

Table. 7 Accuracies with various number of epochs

Number of Epochs	Accuracy
10	0.460
20	0.493
30	0.490
50	0.472
80	**0.505**
100	0.500

5) Experiment 4: Instances Reduction

The results from Experiments 1 to 3 were still not very convincing as the deep learner was only able to achieve the best accuracy of very slightly over 50.0%. The original dataset has originally 1608 songs with 40 dimensions which represent four acoustic features of music. After pre-processing, the new dataset has increased both in the number of columns and rows, which now has 26914 instances with 288 features. Instances reduction is performed in this experiment with respect to Russell's Four Quadrant Circumplex Model. Revisiting back Russell's model, no emotion annotations are ever located in the central (middle) region of the overall 2-axes graph:the emotion annotations are all essentially distributed on the outermost edges of each of the quadrants. Thus, in this instance reduction experiment, the modified dataset will attempt to mimic the actual regions of annotated emotions as per the graph of Russell's model. A threshold, x^2, will be set whereby the value will come in a form of square. In other words, instances that fall within this threshold will be removed from the dataset. Three values are used in this experiment, which are 0.2, 0.5 and 0.8. Using the pre-processed dataset, the instances in the dataset is reduced based on the threshold values. The results are shown in Table 8 below. The trained deep learning model showed the highest accuracy at 57.5% with 0.5 set as the instance reduction threshold. This instance reduction approach produced a very significant improvement in the prediction accuracy of the deep learning model where the classification performance has increased by 7%. Hence, this shows that instance reduction to use only the outermost-lying training records is important in music emotion recognition.

6) Experiment 5: Hyper-Parameters Tuning

In order to reduce the issue of overfitting, regularization techniques can be implemented in the deep learning model. Two parameters are considered in this context, namely the input dropout ratio, as well as the l1 and l2 regularization parameters. Previous best parameter values found so far for the deep learner is fixed and used in the experiments, which are a hidden layer architecture of 500-100-50, 80 epochs, ReLU as activation function and dataset with excluded instances located in the middle with $x^2 = 0.5$.

Table. 8 Results of experiments on instances reduction

Threshold(x^2)	Hidden layer	Number of epochs	Accuracy
0.2	500,50 500,100,50	10 80	0.461 0.543
0.5	500,50 500,100,50	10 80	0.456 **0.575**
0.8	500,50 500,100,50	10 80	0.452 0.556

a) Experiment 5a: Input Dropout Ratio

In the case of input dropout ratio, nine different values are tested using grid search, where the values are {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}. Table 9 below shows the accuracies for each setting. The initial value used in preliminary experiment is 0.5, and the results showed that input dropout ratio of 0.1 resulted in the highest accuracy, which is 60.6%. Compared to default setting of 0.5, the prediction accuracy of the deep learning model has increased by 3.1%.

Table. 9 Table of results with various initial dropout ratio

Dropout ratio	Accuracy
0.1	**0.606**
0.2	0.601
0.3	0.596
0.4	0.589
0.5	0.575
0.6	0.572
0.7	0.505
0.8	0.457
0.9	0.457

b) Experiment 5b: ℓ1 and ℓ2 Regularization

Another important hyper-parameter to be tuned in deep learner is regularization. Experiments on ℓ1 and ℓ2 regularization parameters were carried out using the best settings found thus far. The values used for tuning are {2e-5,3e-5,4e-5,5e-5} for both parameters, which give 16 unique combinations of these hyper-parameters. The results are shown in Table 10 above. The best setting with ℓ1 = 4e-5 and ℓ2 = 3e-5 provided the best prediction accuracy of 61.7%.

Table. 10 Results of 16 combinations with different ℓ1 and ℓ2 values

ℓ1	ℓ2	Accuracy
2e-5	2e-5	0.601
	3e-5	0.609
	4e-5	0.607
	5e-5	0.607
3e-5	2e-5	0.598
	3e-5	0.609
	4e-5	0.602
	5e-5	0.610
4e-5	2e-5	0.617
	3e-5	**0.617**
	4e-5	0.607
	5e-5	0.609
5e-5	2e-5	0.595
	3e-5	0.608
	4e-5	0.604
	5e-5	0.610

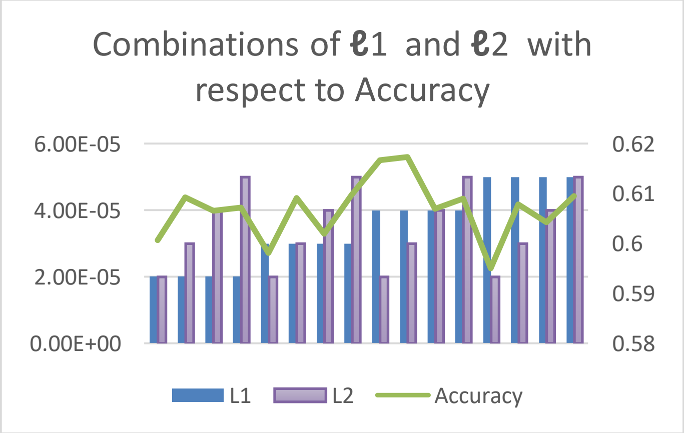


Fig. 9 Line graph of combinations of ℓ1 and ℓ2 and respective accuracy

Table. 12 Parameters settings based on each experimented result

Hidden Layers	Number of Epochs	Initial Dropout Ratio	ℓ1	ℓ2	Accuracy
500-100-50	80	0.1	4e-5	3e-5	**0.617**

V. CONCLUSION

This study firstly investigated the deep tuning of deep learning neural networks used for the task of emotion prediction in a virtual reality environment setting. Three deep parameters were tested, namely input dropout ratio, L1 regularization and L2 regularization. Compared to shallow tuning, the results have shown that deep tuning is able to increase the prediction accuracy to above 95% from 88%. Smaller values for input dropout ratio appeared to generate better prediction accuracies compared to larger setting. However for L1 and L2 regularization, there appears to be a sweet spot without any obvious trends in terms of finding good settings for generating superior prediction results. For future work, we intend to augment the current EEG-based prediction with inertial sensing data obtained from accelerometer sensors coupled with gyroscopic data of the user's head positions while experiencing the VR stimuli.

This study secondly investigated deep learning for music emotion recognition. Starting out with an average accuracy of 46.0% using the default deep learning settings of 10 epochs, hidden layer of 500-50 architecture, dropout ratio of 0.5 and activation function of Rectified Linear Unit (ReLU),subsequent experiments in tuning the various parameters of the deep learning architecture progressively improved the accuracy at each stage of hand-tuning. Optimizing the hidden layer architecture improved the accuracy slightly to 49.3%. The greatest improvement to 57.5% was achieved through the introduction of a novel instance reduction parameter to enable the input feature space to be focused around the outermost edges of Russell's Circumplex Model of Emotions. Futher tuning of the input dropout ratio regularlization terms further improved the best final prediction accuracy of 61.7%. For future work, instead of using Russell's model, other emotion model could be used, for example those that include the utilization of dominance information.

ACKNOWLEDGMENT

This work was supported in part by the research grant ref: IF0318M1003 received from the Ministry of Energy, Science, Technology, Environment & Climate Change (MESTECC), Malaysia.

REFERENCES

1. Hazarika, D., Gorantla, S., Poria, S. and Zimmermann, R., 2018, April. Self-attentive feature-level fusion for multimodal emotion detection. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 196-201). IEEE.
2. Richer, R., Zhao, N., Amores, J., Eskofier, B.M. and Paradiso, J.A., 2018, July. Real-time Mental State Recognition using a Wearable EEG. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 5495-5498). IEEE.
3. Cook, D., 2016. Practical machine learning with H2O: powerful, scalable techniques for deep learning and AI. O'Reilly Media, Inc.
4. Mercer, C. and Macaulay, T. 2018. https://www.techworld.com/picture-gallery/data/tech-giants-investing-in-artificial-intelligence-3629737/ (accessed 27 Feb 2019).
5. Rostykus, B. 2017. https://medium.com/@NetflixTechBlog/introducing-vectorflow-fe10d7f126b8 (accessed 27 February 2019).
6. Srengers, J., Martinez, E.J., Simpraga, S., Jansen, F., Vlaskamp, C., Oranje, B., Poil, S.S., Linkenkaer-Hansen, K. and Bruining, H., 2017. O132 An EEG-based decision-support system for diagnosis and prognosis of autism spectrum disorder. Clinical Neurophysiology, 128(9), p.e221.
7. Wang, P., Min, J. and Hu, J., 2018. Ensemble classifier for driver's fatigue detection based on a single EEG channel. IET Intelligent Transport Systems, 12(10), pp.1322-1328.
8. Schwarz, D., Subramanian, V., Zhuang, K. and Adamczyk, C., 2014, September. Educational neurogaming: Eeg-controlled videogames as interactive teaching tools for introductory neuroscience. In Tenth Artificial Intelligence and Interactive Digital Entertainment Conference.
9. Lee, H., Pham, P., Largman, Y., and Ng, A. Y. 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. In Advances in Neural Information Processing Systems, pp. 1096-1104.
10. Shao, X., Xu, C., andKankanhalli, M. S. 2004. Unsupervised classification of music genre using Hidden Markov Model. In 2004 IEEE International Conference on Multimedia and Expo (ICME), Vol. 3, pp. 2023-2026, IEEE.
11. Hamel, P., and Eck, D. 2010. Learning features from music audio with deep belief networks. In ISMIR, Vol. 10, pp. 339-344.
12. Li, T. L., Chan, A. B., and Chun, A. 2010. Automatic musical pattern feature extraction using convolutional neural network. In Proc. Int. Conf. Data Mining and Applications, Vol. 161.
13. Dahl, G. E., Sainath, T. N., and Hinton, G. E. 2013. Improving deep neural networks for LVCSR using rectified linear units and dropout. In 2013 IEEE international conference on acoustics, speech and signal processing, pp. 8609-8613, IEEE.
14. Sigtia, S., and Dixon, S. 2014. Improved music feature learning with deep neural networks. In 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP),pp. 6959-6963, IEEE.
15. Krigolson, O.E., Williams, C.C., Norton, A., Hassall, C.D. and Colino, F.L., 2017. Choosing MUSE: Validation of a low-cost, portable EEG system for ERP research. Frontiers in neuroscience, 11, p.109.
16. Klem, GH; Lüders, HO; Jasper, and HH; Elger, C. 1999. "The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology". Electroencephalography and Clinical Neurophysiology. Supplement. 52: 3–6. PMID 10590970.
17. Amin, A., Gromala, D., Tong, X., and Shaw, C. 2016. Immersion in cardboard VR compared to a traditional head-mounted display. In International Conference on Virtual, Augmented and Mixed Reality, pp. 269-276. Springer, Cham.
18. Che, X., Ip, B., and Lin, L. 2015. A survey of current YouTube video characteristics. IEEE Multimedia.
19. Chen, Y. A., Yang, Y. H., Wang, J. C., and Chen, H. 2015. The AMG1608 dataset for music emotion recognition. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 693-697, IEEE.
20. Han, S., Pool, J., Tran, J., and Dally, W. 2015. Learning both weights and connections for efficient neural network. In Advances in neural information processing systems, pp. 1135-1143.
21. Ng, A. Y. 2004. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In Proceedings of the twenty-first international conference on machine learning, p. 78. ACM.
22. D. Nguyen, B. Widrow, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights", in: Neural Networks, 1990. 1990 IJCNN International Joint Conference on, IEEE, pp. 21-26(1990).
23. P.T. De Boer, D.P. Kroese, S.Mannor, and R.Y. Rubinstein, "A tutorial on the cross-entropy method", Annals of operations research, 134(1):19-67 (2005).
24. V. Nair, G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines", in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807-814 (2010).
25. Shafer, B., Yaghouby, F. and Vasudevan, S., 2018, July. Short-Time Fourier Transform Based Spike Detection of Spontaneous Peripheral Nerve Activity. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 2418-2421). IEEE.
26. Verma, R. and Dekar, R., 2018. Sleep Disorder Detection by Welch Algorithm Based PSD Analysis on EEG Signals. Sleep, 5(06).