Emotional Recognition Based on EEG Signals Comparing Long-term and Short-term Memory with Gated Recurrent Unit Using Batch Normalization

Yunfei Guo[1,2,a], Wenjun Liu[1,b], Dapeng Wei[1,2,c], Qiaosong Chen[1,d]

[1]School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China

[2]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China

[a]guoyunfei.2007@163.com, [b]2662052636@qq.com, [c]dpwei@cigit.ac.cn, [d]chenqs@cqupt.edu.cn

Keywords: EEG; emotion recognition; LSTM; GRU; batch normalization

Abstract: Expression recognition is the development direction for improving human-computer interaction. At the same time, Electroencephalo-gram(EEG) signals provide us with a way to quantify changes in human emotions. The identification of human emotions through the use of multimodal data sets based on EEG signals is a convenient and safe solution. Using deep learning for expression recognition is a new direction for the development of current emotion recognition. Since EEG signals are biomass signals with temporal characteristics, the use of recurrent neural networks to identify and classify EEG signals has certain advantages. Long-term and Short-term Memory Networks (LSTM) is an important representative of recurrent neural networks, and has achieved good recognition results in the classification and recognition of EEG signals. Gated Recurrent Unit (GRU) is a simpler algorithm than the structure of long-term and short-term memory. We use a gated loop unit with batch normalization for the classification of EEG signals. On the public dataset DEAP, GRU with batch normalization added a better recognition rate for arousal and valence than LSTM.

1. Introduction

The first recorded EEG activity in humans was done by Berger, who used two needle-shaped platinum electrodes to insert into the cerebral cortex of the patient's skull to record EEG signals [1]. In the subsequent experiments, Berger confirmed that the electrodes were placed in the cerebral cortex but the electrodes were placed on the scalp outside the skull. EEG signals were also collected. This finding laid the foundation for future clinical medical EEG techniques. The current rapid development of non-invasive brain-computer interface technology is also due to this fact.

Traditional human-computer interaction systems use keyboards, mice and some sensors as the main input interfaces. These technologies ignore the important emotional information and psychological activities of many users. Being able to objectively describe the changes of inner feelings has always been a major goal of human beings in exploring their own mysteries. With the development of artificial intelligence technology, people have put forward new requirements for a new type of human-computer interaction system that could use different expression strategies according to users' emotions. In this paper, emotional computing will provide important support for the development of new brain-computer interface technology and even human-computer interaction system[2].

Emotional analysis based on image recognition is an important development area for deep learning. But their development will be hindered by the qualities of human beings. Machine vision-based expression recognition technology is unable to recognize human invisible attention transfer and abstract logical thinking process. And because humans can consciously control their own tone, note, body movements and expressions, machine vision technology may be misled by human conscious control, which leads to a decline in recognition accuracy.

Since human conscious self-control will be expressed through EEG signals, the analysis of human emotions through EEG signals will greatly preserve the information of human psychological activities, which is an important factor in the introduction of EEG signals in the new human-computer interaction system. Schachter-Singer theory provides us with a theoretical model of emotions: emotional experience is determined by both physical awakening and understanding of arousal [3]. Physiological arousal as a low-level neuroreflex activity can be collected by physiological signal sensors, but the understanding of physiological arousal is advanced brain activity, which can be recorded by EEG signals.

According to the different EEG patterns, R. Jung classified normal human EEG into four categories: alpha band (8-12 Hz), beta band (12-30 Hz), delta band (0.5-3 Hz) and theta band (3-8 Hz)[4]. The analysis of these four bands is an important method to understand human emotions and emotional changes. This study is mainly based on the analysis of the above four bands of EEG signals to identify the emotions of different testers.

In the process of transition from human driving to automatic driving, the driver's emotion is monitored in real time to determine the driver's safe driving degree, which can be used as a switching threshold for human control driving to automatic driving. Hazardous equipment operators can alert them to safe production through emotional recognition. Teachers can effectively judge students' learning status through emotional recognition. These application areas will provide numerous application possibilities for emotion recognition based on EEG.

The training of deep learning network needs huge data sets to support the training of deep learning network. DEAP [5], as an open multi-modal biological signal data set, provides sufficient data for our training in-depth learning network. The data set includes 32 subjects' EEG, ECG, body surface temperature and other physiological parameters during watching a music video. The signals in alpha, beta and gamma bands are collected by 32-channel EEG signal acquisition system. In order to achieve three binary classifications: low/high arousal, low/high valence, and low/high liking, the researchers of this data set use Fisher linear discriminant to select features, and then use Gauss Naive Bayesian algorithm to classify features.

2. Data Set

To create an adaptive music video recommendation system, Sander Koelstra [5] and their colleagues recorded multiple physiological signals from 32 testers aged 19 to 37 [6]. Twenty-two of the 32 participants also recorded their positive facial video. The sampling frequency of the EEG signal and the multimodal biosignal is 512 Hz. The effect of removing eye artefacts from eye movement is achieved by using the blind source separation technique. The EEG channels were reordered so that they all follow the Geneva order as above. The data was segmented into 60 second trials and a 3 second pre-trial baseline removed. The EEG channels were reordered so that they all followed the Geneva order above. These data were divided into 60 trials and 3 trials before the baseline was removed.

TABLE I. DEAP Arrays of Each Participant

Array name	*Array shape*	*Array contents*
Data	40 x 40 x 8064	video/trial x channel x data
Labels	40 x 4	video/trial x label (valence, arousal, dominance, liking)

As shown in Table I, each participant's file contains two arrays. Each participant has an array of 40 viewing videos x 40 (EEG + peripheral) channels × 8064read. In this paper, only EEG signals are used. The 8064 readings for each EEG channel are divided into 12 sections each of which is 5 seconds long and Approximately 21,504 readings.

3. Long Term and Short Term Memory(LSTM)

In order to use the neural network in dealing with the problem of sequence-to-sequence, Recurrent neural network (RNN) algorithm is proposed. But in the process of training, ordinary RNN will face the problem of gradient explosion and gradient disappearance in long time span sequence data. Hochreiter and his colleagues developed the Long-term and Short-term Memory Networks (LSTM) [7] in 1997. LSTM has the ability to remove or increase information to the cellular state by carefully designed structures called gates. Gates are a way to allow information to pass selectively. This algorithm uses gated cell instead of RNN cell to eliminate the problem of gradient explosion and gradient disappearance. In LSTM algorithm, LSTM cell is used to strategically forget some information so as to achieve better results.

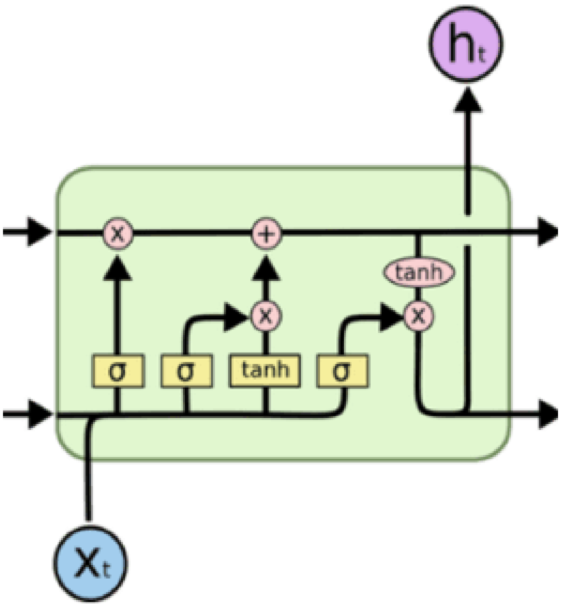


Fig.1. LSTM cell architecture

The core idea of the LSTM model (Figure 1) is the state of the cell. Cell state is determined by various control gates to remember the memory and forgetting. The forgetting gate, the external input gate, and the output gate are the three operational core units that determine the "cell state." Forgetting the door through the sigmoid layer to achieve "cell state" whether to forget some information[8].

The first step in the LSTM model is to use the Forgotten Gate to determine what information the cell loses. The forgotten gate layer reads the current input xt and the previous neuron information ht-1, and then ft decides to discard the information. As shown in Equation 1:

$$f_t = \sigma\left(W_f \bullet \left[h_{t-1}, x_t\right] + b_f\right) \tag{1}$$

The second step of the LSTM model is to determine the new information stored in the cell state. This operation consists of two parts. The sigmoid layer acts as an external input gate to control which values need to be updated, as shown in Equation 2. At the same time, a new candidate value vector is created using the tanh layer to be added to the cell state, as shown in Equation 3.

$$i_t = \sigma\left(W_i \bullet \left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

$$\tilde{C}_t = tanh\left(W_c \bullet \left[h_{t-1}, x_t\right] + b_C\right) \tag{3}$$

The third step is to update the original cell state. C_{t-1} is updated to C_t, and the original cell state is multiplied by ft to determine the information that needs to be discarded. Then add new candidate values, as shown in Equation 4.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

The last step is to determine the output information based on the updated cell status. First, the sigmoid layer is run to determine which part of the cell state can be output, as shown in Equation 5. Second, the state of the cells treated with tanh is multiplied by the output of the sigmoid layer. Finally, we output the above information to determine the output, as shown in Equation 6.

$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right) \tag{5}$$

$$h_t = o_t * tanh\left(C_t\right) \tag{6}$$

Where tanh is a hyperbolic tangent activation function that pushes data into the range of (-1,1). Wf, Wi, Wc, and Wo are weight matrices. σ is the sigmoid activation function that maps the data to the range of (0,1).

4. Gated Recurrent Unit(GRU)

In order to enable each loop unit to adaptively capture dependencies at different time scales, in 1995, Cho et al. proposed the Gated Recurrent Unit (GRU)[9]. Compared with LSTM, the structure of GRU is relatively simple.

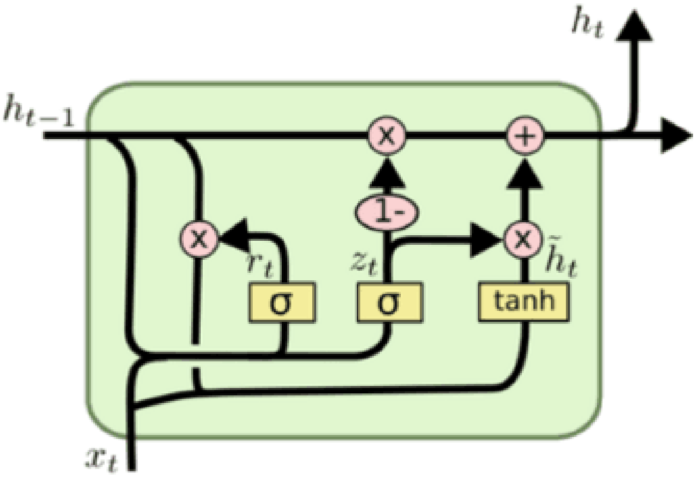


Fig.2. GRU cell architecture

GRU is similar to LSTM in that it has a gated unit with information function in the regulating unit. The difference between GRU and LSTM is that GRU combines forgetting gate and input gate into a single update gate, which can control cell state and hidden state. Fig. 2 shows the unit model of GRU.

The GRU activation factor at time t is a linear interpolation of the previous activation factor and the candidate activation factor, as shown in Equation 7.

$$h_t^j = \left(1 - z_t^j\right) h_{t-1}^j + z_t^j \tilde{h}_t^j \tag{7}$$

The update gate decision unit updates the number and content of its own incentive factors, as shown in Equation 8.

$$z_t^j = \sigma\left(W_z x_t + U_z h_{t-1}\right)^j \tag{8}$$

The process of linear summation between the existing state and the new calculated state is similar to that of the LSTM unit. However, the GRU does not have any mechanism to control the extent of its own state exposure, and each iteration exposes the entire state.

The candidate activation factor is calculated similarly to the traditional periodic unit (as shown in Equation 9) and Equation 10[10] .

$$h_t = g\left(W\mathbf{x}_t + U\mathbf{h}_{t-1}\right) \qquad (9)$$

$$\tilde{h}_t^j = tanh\left(W\mathbf{x}_t + U\left(\mathbf{r}_t \odot \mathbf{h}_{t-1}\right)\right)^j \qquad (10)$$

The reset gate has a similar calculation to the update gate, as shown in Equation 11.

$$r_t^j = \sigma\left(W_r\mathbf{x}_t + U_r\mathbf{h}_{t-1}\right)^j \qquad (11)$$

5. Proposed Method

Because the deep neural network activates the input value before the nonlinear transformation (that is, x=WU+B, U is the input). As the network depth deepens or during the training process, its distribution gradually shifts or changes, so the training The convergence is slow, generally the overall distribution gradually approaches the upper and lower limits of the value interval of the nonlinear function (for the Sigmoid function, it means that the activation input value WU+B is a large negative or positive value), so this leads to The gradient of the low-level neural network disappears during backpropagation, which is the essential reason for the slower convergence of the training deep neural network. Batch normalization(BN)[11] is to force the distribution of the input value of any neuron in each layer of neural network back to the standard normal distribution with a mean of 0 variance, which is to force the more and more partial distribution to be pulled back. Compare the standard distribution so that the activation input value falls in the region where the nonlinear function is sensitive to the input, so that small changes in the input will result in a large change in the loss function, thus making the gradient larger and avoiding the gradient disappearance problem, and Larger gradients mean faster learning convergence and can greatly speed up training.

The x(k) of a neuron in the t-layer is not the original input, or the output of each neuron in the T-1 layer, but the linear activation x=WU+B of the neuron in the t-layer, where the U is the output of the neuron in the T-1 layer. The meaning of transformation is that the original activation x corresponding to a neuron is transformed by subtracting the mean E (x) of M activation x obtained from m instances in mini-Batch and dividing it by the variance Var (x).

The method of batch normalization under Mini-Batch SGD can be explained by the following structure[12]. Assume that for a deep neural network, the two layers are as shown in fig.3.

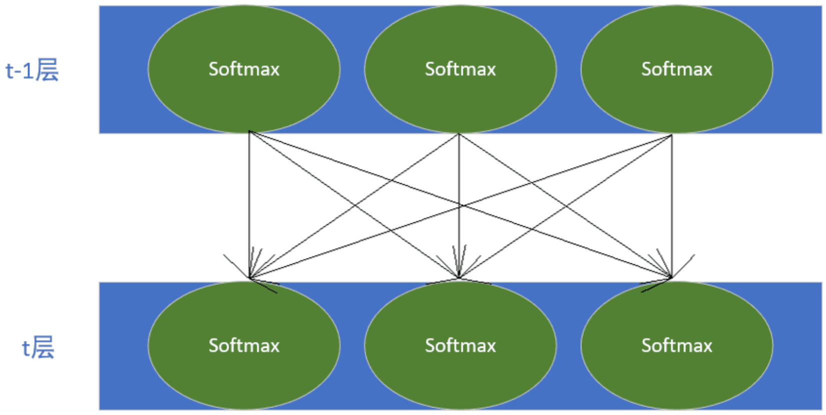


Fig.3. Two-layer network structure

To make a BN for the activation value of each hidden layer neuron, it is conceivable that each hidden layer is further added with a layer of BN operation layer, which is located after the X=WU+B activation value is obtained, before the nonlinear function transformation, such as Figure 4 shows

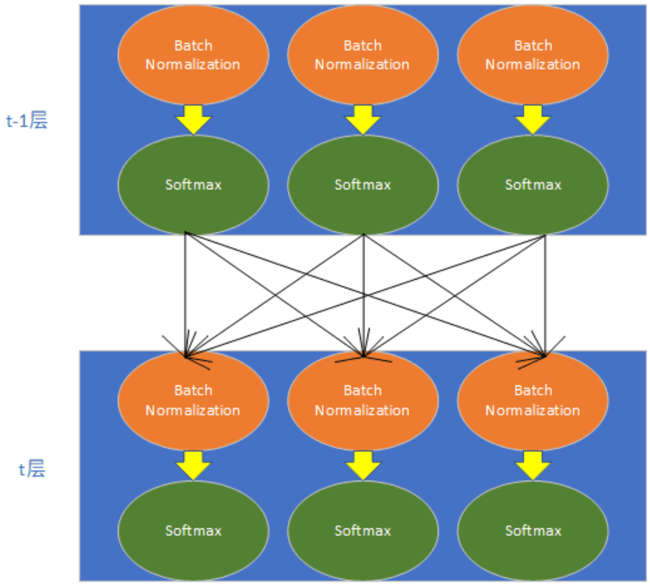
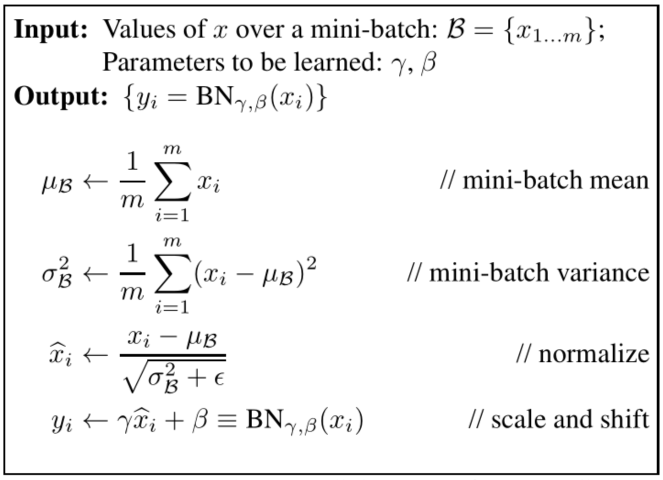


Fig.4. Two-layer network structure with batch normalization

For the Mini-Batch SGD, a training process contains m training instances, and the specific batch normalization operation is performed for the activation value of each neuron in the hidden layer, as shown in Equation 12.

$$\hat{X}^{(k)} = \frac{x^{(k)} - E\left[x^{(k)}\right]}{\sqrt{Var\left[x^{(k)}\right]}} \tag{12}$$

After this transformation, the activation X of a neuron forms a normal distribution with a mean value of 0 and a variance of 1. The purpose is to pull the value to the linear region of the subsequent non-linear transformation, increase the derivative value, enhance the fluidity of back-propagation information, and accelerate the convergence speed of training[12].



Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$ // mini-batch mean

$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$ // mini-batch variance

$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$ // normalize

$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i)$ // scale and shift

Fig.5. Batch Normalization Tansfrom

In order to prevent this, each neuron adds two adjusting parameters (scale and shift), which are learned through training, to inverse the activation of the transformed network and enhance the expressive ability of the network. That is to say, the following scale is applied to the activation of the transformed network. And shift operations, which are actually inverse operations of transformations, are shown in Fig. 5 below.

Each participant's data consists of 8064 readings from 32 EEG channels of a single video. Each video is divided into 12 segments and has a length of 5 seconds, each segment consisting of 672 readings from 32 EEG channels. Use the above data as the input layer for deep learning. The first layer processes the sequenced initial data through the GRU. The second layer forces the data input to the next layer to be returned to the standard normal distribution with a mean of 0 variance of 1 after the batch normalization of the data in the middle layer. The third layer uses Softmax as the excitation function. As is shown in Fig. 6.

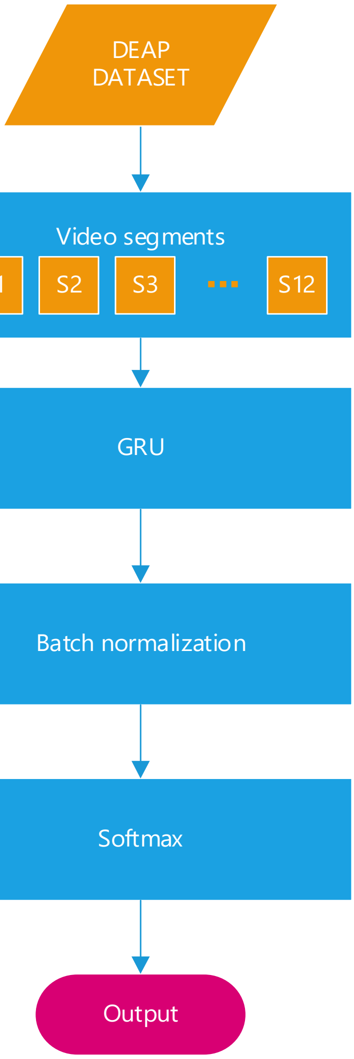


Fig.6. Detailed proposed model.

The model uses four-fold cross-validation and reads 75% of the personal data in the dataset as a training set, using 25% of the data for testing. The Adam optimizer uses a learning rate of 0.001 during the training process. Use Anaconda's Tensorflow framework to implement deep learning methods.

6. Results

Because the DEAP data set divides the average accuracy of all participants into the following three categories: arousal, valence and liking. In this study, LSTM and batch normalization were used to optimize the GRU to compare the recognition accuracy of arousal and valence respectively. The experimental method is the same as the previous description of the DEAP data set.

In the identification of the arousal state, the following different algorithms are tested using the data set DEAP, and the GRU optimized using Batch Normalization has higher accuracy, as shown in Table II.

TABLE II. Arousal accuracy

Name	***Smoothed***	***Value***	***Step***	***Relative***
GRU	0.8556	0.8557	10k	20m39s
BN-GRU	0.8559	0.8561	10k	20m36s
LSTM	0.8474	0.8476	10k	16m46s
BN-LSTM	0.8478	0.8480	10k	16m43s

In the recognition of Valence status, the following different algorithms are tested with DEAP data set. The GRU optimized by Batch Normalization also has higher accuracy, as shown in Table III.

TABLE III. Valence accuracy

Name	***Smoothed***	***Value***	***Step***	***Relative***
GRU	0.8664	0.8665	10k	21m8s
BN-GRU	0.8668	0.8669	10k	21m6s
LSTM	0.8503	0.8503	10k	16m28s
BN-LSTM	0.8507	0.8508	10k	16m26s

7. Conclusion

In the research of emotional recognition using EEG signals, GRU algorithm optimized by batch normalization is more accurate than LSTM algorithm optimized by batch normalization. The above results show that the proposed algorithm is a better way to recognize emotions in EEG-based multi-modal biological information. Because it shows higher quasi-curvature than the traditional single algorithm.

Acknowledgments

I am grateful to Professor Wei Dapeng of the Chongqing Institute of Green and Intelligent Techno logy of the Chinese Academy of Sciences as my correspondent. I also thank the students and professors of the School of Bioinformatics and the School of Computer Science and Technology of Chongqing University of Posts and Telecommunications for their help. This research is supported by Youth Innovation Promotion Association of CAS (2015316), Project of Chongqing brain science Collaborative Innovation Center, the Chongqing Research Program of Basic Research and Frontier Technology (No. cstc2015jcyjA50018), Project of CAS Western Young Scholar, Project of CQ CSTC (cstc2017zdcy-zdyfX0001).

References

[1] H. Berger, "Zur Innervation der Pia mater und der Gehirngefäße," *Archiv Für Psychiatrie Und Nervenkrankheiten,* vol. 70, no. 1, pp. 216-220, 1924.

[2] C. F. Moss and S. R. Sinha, "Neurobiology of echolocation in bats," *Current Opinion in Neurobiology,* vol. 13, no. 6, pp. 751-758, 2003.

[3] L. E. P. WARD M. WINTON, AND ROBERT M. KRAUSS, *Facial and Autonomic Manifestations of the Dimensional Structure of Emotion*. February 10, 1981 pp. 1138 -1140

[4] G. Li, D. Zhang, S. Wang, and Y. Y. Duan, "Novel passive ceramic based semi-dry electrodes for recording electroencephalography signals from the hairy scalp," *Sensors & Actuators B Chemical,* vol. 237, pp. 167-178, 2016.

[5] S. M. Sander Koelstra, Christian Mu¨hl, Mohammad Soleymani, Student Member, Jong-Seok Lee, Member, Ashkan Yazdani, Touradj Ebrahimi, Member, Thierry Pun, Member, Anton Nijholt, Member, Ioannis Patras, Member, *DEAP: A Database for Emotion Analysis using Physiological*

Signals. 2012.

[6] A. A. F. Salma Alhagry, Reda A. El-Khoribi, "Emotion_Recognition_based_on_EEG_using_LSTM" *International Journal of Advanced Computer Science and Applications,* 2017.

[7] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* vol. 9, no. 8, p. 1735, 1997.

[8] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG," *IEEE Trans Neural Syst Rehabil Eng,* vol. 25, no. 11, pp. 1998-2008, Nov 2017.

[9] J. C. C. G. K. C. Y. Bengio, *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. 2014.

[10] D. Bahdanau, *Neural Machine Translation by Jointly Learning to Align and Translate*. 2014.

[11] N. B. Tim Cooijmans, César Laurent, Çaglar Gülçehre & Aaron Courville, *RECURRENT BATCH NORMALIZATION*. 2017.

[12] S. I. C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015.