Emotion Recognition using Multimodal Residual LSTM Network

Jiaxin Ma*
OMRON SINIC X Corporation
Tokyo, Japan
jiaxin.ma@sinicx.com

Hao Tang*
Shanghai Jiao Tong University
Shanghai, China
silent56@sjtu.edu.cn

Wei-Long Zheng
Massachusetts General Hospital, Harvard Medical School
Boston, MA, US
wzheng8@mgh.harvard.edu

Bao-Liang Lu†
Shanghai Jiao Tong University
Shanghai, China
bllu@sjtu.edu.cn

ABSTRACT

Various studies have shown that the temporal information captured by conventional long-short-term memory (LSTM) networks is very useful for enhancing multimodal emotion recognition using encephalography (EEG) and other physiological signals. However, the dependency among multiple modalities and high-level temporal-feature learning using deeper LSTM networks is yet to be investigated. Thus, we propose a multimodal residual LSTM (MMResLSTM) network for emotion recognition. The MMResLSTM network shares the weights across the modalities in each LSTM layer to learn the correlation between the EEG and other physiological signals. It contains both the spatial shortcut paths provided by the residual network and temporal shortcut paths provided by LSTM for efficiently learning emotion-related high-level features. The proposed network was evaluated using a publicly available dataset for EEG-based emotion recognition, DEAP. The experimental results indicate that the proposed MMResLSTM network yielded a promising result, with a classification accuracy of 92.87% for arousal and 92.30% for valence.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Mathematics of computing** → *Time series analysis*; • **Information systems** → Multimedia information systems.

KEYWORDS

Multimodal emotion recognition; long-short-term memory network; electroencephalography

*Both authors contributed equally to this research.
†Corresponding author.
The experiment conducted on the DEAP database was performed at the university, and is not for professional use.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350871

ACM Reference Format:
Jiaxin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu†. 2019. Emotion Recognition using Multimodal Residual LSTM Network. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3343031.3350871

1 INTRODUCTION

Emotion recognition is integral to effective reasoning, planning, and the execution of certain tasks. Automatic emotion recognition plays an important role in human-computer interaction systems for more affective and personal interactions. Inspired by the great success of deep neural networks in numerous recognition tasks, several deep learning architectures have been proposed to improve the performance of multimodal emotion recognition using electroencephalographic (EEG) signals among other physiological signals, including auto-encoders, convolutional neural networks (CNNS), and recurrent neural networks (RNNs) [18, 19, 27].

RNNs, unlike other feed-forward networks, can extract complex temporal information from sequences of various lengths and have drawn increasing attention in the field of multimodal emotion recognition using EEG and other physiological signals. Various studies have demonstrated that considering temporal information improves the performance of RNNs as a temporal feature extractor module [15, 32]. However, the temporal correlation information among multiple modalities, together with high-level temporal feature learning using deeper RNNs, is yet to be investigated.

The temporal correlation among the different modalities is important, because emotions are manifested via various internal physiological responses and external behaviors over time. Although the signals in multiple modalities are collected from various positions using different devices with different reflection latencies, they reflect the same emotion changes with temporal evolution.

The temporal correlation among different modalities is valuable for extracting emotion-related information and filtering unrelated noise, as an EEG signal is typically characterized by a low signal-to-noise ratio (SNR). We propose the construction of neural networks that can capture the temporal correlation among various modalities and achieve a better generalization for modeling emotion evolution.

Moreover, owing to the complexity of human emotions, useful temporal features, as well as the correlation among modalities, could be a highly complex non-linear function of input sequences. Studies in various fields have indicated that deep neural networks

(DNNs) can learn complex high-level features, consequently outperforming the shallow ones [10, 31]. With the discovery of residual learning [11] and normalization of neuron activities [13, 17], DNNs can converge more efficiently, thus optimizing the time cost of training. Similarly to residual learning [11], we hypothesize that deep RNNs with residual connections can learn more complex high-level features for emotion recognition, consequently outperforming other neural network architectures.

There are various multimodal fusion approaches such as training one common layer for the subnetworks of the multiple modalities [19], utilizing disjoint layers for different modalities [32], and combining multiple modalities by feature concatenation or ensemble [21]. However, to the best of our knowledge, previous works have not explicitly modeled the temporal correlation among multiple modalities for multimodal emotion recognition.

The LSTM [12], a gated variant of RNN, alleviates gradient explosion or vanishing problems by utilizing multiplicative gating operations and shortcut paths through time. In our multimodal residual LSTM (MMResLSTM) network, the temporal correlation among modalities is explicitly learned by sharing the weights across modalities. This architecture was first proposed for speaker identification in videos, and it significantly improved the robustness of the voice data to content quality degradations and non-auditory distractors [26]. Because all the modalities cooperate and compete with each other when learning the shared weights, the network tends to accumulate temporal correlations across the modalities. Moreover, by sharing the weights across the time steps in the LSTM, learning is robust to the different emotion reflection latencies of the different modalities, and can learn the correlation using the different time-step biases.

Residual learning [11] and layer normalization [17] are further utilized to build four residual LSTM layers for complex high-level learning of temporal features. Although there are other techniques for building DNNs through spatial shortcut paths [29, 40], it has been demonstrated that the residual networks yield efficient training and interpretation without extra parameters. Consequently, the residual LSTM network contains both temporal and spatial shortcuts provided by the LSTM and the residual network, respectively, for effectively training the deep RNNs on long sequences.

We conducted experiments using the plain deep LSTM network, residual LSTM network, and MMResLSTM network on the DEAP dataset. The results indicated that under different parameter settings, the MMResLSTM network can stably outperform the other LSTM structures. In a comparison with the state-of-the-art methods on the same dataset, the MMResLSTM also achieved promising results.

2 RELATED WORK

There are six main public EEG datasets for emotion recognition tasks: DEAP (2012) [16], MAHNOB-HCI (2012) [28], SEED (2015) [38], HR-EEG4EMO (2017) [4], DREAMER (2018) [14], and SEED-IV (2019) [37]. Previous studies on DEAP explored different aspects of feature extraction, experiment settings, and modalities. Aside from the choice of classification model, the most common differences among the emotion recognition studies relate to the process of feature extraction. The majority of prior studies did not directly utilize raw EEG signals; rather, they extracted the features first. The process of feature extraction can be regarded an effective way to reduce the dimensions of the data. In a typical time-series signal, the conventional EEG features include time-domain, frequency-domain, and time-frequency-domain features. Wang and colleagues investigated three kinds of feature extraction methods for recognizing two emotions from EEG data [35]. Pandey *et al.*, using the discrete wavelet coefficient features with a multilayer perceptron (MLP) model on a single-channel EEG, achieved 58.50% classification accuracy [22]. Garcia *et al.* deploys multiple entropy features with a support vector machine (SVM) classifier [9]. Using the empirical mode decomposition method, Zhuang *et al.* automatically decompose the EEG signals into intrinsic mode function (IMF) features.

However, feature extraction always requires prior knowledge of the specific task, and designing the proper features for a new task may be more time consuming than designing the model itself. Nowadays, DNN models are powerful enough to automatically extract features during the learning process; furthermore, there is the emerging possibility of using raw signals as the input and an "end-to-end" approach to learning the task. For example, Fourati *et al.* use an echo state network (ESN) wherein the recurrent layers enable the projection of the raw EEG data into a high-dimensional state space [8]. Alhagry *et al.* obtain a satisfactory result using a 2-layer plain LSTM with raw EEG input [1].

Existing works conducted on the DEAP dataset utilize different experiment settings. Although the DEAP dataset has four emotion labels: arousal, valence, dominance, and liking, not all the labels are used in related studies. Liu *et al.* [20] combine the indices of arousal, valence, and dominance to obtain a space that contains eight emotions, and classify the eight-class task using SVM. Zheng *et al.* [39] investigate the four quadrants of the valence-arousal space (low arousal-low valence, high arousal-low valence, low arousal-high valence, and high arousal-high valence), and solve the four-class task using the graph regularized extreme learning machine (GELM) method. Both studies obtain approximately 70% accuracy on their multinomial classification tasks. Most existing studies on the DEAP dataset have considered the emotion recognition problem as two independent binomial classification tasks (low/high arousal and low/high valence), and report the accuracy of arousal and valence detection, respectively.

Emotion recognition has been performed using various modalities, such as facial expressions, voice, EEG, pupillary diameter (PD), electrooculography (EOG) [5]. Furthermore, while the abovementioned studies focus solely on single EEG modality, multi-modal approaches have also been widely implemented for emotion recognition [18, 19, 21, 32, 36]. Lu *et al.* apply a fuzzy integral fusion strategy to combine EEG and eye movement features on the SEED dataset [21]. Lin *et al.* transform EEG into images and extract the hand-crafted features of other peripheral physiological signals to train a deep CNN [18]. Liu *et al.* train a bimodal autoencoder network, and they achieved the mean accuracy of 91.0% and 83.6% on the SEED and DEAP datasets, respectively [19]. As mentioned earlier, these works do not explicitly model the temporal correlation among the multiple modalities for multimodal emotion recognition; rather, their approaches are based on common layers, feature concatenation, or decision ensemble. Thus, the proposed model aims to

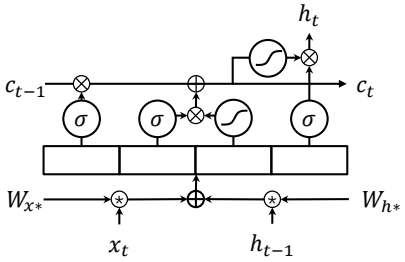

Figure 1: Conventional LSTM

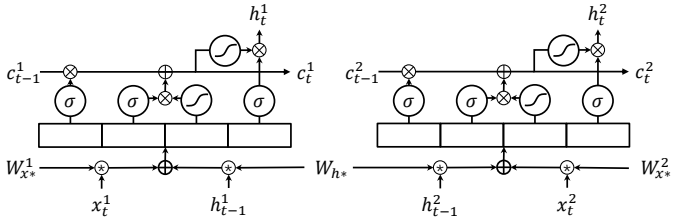

Figure 2: Multimodal LSTM with bimodal inputs

improve on previous studies by using a deep LSTM with temporal weights shared across the multiple modalities.

3 METHODOLOGY

3.1 Long-Short-Term Memory (LSTM)

LSTM [12], as a popular variant of RNN [10, 34], has demonstrated its effectiveness for extracting temporal information from long biosignals [3, 7, 30, 32]. It contains cell states c_t that are propagated through time and are responsible for storing temporal information, and data-driven gates f_t, i_t, o_t that control the process of forgetting, remembering, and outputting information, respectively. The formulas (excluding the bias terms) are as follows:

$$\tilde{c}_t = \tanh(W_{hg} * h_{t-1} + W_{xg} * x_t), \quad (1)$$
$$f_t = \sigma(W_{hf} * h_{t-1} + W_{xf} * x_t), \quad (2)$$
$$i_t = \sigma(W_{hi} * h_{t-1} + W_{xi} * x_t), \quad (3)$$
$$o_t = \sigma(W_{ho} * h_{t-1} + W_{xo} * x_t), \quad (4)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \quad (5)$$
$$h_t = o_t \odot \tanh(c_t). \quad (6)$$

where W_{h*}, including W_{hf}, W_{hi}, W_{hg}, W_{ho}, are the weight matrices of the previous time step's hidden states; W_{x*}, including W_{xf}, W_{xi}, W_{xg}, W_{xo}, are the weight matrices of the current time step's input, and σ represents the sigmoid function. The operator $*$ in the equations denotes the matrix multiplication, while $\odot$ represents the element-wise multiplication.

3.2 Multimodal LSTM Network

Conventionally, for achieving multimodal emotion recognition using EEG and other physiological signals, the multimodal architectures either build parallel LSTMs for the different modalities [15], or directly concatenate the data of multiple modalities to produce a larger input [32]. The first type of architecture contains parallel LSTMs for different modalities, and it is more efficient for training, because each LSTM focuses on inputs with similar properties; it is also more interpretable, because information from different modalities is stored separately. Nevertheless, this architecture makes no provision for correlation learning among the multiple modalities, because the LSTMs are completely independent of each other. The second type of architecture feeds the LSTM with the concatenation of multiple modalities; although the networks are free to simultaneously access the information of multiple modalities, the cross-modality correlation is not explicitly learned. There is no distinction between intra-modality correlation and cross-modality correlation in LSTM, because the inputs of the multiple modalities are treated similarly. In addition, without restricting the structure of the LSTM weights, the second architecture tends to simultaneously address the differences and correlation among the modalities, as well as the requirement for useful emotion-related features; thus, it is more prone to overfitting.

The multimodal LSTM network explicitly learns the correlation among the modalities by sharing the weights across the modalities. As shown in Fig. 2, the network builds an LSTM for each modality while sharing the weights W_{h*}. The formulas, excluding the bias terms, are as follows:

$$\tilde{c}_t^s = \tanh(W_{hg} * h_{t-1}^s + W_{xg}^s * x_t^s), \quad (7)$$
$$f_t^s = \sigma(W_{hf} * h_{t-1}^s + W_{xf}^s * x_t^s,) \quad (8)$$
$$i_t^s = \sigma(W_{hi} * h_{t-1}^s + W_{xi}^s * x_t^s), \quad (9)$$
$$o_t^s = \sigma(W_{ho} * h_{t-1}^s + W_{xo}^s * x_t^s), \quad (10)$$
$$c_t^s = f_t^s \odot c_{t-1}^s + i_t^s \odot \tilde{c}_t^s, \quad (11)$$
$$h_t^s = o_t^s \odot \tanh(c_t^s), \quad (12)$$

where the superscript ^{s} indicates each modality in the input sequences and the subscript $_t$ indicates the time step.

Each modality in the multimodal LSTM network has its own weights W_{x*}^s, hidden states h_t^s, and cell states c_t^s. Therefore, the network is allowed to behave uniquely, and learn separate temporal features for different modalities. This is convenient for handling asynchronous emotion reflection among the modalities, where the strategies of different gates are required. It also provides more interpretation and training efficiency by concentrating on one modality's input. When building the deep multimodal LSTM network, exclusive states preserve the multimodal property, thereby facilitating complex correlation learning among the modalities.

At the same time, the cross-modal correlation is explicitly learned by sharing the weights W_{h*}. Both the previous hidden states h_{t-1} and the current input x_t are involved in the calculation of the gate output and cell-state candidates that controls the process of selectively remembering and forgetting information through time. W_{h*}, and not W_{x*}, is shared because, compared with x_t, h_{t-1} contains information from previous time steps, and is more useful for learning temporal property. By sharing W_{h*} across the modalities, multimodal LSTM layers can interact with each other during the learning processes. Provided that W_{h*} are also shared across the

time steps, they play an important role in learning the temporal correlation among the different modalities.

3.3 Residual Learning

The concept of residual learning was introduced in [11] for training ultra-deep CNNs for image recognition. Residual learning provides reference for the representation learning of higher layers, and explicitly reformulates the layers' objective as approximate residual functions. The formula can be expressed as follows:

$$y = F(x, W) + x, \tag{13}$$

where x, y are the input and output vectors of the layers under consideration, and $F(x, W)$ represents the residual function learned by the corresponding layers.

Following reformulation, the output becomes the linear combination of the input and a non-linear residual, which provides a shortcut across the layers for training the deep networks more effectively. Although increasing the number of layers can cause the problem of vanishing gradients, residual learning makes the learning identity function possible, regardless of the number of layers simply by setting the residual $F(x, W)$ to zero.

3.4 Layer Normalization

Layer normalization [17] was proposed for reducing the training time of deep RNNs by normalizing the activities of the neurons. It is effective for stabilizing the hidden-state dynamics in RNNs. The layer normalization re-centers and re-scales the neurons' activations using extra normalization terms as follows:

$$\mu_t = \frac{1}{H}\sum_{i=1}^{H}(h_t)_i, \tag{14}$$

$$\delta_t = \sqrt{\frac{1}{H}\sum_{i=1}^{H}((h_t)_i - \mu_t)^2}, \tag{15}$$

$$y_t = f(\frac{g}{\delta_t} \odot (h_t - \mu_t) + b), \tag{16}$$

where $(h_t)_i$ denotes the hidden state of the ith neuron in the LSTM, the subscript $_t$ indicates the time steps, and the subscript $_i$ indicates the neurons in each LSTM layer. g, b are trainable weights of the same shape as h_t which are responsible for scaling and centering the input of the activation function f.

3.5 Multimodal Residual LSTM Network

For complex high-level temporal-feature learning with explicit correlation control, the MMResLSTM network contains multiple LSTM layers (we used four layers in this study) for each modality, and shares the weights W_{h*} across each LSTM layer to learn the temporal correlation among the modalities. Residual learning and layer normalization are also deployed for efficient training. Next, high level representations of multiple modalities are concatenated to predict the emotion labels using a fully connected (dense) layer with softmax activation. The structure of a four-layer MMResLSTM network is shown in Fig. 3.

In the MMResLSTM layers, the recurrent connections of the LSTM represent the feed-back of the hidden states for future time-step calculation, and the cross-layer skipping connections represent the identity mapping in residual learning. W_{h*} are shared across the modalities for cross-modality correlation learning. In consonance with [23], dropout is applied to the forward connections to reduce overfitting.

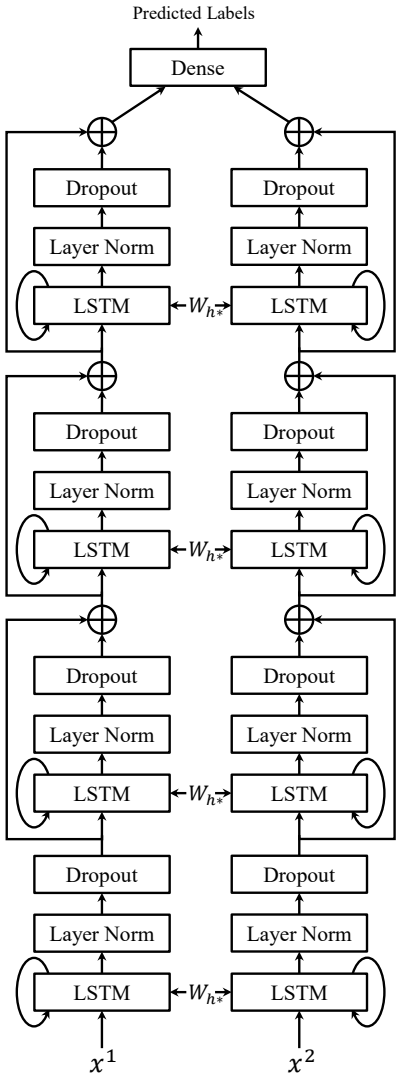

Figure 3: The MMResLSTM network for bimodal input, where x_1, x_2 denotes the input sequences of two modalities, and W_{h*} denotes the shared weights across modalities in each layer.

4 EXPERIMENTS

4.1 Datasets

In this study, we evaluated the model's performance using the DEAP dataset. The DEAP dataset [16] contains EEG and peripheral physiological signals (PPS) that include EOG and electromyographic (EMG) data. In this dataset, 32 subjects watched 40 emotion-stimulating video clips. Each video was one-minute long. On a scale

of 1-9, personal ratings 9 were provided in four emotional dimensions: arousal, valence, dominance, and liking. Similar to many state-of-the-art studies, we evaluated our model on the dimensions of arousal and valence, thereby reformulating the problem into two single-trial binary classification problems, thus, we undertook the classification task of low/high arousal and low/high valence in each experiment, with 5 being the threshold value. The model was trained and tested individually for each subject, thus making our approach subject-dependent. This choice was predicated on two factors: 1) According to our review of recent studies on arousal-valence tasks, scant subject-independent studies exist. A subject-dependent approach enables more comparisons and 2) the subject-independent approach is likely to raise personal privacy concerns, as it necessitates collecting private EEG data and building a large emotion database.

4.2 Feature Extraction and Implementation

As mentioned previously, one important feature of DNN approaches is the end-to-end approach in which the conventional feature extraction process is usually implicitly implemented. In our implementation of the MMResLSTM network, we utilize the downloadable DEAP data (preprocessed) directly without any feature extraction. Raw EEG signals and PPS were down-sampled to 128 Hz. For the EEG signals, the EOG artifacts were removed and a bandpass frequency filter ranging from 4.0-45.0 Hz was applied. The EEG signals were also averaged to the common reference. The preprocessing details are stated in [16].

The input signals obtained from each subject (40 63-s video clips) were divided into 1-s non-overlapping intervals, shuffled, and then fed into the LSTM network. The total dimensions of the input intervals were formed by thirty-two channels of EEG signal and eight channels of PPS. Because the sample rate was 128 Hz, each input of the LSTM network consisted of 128×32 for EEG and 128×8 for PPS.

To compare the performance of the proposed MMResLSTM network with that of the state-of-the-art methods based on feature extraction, we also implemented the following feature extraction routine that utilized the wavelet entropy and wavelet energy features [2, 6, 33]. First, *db*4 discrete wavelet transform was applied to each 4-s signal of each channel with a 2-s overlap to produce a series of wavelet coefficients. These level 1-4 wavelet coefficients approximately represent the components of the gamma, beta, alpha, and theta bands, respectively. Second, the wavelet coefficients were used for calculating the wavelet entropy and wavelet energy as follows:

$$entropy = -\sum_{k=1}^{N} D_j(k)^2 log D_j(k)^2, k = 1...N, \quad (17)$$

$$energy = \sum_{k=1}^{N} D_j(k)^2, k = 1...N, \quad (18)$$

where $D(k)$ represents the k-th wavelet coefficient, and subscription j indicates the level of wavelet decomposition ($j = 1, 2, 3, 4$). The extracted wavelet entropy and wavelet energy features were then used to train an SVM classifier.

Table 1: Comparison of the plain LSTM, residual LSTM, and MMResLSTM

method	layer No.	arousal (%)	valence (%)
LSTM	4	91.00±4.10	89.04±5.79
residual LSTM	4	92.34±2.15	91.66±1.69
MMResLSTM	2	92.54±2.18	91.85±1.81
	3	92.65±2.50	92.15±1.61
	4	**92.87±2.11**	**92.30±1.55**

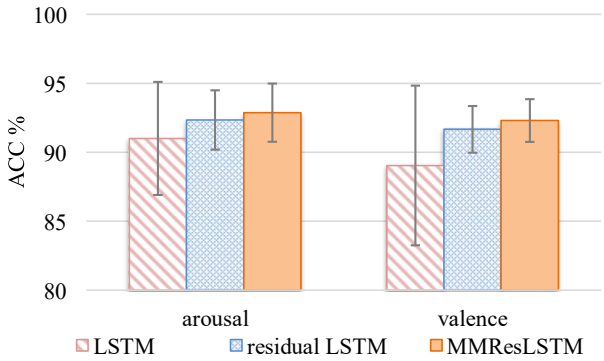

Figure 4: Comparison results of the plain LSTM, residual LSTM, and MMResLSTM (all with 4 layers)

All the methods were implemented using Python and TensorFlow[1].

4.3 Results

4.3.1 Effect of residual and weight-sharing terms. First, we conducted experiments to compare the plain LSTM network, LSTM with residual terms, and the proposed MMResLSTM to explore the effect of the residual and weight-sharing terms. We experimented with the combinations of the following parameters:

- Node number: 32, 64, 128
- Layer number: 1, 2, 3, 4
- Learning rate: 1e-2, 1e-3, 3e-3, 1e-4
- L2 regularization: 1e-2, 1e-4, 1e-6

Following experiments, the optimal parameter settings were determined as follows for all the LSTM models: The LSTM node number (for both modalities) was set to 128, layer number, 4, dropout ratio, 0.5, and the learning rate, 3e-3; The l2 regularization was 1e-6 for the arousal task and 1e-2 for the valence task and the maximum number of epochs was 2000. The Adam optimization algorithm was used to train networks. Furthermore, training was terminated early, that is, when the accuracy had not increased 0.1% in the previous 120 epochs. Five-fold cross validation (with randomly shuffled data) was applied to each subject, and the average classification accuracy over subjects and folds was calculated.

The results for all the thirty-two subjects are shown in Table 1, Fig. 4, and Fig. 5. From Table 1 and Fig. 4, it can be observed that,

[1]https://www.tensorflow.org/

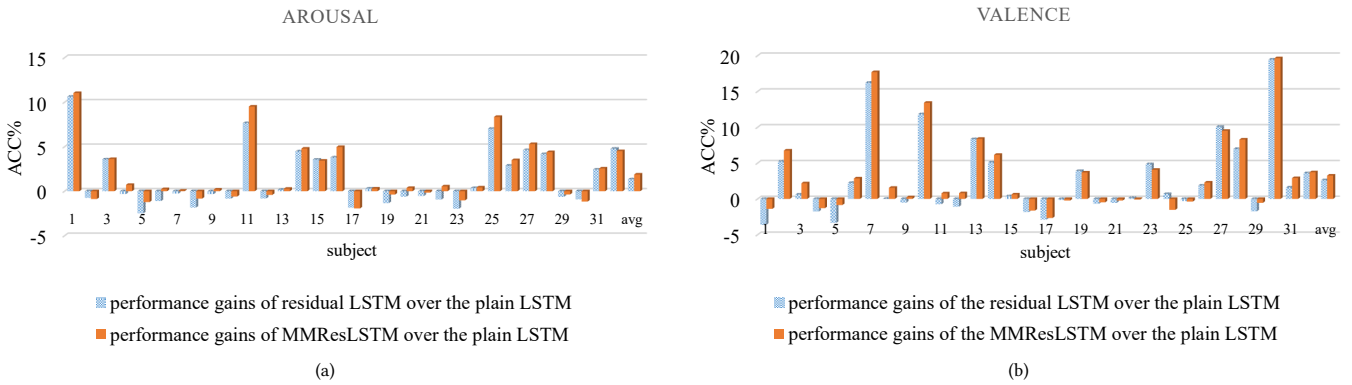

Figure 5: The results of the performance gains of the residual LSTM over the plain LSTM, and the performance gains of the MMResLSTM over the plain LSTM, respectively, on the (a) arousal and (b) valence classification task, for 32 individual subjects and their average

first, in both the arousal and valence tasks, the MMResLSTM outperformed the other two models. Second, the results obtained with the plain LSTM have very large deviations in both tasks. The results were stabilized as a result of the effect of the residual terms, and the average accuracies were increased. Third, the performance of MMResLSTM increased with the increase of layer number. Compared with the residual LSTM, even the 2-layer MMResLSTM yielded slightly greater accuracies. This result implies that the advantages of learning the correlation by sharing weights across the modalities were significant.

When, as shown in Fig. 4, the three methods have overlapping (similar) accuracy regions, it is worth noting that the accuracies of the different methods should be compared subject by subject, not group by group. We conducted a paired-t test to demonstrate the statistical difference among the plain LSTM, residual LSTM, and MMResLSTM. Consequently, all the pairs (plain vs. residual, plain vs. MMRes, and residual vs. MMRes) in both the arousal and valence tasks have $p < 0.05$, which means the differences between the pairs are statistically significant.

The improvements attributable to the residual terms and weight-sharing terms can be observed more clearly in Fig. 5, where the individual results for all the thirty-two subjects are shown. In the figure, the dotted bars (blue) indicate the performance gains of the residual LSTM over the plain LSTM, and the solid bars (orange) indicate the performance gains of the MMResLSTM over the plain LSTM. In some cases, merely applying the residual terms boosted the accuracy significantly, and the weight-sharing terms resulted in additional performance gain (e.g., #11 in Fig. 5(a), #7 and #10 in Fig. 5(b)). Further more, sometimes when the residual LSTM performed worse than the plain LSTM, sharing the weights compensated the loss (e.g., #5 in both Fig. 5(a) and Fig. 5(b)). It is true that occasionally, even the MMResLSTM failed to improve the performance (e.g., #24 in Fig. 5(b)). This is mainly attributable to the use of the same set of hyper-parameters for all the 32 subjects, rather than optimizing these hyper-parameters for each subject. This was to reduce overfitting, and validate the robustness of the network without optimizing against each individual. As long as the proposed method exhibits a superior performance overall, its underperformance in comparison to the plain LSTM in minor cases is insignificant.

4.3.2 Comparison with other methods. Next, we conducted experiments to compare the performance of the proposed method with some feature-extraction-based and end-to-end state-of-the-art methods.

Table 2: Comparison of the accuracies of the proposed MMResLSTM method, the reference SVM method, with the reported accuracies of representative subject-dependent approaches. In some studies, standard deviations were not provided; thus, only the detection accuracies of arousal and valence are listed.

method	feature	modality	arousal (%)	valence (%)
MESAE [36]	statistical features	EEG+PPS	84.18	83.04
CNN [18]	statistical features	EEG+PPS	87.30	85.50
k-NN [24]	statistical features	EEG	89.84	89.61
LSTM [32]	differential entropy	EEG+PPS	83.23±2.61	83.82±5.01
DCCA [25]	differential entropy	EEG+PPS	84.33	85.62
BDAE [19]	raw signal	EEG+PPS	80.5	85.2
SVM	entropy & energy	EEG	86.80±9.58	86.41±9.40
		PPS	72.59±9.73	74.44±9.84
		EEG+PPS	88.15±9.11	89.53±7.62
MM-ResLSTM	raw signal	EEG+PPS	**92.87±2.11**	**92.30±1.55**

In Table 2, we collected the results obtained for the DEAP dataset in recent studies. These studies use various approaches of feature extraction and classification models; however, they were all

subject-dependent, and most of them adopted multiple modalities (32-channel EEG plus 8-channel PPS). Under these similar experiment settings, the proposed MMResLSTM outperformed these other methods in terms of accuracy.

To further validate the proposed method, we also implemented a reference method that used the wavelet entropy and wavelet energy as features. These features have been reported to be effective for emotion recognition [2, 6, 33]. SVM (with RBF kernel) is a frequently used benchmark; thus, it was chosen as the classifier. The other settings remained similar to our proposed method (subject dependent, 5-fold cross validation with shuffled data).

We tested the method on varying modality settings (only EEG, only PPS, and concatenated EEG and PPS); greater accuracies were achieved with EEG + PPS than the other two variations, which indicates the value of using information extracted from multiple modalities. Even so, the approach incorporating feature extraction and the SVM classifier did not outperform the proposed MMResLSTM method. Although the MMResLSTM does not incorporate any feature extraction technique, its deep residual structure learns the complex high-level features, and the temporal correlation between the two modalities is learnt through weight sharing.

5 CONCLUSION

Biosignals from different modalities contain different aspects of human emotions. Finding the correlation between the information from these modalities and integrating them could improve the recognition of human emotions. In this study, we proposed the MMResLSTM, a multimodal deep LSTM network reinforced by residual learning and weight sharing. The network functions in an end-to-end way to implicitly extract high-level temporal features. Experiments on the DEAP dataset showed that the proposed method classified arousal and valence with an accuracy of 92.87% and 92.30%, respectively, thereby outperforming the state-of-the-art methods.

In future, we plan to extend the proposed network to additional applications that simultaneously utilize EEG and other modalities, such as the data analysis of sleep stages, labor concentration, and driving fatigue, possibly undertaking the more challenging task of utilizing a totally different signal (e.g. video) as the secondary modality.

6 ACKNOWLEDGMENTS

This work was supported in part by the grants from the National Key Research and Development Program of China (Grant No. 2017-YFB1002501), the National Natural Science Foundation of China (Grant No. 61673266), and the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] Salma Alhagry, Aly Aly, and Reda El-Khoribi. 2017. Emotion Recognition based on EEG using LSTM Recurrent Neural Network. *International Journal of Advanced Computer Science and Applications* 8, 10 (2017). https://doi.org/10.14569/IJACSA.2017.081046

[2] Mouhannad Ali, Fadi Al Machot, Ahmad Haj Mosa, and Kyandoghere Kyamakya. 2016. A Novel EEG-based Emotion Recognition Approach for e-Healthcare Applications. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing (SAC '16)*. ACM, New York, NY, USA, 162–164. https://doi.org/10.1145/2851613.2851916

[3] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. 2016. Learning representations from EEG with deep recurrent-convolutional neural networks. In *International Conference on Learning Representations (ICLR)*.

[4] H. Becker, J. Fleureau, P. Guillotel, F. Wendling, I. Merlet, and L. Albera. 2018. Emotion recognition based on high-resolution EEG recordings and reconstructed brain sources. *IEEE Transactions on Affective Computing* (2018), 1–1. https://doi.org/10.1109/TAFFC.2017.2768030

[5] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI '04)*. ACM, New York, NY, USA, 205–211. https://doi.org/10.1145/1027933.1027968

[6] H. Candra, M. Yuwono, R. Chai, A. Handojoseno, I. Elamvazuthi, H. T. Nguyen, and S. Su. 2015. Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 7250–7253. https://doi.org/10.1109/EMBC.2015.7320065

[7] Weitong Chen, Sen Wang, Xiang Zhang, Lina Yao, Lin Yue, Buyue Qian, and Xue Li. 2018. *EEG-based Motion Intention Recognition via Multi-task RNNs*. 279–287. https://doi.org/10.1137/1.9781611975321.32 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611975321.32

[8] Rahma Fourati, Boudour Ammar, Chaouki Aouiti, Javier Sanchez-Medina, and Adel M. Alimi. 2017. Optimized Echo State Network with Intrinsic Plasticity for EEG-Based Emotion Recognition. In *Neural Information Processing*, Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, and El-Sayed M. El-Alfy (Eds.). Springer International Publishing, Cham, 718–727. https://doi.org/10.1007/978-3-319-70096-0_73

[9] Beatriz García-Martínez, Arturo Martínez-Rodrigo, Roberto Zangróniz, José Manuel Pastor, and Raúl Alcaraz. 2017. Symbolic Analysis of Brain Dynamics Detects Negative Stress. *Entropy* 19, 5 (2017). https://doi.org/10.3390/e19050196

[10] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 6645–6649.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[13] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International International Conference on Machine Learning - Volume 37 (ICML'15)*. JMLR.org, 448–456. http://dl.acm.org/citation.cfm?id=3045118.3045167

[14] S. Katsigiannis and N. Ramzan. 2018. DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices. *IEEE Journal of Biomedical and Health Informatics* 22, 1 (Jan 2018), 98–107. https://doi.org/10.1109/JBHI.2017.2688239

[15] B. H. Kim and S. Jo. 2018. Deep Physiological Affect Network for the Recognition of Human Emotions. *IEEE Transactions on Affective Computing* (2018), 1–1. https://doi.org/10.1109/TAFFC.2018.2790939

[16] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2012. DEAP: A Database for Emotion Analysis using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (Jan 2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[17] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[18] Wenqian Lin, Chao Li, and Shouqian Sun. 2017. Deep Convolutional Neural Network for Emotion Recognition Using EEG and Peripheral Physiological Signal. In *Image and Graphics*, Yao Zhao, Xiangwei Kong, and David Taubman (Eds.). Springer International Publishing, Cham, 385–394. https://doi.org/10.1007/978-3-319-71589-6_33

[19] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. 2016. Emotion Recognition Using Multimodal Deep Learning. In *Proceedings of the 23rd International Conference on Neural Information Processing - Volume 9948*. Springer-Verlag, Berlin, Heidelberg, 521–529. https://doi.org/10.1007/978-3-319-46672-9_58

[20] Y. Liu and O. Sourina. 2013. EEG Databases for Emotion Recognition. In *2013 International Conference on Cyberworlds*. 302–309. https://doi.org/10.1109/CW.2013.52

[21] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. 2015. Combining Eye Movements and EEG to Enhance Emotion Recognition. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, 1170–1176. http://dl.acm.org/citation.cfm?id=2832249.2832411

[22] Pallavi Pandey and K. R. Seeja. 2019. Emotional State Recognition with EEG Signals Using Subject Independent Approach. In *Data Science and Big Data Analytics*, Durgesh Kumar Mishra, Xin-She Yang, and Aynur Unal (Eds.). Springer Singapore, Singapore, 117–124. https://doi.org/10.1007/978-981-10-7641-1_10

[23] V. Pham, T. Bluche, C. Kermorvant, and J. Louradour. 2014. Dropout Improves Recurrent Neural Networks for Handwriting Recognition. In *2014 14th International Conference on Frontiers in Handwriting Recognition*. 285–290. https:

//doi.org/10.1109/ICFHR.2014.55

[24] L. Piho and T. Tjahjadi. 2018. A mutual information based adaptive windowing of informative EEG for emotion recognition. *IEEE Transactions on Affective Computing* (2018), 1–1. https://doi.org/10.1109/TAFFC.2018.2840973

[25] Jie-Lin Qiu, Wei Liu, and Bao-Liang Lu. 2018. Multi-view Emotion Recognition Using Deep Canonical Correlation Analysis. In *Proceedings of the 25th International Conference on Neural Information Processing (ICONIP'18)*. 221–231. https://doi.org/10.1007/978-3-030-04221-9_20

[26] Jimmy Ren, Yongtao Hu, Yu-Wing Tai, Chuan Wang, Li Xu, Wenxiu Sun, and Qiong Yan. 2016. Look, Listen and Learn — a Multimodal LSTM for Speaker Identification. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16)*. AAAI Press, 3581–3587. http://dl.acm.org/citation.cfm?id=3016387.3016407

[27] M. Soleymani, S. Asghari-Esfeden, M. Pantic, and Y. Fu. 2014. Continuous emotion detection using EEG signals and facial expressions. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME.2014.6890301

[28] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing* 3, 1 (Jan 2012), 42–55. https://doi.org/10.1109/T-AFFC.2011.25

[29] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training Very Deep Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'15)*. MIT Press, Cambridge, MA, USA, 2377–2385. http://dl.acm.org/citation.cfm?id=2969442.2969505

[30] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. 2017. DeepSleepNet: a Model for Automatic Sleep Stage Scoring based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 11 (Nov 2017), 1998–2008. https://doi.org/10.1109/TNSRE.2017.2721116

[31] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[32] Hao Tang, Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. 2017. Multimodal Emotion Recognition Using Deep Neural Networks. In *Neural Information Processing*, Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, and El-Sayed M. El-Alfy (Eds.). Springer International Publishing, Cham, 811–819. https://doi.org/10.1007/978-3-319-70093-9_86

[33] Gyanendra K. Verma and Uma Shanker Tiwary. 2017. Affect Representation and Recognition in 3D Continuous Valence–arousal–dominance Space. *Multimedia Tools Appl.* 76, 2 (Jan. 2017), 2159–2183. https://doi.org/10.1007/s11042-015-3119-y

[34] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935

[35] Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu. 2014. Emotional state classification from EEG data using machine learning approach. *Neurocomputing* 129 (Apr. 2014), 94–106. https://doi.org/10.1016/j.neucom.2013.06.046

[36] Zhong Yin, Mengyuan Zhao, Yongxiong Wang, Jingdong Yang, and Jianhua Zhang. 2017. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer Methods and Programs in Biomedicine* 140 (2017), 93–110. https://doi.org/10.1016/j.cmpb.2016.12.005

[37] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. 2018. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics* 49, 3 (Feb. 2018), 1110–1122. https://doi.org/10.1109/TCYB.2018.2797176

[38] Wei-Long Zheng and Bao-Liang Lu. 2015. Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development* 7, 3 (Sep. 2015), 162–175. https://doi.org/10.1109/TAMD.2015.2431497

[39] Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. 2018. Identifying Stable Patterns over Time for Emotion Recognition from EEG. *IEEE Transactions on Affective Computing* (2018), 1–1. https://doi.org/10.1109/TAFFC.2017.2712143

[40] Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutník, and Jürgen Schmidhuber. 2017. Recurrent Highway Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Doina Precup and Yee Whye Teh (Eds.), Vol. 70. PMLR, International Convention Centre, Sydney, Australia, 4189–4198. http://proceedings.mlr.press/v70/zilly17a.html