Multimodal Emotion Recognition Using Deep Neural Networks

Hao Tang[1], Wei Liu[1], Wei-Long Zheng[1], and Bao-Liang Lu[1,2,3(✉)]

[1] Department of Computer Science and Engineering,
Center for Brain-like Computing and Machine Intelligence, Shanghai, China
{silent56,liuwei-albert,weilong}@sjtu.edu.cn
[2] Key Laboratory of Shanghai Education Commission for Intelligent
Interaction and Cognitive Engineering, Shanghai, China
[3] Brain Science and Technology Research Center,
Shanghai Jiao Tong University, Shanghai, China
bllu@sjtu.edu.cn

Abstract. The change of emotions is a temporal dependent process. In this paper, a Bimodal-LSTM model is introduced to take temporal information into account for emotion recognition with multimodal signals. We extend the implementation of denoising autoencoders and adopt the Bimodal Deep Denoising AutoEncoder modal. Both models are evaluated on a public dataset, SEED, using EEG features and eye movement features as inputs. Our experimental results indicate that the Bimodal-LSTM model outperforms other state-of-the-art methods with a mean accuracy of 93.97%. The Bimodal-LSTM model is also examined on DEAP dataset with EEG and peripheral physiological signals, and it achieves the state-of-the-art results with a mean accuracy of 83.53%.

Keywords: Multimodal emotion recognition · EEG · Deep neural networks · LSTM

1 Introduction

Automatic emotion recognition has drawn increasing attention due to its potential applications to human computer interaction. There are many modalities that contain emotion information, such as facial expression, voice, electroencephalography (EEG), eletrocardiogram (ECG), pupillary diameter (PD), and so on. However, since emotions are complex and associated with many nonverbal cues, it's difficult to recognize emotions robustly based on a single modality. Saneiro *et al.* detected emotions in educational scenarios from facial expressions and body movements [8]. Koelstra *et al.* built an emotion recognition system based on EEG and peripheral physiological signals [5]. Lu *et al.* used both EEG signals and eye movement signals to recognize three types of emotions and revealed that EEG features and eye movement features were complementary to emotion recognition [7]. Liu *et al.* furthermore used Bimodal Deep AutoEncoder to

D. Liu et al. (Eds.): ICONIP 2017, Part IV, LNCS 10637, pp. 811–819, 2017.
https://doi.org/10.1007/978-3-319-70093-9_86

extract high level representation features and achieved competitive results on both SEED[1] and DEAP[2] datasets [6].

Most of the existing methods treated features at each time step as independent samples, and ignored the temporal dependency property of emotions [11]. Recurrent Neural Networks (RNNs) are powerful tools for modeling sequential data and have the ability to extract temporal information from input signals. Moreover, Long Short Term Memory (LSTM) neural network [4], which is a gated RNN with linear self-loop, has been successfully used to capture temporal dependency property in many fields, such as speech recognition [13] and machine translation [12]. In this paper, to reveal the effect of temporal information in emotion recognition, we introduced a Bimodal-LSTM (Long Short Term Memory) model, which could use both the temporal information and the frequency-domain information to discriminate emotion states. Specifically, the Bimodal-LSTM model consists of two LSTM encoders, for features from EEG and other modalities respectively, and one classification layer. We also extended the implementation of denoising autoencoders in the field of multimodal emotion recognition and introduced the Bimodal Deep Denoising AutoEncoder (BDDAE) model. We evaluated our proposed models on two public multimodal datasets called SEED and DEAP for emotion recognition and achieved the state-of-the-art performance.

2 Bimodal Deep Denoising AutoEncoders

2.1 Denoising Autoencoders

An autoencoder is an unsupervised model, which can be used for dimensionality reduction, data compression, and feature learning [2,3]. Classical autoencoders map the input to its hidden representation with an encoder function and then use a decoder function to map the hidden representation to the reconstruction of input. The reconstruction errors are minimized to train autoencoders.

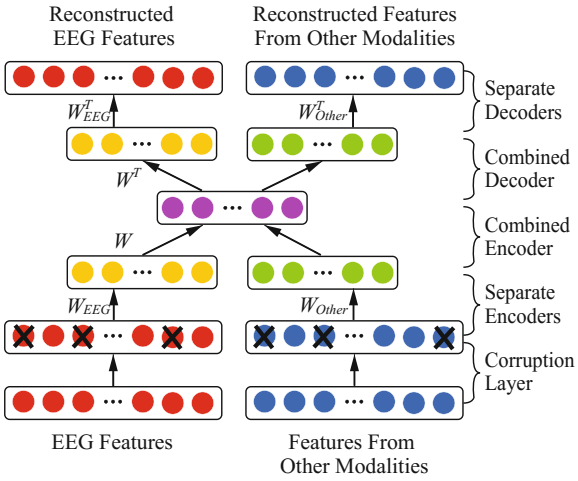
The denoising autoencoder, which is an extension of the classical autoencoder, reconstructs the input from a corrupted version of it [10]. It can prevent the autoencoder from learning the identity function when the encoder and decoder are given too much capacity. And denoising autoencoders can learn more robust hidden representation.

The Bimodal Deep Denoising AutoEncoder (BDDAE) model consists of two networks, the autoencoder network and the classifier network. The autoencoder network is used to pre-train the encoders' weights. And the classifier network predicts emotion labels using EEG features and other modalities' features.

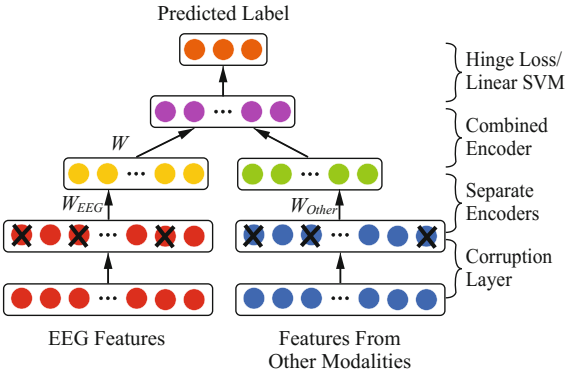
The autoencoder network of BDDAE, as illustrated in Fig. 1(a), contains one corruption layer, three encoders, and three decoders. The corruption layer randomly sets some of the inputs to zeros according to the dropout probability. The encoders and decoders, whose form is an affine mapping followed by a sigmoid

[1] http://bcmi.sjtu.edu.cn/~seed/.
[2] http://www.eecs.qmul.ac.uk/mmv/datasets/deap/.

function, are mirror images of each other. There are two encoders for EEG features and other modalities' features, respectively. The encoded features are then concatenated together, and another encoder is used to extract the combined high-level features. Mean squared error criterion is used to train the network.



(a) The autoencoder network



(b) The classifier network

Fig. 1. The structure of BDDAE's autoencoder network and classifier network

The classifier network of BDDAE is depicted in Fig. 1(b). It also contains three encoders, which are the same as the autoencoder network. Moreover, the last layer can be considered as a linear kernel Support Vector Machine (SVM), since the loss function the network uses is L2-regularized hinge loss [9].

2.2 Training

We firstly pre-trained the encoders' weights, W_{EEG}, W_{Other}, W by training the autoencoder network. Moreover, to adapt encoders to the specific task (emotion recognition), we attached the SVM layer to the encoder layers and trained the classifier network. In detail, the pre-trained encoders' weight matrices as in Fig. 1(a) were used to initialize the corresponding encoders' weight matrices in the classifier network, as in Fig. 1(b). The encoder layers' learning rate was set to one percent of the last classification layer's learning rate, so that the classification layer was trained mostly, while the encoder layers were fine-tuned.

3 Bimodal-LSTM

3.1 LSTM Neural Networks

To incorporate the temporal dependency information of features, we introduce Long Short Term Memory (LSTM) neural networks as a temporal encoder. LSTM neural network, which is a RNN using LSTM blocks, can prevent the vanishing (and exploding) gradient problem [1] and has the ability to learn information from long sequences. Each LSTM block contains memory cell states c_t propagated over time. At every time step, the states of memory cells c_t are updated according to the input of current time step x_t and the output of the previous time step h_{t-1} as follows:

$$\begin{aligned} f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) \\ g_t &= \tanh(W_g[h_{t-1}, x_t] + b_g) \\ c_t &= c_{t-1} * f_t + i_t * g_t, \end{aligned} \tag{1}$$

where σ denotes the sigmoid function, f_t, i_t denotes the forget gate and input gate, g_t denotes the candidate of cell states, W_f, W_i, W_g denotes the weight matrices, b_f, b_i, b_g denotes the biases and c_t and h_t are the memory cell states and the output of LSTM block, respectively. The forget gate controls the process of forgetting information by multiplying the cell states by real numbers between zero and one. Similarly, the input gate controls the process of remembering information. The output of LSTM blocks h_t is a filtered version of memory cell states, as follows:

$$\begin{aligned} o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(c_t). \end{aligned} \tag{2}$$

where W_o and b_o denotes the weight matrix and the bias for the output gate o_t. The output gate controls the process of output.

As depicted in Fig. 2, the Bimodal-LSTM network contains one classification layer and two LSTM encoders. Two LSTM encoders are for EEG features and other modalities' features, respectively. The network also uses L2-regularized

hinge loss as the objective function to minimize, so the classification layer can be considered as a linear kernel SVM. Dropout is applied to the output of LSTM blocks to obtain more robustness.

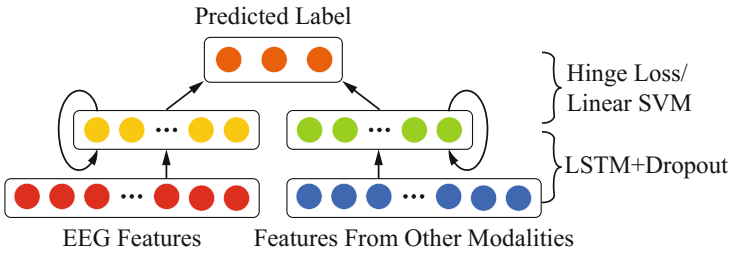


Fig. 2. The structure of Bimodal-LSTM

3.2 Training

We firstly trained the network thoroughly using Adam optimization algorithm. To further minimize the loss of the network after training it thoroughly, we trained a SVM classifier, which minimized the same loss function as the Bimodal-LSTM network. In detail, we used the trained LSTM encoders to extract high-level features from EEG features and features from other modalities at each time step. And then the extracted features were multiplied by the dropout probability to simulate the effect of dropout layer at test time. We used the liblinear package[3] to implement the SVM classifier and trained it using the scaled high-level features.

To minimize the same loss function as in the Bimodal-LSTM network, we optimized the primal problem by setting the option '$-s$' to 2 and set the cost of the SVM classifier equal to $\frac{1}{2\lambda}$, where λ denotes the L2 regularization strength used when training the network thoroughly. After training the SVM classifier, we copied the trained weights back into the last classification layer of the Bimodal-LSTM network to produce the final classifier.

4 Experiment Settings

4.1 The Datasets

The SEED dataset contains EEG signals and eye movement signals of three emotions (positive, neutral, and negative) from 15 subjects . All subjects were watching 15 four-minute-long emotional movie clips while their signals were collected. The EEG signals were recorded with ESI NeuroScan System at a sampling rate of 1000 Hz with a 62-channel electrode cap. The eye movement signals were recorded with SMI ETG eye tracking glasses. To compare with our previous work

[3] http://www.csie.ntu.edu.tw/~cjlin/liblinear/.

[6,7], we used the same data, which contained 27 experiments from 9 subjects. Signals recorded while the subject watching the first 9 movie clips were used as training datasets for each experiment and the rest were used as test datasets.

The DEAP dataset contains EEG signals and peripheral physiological signals of 32 participants. Signals were collected while participants were watching one-minute-long emotional music videos. And participants were asked to rate the levels of arousal, valence, like/dislike, dominance and familiarity for each video. We chose 5 as the threshold to divide the trials into two classes according to the rated levels of arousal and valence. Then the tasks can be treated as two binary classification problems, namely high or low arousal and valence. We used 10-fold cross validation to compare with [6,15].

4.2 Feature Extraction

For SEED dataset, we extracted Differential Entropy (DE) features from each EEG signal channel in five frequency bands: δ (1–4 Hz), θ (4–8 Hz), α (8–14 Hz), β (14–31 Hz), and γ (31–50 Hz). The size of Hanning window used when extracting EEG features was 4 s. At each time step, there were totally 310 (5 bands $\times$ 62 channels) dimensions for EEG features. As for eye movement data, the same features as in [6] were used. There were totally 41 dimensions including both Power Spectral Density (PSD) and DE features of pupil diameters at each time step. The features were rescaled between 0 and 1 when used as the inputs of the BDDAE model. Before training the BLSTM model, the features were normalized to zero mean and unit variance and then split into small sequences of length 60.

For DEAP dataset, we extracted DE features from EEG signals in four frequency bands: θ (4–8 Hz), α (8–14 Hz), β (14–31 Hz), and γ (31–50 Hz), since a bandpass frequency filter from 4 - 45 Hz was applied during pre-processing. The size of Hanning windows was 2 s. Then there were totally 128 (4 bands $\times$ 32 channels) dimensions of extracted 32-channel EEG features. As for peripheral physiological signals, time-domain features were extracted to describe the signals in different perspective, including maximum value, minimum value, mean value, standard deviation, variance and squared sum. So there were totally 48 (6 features $\times$ 8 channels) dimensions of extracted peripheral physiological features. Features were rescaled to $[0, 1]$ before fed into the BDDAE model. And we also standardized the features and split them into sequences of length 5 to train the Bimodal-LSTM model.

4.3 Parameter Details

We trained different models for different experiments. For each experiment, we randomly selected some sets of hyper-parameters within a given range to train the model. The hyper-parameters of the BDDAE model include the hidden units' number of three encoders, the dropout probability, the L2 regularization strength, and the learning rate for the autoencoder network and the classifier

network. The hyper-parameters and their corresponding range of the Bimodal-LSTM model for SEED and DEAP datasets are shown in Table 1.

Both models were implemented using tensorflow[4]. All weights were initialized from a Gaussian distribution with a mean of 0 and a standard deviation of 0.001. All biases were initialized to zero. And the initial hidden states and cell states of LSTM blocks were set to zero. The Adam optimization algorithm was used to train networks. EarlyStopping was also adopted to stop training when the accuracy had not increased 0.1% in the last 120 epochs.

Table 1. The hyper-parameters and their corresponding range of the Bimodal-LSTM model for SEED and DEAP datasets

Model	Hyper-parameter	SEED range	DEAP range
Bimodal-LSTM	EEG hidden size	16 to 256	32 to 256
	Other modalities' hidden size	8 to 64	16 to 256
	Dropout probability	0.3 to 0.99	0.2 to 0.9
	$\log_{10}$(L2 regularization strength)	-9 to 0	-4.5 to -1
	$\log_{10}$(learning rate)	-4 to -1.5	-2.5 to -0.5

5 Experiment Results

For SEED dataset, we randomly selected 200 sets of hyper-parameters within a given range for each experiment. We compared our models with two other state-of-the-art approaches [6,14] and the baseline method, which uses SVM directly as the classifier. As shown in Fig. 3, Bimodal-LSTM achieves the best accuracy (93.97%), which is about 2% points higher than the state-of-the-art approaches, and the smallest standard deviation (7.03%).

For DEAP dataset, we randomly selected 15 sets of hyper-parameters and tuned the parameters using 10-fold cross validation. The Bimodal-LSTM model was compared with one baseline method and two state-of-the-art approaches [6,15]. Liu *et al.* used Bimodal Deep AutoEncoder to extract high level features and used the preprocessed data as inputs. Yin *et al.* proposed a multiple-fusion-layer based ensemble classifier of stacked autoencoder to recognize emotions, and also estimated the accuracy by 10-fold cross validation. The baseline method used the same features as the Bimodal-LSTM model and used linear kernel SVM as the classifier. As shown in Table 2, Bimodal-LSTM obtains state-of-the-art performance on both arousal and valence classification tasks, with the mean accuracies of 83.23% and 83.83%, respectively.

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

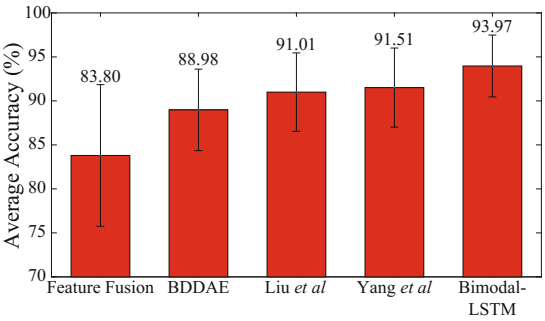


Fig. 3. Results of different models on SEED dataset. Feature Fusion denotes the model using directly concatenated features as inputs and using SVM with a radial basis function (RBF) kernel as the classifier. Liu *et al.* denotes the best result in [6], which uses the Bimodal Deep AutoEncoder model. And Yang *et al.* denotes the best result in [14].

Table 2. Average accuracies (%) and standard deviations of different approaches on DEAP dataset

	Feature fusion	Liu *et al.* [6]	Yin *et al.* [15]	Bimodal-LSTM
Arousal (%)	65.43/7.79	80.5/-	84.18/-	83.23/2.61
Valence (%)	65.29/7.93	85.2/-	83.04/-	83.82/5.01

6 Conclusion

In this paper, we have introduced two models to predict emotions based on EEG features and features from other modalities. The first is an extension of denoising autoencoders, called BDDAE, and the second is the Bimodal-LSTM model, which can use both the temporal information and frequency-domain information of features. Compared with other existing methods, the Bimodal-LSTM model has achieved the best performance with a mean accuracy of 93.97% on SEED dataset. For DEAP dataset, the Bimodal-LSTM model has achieved the state-of-the-art results with mean accuracies of 83.23% and 83.83% for arousal and valence classification tasks, respectively.

Acknowledgments. This work was supported in part by grants from the National Key Research and Development Program of China (Grant No. 2017YFB1002501), the National Natural Science Foundation of China (Grant No. 61673266), the Major Basic Research Program of Shanghai Science and Technology Committee (Grant No. 15JC1400103), ZBYY-MOE Joint Funding (Grant No. 6141A02022604), and the Technology Research and Development Program of China Railway Corporation (Grant No. 2016Z003-B).

References

1. Bengio, Y., Simard, P.Y., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**(2), 157–166 (1994)
2. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
3. Hinton, G.E., Zemel, R.S.: Autoencoders, minimum description length and helmholtz free energy. In: NIPS, pp. 3–10 (1994)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
5. Koelstra, S., Yazdani, A., Soleymani, M., Mühl, C., Lee, J., Nijholt, A., Pun, T., Ebrahimi, T., Patras, I.: Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos. In: Yao, Y., Sun, R., Poggio, T., Liu, J., Zhong, N., Huang, J. (eds.) BI 2010. LNCS, vol. 6334, pp. 89–100. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15314-3_9
6. Liu, W., Zheng, W.L., Lu, B.L.: Emotion recognition using multimodal deep learning. In: Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M., Liu, D. (eds.) ICONIP 2016. LNCS, vol. 9948, pp. 521–529. Springer, Cham (2016). doi:10.1007/978-3-319-46672-9_58
7. Lu, Y., Zheng, W.L., Li, B., Lu, B.L.: Combining eye movements and EEG to enhance emotion recognition. In: IJCAI, pp. 1170–1176 (2015)
8. Saneiro, M., Santos, O.C., Salmeronmajadas, S., Boticario, J.G.: Towards emotion detection in educational scenarios from facial expressions and body movements through multimodal approaches. Sci. World J. **2014**, 484873 (2014)
9. Tang, Y.: Deep learning using linear support vector machines. Workshop on Representational Learning, ICML (2013)
10. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.: Extracting and composing robust features with denoising autoencoders. In: ICML, pp. 1096–1103 (2008)
11. Wang, X.W., Nie, D., Lu, B.L.: Emotional state classification from eeg data using machine learning approach. Neurocomputing **129**, 94–106 (2014)
12. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint (2016). arXiv:1609.08144
13. Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G.: The microsoft 2016 conversational speech recognition system. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5255–5259. IEEE (2017)
14. Yang, Y., Wu, Q.J., Zheng, W.L., Lu, B.L.: EEG-based emotion recognition using hierarchical network with subnetwork nodes. IEEE Trans. Cogn. Dev. Syst. (2017). doi:10.1109/TCDS.2017.2685338
15. Yin, Z., Zhao, M., Wang, Y., Yang, J., Zhang, J.: Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. Comput. Methods Prog. Biomed. **140**, 93–110 (2017)