Emotion Recognition Using Multimodal Deep Learning

Wei Liu[1], Wei-Long Zheng[1], and Bao-Liang Lu[1,2,3,*]

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

Abstract. To enhance the performance of affective models and reduce the cost of acquiring physiological signals for real-world applications, we adopt multimodal deep learning approach to construct affective models with SEED and DEAP datasets to recognize different kinds of emotions. We demonstrate that high level representation features extracted by the Bimodal Deep AutoEncoder (BDAE) are effective for emotion recognition. With the BDAE network, we achieve mean accuracies of 91.01% and 83.25% on SEED and DEAP datasets, respectively, which are much superior to those of the state-of-the-art approaches. By analysing the confusing matrices, we found that EEG and eye features contain complementary information and the BDAE network could fully take advantage of this complement property to enhance emotion recognition.

1 Introduction

Nowadays, many human machine interface (HMI) products are used by ordinary people and more HMI equipments will be needed in the future. Since emotional functions of HMI products play an important role in our daily life, it is necessary for HMI equipments to be able to recognize humans emotions automatically.

Many researchers studied emotion recognition from EEG. Liu *et al.* used fractal dimension based algorithm to recognize and visualize emotions in real time [1]. Li and Lu used EEG signals of gamma band to classify two kinds of emotions, and their results showed that gamma band was suitable for emotion recognition [2]. Duan *et al.* found that differential entropy features are more suited for emotion recognition tasks [3]. Wang *et al.* compared three different kinds of EEG features and proposed a simple approach to track the trajectory of emotion changes with time [4]. Zheng and Lu employed deep neural network to classify EEG signals and examined critical bands and channels of EEG for emotion recognition [5].

To fully use information from different modalities, Yang *et al.* proposed an auxiliary information regularized machine, which treats different modalities with different strategies [6]. In [7], the authors built a single modal deep autoencoder and a bimodal deep autoencoder to generate shared representations of images and audios. Srivastava

* corresponding author

and Salakhutdinov extended the methods developed in [7] to bimodal deep Boltzmann machines to handle multimodal deep learning problems [8].

As for multimodal emotion recognition, Verma and Tiwary carried out emotion classification experiments with EEG singals and peripheral physiological signals [9]. Lu *et al.* used two different fusion strategies for combining EEG and eye movement data: feature level fusion and decision level fusion [10]. Liu *et al.* employed bimodal deep autoencoders to fuse different modalities and the authors tested the framework on multimodal facilitation, unimodal enhancement, and crossmodal learning tasks [11].

To our best knowledge, there is no research work reported in the literature to deal with emotion recognition from multiple physiological signals using multimodal deep learning algorithms. In this paper, we propose a novel multimodal emotion recognition method using multimodal deep learning techniques. In Section 2, we will introduce the bimodal deep autoencoder. Section 3 presents data pre-proessing, feature extraction and experiment settings. The experiment results are described in Section 4. Following discusses in Section 5, conclusions and future work are in Section 6.

2 Multimodal Deep Learning

2.1 Restricted Boltzmann Machine

A restricted Boltzmann machine (RBM) is an undirected graph model, which has a visible layer and a hidden layer. Connections exist only between visible layer and hidden layer and there is no connection either in visible layer or in hidden layer. Assuming visible variables $\mathbf{v} \in \{0,1\}^M$ and hidden variables $\mathbf{h} \in \{0,1\}^N$, we have the following energy function E:

$$E(\mathbf{v},\mathbf{h};\theta) = -\sum_{i=1}^{M}\sum_{j=1}^{N} W_{ij} v_i h_j - \sum_{i=1}^{M} b_i v_i - \sum_{j=1}^{N} a_j h_j \tag{1}$$

where $\theta = \{\mathbf{a}, \mathbf{b}, \mathbf{W}\}$ are parameters, W_{ij} is the symmetric weight between visible unit i and hidden unit j, and b_i and a_j are bias terms of visible unit and hidden unit, respectively. With energy function, we can get the joint distribution over the visible and hidden units:

$$\begin{aligned} p(\mathbf{v},\mathbf{h};\theta) &= \frac{1}{\mathcal{Z}(\theta)}\exp(E(\mathbf{v},\mathbf{h};\theta)) \\ \mathcal{Z}(\theta) &= \sum_{\mathbf{v}}\sum_{\mathbf{h}}\exp(E(\mathbf{v},\mathbf{h};\theta)) \end{aligned} \tag{2}$$

where $\mathcal{Z}(\theta)$ is the normalization constant. Given a set of visible variables $\{\mathbf{v}_n\}_{n=1}^{N}$, the derivative of log-likelihood with respect to weight $\mathbf{W}$ can be calculated from Eq. (2):

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \log p(\mathbf{v}_n;\theta)}{\partial W_{ij}} = \mathbb{E}_{P_{data}}[v_i h_j] - \mathbb{E}_{P_{model}}[v_i h_j]$$

Various algorithms can be used to train a RBM, such as Contrastive Divergence (CD) algorithm [12]. In this paper, Bernoulli RBM is used. We treat the visual layer as the probabilities and we use CD algorithm to train RBMs.

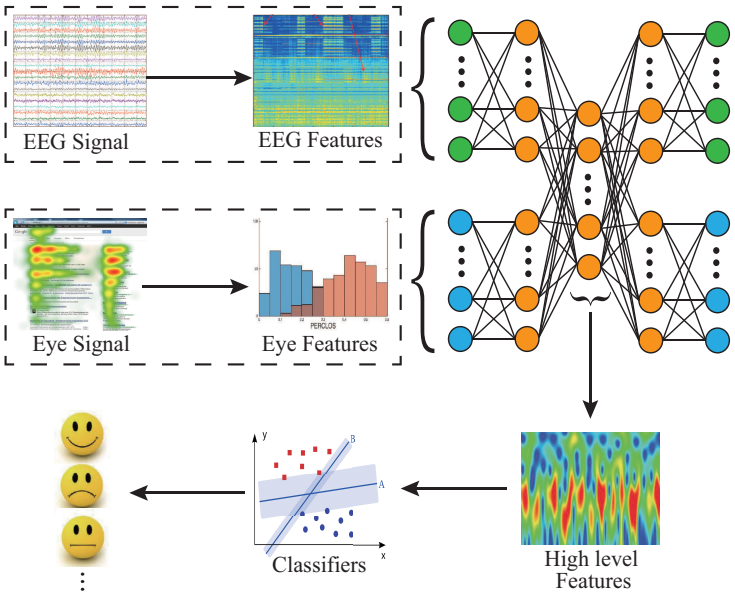

Fig. 1: The proposed multimodal emotion recognition framework. Here the BDAE network is used to generate high level features from low level features or original data and a linear SVM is trained with extracted high level features.

2.2 Model construction

The proposed multimodal emotion recognition framework using deep learning is depicted in Figure 1. There are three steps in total. The first step is to train the BDAE network. We call this step feature selection. The second step is supervised training, and we use the extracted high level features to train a linear SVM classifier. And the last step is a testing process, from which the recognition results are produced.

The BDAE training procedures, including encoding part and decoding part, are shown in Figure 2. In encoding part, we first train two RBMs for EEG features and eye movement features, respectively. As shown in Figure 2(a), EEG RBM is for EEG features and eye RBM is for eye movement features. Hidden layers are indicated by h_{EEG} and h_{Eye}, and W_1, W_2 are the corresponding weight matrices. After training these two RBMs, hidden layers, h_{EEG} and h_{Eye}, are concatenated together. The concatenated layer is used as the visual layer of an upper RBM, as depicted in Figure 2(b). Figure 2(c) shows the decoding part. When unfolding the stacked RBMs to reconstruct input features, we keep the weight matrices tied, and W_1, W_2, and W_3 and W_1^T, W_2^T, and W_3^T in Figure 2(c) are tied weights. At last, we used unsupervised back-propagation algorithm to fine-tune the weights and bias.

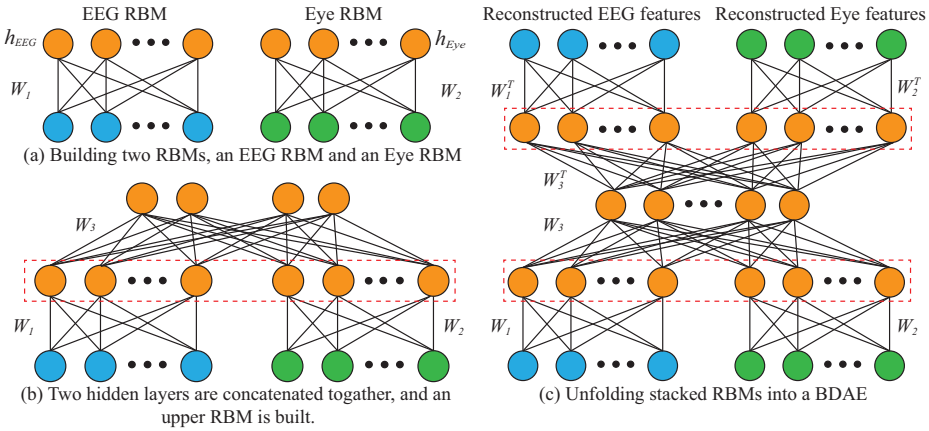

Fig. 2: The structure of Bimodal Deep AutoEncoder.

3 Experiment settings

3.1 The Datasets

The SEED dataset[4], which was first introduced in [5], contains EEG signals and eye movement signals of three different emotions (positive, negative, and neutral). These signals are collected from 15 subjects during watching emotional movie clips. There are 15 movie clips and each clip lasts about 4 minutes long. The EEG signals, recorded with ESI NeuroScan System, are of 62 channels at a sampling rate of 1000 Hz and the eye movement signals, collected with SMI ETG eye tracking glasses, contain information about blink, saccade fixation, and so on. In order to compare our proposed method with the existing approach [10], we use the same data as in [10], that is, 27 data files from 9 subjects. For every data file, the data from the subjects watching the first 9 movie clips are used as training samples and the rest ones are used as test samples.

The DEAP dataset was first introduced in [13]. The EEG signals and peripheral physiological signals of 32 participants were recorded when they were watching music videos. The dataset contains 32 channel EEG signals and 8 peripheral physiological signals. The emotional music videos include 40 one-minute long small clips and subjects were asked to do self-assessment by assigning values from 1 to 9 to five different status, namely, valence, arousal, dominance, liking, and familiarity. In order to compare the performance of our proposed method with previous results in [14] and [15], we did not take familiarity into consideration. We divided the trials into two different classes according to the assigned values. The threshold we chose is 5, and the tasks can be treated as four binary classification problems, namely, high or low valence, arousal, dominance and liking. Among all of the data, 90% samples were used as training data and the rest 10% samples were used as test data.

[4] http://bcmi.sjtu.edu.cn/~seed/index.html

3.2 Feature Extraction

For SEED dataset, both Power Spectral Density (PSD) and Differential Entropy (DE) features were extracted from EEG data. These two kinds of features contain five frequency bands: δ (1–4 Hz), θ (4–8 Hz), α (8–14 Hz), β (14–31 Hz), and γ (31–50 Hz). For every frequency band, the extracted features are of 62 dimensions and there are 310 dimensions for EEG features in total. As for eye movement data, we used the same features as in [10], and there are 41 dimensions in total including both PSD and DE features. The extracted EEG features and eye movement features were then rescaled to [0,1] and the rescaled features were used as the inputs of BDAE network.

For DEAP dataset, we used the downloaded preprocessed data directly as the inputs of BDAE network to generate shared representations of EEG signals and peripheral physiological signals. First, the EEG signals and peripheral physiological signals were separated and then the signals were segmented into 63 seconds. After segmentation, different channel data of the same time period (one second) are combined to form the input signals of BDAE network. And then, shared representation features were generated by the BDAE network.

3.3 Classification

The shared representation features generated by BDAE network are used to train a linear SVM classifier. Because of the variance between EEG signals collected from different people at different time, the model is data-specified, which means that we will build a model for each data and there are 27 models built for SEED dataset and 32 models for DEAP dataset. Network parameters, including hidden neuron numbers, epoch numbers, and learning rate, are chosen with grid searching.

4 Results

We compare our model with two other experimental settings. When only single modality is available, we classify different emotions with PSD and DE features by SVM. When multimodal information is available, features of different modalities are linked directly and different emotions are recognized with the concatenated features by SVM.

SEED results Figure 3 shows the summary of multimodal facilitation experiment results. As can be seen from Figure 3, the BDAE model has the best accuracy (91.01%) and the smallest standard deviation (8.91%).

Table 1 is the detailed experimental results of the BDAE model. The last column means that we linked all five frequency bands of EEG features and eye movement features directly. We examined the BDAE model three times and the recognition accuracies shown in Table 1 were averaged.

DEAP results In the literature, Rozgic *et al.* treated the EEG signals as a sequence of overlapping segments and a novel non-parametric nearest neighbor model was employed to extract response-level feature from these segments [14]. Li *et al.* used Deep

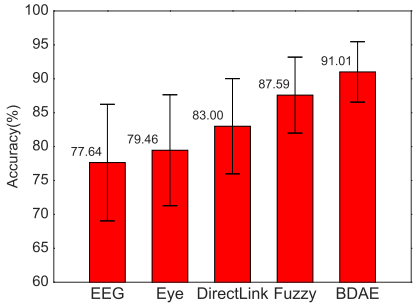

Fig. 3: Multimodal facilitation results on SEED dataset. Here the first two bars denote single modality, the rest bars denote multimodal with different fusion strategies and the fourth Fuzzy bar denotes the best result in [10].

Table 1: Accuracies of BDAE model on SEED dataset (%).

Feature		δ+eye	θ+eye	α+eye	β+eye	γ+eye	All
PSD	Ave.	**85.12**	83.89	83.18	83.23	82.92	85.10
	Std.	**11.09**	13.13	12.68	13.65	13.59	11.82
DE	Ave.	85.41	84.64	84.58	86.55	88.01	**91.01**
	Std.	14.03	11.03	12.78	10.48	10.25	**8.91**

Table 2: Comparison of six different approaches on DEAP dataset (Accuracy, %).

Method	Valence	Arousal	Dominance	Liking
EEG only	52.6	53.01	55.0	55.0
Others only	63.9	59.6	62.5	60.7
Linking	61.5	58.6	59.7	60.0
Rozgic *et al* [14]	76.9	69.1	73.9	75.3
Li *et al* [15]	58.4	64.3	65.8	66.9
Our Method	**85.2**	**80.5**	**84.9**	**82.4**

Belief Network (DBN) to automatically extract high-level features from raw EEG signals [15].

The experimental results on the DEAP dataset are shown in Table 2. Besides baselines mentioned above, we also compared the BDAE results with results in [15] and [14]. As can be seen from Table 2, the BDAE model improved recognition accuracies in all classification tasks.

5 Discussion

From the experimental results, we have demonstrated that the BDAE network can be used to extract shared representations from different modalities and the extracted features have better performance than other features.

Table 3: Confusing matrices of single modality and different modality merging methods

(a) EEG

	Positive	Neutral	Negative
Positive	**93.72%**	0.94%	5.34%
Neutral	5.56%	**81.35%**	13.09%
Negative	14.24%	29.49%	**56.27%**

(b) Eye

	Positive	Neutral	Negative
Positive	**81.92%**	7.41%	10.67%
Neutral	14.81%	**74.08%**	11.11%
Negative	9.38%	11.59%	**79.03%**

(c) Linking

	Positive	Neutral	Negative
Positive	**93.69%**	3.42%	2.89%
Neutral	7.06%	**77.62%**	15.32%
Negative	6.11%	16.72%	**77.17%**

(d) BDAE

	Positive	Neutral	Negative
Positive	**99.03%**	0.00%	0.97%
Neutral	3.70%	**90.26%**	6.04%
Negative	11.25%	3.57%	**85.18%**

From Table 3(a), we can see that EEG features are good for positive emotion recognition but are not good for negative emotions. As a complement, eye features have advantage in negative emotion recognition which can be seen from Table 3(b). When linking EEG and eye features directly, positive emotion accuracy is improved compare with situation where only eye features exist and negative emotion accuracy is also enhanced compared with when only EEG features are used. The BDAE framework achieves an even better result. The BDAE model has the highest accuracies in all three kinds of emotions, indicating that the BDAE model can fully use both EEG features and eye features.

6 Conclusions and future work

This paper has shown that the shared representations extracted from the BDAE model are good features to discriminate different emotions. Compared with other existing feature extraction strategies, the BDAE model is the best with accuracy of 91.01% on SEED dataset. For DEAP dataset, the BDAE network largely improves recognition accuracies on all four binary classification tasks. We analysed the confusing matrices of different methods and found that EEG features and eye features contain complementary information. The BDAE framework could fully take advantage of the complementary property between EEG and eye features to improve emotion recognition accuracies.

Our future work will focus on invesgating the complementarity between EEG features and eye movement features and explaining the mechanism of multimodal deep learning for emotion recognition from EEG and other physiological signals.

Acknowledgment

This work was supported in part by the grants from the National Natural Science Foundation of China (Grant No.61272248), the National Basic Research Program of China

(Grant No.2013CB329401) and the Major Basic Research Program of Shanghai Science and Technology Committee (15JC1400103).

References

1. Yisi Liu, Olga Sourina, and Minh Khoa Nguyen. Real-time EEG-based human emotion recognition and visualization. In *2010 International Conference on Cyberworlds*, pages 262–269. IEEE, 2010.
2. Mu Li and Bao-Liang Lu. Emotion classification based on gamma-band eeg. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. EMBC 2009.*, pages 1223–1226. IEEE, 2009.
3. Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering*, pages 81–84. IEEE, 2013.
4. Xiao-Wei Wang, Dan Nie, and Bao-Liang Lu. Emotional state classification from eeg data using machine learning approach. *Neurocomputing*, 129:94–106, 2014.
5. Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, 7(3):162–175, 2015.
6. Yang Yang, Han-Jia Ye, De-Chuan Zhan, and Yuan Jiang. Auxiliary information regularized machine for multiple modality feature learning. In *IJCAI'15*, pages 1033–1039. AAAI Press, 2015.
7. Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML'11*, pages 689–696, 2011.
8. Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *The Journal of Machine Learning Research*, 15(1):2949–2980, 2014.
9. Gyanendra K Verma and Uma Shanker Tiwary. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*, 102:162–172, 2014.
10. Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. Combining eye movements and EEG to enhance emotion recognition. In *IJCAI'15*, pages 1170–1176, 2015.
11. Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using multimodal deep learning. *arXiv preprint arXiv:1602.08225*, 2016.
12. Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.
13. Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.
14. Viktor Rozgic, Shiv N Vitaladevuni, and Ranga Prasad. Robust EEG emotion classification using segment level decision fusion. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1286–1290. IEEE, 2013.
15. Xiang Li, Peng Zhang, Dawei Song, Guangliang Yu, Yuexian Hou, and Bin Hu. EEG based emotion identification using unsupervised deep feature learning. In *SIGIR2015 Workshop on Neuro-Physiological Methods in IR Research*, August 2015.