Semi-supervised Bayesian Deep Multi-modal Emotion Recognition

Changde Du[1,4], Changying Du[2,3,⋆], Jinpeng Li[1,4], Wei-long Zheng[5], Bao-liang Lu [5], and Huiguang He[1,4,6,⋆]

[1]Research Center for Brain-Inspired Intelligence,
Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China
[2]Laboratory of Parallel Software and Computational Science,
Institute of Software, CAS, Beijing, China
[3]Qihoo 360 Search Lab
[4]University of Chinese Academy of Sciences, Beijing, China
[5]Center for Brain-like Computing and Machine Intelligence,
Shanghai Jiao Tong University, Shanghai, China
[6]Center for Excellence in Brain Science and Intelligence Technology, CAS, Beijing, China

Abstract. In emotion recognition, it is difficult to recognize human's emotional states using just a single modality. Besides, the annotation of physiological emotional data is particularly expensive. These two aspects make the building of effective emotion recognition model challenging. In this paper, we first build a multi-view deep generative model to simulate the generative process of multi-modality emotional data. By imposing a mixture of Gaussians assumption on the posterior approximation of the latent variables, our model can learn the shared deep representation from multiple modalities. To solve the labeled-data-scarcity problem, we further extend our multi-view model to semi-supervised learning scenario by casting the semi-supervised classification problem as a specialized missing data imputation task. Our semi-supervised multi-view deep generative framework can leverage both labeled and unlabeled data from multiple modalities, where the weight factor for each modality can be learned automatically. Compared with previous emotion recognition methods, our method is more robust and flexible. The experiments conducted on two real multi-modal emotion datasets have demonstrated the superiority of our framework over a number of competitors.

1 Introduction

With the development of human-computer interaction, emotion recognition has become increasingly important. Since human's emotion contains many nonverbal cues, various modalities ranging from facial expressions, voice, Electroencephalogram (EEG), eye movements to other physiological signals can be used as the indicators of emotional states [?]. In real-world applications, it is difficult to recognize the emotional states using just a single modality, because signals from different modalities represent different aspects of emotion and provide complementary information. Recent studies show that integrating multiple modalities can significantly boost the emotion recognition accuracy [?,?,?,?,?,?,?].

⋆ Corresponding author. E-mail: changying@iscas.ac.cn, huiguang.he@ia.ac.cn

The most successful approach to fuse the information from multiple modalities is based on deep multi-view representation learning [?,?,?,?,?]. For example, [?] proposed to learn a joint density model for emotion analysis with a multi-modal Deep Boltzmann Machine (DBM) [?]. This multi-modal DBM is exploited to model the joint distribution over visual, auditory, and textual features. [?] proposed a multi-modal emotion recognition method by using multi-modal Deep Autoencoders (DAE) [?], in which the joint representations of EEG and eye movement signals were extracted. Nevertheless, there are still limitations with these deep multi-modal emotion recognition methods, e.g., their performances depend on the amount of labeled data.

By using the modern sensor equipments, we can easily collect massive physiological signals, which are closely related to people's emotional states. Despite the convenience of data acquisition, the data labeling procedure requires lots of manual efforts. Therefore, in most cases only a small set of labeled samples is available, while the majority of whole dataset is left unlabeled. Traditional emotion recognition approaches only utilized the limited amount of labeled data, which may result in severe overfitting. The most attractive way to deal with this issue is based on Semi-supervised Learning (SSL), which builds more robust model by exploiting both labeled and unlabeled data [?,?,?].

Amongst existing SSL approaches, the most competitive one is based on deep generative models, which employs the Deep Neural Networks (DNNs) to learn discriminative features and casts the semi-supervised classification problem as a specialized missing data imputation task. [?] and [?] have shown that deep generative models and approximate Bayesian inference exploiting recent advances in scalable variational methods [?,?] can provide state-of-the-art performance for semi-supervised classification. Though the Variational Autoencoder (VAE) framework [?] has shown great advantages in SSL, its potential merits remain under-explored. For example, until recently, there was no successful multi-view extension for it. The main difficulty lies in its inherent assumption that the posterior approximation should be conditioned on the data point, which is natural to single-view data but becomes problematic for multi-view case.

In this paper, we propose a novel semi-supervised multi-view deep generative framework for multi-modal emotion recognition. Our framework combines the advantages of deep multi-view representation learning and Bayesian modeling, thus it has sufficient flexibility and robustness in learning joint features and classifier. Our main contributions can be summarized as follows.

- We propose a multi-view extension for VAE by imposing a mixture of Gaussians assumption on the posterior approximation of the latent variables. For multi-view learning, this is critical for fully exploiting the information from multiple views.
- We introduce a semi-supervised multi-modal emotion recognition framework based on multi-view VAE. Our framework can leverage both labeled and unlabeled samples from multiple modalities and the weight factor for each modality can be learned automatically, which is critical for building a robust emotion recognition system.
- We demonstrate the superiority of our framework and provide insightful observations on two real multi-modal emotion datasets.

2 Related Work

The key in multi-modal emotion recognition is the fusion of information from multiple modalities. There are mainly two types of information fusion studied by researchers: feature-level fusion and decision-level fusion. Feature-level fusion [**?**,**?**,**?**,**?**] fuses the features extracted from various modalities such as visual features, audio features, text features, etc., as a general feature vector and the combined features are sent for analysis. In decision-level fusion [**?**,**?**,**?**,**?**], the features of each modality are examined and classified independently and the results are fused to obtain the final decision. Most of the existing feature-level fusion based emotion recognition models treat each modality equally thus have little flexibility. As shown in the next section, our model can learn the weight factor for each modality automatically, which is more flexible in feature fusion.

Recently, many emotion recognition researches [**?**,**?**,**?**,**?**,**?**,**?**] have been conducted by using various physiological signals such as EEG and eye movements. For example, [**?**] used the features from peripheral physiological signals to represent neutral, fear, and sadness responses to movie excerpts. [**?**] collected EEG and eye movement signals from 15 participants and classified their response to emotional videos into three affective states: positive, neutral, and negative. In this paper, we also focus on using physiological signals to conduct multi-modal emotion recognition.

Though multi-modal approaches have been widely implemented for emotion recognition [**?**,**?**,**?**,**?**,**?**,**?**], very few of them explored SSL simultaneously. To the best of our knowledge, only [**?**] proposed an enhanced multi-modal co-training algorithm for semi-supervised emotion recognition, but its shallow structure is hard to capture the high-level correlation between different modalities. On the other hand, the auto-encoding variational Bayesian learning [**?**] has shown great advantages in semi-supervised classification tasks [**?**,**?**,**?**], but their single-view design can't effectively deal with multi-view data.

3 Multi-view Variational Autoencoder for Semi-supervised Emotion Recognition

The VAE framework has recently been introduced as a robust model for latent feature learning [**?**,**?**]. However, the single-view architecture in VAE can't effectively deal with multi-view data. In this section, we first build a multi-view VAE, which can learn the shared deep representation from multi-view data. And then, we extend it to the semi-supervised scenario. Assume we are faced with multi-view data that appears as pairs $(\mathcal{X},\ y) = (\{\mathbf{x}^{(v)}\}_{v=1}^{V},\ y)$, with observation $\mathbf{x}^{(v)}$ from the v-th view and the corresponding class label y.

3.1 Multi-view Variational Autoencoder

DNN-parameterized Likelihoods We assume the latent variable $\mathbf{z}$ can generate multi-view features $\{\mathbf{x}^{(v)}\}_{v=1}^{V}$. Specifically, we assume $\mathbf{z}$ generates $\mathbf{x}^{(v)}$ for any $v \in \{1, ..., V\}$, with the following generative model (cf. Fig. 1a):

$$p_{\theta^{(v)}}(\mathbf{x}^{(v)}|\mathbf{z}) = f(\mathbf{x}^{(v)}; \mathbf{z}, \theta^{(v)}), \tag{1}$$

where $f(\mathbf{x}^{(v)}; \mathbf{z}, \theta^{(v)})$ is a suitable likelihood function (e.g. a Gaussian for continuous observation or Bernoulli for binary observation), which is formed by a non-linear transformation of the latent variable $\mathbf{z}$. This non-linear transformation is essential to allow for higher moments of the data to be captured by the density model, and we choose these non-linear functions to be DNNs, referred to as the generative networks, with parameters $\{\theta^{(v)}\}_{v=1}^{V}$. Note that, the likelihoods for different data views are assumed to be independent of each other, with different nonlinear transformations.

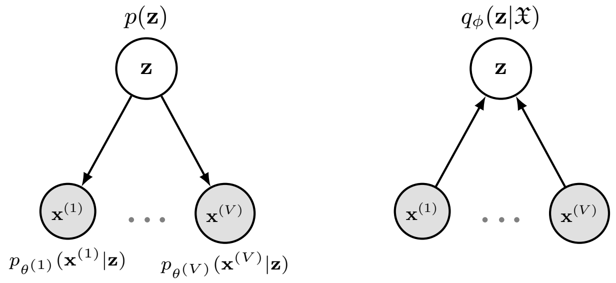
The Bayesian Canonical Correlation Analysis (CCA) model [**?**] can be seen as a special case of our model, where linear shallow transformations were used to generate each data view and only two different views were considered. [**?**] used a similar deep nonlinear generative process as ours to construct deep Bayesian CCA model, but during inference they construct the variational posterior approximation from just one view and ignore the rest one. Such a choice is convenient for inference and computation, but only seeks suboptimal solutions as it doesn't fully exploit the data. As shown in the following, we assume the variational approximation to the posterior of latent variables to be a mixture of Gaussians, utilizing information from multiple views.

Gaussian Prior and Mixture of Gaussians Posterior Typically, both the prior $p(\mathbf{z})$ and the approximate posterior $q_\phi(\mathbf{z}|\mathcal{X})$ are assumed to be Gaussian distributions [**?**,**?**] in order to maintain mathematical and computational tractability. Although this assumption has leaded to favorable results on several tasks, it is clearly a restrictive and often unrealistic assumption. Specifically, the choice of a Gaussian distribution for $p(\mathbf{z})$ and $q_\phi(\mathbf{z}|\mathcal{X})$ imposes a strong uni-modal structure assumption on the latent space. However, for data distributions that are strongly multi-modal, the uni-modal Gaussian assumption inhibits the model's ability to extract and represent important structure in the data. To improve the flexibility of the model, one way is to impose a mixture of Gaussians assumption on $p(\mathbf{z})$. However, it has the risk of creating separate "islands" of discontinuous manifolds that may break the meaningfulness of the representation in the latent space.

To learn more powerful and expressive models – in particular, models with multi-modal latent variable structures for multi-modal emotion recognition applications – we seek a mixture of Gaussians for $q_\phi(\mathbf{z}|\mathcal{X})$, while preserving $p(\mathbf{z})$ as a standard Gaussian. Thus (cf. Fig. 1b),

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}\left(\mathbf{z}|\mathbf{0}, \mathbf{I}\right), \\ q_\phi(\mathbf{z}|\mathcal{X}) &= \sum_{v=1}^{V} \lambda^{(v)} \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)}),\ \boldsymbol{\Sigma}_{\phi^{(v)}}(\mathbf{x}^{(v)})\right), \end{aligned} \tag{2}$$

where the mean $\boldsymbol{\mu}_{\phi^{(v)}}$ and the covariance $\boldsymbol{\Sigma}_{\phi^{(v)}}$ are nonlinear functions of the observation $\mathbf{x}^{(v)}$, with variational parameter $\boldsymbol{\phi}^{(v)}$. As in our generative model, we choose these nonlinear functions to be DNNs, referred to as the inference networks. $\lambda^{(v)}$ is the non-negative normalized weight factor for the v-th view, i.e., $\lambda^{(v)} > 0$ and $\sum_{v=1}^{V} \lambda^{(v)} = 1$. In [**?**], Gershman et al. simply assumed the variational distribution to be a uniformly weighted Gaussian mixture, which treats each component equally and loses flexibility.


$p(\mathbf{z})$
$\mathbf{z}$
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(V)}$
$p_{\theta^{(1)}}(\mathbf{x}^{(1)}|\mathbf{z})$
$p_{\theta^{(V)}}(\mathbf{x}^{(V)}|\mathbf{z})$
$q_\phi(\mathbf{z}|\mathcal{X})$
$\mathbf{z}$
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(V)}$

(a) Generative model P (b) Inference model Q

Fig. 1. Graphical model of the multi-view VAE, where $\mathcal{X} = \{\mathbf{x}^{(v)}\}_{v=1}^{V}$.

Instead of treating each view equally, our non-uniformly weighted Gaussian mixture assumption can weight each view automatically, which is useful to identify the importance of each view. By conditioning the posterior approximation on the data point, we avoid variational parameters per data point, instead only requiring to fit global variational parameters. Note that, our mixed Gaussian assumption on the variational approximation distinguishes our method from all existing ones using the auto-encoding variational framework [**?**,**?**,**?**,**?**,**?**,**?**]. For multi-view learning, this is critical for fully exploiting the information from multiple views.

3.2 Semi-supervised Emotion Recognition

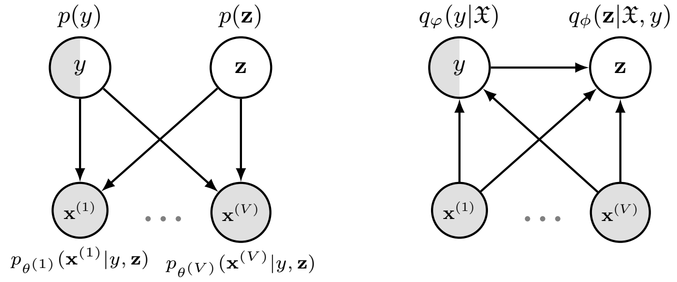
In semi-supervised classification, only a subset of the samples have corresponding class labels, and we focus on using the multi-view VAE to build a model (semiMVAE) that learns classifier from both labeled and unlabeled multi-view data. Since the emotional data is continuous, we choose the Gaussian likelihoods. Then the generative model P is defined as $p(y)p(\mathbf{z})\prod_{v=1}^{V} p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z})$ (cf. Fig. 2a):

$$\begin{aligned} p(y) &= \mathrm{Cat}\left(y|\boldsymbol{\pi}\right), \\ p(\mathbf{z}) &= \mathcal{N}\left(\mathbf{z}|\mathbf{0},\mathbf{I}\right), \\ p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z}) &= \mathcal{N}\left(\boldsymbol{\mu}_{\theta^{(v)}}(y,\mathbf{z}),\ \mathrm{diag}(\boldsymbol{\sigma}^2_{\theta^{(v)}}(y,\mathbf{z}))\right), \end{aligned} \tag{3}$$

where $\mathrm{Cat}(\cdot)$ denotes the categorical distribution, y is treated as a latent variable for the unlabeled data points, and the mean $\boldsymbol{\mu}_{\theta^{(v)}}$ and variance $\boldsymbol{\sigma}^2_{\theta^{(v)}}$ are nonlinear functions of y and $\mathbf{z}$, with parameter $\theta^{(v)}$. The inference model Q is defined as $q_\varphi(y|\mathcal{X})q_\phi(\mathbf{z}|\mathcal{X},y)$ (cf. Fig. 2b):

$$\begin{aligned} q_\varphi(y|\mathcal{X}) &= \mathrm{Cat}\left(y|\boldsymbol{\pi}_\varphi(\mathcal{X})\right), \\ q_\phi(\mathbf{z}|\mathcal{X},y) &= \sum_{v=1}^{V} \lambda^{(v)} \mathcal{N}\left(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)},y),\ \boldsymbol{\Sigma}_{\phi^{(v)}}(\mathbf{x}^{(v)},y)\right), \end{aligned} \tag{4}$$

where $q_\phi(\mathbf{z}|\mathcal{X},y)$ is assumed to be a mixture of Gaussians to combine the information from multiple data views. Intuitively, $q_\phi(\mathbf{z}|\mathcal{X},y)$, $p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z})$ and $q_\varphi(y|\mathcal{X})$ correspond to the encoder, the decoder and the classifier, respectively.


$p(y)$
$p(\mathbf{z})$
y
$\mathbf{z}$
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(V)}$
$p_{\theta^{(1)}}(\mathbf{x}^{(1)}|y,\mathbf{z})$
$p_{\theta^{(V)}}(\mathbf{x}^{(V)}|y,\mathbf{z})$
$q_{\varphi}(y|\mathcal{X})$
$q_{\phi}(\mathbf{z}|\mathcal{X},y)$
y
$\mathbf{z}$
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(V)}$

(a) Generative model P (b) Inference model Q

Fig. 2. Graphical model of the semiMVAE for semi-supervised multi-view learning, where $\mathcal{X} = \{\mathbf{x}^{(v)}\}_{v=1}^{V}$.

For brevity, we omit the explicit dependencies on $\mathbf{x}^{(v)}$, y and $\mathbf{z}$ for the moment variables mentioned above hereafter. In principle, $\boldsymbol{\mu}_{\theta^{(v)}}$, $\boldsymbol{\sigma}^2_{\theta^{(v)}}$, $\boldsymbol{\pi}_{\varphi}$, $\boldsymbol{\mu}_{\phi^{(v)}}$ and $\boldsymbol{\Sigma}_{\phi^{(v)}}$ can be implemented by various DNN models, e.g., Multiple Layer Perceptrons (MLP) and Convolutional Neural Networks (CNN).

3.3 Variational Lower Bound

The variational lower bound on the marginal likelihood for a single labeled data point is

$$\begin{aligned}
\log p_\theta(\mathcal{X}, y) &= \log \int_{\mathbf{z}} p_\theta(\mathcal{X}, y, \mathbf{z})\, d\mathbf{z} \\
&\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X},y)}\left[\log \frac{p_\theta(\mathcal{X}, y, \mathbf{z})}{q_\phi(\mathbf{z}|\mathcal{X}, y)}\right] \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X},y)}\left[\sum_{v=1}^{V} \log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z}) + \log p(y) + \log p(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathcal{X}, y)\right] \\
&\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X},y)}\left[\sum_{v=1}^{V} \log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z}) + \log p(y) + \log p(\mathbf{z})\right] \\
&\quad - \sum_{v=1}^{V} \lambda^{(v)} \cdot \log\left(\sum_{l=1}^{V} \lambda^{(l)} \cdot \omega_{v,l}\right) \\
&\equiv -\mathcal{L}(\mathcal{X}, y),
\end{aligned} \tag{5}$$

where $\omega_{v,l} = \mathcal{N}\big(\boldsymbol{\mu}_{\phi^{(v)}}|\boldsymbol{\mu}_{\phi^{(l)}},\ \boldsymbol{\Sigma}_{\phi^{(v)}} + \boldsymbol{\Sigma}_{\phi^{(l)}}\big)$. It should be noted that, the Shannon entropy $\mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X},y)}[-\log q_\phi(\mathbf{z}|\mathcal{X}, y)]$ is hard to compute analytically, and we have used the Jensen's inequality to derive a lower bound of it:

$$
\begin{aligned}
&\mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X},y)}[-\log q_\phi(\mathbf{z}|\mathcal{X},y)] \\
&= -\int q_\phi(\mathbf{z}|\mathcal{X},y)\log q_\phi(\mathbf{z}|\mathcal{X},y)\; d\mathbf{z} \\
&= -\sum_{v=1}^{V}\lambda^{(v)}\cdot\int \mathcal{N}\big(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(v)}},\ \boldsymbol{\Sigma}_{\phi^{(v)}}\big)\log\sum_{l=1}^{V}\lambda^{(l)}\cdot\mathcal{N}\big(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(l)}},\ \boldsymbol{\Sigma}_{\phi^{(l)}}\big)\; d\mathbf{z} \\
&\geq -\sum_{v=1}^{V}\lambda^{(v)}\cdot\log\sum_{l=1}^{V}\lambda^{(l)}\cdot\int \mathcal{N}\big(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(v)}},\ \boldsymbol{\Sigma}_{\phi^{(v)}}\big)\cdot\mathcal{N}\big(\mathbf{z}|\boldsymbol{\mu}_{\phi^{(l)}},\ \boldsymbol{\Sigma}_{\phi^{(l)}}\big)\; d\mathbf{z} \\
&= -\sum_{v=1}^{V}\lambda^{(v)}\cdot\log\sum_{l=1}^{V}\lambda^{(l)}\cdot\mathcal{N}\big(\boldsymbol{\mu}_{\phi^{(v)}}|\boldsymbol{\mu}_{\phi^{(l)}},\ \boldsymbol{\Sigma}_{\phi^{(v)}}+\boldsymbol{\Sigma}_{\phi^{(l)}}\big) \\
&= -\sum_{v=1}^{V}\lambda^{(v)}\cdot\log\bigg(\sum_{l=1}^{V}\lambda^{(l)}\cdot\omega_{v,l}\bigg),
\end{aligned}
$$

where we have used the fact that the convolution of two Gaussians is another Gaussian.

For unlabeled data, we further introduce the variational distribution $q_\varphi(y|\mathcal{X})$ for y:

$$
\begin{aligned}
\log p_\theta(\mathcal{X}) &= \log\int_{\mathbf{z}}\int_{y} p_\theta(\mathcal{X},y,\mathbf{z})\; dy\; d\mathbf{z} \\
&\geq \mathbb{E}_{q_{\varphi,\phi}(y,\mathbf{z}|\mathcal{X})}\left[\log\frac{p_\theta(\mathcal{X},y,\mathbf{z})}{q_{\varphi,\phi}(y,\mathbf{z}|\mathcal{X})}\right] \\
&= \mathbb{E}_{q_\varphi(y|\mathcal{X})}\big[-\mathcal{L}(\mathcal{X},y)-\log q_\varphi(y|\mathcal{X})\big] \\
&\equiv -\mathcal{U}(\mathcal{X}),
\end{aligned}
\tag{6}
$$

with $q_{\varphi,\phi}(y,\mathbf{z}|\mathcal{X}) = q_\varphi(y|\mathcal{X})q_\phi(\mathbf{z}|\mathcal{X},y)$. The objective function for the entire dataset is now:

$$
\mathcal{J} = \sum_{(\mathcal{X},y)\in S_l}\mathcal{L}(\mathcal{X},y) + \sum_{\mathcal{X}\in S_u}\mathcal{U}(\mathcal{X})\,,
\tag{7}
$$

where S_l and S_u are labeled and unlabeled dataset, respectively. The classification accuracy can be improved by introducing an explicit classification loss for labeled data. The extended objective function is:

$$
\mathcal{F} = \mathcal{J} + \alpha\cdot\sum_{(\mathcal{X},y)\in S_l}\left[-\log q_\varphi(y|\mathcal{X})\right],
\tag{8}
$$

where the hyper-parameter α is a weight between generative and discriminative learning. We set $\alpha = \beta\cdot(N_l + N_u)$, where β is a scaling constant, and N_l and N_u are the numbers of labeled and unlabeled data points in one minibatch, respectively. Note that, the classifier $q_\varphi(y|\mathcal{X})$ is also used at test phase for the prediction of unseen data.

3.4 Optimization

Eq. (8) provides a unified objective function for optimizing the parameters of encoder, decoder and classifier networks. This optimization can be done jointly, without resort to the variational EM algorithm, using the stochastic backpropagation technique [**?**,**?**].

Reparameterization Trick The reparameterization trick is a vital component of the algorithm, because it allows us to easily take the derivative of $\mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X},y)}[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z})]$ with respect to the variational parameters ϕ. However, the use of a mixture of Gaussians for the variational distribution $q_\phi(\mathbf{z}|\mathcal{X},y)$ makes the application of reparameterization trick challenging. It can be shown that, for any $v \in \{1,...,V\}$, $\mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X},y)}[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z})]$ can be rewritten, using the location-scale transformation for the Gaussian distribution, as:

$$\begin{aligned}&\mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X},y)}[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z})]\\&=\sum_{l=1}^{V}\lambda^{(l)}\mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon}^{(l)}|\mathbf{0},\mathbf{I})}\big[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\boldsymbol{\mu}_{\phi^{(l)}}+\mathbf{R}_{\phi^{(l)}}\boldsymbol{\epsilon}^{(l)})\big],\end{aligned}\tag{9}$$

where $\mathbf{R}_{\phi^{(l)}}\mathbf{R}_{\phi^{(l)}}^\top=\boldsymbol{\Sigma}_{\phi^{(l)}}$ and $l\in\{1,...,V\}$.

Gradients of the Objective While the expectations on the right hand side of Eq. (9) still cannot be solved analytically, their gradients w.r.t. $\theta^{(v)}$, $\phi^{(l)}$ and $\lambda^{(l)}$ can be efficiently estimated using the following Monte-Carlo estimators,

$$\begin{aligned}&\frac{\partial}{\partial\theta^{(v)}}\mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X},y)}[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z})]\\&=\sum_{l=1}^{V}\lambda^{(l)}\mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon}^{(l)}|\mathbf{0},\mathbf{I})}\left[\frac{\partial}{\partial\theta^{(v)}}\log p_{\theta^{(v)}}\big(\mathbf{x}^{(v)}|y,\mathbf{z}^{(l)}\big)\right]\\&\approx\frac{\lambda^{(l)}}{T}\sum_{t=1}^{T}\sum_{l=1}^{V}\frac{\partial}{\partial\theta^{(v)}}\log p_{\theta^{(v)}}\big(\mathbf{x}^{(v)}|y,\mathbf{z}^{(l,t)}\big),\end{aligned}\tag{10}$$

$$\begin{aligned}&\frac{\partial}{\partial\phi^{(l)}}\mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X},y)}[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z})]\\&=\lambda^{(l)}\frac{\partial}{\partial\phi^{(l)}}\mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon}^{(l)}|\mathbf{0},\mathbf{I})}\Big[\frac{\partial}{\partial\mathbf{z}^{(l)}}\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z}^{(l)})\cdot\Big(\frac{\partial\boldsymbol{\mu}_{\phi^{(l)}}}{\partial\phi^{(l)}}+\frac{\partial\mathbf{R}_{\phi^{(l)}}}{\partial\phi^{(l)}}\boldsymbol{\epsilon}^{(l)}\Big)\Big]\\&\approx\frac{\lambda^{(l)}}{T}\sum_{t=1}^{T}\frac{\partial}{\partial\mathbf{z}^{(l,t)}}\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z}^{(l,t)})\Big(\frac{\partial\boldsymbol{\mu}_{\phi^{(l)}}}{\partial\boldsymbol{\phi}^{(l)}}+\frac{\partial\mathbf{R}_{\phi^{(l)}}}{\partial\phi^{(l)}}\boldsymbol{\epsilon}^{(l,t)}\Big),\end{aligned}\tag{11}$$

$$\begin{aligned}&\frac{\partial}{\partial\lambda^{(l)}}\mathbb{E}_{q_\phi(\mathbf{z}|\mathcal{X},y)}[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z})]\\&=\mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon}^{(l)}|\mathbf{0},\mathbf{I})}\big[\log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|y,\mathbf{z}^{(l)})\big]\\&\approx\frac{1}{T}\sum_{t=1}^{T}\log p_{\theta^{(v)}}\big(\mathbf{x}^{(v)}|y,\mathbf{z}^{(l,t)}\big),\end{aligned}\tag{12}$$

where $\mathbf{z}^{(l)}$ is evaluated at $\mathbf{z}^{(l)} = \boldsymbol{\mu}_{\phi^{(l)}} + \mathbf{R}_{\phi^{(l)}}\boldsymbol{\epsilon}^{(l)}$ and $\mathbf{z}^{(l,t)} = \boldsymbol{\mu}_{\phi^{(l)}} + \mathbf{R}_{\phi^{(l)}}\boldsymbol{\epsilon}^{(l,t)}$ with $\boldsymbol{\epsilon}^{(l,t)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In practice, it suffices to use a small T (e.g. $T = 1$) and then estimate the gradient using minibatches of data points. Though the above Monte-Carlo estimators could have large variances if a small T is used, the experimental results show that it suffices to obtain good performance[1]. The same observation can be found in previous works [**?**,**?**]. Furthermore, we use the same random numbers $\boldsymbol{\epsilon}^{(l,t)}$ for all estimators to have lower variances. The gradient w.r.t. φ is omitted here, since it can be derived straightforwardly by using traditional reparameterization trick [**?**].

The gradients of the objective function for semiMVAE (Eq. (8)) can then be computed by a direct application of the chain rule and estimators presented above. During optimization we can use the estimated gradients in conjunction with standard stochastic gradient based optimization methods such as SGD, RMSprop or Adam [**?**]. Overall, the model can be trained with reparameterization trick for backpropagation through the mixed Gaussian latent variables. We summarize the training procedure of semiMVAE in Algorithm 1.

Algorithm 1 Semi-supervised multi-view VAE (semiMVAE)

```
Input:
  – the labeled dataset $S_l$ and the unlabeled dataset $S_u$
  – the types and structures of the networks in multiple views
  – the scaling constant $\beta$
  – the learning rate $\eta$
  – the size of minibatch
  – the maximal number of epochs $MaxEpoch$
Output: the encoder parameters $\{\phi^{(v)}\}_{v=1}^{V}$, the decoder parameters $\{\theta^{(v)}\}_{v=1}^{V}$, the classifier parameters $\varphi$ and the weight factors $\{\lambda^{(v)}\}_{v=1}^{V}$
1. Initialize the model parameters $\{\phi^{(v)}\}_{v=1}^{V}$, $\{\theta^{(v)}\}_{v=1}^{V}$, $\varphi$ and $\{\lambda^{(v)}\}_{v=1}^{V}$
2. for $epoch = 1$ to $MaxEpoch$ do
3.     Get random minibatch from dataset
4.     Calculate the gradients $\frac{\partial \mathcal{F}}{\partial \phi^{(v)}}$, $\frac{\partial \mathcal{F}}{\partial \theta^{(v)}}$, $\frac{\partial \mathcal{F}}{\partial \varphi}$ and $\frac{\partial \mathcal{F}}{\partial \lambda^{(v)}}$ for any $v \in \{1, ..., V\}$
5.     Update model parameters according to gradients for any $v \in \{1, ..., V\}$:
           $\phi^{(v)} = \phi^{(v)} - \eta \cdot \frac{\partial \mathcal{F}}{\partial \phi^{(v)}}$    $\theta^{(v)} = \theta^{(v)} - \eta \cdot \frac{\partial \mathcal{F}}{\partial \theta^{(v)}}$
           $\varphi = \varphi - \eta \cdot \frac{\partial \mathcal{F}}{\partial \varphi}$    $\lambda^{(v)} = \lambda^{(v)} - \eta \cdot \frac{\partial \mathcal{F}}{\partial \lambda^{(v)}}$
6. end for
7. Output $\{\phi^{(v)}\}_{v=1}^{V}$, $\{\theta^{(v)}\}_{v=1}^{V}$, $\varphi$ and $\{\lambda^{(v)}\}_{v=1}^{V}$.
```

[1] In the experiments, we have also tried larger values of T ($T = 5$ and $T = 10$), but we didn't observe obvious performance improvement. For efficiency, we set $T = 1$ for all experiments.

4 Experiments

In this section, we present extensive experimental results to demonstrate the effectiveness of the proposed semi-supervised multi-view framework for emotion recognition.

4.1 Experimental Testbed and Setup

Data description There are lots of multi-modal emotion recognition datasets, such as IEMOCAP[**?**], YouTube[**?**], DEAP [**?**], DECAF[**?**], SEED [**?**], etc., which can be used to evaluate the performance of various emotion recognition models (see a latest review in [**?**]). In our experiments, we focused on the EEG-based multi-modal emotion recognition, thus the SEED[2] and DEAP[3] datasets were used to demonstrate the effectiveness of our framework.

The SEED dataset contains EEG and eye movement signals from 15 subjects during watching 15 movie clips, where each movie clip lasts about 4 minutes long. The EEG signals were recorded from 62 channels and the eye movement signals contained information about blink, saccade fixation and so on. The EEG signals was collected from all 15 subjects, while only 9 subjects' eye movement signals have been successfully recorded. Because we focused on multi-modal emotion recognition, we used the EEG and eye movement data from 9 subjects across 3 sessions, totally 27 data files. For each data file, data from watching the 1-9 movie clips were used as training set, while data from watching the 10-12 movie clips were used as validation set and the rest (13-15) were used as testing set.

The DEAP dataset contains EEG and peripheral physiological signals of 32 participants. Signals were recorded when they were watching 40 one-minute duration music videos. The EEG signals were recorded from 32 channels, whereas the peripheral physiological signals were recorded from 8 channels. The participants, using values from 1 to 9, rated each music video in terms of the levels of valence, arousal and so on. In our experiment, the valence-arousal space was divided into four quadrants according to the ratings. The threshold we used was 5, leading to four classes of data. Considering the fuzzy boundary of emotions and the variations of participants' ratings possibly associated with individual difference in rating scale, we discarded the samples whose ratings of arousal and valence are between 3 and 6. The dataset was randomly divided into 10-folds, where 8 folds for training, one fold for validation and the last fold for testing. The size of testing set is relative small, because some graph-based semi-supervised baselines are hard to deal with large dataset.

Feature selection For SEED dataset, [**?**] have extracted the Differential Entropy (DE) features and 33 eye movement features from EEG and eye movement signals, respectively. We also used these features in our experiments. For DEAP dataset, we extracted the DE features from EEG and peripheral physiological signals. The DE features can be calculated in four frequency bands: theta (4-8Hz), alpha (8-14Hz), beta (14-31Hz), and gamma (31-45Hz), and we used all band's features. The details of the data used in our experiments were summarized in Table 1.

[2] http://bcmi.sjtu.edu.cn/%7Eseed/index.html

[3] http://www.eecs.qmul.ac.uk/mmv/datasets/deap/download.html

Table 1. The details of the datasets used in our experiments.

Datasets	#Instances	#Features	#Training	#Validation	#Testing	#Classes
SEED	22734	310(EEG), 33(Eye)	13473	4725	4536	3
DEAP	21042	128(EEG), 32(Phy.)	16834	2104	2104	4

Compared methods We compared our semiMVAE with a broad range of solutions, including supervised learning, transductive and inductive semi-supervised learning. We briefly summarize the various baselines in the following.

- **MAE**: the multi-view extension of deep autoencoders, which can be used to extract the joint representations from multiple modalities [**?**].
- **DCCA**: the full deep neural network extension of Canonical Correlation Analysis (CCA). DCCA can learn deep nonlinear mappings of two views, which are maximally correlated [**?**].
- **DCCAE**: a deep multi-view representation learning model which combines the advantages of the DCCA and deep autoencoders. In particular, DCCAE consists of two autoencoders and optimizes the combination of canonical correlation between the learned bottleneck representations and the reconstruction errors of the autoencoders [**?**].
- **AMMSS**: a graph-based multi-view semi-supervised classification algorithm, which can integrate heterogeneous features from both labeled and unlabeled data [**?**].
- **AMGL**: a latest auto-weighted multiple graph learning framework, which can be applied to multi-view semi-supervised classification task [**?**].
- **semiVAE**: a single-view semi-supervised deep generative model proposed in [**?**]. We evaluate semiVAE's performance for each modality and the concatenation of all modalities, respectively.

For MAE, DCCA and DCCAE, we used the Support Vector Machines[4] (SVM) and transductive SVM[5] (TSVM) for supervised learning and transductive semi-supervised learning, respectively.

Parameter setting For semiMVAE, we considered multiple layer perceptrons as the type of inference and generative networks. On both datasets, we set the structures of the inference and generative networks for each view as '100-50-30' and '30-50-100', respectively. We used the Adam optimizer [**?**] with a learning rate $\eta = 3 \times 10^{-4}$ in training. The scaling constant β was selected from $\{0.1, 0.5, 1\}$ throughout the experiments. The weight factor for each view was initialized with $\lambda^{(v)} = 1/V$, where V is the number of views. For MAE, DCCA and DCCAE, we considered the same setups (network structure, learning rate, etc.) as our semiMVAE. For AMMSS, we tuned the parameters as suggested in [**?**]. For AMGL and semiVAE, we used their default settings.

4.2 Performance Evaluation

To simulate semi-supervised learning scenario, on both datasets, we randomly labeled different proportions of samples in the training set, and remained the rest samples in the

[4] http://www.csie.ntu.edu.tw/%7Ecjlin/liblinear/.

[5] http://svmlight.joachims.org/.

training set unlabeled. For transductive semi-supervised learning, we trained models on the dataset consisting of the testing data and labeled data belonging to training set. For inductive semi-supervised learning, we trained models on the entire training set consisting of the labeled and unlabeled data. For supervised learning, we trained models on the labeled data belonging to training set, and test their performance on the testing set. Table 2 presents the classification accuracies of all methods on SEED and DEAP datasets. The proportions of labeled samples in the training set vary from 1% to 3%. Several observations can be drawn as follows.

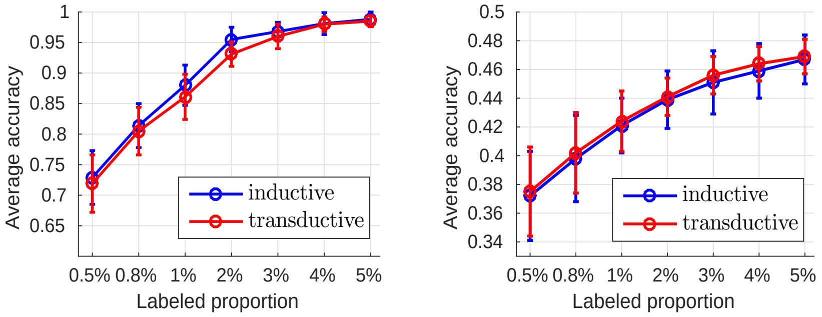
Table 2. Comparison with several supervised and semi-supervised methods on SEED and DEAP datasets with few labels. Results (mean±std) were averaged over 20 independent runs.

SEED data	Algorithms	1% labeled	2% labeled	3% labeled
Supervised learning	MAE+SVM	.814±.031	.896±.024	.925±.024
	DCCA+SVM	.809±.035	.891±.035	.923±.028
	DCCAE+SVM	.819±.036	.893±.034	.923±.027
Transductive semi-supervised learning	AMMSS	.731±.055	.839±.036	.912±.018
	AMGL	.711±.047	.817±.023	.886±.028
	MAE+TSVM	.818±.035	.910±.025	.931±.026
	DCCA+TSVM	.811±.031	.903±.024	.928±.021
	DCCAE+TSVM	.823±.040	.907±.027	.929±.023
	semiMVAE	**.861**±.037	**.931**±.020	**.960**±.021
Inductive semi-supervised learning	semiVAE (Eye)	.753±.024	.849±.055	.899±.049
	semiVAE (EEG)	.768±.041	.861±.040	.919±.026
	semiVAE (Concat.)	.803±.035	.876±.043	.926±.044
	semiMVAE	**.880**±.033	**.955**±.020	**.968**±.015

DEAP data	Algorithms	1% labeled	2% labeled	3% labeled
Supervised learning	MAE+SVM	.353±.027	.387±.014	.411±.016
	DCCA+SVM	.359±.016	.400±.014	.416±.018
	DCCAE+SVM	.361±.023	.403±.017	.419±.013
Transductive semi-supervised learning	AMMSS	.303±.029	.353±.024	.386±.014
	AMGL	.291±.027	.341±.021	.367±.019
	MAE+TSVM	.376±.025	.403±.031	.417±.026
	DCCA+TSVM	.379±.021	.408±.024	.421±.017
	DCCAE+TSVM	.384±.022	.412±.027	.425±.021
	semiMVAE	**.424**±.020	**.441**±.013	**.456**±.013
Inductive semi-supervised learning	semiVAE (Phy.)	.366±.024	.389±.048	.402±.034
	semiVAE (EEG)	.374±.019	.397±.013	.407±.016
	semiVAE (Concat.)	.383±.019	.404±.016	.416±.012
	semiMVAE	**.421**±.019	**.439**±.025	**.451**±.022

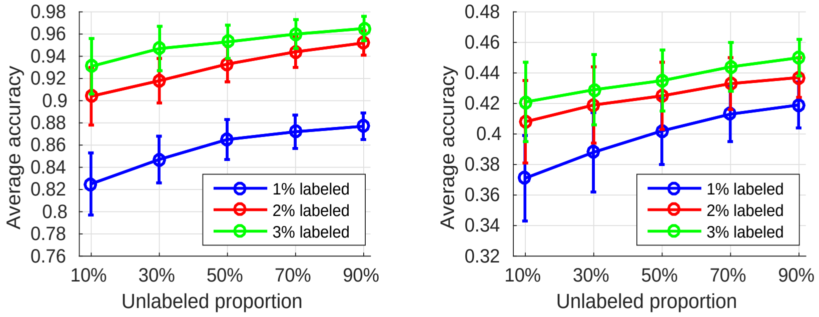
First, the average accuracy of semiMVAE significantly surpasses the baselines in all cases. Second, by examining semiMVAE against supervised learning approaches trained on very limited labeled data, we can find that semiMVAE always outperforms them. This encouraging result shows that semiMVAE can effectively leverage the useful information from unlabeled data. Third, multi-view semi-supervised algorithms AMMSS and AMGL perform worst in all cases. We attribute this to the fact that graph-based shallow models AMMSS and AMGL can't extract the deep features from the original data. Fourth, the performances of three TSVM based semi-supervised methods are moderate. Although MAE+TSVM, DCCA+TSVM and DCCAE+TSVM can also integrate multi-modality information from unlabeled samples, their two-stage learning can't obtain the global optimal model parameters. Finally, compared with the single-view semi-supervised method semiVAE, our multi-view method is more effective in integrating multiple modalities.

It also should be noted that the classification accuracy is quite different between the two datasets. As it is, the stimuli used in SEED are movie clips showing strong emotional expressions, thus the evoked EEG/eye movement signals have relatively clear categories. In contrast, the stimuli used in DEAP are music video clips showing blurry emotional expressions, thus the evoked EEG/physiological signals are difficult to be classified.



(a) SEED dataset (b) DEAP dataset

Fig. 3. semiMVAE's performance with different proportions of labeled samples in the training set.

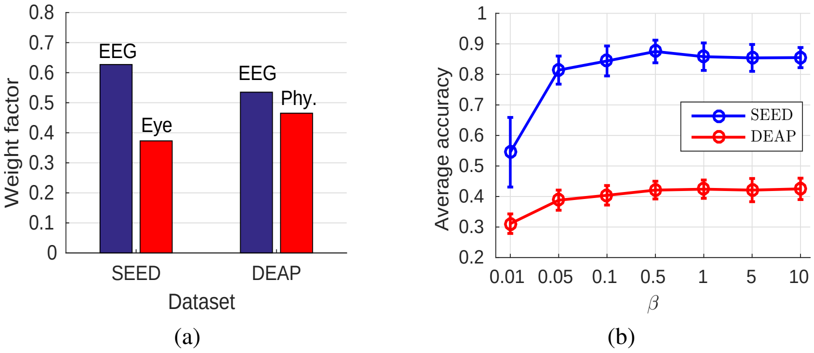


(a) SEED dataset (b) DEAP dataset

Fig. 4. Inductive semiMVAE's performance with different proportions of unlabeled samples in the training set.

The proportion of labeled and unlabeled samples in the training set will affect the performance of semi-supervised models. Figs. 3 and 4 show the changes of semiMVAE's average accuracy on both datasets with different proportions of labeled and unlabeled samples in the training set. We can observe that both labeled and unlabeled samples can effectively boost the classification accuracy of semiMVAE.

Instead of treating each modality equally, our semiMVAE can weight each modality and perform classification simultaneously. Fig. 5a shows the learned weight factors by inductive semiMVAE on SEED and DEAP datasets (1% labeled). From it, we can observe that EEG modality has the highest weight on both datasets, which is consistent with single modality's performance of semiVAE shown in Table 2 and the results in previous work [**?**].



(a) (b)

Fig. 5. (a) Learned weight factors by inductive semiMVAE. (b) The impact of scaling constant β.

The scaling constant β controls the weight of discriminative learning in semiMVAE. Fig. 5b shows the performance of inductive semiMVAE with different β values (1% labeled). From it, we can find that the scaling constant β can be chosen from $\{0.1, 0.5, 1\}$, where semiMVAE achieves good results.

5 Conclusion

This paper proposes a semi-supervised multi-view deep generative framework for emotion recognition, which can leverage both labeled and unlabeled data from multiple modalities. The key to our framework are two parts: 1) multi-view VAE can fully integrate the information from multiple modalities and 2) semi-supervised learning can overcome the labeled-data-scarcity problem. Experimental results on two real multimodal emotion datasets demonstrate the effectiveness of our approach. Actually, they are many real applications containing multiple views, e.g., the human action recognition, where we can collect the data from multiple cameras, and each camera can be regarded as one view. Other applications include multimedia analysis, neural decoding, disease diagnosis, etc. In principle, our semi-supervised multi-view framework can be applied to all of these problems, with potentially different deep neural network types for different modalities.

6 Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 91520202, 61602449) and Youth Innovation Promotion Association CAS.

References

1. Abadi, M.K., Subramanian, R., Kia, S.M., Avesani, P., Patras, I., Sebe, N.: Decaf: Meg-based multimodal database for decoding affective physiological responses. IEEE Transactions on Affective Computing 6(3), 209–222 (2015)
2. Alam, F., Riccardi, G.: Predicting personality traits using multimodal information. In: Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition. pp. 15–18. ACM (2014)
3. Andrew, G., Arora, R., Bilmes, J.A., Livescu, K.: Deep canonical correlation analysis. In: ICML. pp. 1247–1255 (2013)
4. Burda, Y., Grosse, R., Salakhutdinov, R.: Importance weighted autoencoders. In: ICLR (2016)
5. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation 42(4), 335 (2008)
6. Cai, G., Xia, B.: Convolutional neural networks for multimedia sentiment analysis. In: National CCF Conference on Natural Language Processing and Chinese Computing. pp. 159–167. Springer (2015)
7. Cai, X., Nie, F., Cai, W., Huang, H.: Heterogeneous image features integration via multimodal semi-supervised learning model. In: ICCV. pp. 1737–1744 (2013)
8. Calvo, R.A., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing 1(1), 18–37 (2010)
9. Chandar, S., Khapra, M.M., Larochelle, H., Ravindran, B.: Correlational neural networks. Neural computation 28(2), 257–285 (2016)
10. Dobrišek, S., Gajšek, R., Mihelič, F., Pavešić, N., Štruc, V.: Towards efficient multi-modal emotion recognition. International Journal of Advanced Robotic Systems 10(1), 53 (2013)
11. Gershman, S., Hoffman, M., Blei, D.: Nonparametric variational inference. In: ICML (2012)
12. Glodek, M., Reuter, S., Schels, M., Dietmayer, K., Schwenker, F.: Kalman filter based classifier fusion for affective state recognition. In: International Workshop on Multiple Classifier Systems. pp. 85–94. Springer (2013)
13. Jia, X., Li, K., Li, X., Zhang, A.: A novel semi-supervised deep learning framework for affective state recognition on EEG signals. In: International Conference on Bioinformatics and Bioengineering (BIBE). pp. 30–37. IEEE (2014)
14. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: NIPS. pp. 3581–3589 (2014)
16. Kingma, D.P., Salimans, T., Welling, M.: Improving variational inference with inverse autoregressive flow. In: NIPS (2016)
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: ICLR (2014)
18. Klami, A., Virtanen, S., Kaski, S.: Bayesian canonical correlation analysis. Journal of Machine Learning Research 14(1), 965–1003 (2013)
19. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis; using physiological signals. IEEE Transactions on Affective Computing 3(1), 18–31 (2012)
20. Kolodyazhniy, V., Kreibig, S.D., Gross, J.J., Roth, W.T., Wilhelm, F.H.: An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions. Psychophysiology 48(7), 908–922 (2011)

21. Liu, W., Zheng, W.L., Lu, B.L.: Multimodal emotion recognition using multimodal deep learning. arXiv preprint arXiv:1602.08225 (2016)
22. 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)
23. Maaløe, L., Sønderby, C.K., Sønderby, S.K., Winther, O.: Auxiliary deep generative models. In: ICML. pp. 1445–1453 (2016)
24. Monkaresi, H., Hussain, M.S., Calvo, R.A.: Classification of affects using head movement, skin color features and physiological signals. In: IEEE International Conference on Systems, Man, and Cybernetics. pp. 2664–2669. IEEE (2012)
25. Morency, L.P., Mihalcea, R., Doshi, P.: Towards multimodal sentiment analysis: Harvesting opinions from the web. In: Proceedings of the 13th international conference on multimodal interfaces. pp. 169–176. ACM (2011)
26. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML. pp. 689–696 (2011)
27. Nie, F., Li, J., Li, X., et al.: Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In: IJCAI. pp. 1881–1887 (2016)
28. Pang, L., Zhu, S., Ngo, C.W.: Deep multimodal learning for affective analysis and retrieval. IEEE Transactions on Multimedia 17(11), 2008–2020 (2015)
29. Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion 37, 98–125 (2017)
30. Poria, S., Cambria, E., Hussain, A., Huang, G.B.: Towards an intelligent framework for multimodal affective data analysis. Neural Networks 63, 104–116 (2015)
31. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: NIPS. pp. 1278–1286 (2014)
32. Rosas, V.P., Mihalcea, R., Morency, L.P.: Multimodal sentiment analysis of spanish online videos. IEEE Intelligent Systems 28(3), 38–45 (2013)
33. Schels, M., Kächele, M., Glodek, M., Hrabal, D., Walter, S., Schwenker, F.: Using unlabeled data to improve classification of emotional states in human computer interaction. Journal on Multimodal User Interfaces 8(1), 5–16 (2014)
34. Serban, I.V., Ororbia, I., Alexander, G., Pineau, J., Courville, A.: Multi-modal variational encoder-decoders. arXiv preprint arXiv:1612.00377 (2016)
35. Soleymani, M., Asghari-Esfeden, S., Fu, Y., Pantic, M.: Analysis of EEG signals and facial expressions for continuous emotion detection. IEEE Transactions on Affective Computing 7(1), 17–28 (2016)
36. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O.: Ladder variational autoencoders. In: NIPS. pp. 3738–3746 (2016)
37. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. Journal of Machine Learning Research 15, 2949–2980 (2014)
38. Verma, G.K., Tiwary, U.S.: Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. NeuroImage 102, 162–172 (2014)
39. Wang, S., Zhu, Y., Wu, G., Ji, Q.: Hybrid video emotional tagging using users' EEG and video content. Multimedia tools and applications 72(2), 1257–1283 (2014)
40. Wang, W., Arora, R., Livescu, K., Bilmes, J.A.: On deep multi-view representation learning. In: ICML. pp. 1083–1092 (2015)
41. Wang, W., Yan, X., Lee, H., Livescu, K.: Deep variational canonical correlation analysis. arXiv: 1610.03454 (2016)
42. Zhang, Z., Ringeval, F., Dong, B., Coutinho, E., Marchi, E., Schüller, B.: Enhanced semi-supervised learning for multimodal emotion recognition. In: ICASSP. pp. 5185–5189. IEEE (2016)