A Multimodal Deep Regression Bayesian Network for Affective Video Content Analyses

Quan Gan[1], Shangfei Wang[*,1], Longfei Hao[1], and Qiang Ji[2]
[1]University of Science and Technology of China, Hefei, Anhui, China
[2]Rensselaer Polytechnic Institute, Troy, NY 12180, USA
gqquan@mail.ustc.edu.cn, sfwang@ustc.edu.cn, hlf101@mail.ustc.edu.cn, qji@ecse.rpi.edu

Abstract

The inherent dependencies between visual elements and aural elements are crucial for affective video content analyses, yet have not been successfully exploited. Therefore, we propose a multimodal deep regression Bayesian network (MMDRBN) to capture the dependencies between visual elements and aural elements for affective video content analyses. The regression Bayesian network (RBN) is a directed graphical model consisting of one latent layer and one visible layer. Due to the explaining away effect in Bayesian networks (BN), RBN is able to capture both the dependencies among the latent variables given the observation and the dependencies among visible variables. We propose a fast learning algorithm to learn the RBN. For the MMDRBN, first, we learn several RBNs layer-wisely from visual modality and audio modality respectively. Then we stack these RBNs and obtain two deep networks. After that, a joint representation is extracted from the top layers of the two deep networks, and thus captures the high order dependencies between visual modality and audio modality. In order to predict the valence or arousal score of video contents, we initialize a feed-forward inference network from the MMDRBN whose inference is intractable by minimizing the KullbackCLeibler (KL)divergence between the two networks. The back propagation algorithm is adopted for fine-tuning the inference network. Experimental results on the LIRIS-ACCEDE database demonstrate that the proposed MMDRBN successfully captures the dependencies between visual and audio elements, and thus achieves better performance compared with state-of-the-art work.

1. Introduction

Recent years have seen increasingly big amount of video data, in particular user-created videos, with the rapid development in consumer electronics and the proliferation of mobile devices. For example, there are tens of thousands of videos uploaded to YouTube each day. In addition to the exponential growth in digital video repositories, users' purposes for consuming videos are also evolving due to the popularity of social networks. Emotional decisions are more and more widely utilized when querying and browsing video databases. Both the significant increase in video data size and the change in users' video viewing purposes demand a new way to effectively organize the videos to better meet users' specific needs. Therefore, affective video content analyses have attracted increasing attentions.

Current study of affective video content analyses can be categorized into two groups: direct approaches and implicit approaches. Direct video affective content analyses assign emotion tags to videos from the visual and audio features of videos. Implicit video affective content analyses, on the other hand, infer videos' emotion tags from a user's spontaneous nonverbal response while watching the videos [40]. Compared with implicit video affective content analyses, the direct approach is more practical, since it requires only video data without users' spontaneous nonverbal behavior signals. Therefore, this paper focuses on direct approaches of affective video content analyses.

The video content can be captured by various visual and audio features. The main stream of current affective video content analyses adopt either feature-level fusion or decision-level fusion to integrate visual and audio features. The former simply concatenates all the visual and audio features to one feature vector as the input to a classifier, and the later directly combines the classification results from visual features and audio features. These kinds of data fusion may not successfully model the inherent dependencies among visual features and audio features for video emotion tagging, since the dependencies between visual content and audio content as well as the influence of video content on users' emotions are very complex. Only very recently, Pang et al. [29] proposed to use multimodal deep Boltzmann machines (MMDBMs) to learn a joint density model over visual, auditory, and textual modalities for emo-

[*] This is the corresponding author.

tion tagging. Their experiments of affective analyses and retrieval on web videos and images demonstrated that the MMDBM can capture the non-linear and complex correlations among different modalities in a joint space for better affective analyses and retrieval. However, as an undirected graphic model, the deep Boltzmann machine assumes the latent nodes are independent of each other. It inevitably weakens the representation power of the MMDBM. Therefore, in this paper, we propose a new multimodal learning method, multimodal deep regression Bayesian network (MMDRBN), to construct the high-level joint representation of visual and audio modalities for emotion tagging. Then the MMDRBN is transformed into an inference network by minimizing the KL-divergence. After that, the inference network is used to predict discrete or continuous affective scores from video content. Experimental results on the LIRIS-ACCEDE database demonstrate the advantages of our proposed method.

2. Related Work

2.1. Affective Video Content Analyses

The video content consists of both visual and audio signals. The mainstream of current affective video content analyses first extracts several visual and audio features to characterize the video content, and then concatenates them to feed into a general purpose classifier or regressor for emotion classification or regression. For example, Baveye et al. [6] and Zhang et al. [46] used support vector regression (SVR) to predict the valence and arousal of video clips from a large number of visual and audio features. Canini et al. [7] mapped visual and audio features to natural, temporal and energetic dimensions using SVR, polynomial regression, and neural network. Wang et al. [39] adopted the support vector machine (SVM) to recognize anger, sadness, surprise, happiness, disgust and neutral from visual and audio features. In addition to feature-level fusion, some work adopts decision-level fusion to integrate visual features and audio features for emotion classification or regression from videos. For example, Hanjalic and Xu [15] defined the arousal curve and valence curve as the linear combination of motion component, rhythm component and sound energy component. Jiang et al. [19] adopted kernel-level fusion to linearly combine kernels computed on the individual features for video emotion tagging. Acar et al. [1] used convolutional neural networks (CNNs) to learn mid-level representations from low-level visual features and audio features respectively, and employed three multi-class SVM to classify video clips into four affective categories through decision-level fusion. In addition to performing feature-level and decision-level fusion separately, hybrid methods have also been proposed to combine feature-level and decision-level fusion. For example, Yazdani et al. [43] proposed a hybrid multilevel fusion approach to take advantage of both feature-level fusion and decision-level fusion. Specifically, in addition to the audio and video modalities, a joint audio and video modality derived from feature fusion forms an additional modality. The final decision is granted using the sum rule over the tagging results of the three modalities. A comprehensive survey of current affective video content analyses can be found in [40].

Although feature-level fusion and decision-level fusion can combine the information from visual content and audio content for emotion classification and regression, they can not successfully capture the structures embedded in video content and the inherent interactions among visual content, audio content and emotion labels by simply concatenating all the visual and audio features to one feature vector as the input of a classifier, or directly combining the classification results from visual features and audio features. Multi-view learning or multimodal learning may be a better approach to leverage the dependencies among visual content and audio content for affective video content analyses. However, to the best of our knowledge, little work exploit multi-view learning for affective video content analyses except for Pang et al.'s [29]. Pang et al. [29] proposed to use MMDBM to learn a joint density model over visual, auditory, and textual modalities for emotion tagging. Specifically, the proposed MMDBM first learns middle-level representations from low-level visual features, audio features and textual features using DBM respectively, and then constructed a joint representation from the learned middle-level representations. After that, the final joint representations are used for learning a logistic regression model. Their experiments of affective analyses and retrieval on web videos demonstrated that the MMDBM can capture the non-linear and complex correlations among different modalities in a joint space for better affective analyses and retrieval. However, as a undirected graphic model, deep Boltzmann machine (DBM) assumes the latent nodes are independent of each other given the neighbouring layers. However, latent variables should be dependent to jointly explain the patterns in the input data. The independences inevitably weakens the representation power of the MMDBM.

Instead of using MMDBM, we propose a new multimodal learning method, MMDRBN, to construct the high-level joint representation of visual and audio modalities for affective video content analyses. While similar in the structures to the existing generative deep models, the proposed MMDRBN is fundamentally different from these generative models, since the dependencies among latent nodes are preserved during learning phase by our proposed inference method. Thus, the proposed MMDRBN can successfully capture the inherent dependencies between visual content and audio content, and achieve better performance of emotion recognition and emotion regression compared with state of the art work.

2.2. Multi-view Learning

Multi-view learning has attracted increasing attentions recently, since most data can be split into multiple distinct feature sets naturally or manually. Unlike single view learning, multi-view learning jointly optimizes functions of multiple views and models the inherent dependencies among multiple views for performance improvement. Comprehensive survey on multi-view learning can be found in [37, 42].

One mainstream of multi-view learning algorithms is to find the common spaces of multi views, since every view contains relevant information for classification. Among them, canonical correlation analysis (CCA) is a very popular method, which aims to find two bases, one for each view, that are optimal with respect to maximum correlations [4]. CCA is a way of measuring the linear relationship between two views. Later, kernel canonical correlation analysis (KCCA) is proposed to extend CCA from the linear change to the nonlinear change by leveraging kernel method during CCA transformation [2]. Recently, due to the emergence of big data and success of deep learning, several Deep Neural Network (DNN)-based multi-view representation learning algorithms are proposed. The training criteria of DNN-based multiview representation learning can be classified into two categories: one is to learn representations in two views that are maximally correlated, i.e., deep canonical correlation analysis (DCCA) [41], which is a deep neural network version of CCA [5]. By applying DNNs as a nonlinear function to model real world data accurately through extracting high level representations [16], DCCA aims to learn feature representations of two view which are maximally correlated. DCCA has been proved more accurate than KCCA in nonlinear transformation tasks [5]. The other category of DNN-based multiview representation learning is to learn a compact representation that best reconstructs the inputs, such as multimodal deep belief network (DBN) [28] and MMDBM [36]. Both multimodal DBN and multimodal DBM consists of several layers of restricted Boltzmann machine (RBM), which is an undirected graphical model. Although RBM can effectively capture global dependencies among visible units through introducing hidden units, hidden units are independent to each other given visible units. Introducing dependencies among hidden units will increase the model power in explaining the patterns embedded in the visible units. Unlike RBM, RBN is a directed latent variable model. Through directed links among hidden units and visible units, RBN can capture both the dependencies among the latent variables given the observation and the dependencies among visible variables, and thus better represent the visible units.

However, the inference of RBN is computationally intractable. Common-used variation approximation algorithm can solve the problem but the dependencies among latent variables will be discarded. We propose to solve it by Gibbs sampling so that the dependencies will be preserved to some degree. After several RBNs are learned, they can be used to stack a MMDRBN. The MMDRBN can model the inherent dependencies between visual content and audio content. An inference network is initialized from the MMDRBN by minimizing KL divergence and it is used for affective video content analyses.

In addition, MMDRBN is not limited to capture dependencies between visual content and audio content for affective video content analyses, it could be widely applicable for analysing dependencies among other multiple modalities.

3. Proposed Method

Fig. 1 shows the framework of our proposed method. First, we train a multimodal generative network. It consists of two stacked RBNs that are created for visual and audio modalities respectively. As Fig. 2 shows, the RBN is a kind of Bayesian network and can be used as a directed generative model. The connections of the RBN are top-down, which makes the RBN different from feed-forward neural networks and undirected generative models such as RBMs. Then, a multimodal inference network is initialized at the basis of the generative network by minimizing the KL divergence between them. Finally, the back propagation algorithm is applied to adjust parameters of the inference network for video tagging tasks.

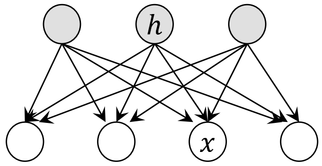


Figure 2. The structure of a RBN.

3.1. Learning a Multimodal Generative Network

3.1.1 Introduction to RBN

The RBN is a special kind of Bayesian network, consisting of one visible layer and one latent layer. Every latent variable connects to every visible variable with a directed edge as shown in Fig. 2. The directed connections bring the RBN model ”explaining away” effect, which makes the latent variables dependent on each other given the visible variables. Therefore, RBN can effectively captures the dependencies among the latent variables. To meet the requirements of our experimental data, in the following part we shall focus on the Gaussian-Bernoulli RBN that takes the continuous input.

Since the RBN is a kind of Bayesian network, chain rule

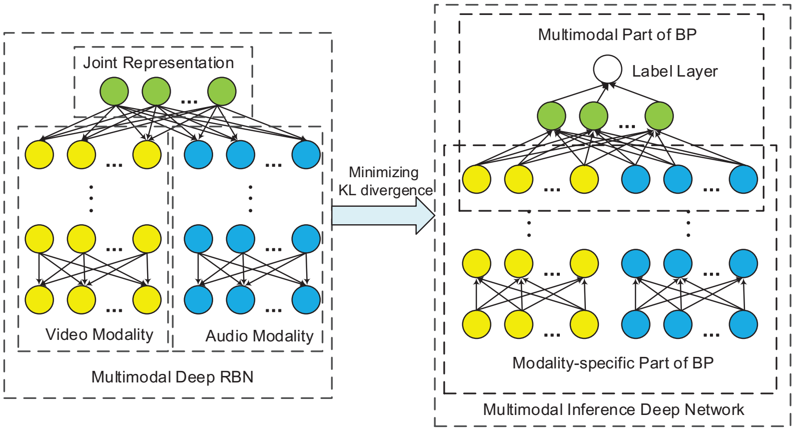


Figure 1. The framework of our proposed method.

also works for it. According to the chain rule, the joint distribution of RBN can be factorized as Eq. 1,

$$P(\boldsymbol{x},\boldsymbol{h}) = \prod_{j=1}^{n_h} P(h_j) \prod_{i=1}^{n_d} P(x_i|\boldsymbol{h})\,. \tag{1}$$

The prior probalility for a latent variable h_j satisfies the Bernoulli distribution, and can be written as Eq. 2,

$$P(h_j) = \sigma(d_j)^h(1-\sigma(d_j))^{1-h} \tag{2}$$

where $\sigma(d_j) = 1/(1+\exp(-d_j))$ and d_j is the bias of the variable h_j.

The conditional probability of a visible variable given all the latent variables is defined as a linear Gaussian, as shown in Eq. 3,

$$P\left(x_i|\boldsymbol{h}\right) \sim \mathcal{N}\left(\boldsymbol{w}_i^T\boldsymbol{h}+b_i,\, \sigma_i\right)\ , \tag{3}$$

The mean of the Gaussian distribution is a linear combination of latent variables. The standard deviation is calculated from visible variables x. w_{ij} is the weight for node h_j and x_i, and b_i is the bias term for x_i.

Thus, the RBN can be viewed as a mixture of Gaussian with the number of components exponential in the number of latent variables.

Combining Eq. 2 and Eq. 3, Eq. 1 can be rewritten as the following formulation,

$$\begin{aligned} P_\Theta(\boldsymbol{x},\boldsymbol{h}) &= \prod_j \frac{\exp(d_j h_j)}{1+\exp(d_j)} \prod_i \mathcal{N}(x_i : \boldsymbol{w}_i^T\boldsymbol{h}+b_i,\, \sigma_i) \\ &= \frac{\exp(-\psi_\Theta(\boldsymbol{x},\boldsymbol{h}))}{(2\pi)^{n_d/2}\prod_i \sigma_i \prod_j\left(1+\exp(d_j)\right)} \end{aligned} \tag{4}$$

where $\Theta = \{\boldsymbol{W},\boldsymbol{\sigma},\boldsymbol{b},\boldsymbol{d}\}$, and

$$\begin{aligned} \psi_\Theta(\boldsymbol{x},\boldsymbol{h}) = \sum_i \frac{(x_i-b_i)^2}{2\sigma_i^2} - \sum_i \frac{x_i-b_i}{\sigma_i^2}\boldsymbol{w}_i^T\boldsymbol{h} \\ + \sum_i \frac{1}{2\sigma_i^2}(\boldsymbol{w}_i^T\boldsymbol{h})^2 - \boldsymbol{d}^T\boldsymbol{h}\,, \end{aligned} \tag{5}$$

Compared with the joint probability of Gaussian-Bernoulli RBMs (GRBMs) as shown in Eq. 6, there are two main differences between them. First, the RBN adopts top-down connections instead of undirected connections, which introduces the extra term $\sum_i \frac{1}{2\sigma_i^2}(\boldsymbol{w_i}\boldsymbol{h})^2$ compared to GRBMs. This term explicitly models the dependencies among latent variables according to the formulation. We hope the dependencies among latent variables can help us better model the patterns embedded in the data. Second, the normalized term of RBN is easy to calculate, while the partition function of RBM is computationally intractable.

$$P_{RBM}(\boldsymbol{x},\boldsymbol{h}) = \frac{1}{Z}\exp(-\psi_{RBM}(\boldsymbol{x},\boldsymbol{h})) \tag{6}$$

$$\psi_{RBM}(\boldsymbol{x},\boldsymbol{h}) = \sum_i \frac{(x_i-b_i)^2}{2\sigma_i^2} - \sum_i \frac{x_i}{\sigma_i^2}\boldsymbol{w}_i^T\boldsymbol{h} - \boldsymbol{d}^T\boldsymbol{h}, \tag{7}$$

3.1.2 An Efficient Learning Method of RBN

In this subsection, we introduce an efficient parameter learning method for RBN based on stochastic approximation procedure (SAP).

Consider Eq. 4, model parameters Θ can be learned by maximizing marginal log-likelihood. Given the training data set $\mathcal{D} = \{\boldsymbol{x}^{(m)}\}_{m=1}^{M}$, the learning target can be presented as following equation.

$$\begin{aligned}\mathcal{L}(\mathcal{D};\Theta) &= \sum_m \log P_\Theta(\boldsymbol{x}^{(m)}) \\ &= \sum_m \log\left(\sum_{\boldsymbol{h}} P_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h})\right).\end{aligned} \tag{8}$$

Usually parameters will be calculated by maximizing the object function through gradient ascent. The equation of parameter gradient is as follows,

$$\nabla_\theta \mathcal{L}(\mathcal{D};\Theta) = \sum_m \sum_h P_\Theta(\boldsymbol{h}|\boldsymbol{x}^{(m)}) \frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h})}{\partial\theta}. \tag{9}$$

According to Eq. 9, we can find that exact gradient is computationally intractable. Due to the top-down connections of the RBN, the posterior probability $P_\Theta(h|x^{(m)})$ is unable to be calculated, as know as the intractable inference issue. Furthermore, the exact gradient needs exponential summations over all possible latent variables h.

Typically, intractable inference can be solved by variational approximation algorithms. Some examples of the variational approximation are the mean field algorithm [32] the wake-sleep algorithm [17], and inference networks [27, 20, 30, 14]. However, such approximations introduce a gap between the true posterior and the approximate ones, since the dependencies are not captured in the approximate distribution.

In this work, we intend to preserve such dependencies by directly using samples from the true posterior probability. Specifically, we employ Gibbs sampling to draw samples for the latent variables. One latent variable are sampled conditioned on all the other variables. Therefore, dependencies are preserved to some degree.

Drawing samples exactly from $P(\boldsymbol{h}|\boldsymbol{x})$ through Gibbs sampling is intractable since the specific form of $P(\boldsymbol{h}|\boldsymbol{x})$ is not available. To solve the problem, we take some approximations during the sampling,

$$P(h|x) = \prod_j P(h_j|h_1,...,h_{j-1},x) \approx \prod_j P(h_j|h_{-j},x). \tag{10}$$

where $h_{-j} = \{h_1,...,h_{j-1},h_{j+1},...,h_{n_h}\}$. Then each latent node is sampled with all others fixed as Eq. 11. Therefore, dependencies among latent variables can be preserved to some extent. The procedure is repeated for several iterations until mixing, and then a sample is collected for updating the parameters.

$$h_j^t \sim P(h_j|\boldsymbol{x},\boldsymbol{h}_{-j}^{t-1}). \tag{11}$$

To address the exponential summation issue, typically Markov Chain Monte Carlo (MCMC) methods are used to estimate the summation using samples. An intuitive estimation is,

$$\nabla_\theta \mathcal{L}(\mathcal{D};\Theta) \approx \frac{1}{n}\sum_m \sum_s \frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m,s)})}{\partial\theta}, \tag{12}$$

where $\boldsymbol{h}^{(m,1)},...,\boldsymbol{h}^{(m,n)}$ are n samples from $P(\boldsymbol{h}|\boldsymbol{x}^{(m)})$. In this work, we employ the stochastic approximation procedure (SAP) framework [31], in which only one sample of the latent variables are used to estimate the gradient, so multiple Gibbs chains are avoided.

Under some mild assumptions, the SAP is guaranteed to converge to a local optimum [45] if the learning rate γ_t satisfies,

$$\begin{aligned}\sum_{t=1}^{\infty}\gamma_t = \infty, \\ \sum_{t=1}^{\infty}\gamma_t^2 < \infty.\end{aligned} \tag{13}$$

The gradient is then estimated as,

$$\nabla_\theta \mathcal{L}(\mathcal{D};\Theta) \approx \sum_m \frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m)})}{\partial\theta}, \tag{14}$$

The derivative has a simple formulation because the energy function is merely a linear function of the parameters,

$$\frac{\partial - E_\Theta(\boldsymbol{x}^{(m)},\boldsymbol{h}^{(m)})}{\partial w_{ij}} = \frac{h_j^{(m)}(x_i^{(m)} - \boldsymbol{w}_i^T\boldsymbol{h}^{(m)})}{\sigma_i^2}. \tag{15}$$

The gradient of other parameters can be derived similarly.

To speed up the learning phase and scale up to large databases in practice, we employ the stochastic gradient ascent algorithm, which estimates the gradient using a mini-batch of training samples. Several passes is made over the training set until convergence.

In Algorithm 1 we present the SAP for learning a RBN.

3.1.3 Stacking RBNs

As shown in the framework of the proposed method, we train two deep networks for visual and audio modalities respectively. Each deep network is constructed by stacking several trained RBNs. The RBNs are layer-wisely trained using the method proposed in Section. 3.1.2. In layer-wise training phase, the output of a lower RBN is used as the input of its upper RBN. After obtaining networks of two modalities, we add a multimodal layer upon the top layer. Specifically, the visible variables of the multimodal layer are concatenation of the top layers of visual and audio

Algorithm 1 Parameter Learning for an RBN.

```
Input database $\mathcal{D} = \{\boldsymbol{x}^{(m)}\}_{m=1}^{M}$;
Output parameters $\Theta = \{\boldsymbol{W}, \boldsymbol{\sigma}, \boldsymbol{b}, \boldsymbol{d}\}$.
Randomly initialize the parameters $\Theta$;
Generate Gibbs samples at time step 0;
while parameters not converged, do
    Randomly choose a batch of data samples $\boldsymbol{x}$;
    Perform Gibbs sampling to obtain one sample of the latent variables for one input data, $\boldsymbol{h}^{(t)} \sim P(\boldsymbol{h}|\boldsymbol{x}, \boldsymbol{h}^{(t-1)})$;
    Compute the gradient using Eq. 15;
    Update the parameters,
    $\theta_t = \theta_{t-1} + \gamma_t \nabla_\theta \mathcal{L}(\boldsymbol{x})$.
end while
```

modality networks. The latent variables of the multimodal layer are seen as the joint representation of two different modalities. After stacking the trained RBNs, we obtain a MMDRBN as shown in the left part of Fig. 1.

3.2. Learning an Inference Network

Since exact inference is intractable for the RBN model due to the edge direction, we will learn an inference network at the basis of the MMDRBN. For approximating the exact inference of RBNs, we introduce an distribution Q_Φ by minimizing the KL divergence between $Q_\Phi(h|x)$ and $P_\Theta(h|x)$. Φ is the set of parameters of Q and P_Θ is the RBN learned with our proposed method. The equation is as follow:

$$\begin{aligned} &KL(Q_\Phi(h|x)||P_\Theta(h|x)) \\ &= \sum_h Q_\Phi(h|x) log \frac{Q_\Phi(h|x)}{P_\Theta(h|x)}. \end{aligned} \tag{16}$$

Distribution Q should be easy to calculate since it will be used to approximate the intractable exact inference $P_\Theta(h|x)$ by minimizing KL divergence. We use gradient descent to optimize the KL divergence and the gradient is given by:

$$\begin{aligned} &\frac{\partial KL(Q_\Phi(h|x)||P_\Theta(h|x))}{\Phi} \\ &= \sum_h \frac{\partial Q_\Phi(h|x)}{\Phi} log Q_\Phi + \sum_h \frac{\partial Q_\Phi(h|x)}{\Phi} \\ &- \sum_h log P_\Theta(x,h) \frac{\partial Q_\Phi(h|x)}{\Phi} \\ &= E((log P_\Theta(x,h) - log Q_\Phi(h|x)) \\ &\times \frac{\partial log Q_\Phi(h|x)}{\partial \Phi}. \end{aligned} \tag{17}$$

Since calculating the expectation in Eq. 17 is time consuming, we will estimate it with sampling-based method. After getting n samples from the distribution $Q_\Phi(h|x)$, the estimation of expectation can be written as:

$$\begin{aligned} &\frac{\partial KL(Q_\Phi(h|x)||P_\Theta(h|x))}{\Phi} \\ &= \frac{1}{n} \sum_{i=1}^{n} ((log P_\Theta(x, h^{(i)}) - log Q_\Phi(h^{(i)}|x)) \\ &\times \frac{\partial log Q_\Phi(h^{(i)}|x)}{\partial \Phi}. \end{aligned} \tag{18}$$

Stochastic gradient ascent algorithm is also adopted here for the same reasons as mentioned in Section. 3.1.2.

3.3. Discriminative Fine-tuning

After initializing an inference network from a MMDRBN, we add a label layer on the top of the inference network. The label layer is used to fine-tune the network through back propagation algorithm. Since the inference network is also multimodal, the application of the back propagation in our network is a little different from the standard back propagation process. It takes two steps to fine-tune the parameters. In step one, the back propagation is applied to the multimodal part in Fig. 1. The top layers of audio and visual modality networks will be treated as a single one in this step. After fine-tuning the multimodal part, the errors will be propagated to the top layers of audio and visual modality networks. Then in step 2, the back-propagated errors will spread from the top layers of the two networks. The errors are used to fine-tune the remaining parts of the two networks respectively through the back propagation algorithm. This inference network can be used for our classification tasks after two steps of fine-tuning.

4. Experiments and Results

4.1. Database

There are several benchmark databases that can be used for affective video content analyses, such as the HUMAINE database [13], the Database for Emotion Analysis using Physiological signals (DEAP) [21], the MAHNOB-HCI database [35] and the LIRIS-ACCEDE database [6]. The HUMAINE database contains 50 video clips; the DEAP database contains 120 one-minute excerpts from music videos; the MAHNOB-HCI database contains 20 emotional videos between 34.9 and 117s long and the LIRIS-ACCEDE database contains 9,800 excerpts extracted from 160 feature films and short films. Thus, the LIRIS-ACCEDE database is the largest video database for video content analyses. Since the proposed MMDRBN requires large amount of samples to achieve better performance and avoid over-fitting, we adopt the LIRIS-ACCEDE database for the evaluation.

The video clips from the LIRIS-ACCEDE database is ranked along the induced arousal and valence axis, initially ranging from 0 to 9,799. Based on these valence and arousal ranks, MediaEval 2015 [34] and MediaEval 2016 [12] proposed classification and regression tasks respectively on the LIRIS-ACCEDE database. MediaEval 2015 rescales the ranks uniformly to a more common [-1,1] range. Then valence or arousal scores are assigned with -1, 0, 1 corresponding to three ranges [-1, -0.15), [-0.15, 0.15] and (0.15, 1]. MediaEval 2016 estimates the absolute affective scores for valence and arousal from the initial ranks using Gaussian process regression models. Then the absolute affective scores are provided as the ground truth for the regression task.

4.2. Experimental Conditions on the LIRIS-ACCEDE Database

Our experiments on LIRIS-ACCEDE database consist of two parts: classification task proposed by MediaEval 2015 and regression task proposed by 2016.

For classification task, we adopt the features proposed in [6]. A set of 17 features is used for valence and 12 features for arousal. Futhermore, we add three visual features and 31-dimensional commonly used audio features that are used in [10] for augmenting the features. For regression task, we remove two features (audio flatness envelope and the slope of the power spectrum) from arousal feature set according to the baseline features of MediaEval 2016.

We adopt 10-fold cross validation in our experiments. The following designs of network are chosen for working well on both classification and regression tasks. The deep RBNs for audio modality and visual modality both consist of two layers of RBNs. For the visual modality, the number of nodes of the first hidden layer is set to 8 and the ones of the second hidden layer is set to 18. For the audio modality, the number of nodes of the first hidden layer is set to 30 and the ones of the second hidden layer is set to 18. The dimension of the joint representation layer is 18. We adopted accuracy as the metric for classification task and Pearson correlation coefficient (PCC) and mean squared error (MSE) for regression task.

Under above data and experimental settings, we design several experiments to demonstrate the effectiveness of our method from several aspects.

To demonstrate the advantage of the proposed method over other multimodal methods, we conduct experiments using CCA, KCCA, LCCA, DCCA, DCCAE and MMDBM on the same experimental conditions. The comparison results are shown in Table. 1. To demonstrate the advantage of the multimodal process in our method, we compare the proposed model with the early fusion and late fusion methods. The comparison results are shown in Table. 2. In early fusion method, we merge the two modalities into one vector, and train a network based on merged feature vectors. The late fusion method learns two separated inference networks using the proposed method. Each network is able to predict the affective label according to the output vector. Then the output vectors of two networks are normalized respectively and combined using the sum rule to predict the affective label. The comparison is also used to demonstrate that the joint representation of the abstracted features has advantages over simply merged features, which is clarified in [36]. Furthermore, we compare our classification and regression results with the state-of-the-art work [44, 38, 11, 22] in Table. 3 and [9, 24, 3, 18] in Table. 4.

Table 1. Comparison with multimodal methods.

	MediaEval 2015		MediaEval 2016			
	Valence	Arousal	Valence		Arousal	
	Acc	Acc	MSE	PCC	MSE	PCC
LCCA	42.16	63.33	0.42	0.20	0.92	0.23
KCCA	42.95	63.32	0.40	0.21	0.89	0.24
DCCA	41.77	63.58	0.39	0.24	0.88	0.26
DCCAE	43.38	63.74	0.41	0.23	0.88	0.27
MMDBM	41.38	63.75	0.39	0.19	0.92	0.20
Ours	**44.26**	**64.30**	**0.33**	**0.39**	**0.77**	**0.42**

Table 2. Comparison with different fusion methods.

	MediaEval 2015		MediaEval 2016			
	Valence	Arousal	Valence		Arousal	
	Acc	Acc	MSE	PCC	MSE	PCC
Early fusion	43.74	63.85	0.34	0.33	0.83	0.31
Late fusion	42.22	63.76	0.36	0.26	0.82	0.32
Ours	**44.26**	**64.30**	**0.33**	**0.39**	**0.77**	**0.42**

Table 3. Comparison with MediaEval 2015 related work.

	Valence	Arousal
MIC-TJU [44]	41.95	55.93
NII-UIT [22]	42.96	55.91
ICL-TUM-PASSAU [38]	41.48	55.72
Fudan-Huawei [11]	41.80	48.80
TCS-ILAB [8]	35.66	48.95
UMons [33]	37.28	52.44
RFA [26]	33.03	45.04
KIT [25]	38.50	51.90
Ours	**44.26**	**64.30**

Table 4. Comparison with MediaEval 2016 related work.

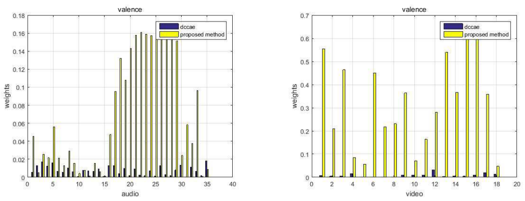
	Valence		Arousal	
	MSE	PCC	MSE	PCC
RUC [9](run 1)	0.218	0.312	1.479	0.405
THU-HCSI [24]	**0.214**	0.296	1.531	0.267
AUTH-SGP [3]	/	0.290	/	0.303
BUL [18]	0.231	0.149	1.413	0.271
Liu's [23]	0.240	0.102	1.185	0.159
Ours	0.33	**0.39**	**0.77**	**0.42**

4.3. Experimental Results and Analysis

The proposed method shows good performance on the LIRIS-ACCEDE database for both classification and regression tasks. For classification, the accuracy is 44.26%

for valence and 64.30% for arousal. For regression, the MSE is 0.33 and PCC is 0.39 in the valence space. As for arousal, we get 0.77 for MSE and 0.42 for PCC. The result of arousal is higher than that of valence for both classification and regression. The reason may be that the inner pattern of arousal data is more helpful for recognition.

The comparison results with other multimodal methods are listed in Table. 1. The performance of the proposed method is the highest among the related work and related methods for classification and regression tasks. CCA, KCCA, LCCA, DCCA and Deep Canonically Correlated Autoencoders (DCCAE) are CCA-related methods which try to find the common space of different modalities. Among them, CCA, KCCA and LCCA directly learn a representation for the input modalities without the layer-wise feature abstraction. The input features contain modality-specific information, which makes it much harder to discover relationships across modalities than relationships among features in the same modality [36]. Our proposed model intends to eliminate the modality-specific information through the deep networks and then learns a joint representation of the high-level features of two modalities, which may be the reason why our model outperforms CCA, KCCA and LCCA methods. DCCA, DCCAE and MMDBM are based on RBMs. DCCA and DCCAE combine the RBM-based deep neural network and CCA method. MMDBM is constructed with DBMs of several modalities. Our experimental results are also higher that those of DCCA, DCCAE and MMDBM since the RBN is able to capture more dependencies from data than the RBM, as we discussed in Section. 3. In order to get in-depth analysis on the comparison results, we visualize and analyse the learned features of our model and DCCAE that performs best among LCCA, KCCA, DCCA and MMDBM. Taking classification task as an example, we visualize the parameters of valence model of the bottom layers for the LRBN model and DAACE model in Fig. 3. In the visual figure, parameters of our model focus on colourfulness, fades and number of scene cut, and these features are closely related to the emotion according to [39]. In the audio figure, we obtain larger parameters on MFCC-related features which are common-used features in affective computing.



(a) (b)

Figure 3. Visualization of parameters for classification task.

Table. 2 shows the comparison with the early fusion and late fusion methods. The two methods are also based on RBNs and learned using the proposed algorithm in Section. 3. The comparison further demonstrates the important role of the joint representation in a multimodal method. The joint representation can model the relationships between the features of different modalities with less differences in modal concepts than merging the raw features directly. From the comparison results, we can infer that representing multimodal data in the same output space is one of the most important steps for improving the model performance.

Table. 3 shows the results of all work published in MediaEval 2015 [34]. Since the used features and experimental settings are different from each other, we only compare with their highest results and the comparison is listed for reference only. However, considering that our features are simplest among all the related work, our highest results show a strong ability of the proposed method to model the data distribution.

Table. 4 shows the results of all work published in MediaEval 2016 [12]. Just as the comparison with MediaEval 2015, the comparisons can be used only for reference due to different experimental data and settings. Our model performs better than all the listed work except for RUC's [9]. RUC achieves nearly the best experimental results among all the participants of MediaEval 2016. Our results is competitive, though the MSE of valence is worse than others'. Take the results of classification and regression into consideration together, our model not only produces good results for video tagging task, but also has a excellent generalization ability for both classification and regression tasks.

5. Conclusion and Future Work

In this work, we propose a fast learning algorithm for the RBN by maximizing the marginal probability of the observed data. The learned RBNs can be used to construct a deep network. For solving the intractable inference of the RBN, we propose learning a feed-forward network from the RBN-based deep network by minimizing KL divergence. Through the proposed method, we learn a multimodal inference network for the video tagging problem. The experimental results and several comparison on LIRIS-ACCEDE database show the advantages of our proposed method. In future work, we plan to apply the RBN to other application area to solve various tasks in the help of its model ability.

Acknowledgment

This work has been supported by the National Science Foundation of China (Grant No.61473270, 61228304, 61175037), the project from Anhui Science and Technology Agency (1508085SMF223), and the National Science Foundation (1539012).

References

[1] E. Acar, F. Hopfgartner, and S. Albayrak. Understanding affective content of music videos through learned representations. In *MultiMedia Modeling*, pages 303–314, 2014. 2

[2] S. Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006. 3

[3] T. Anastasia and H. Leontios. Auth-sgp in mediaeval 2016 emotional impact of movies task. 2016. 7

[4] T. W. Anderson, T. W. Anderson, T. W. Anderson, and T. W. Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958. 3

[5] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1247–1255, 2013. 3

[6] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015. 2, 6, 7

[7] L. Canini, S. Benini, and R. Leonardi. Affective recommendation of movies based on selected connotative features. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):636–647, 2013. 2

[8] R. Chakraborty, A. K. Maurya, M. Pandharipande, E. Hassan, H. Ghosh, and S. K. Kopparapu. Tcs-ilab-mediaeval 2015: Affective impact of movies and violent scene detection. In *MediaEval*, 2015. 7

[9] S. Chen and Q. Jin. Ruc at mediaeval 2016 emotional impact of movies task: Fusion of multimodal features. 7, 8

[10] S. Chen, S. Wang, C. Wu, Z. Gao, X. Shi, and Q. Ji. Implicit hybrid video emotion tagging by integrating video content and users' multiple physiological responses. In *International Conference on Pattern Recognition*, 2016. 7

[11] Q. Dai, R.-W. Zhao, Z. Wu, X. Wang, Z. Gu, W. Wu, and Y.-G. Jiang. Fudan-huawei at mediaeval 2015: Detecting violent scenes and affective impact in movies with deep learning. In *MediaEval Workshop*, 2015. 7

[12] E. Dellandréa, L. Chen, Y. Baveye, M. Sjöberg, C. Chamaret, and E. Lyon. The mediaeval 2016 emotional impact of movies task. In *Proc. of the MediaEval 2016 Workshop, Hilversum, Netherlands*, 2016. 7, 8

[13] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, et al. The humaine database: addressing the collection and annotation of naturalistic and induced emotional data. In *International Conference on Affective Computing and Intelligent Interaction*, pages 488–500. Springer, 2007. 6

[14] K. Gregor, A. Mnih, and D. Wierstra. Deep autoregressive networks. *In Proceedings of the 31st International Conference on Machine Learning*, 2014. 5

[15] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE transactions on multimedia*, 7(1):143–154, 2005. 2

[16] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. 3

[17] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The" wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995. 5

[18] A. Jan, Y. F. A. Gaus, F. Zhang, and H. Meng. Bul in mediaeval 2016 emotional impact of movies task. 7

[19] Y.-G. Jiang, B. Xu, and X. Xue. Predicting emotions in user-generated videos. In *AAAI*, pages 73–79, 2014. 2

[20] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *In Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 5

[21] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. 6

[22] V. Lam, S. Phan, D.-D. Le, S. Satoh, and D. A. Duong. Niiuit at mediaeval 2015 affective impact of movies task. In *MediaEval Workshop*, 2015. 7

[23] Y. Liu, Z. Gu, Y. Zhang, and Y. Liu. Mining emotional features of movies. 7

[24] Y. Ma, Z. Ye, and M. Xu. Thu-hcsi at mediaeval 2016: Emotional impact of movies task. 7

[25] P. Marin Vlastelica, S. Hayrapetyan, M. Tapaswi, and R. Stiefelhagen. Kit at mediaeval 2015-evaluating visual cues for affective impact of movies task. In *MediaEval*, 2015. 7

[26] I. Mironica, B. Ionescu, M. Sjöberg, M. Schedl, and M. Skowron. Rfa at mediaeval 2015 affective impact of movies task: A multimodal approach. In *MediaEval*, 2015. 7

[27] A. Mnih and K. Gregor. Neural variational inference and learning in belief networks. *In Proceedings of the 31st International Conference on Machine Learning*, 2014. 5

[28] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011. 3

[29] L. Pang, S. Zhu, and C.-W. Ngo. Deep multimodal learning for affective analysis and retrieval. *IEEE Transactions on Multimedia*, 17(11):2008–2020, 2015. 1, 2

[30] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1278–1286, 2014. 5

[31] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 5

[32] L. K. Saul, T. Jaakkola, and M. I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4(61):76, 1996. 5

[33] O. Seddati, E. Kulah, G. Pironkov, S. Dupont, S. Mahmoudi, and T. Dutoit. Umons at mediaeval 2015 affective impact of movies task including violent scenes detection. In *MediaEval*, 2015. 7

[34] M. Sjöberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, and L. Chen. The

mediaeval 2015 affective impact of movies task. In *MediaEval 2015 Workshop*, 2015. 7, 8

[35] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012. 6

[36] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012. 3, 7, 8

[37] S. Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013. 3

[38] G. Trigeorgis, E. Coutinho, F. Ringeval, E. Marchi, S. Zafeiriou, and B. Schuller. The icl-tum-passau approach for the mediaeval 2015 affective impact of movies task. In *Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop, satellite of Interspeech*, volume 2015, 2015. 7

[39] H. L. Wang and L.-F. Cheong. Affective understanding in film. *IEEE Transactions on circuits and systems for video technology*, 16(6):689–704, 2006. 2, 8

[40] S. Wang and Q. Ji. Video affective content analysis: a survey of state-of-the-art methods. *IEEE Transactions on Affective Computing*, 6(4):410–430, 2015. 1, 2

[41] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1083–1092, 2015. 3

[42] C. Xu, D. Tao, and C. Xu. A survey on multi-view learning. *CoRR*, abs/1304.5634, 2013. 3

[43] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi. Multimedia content analysis for emotional characterization of music video clips. *EURASIP J. on Image and Video Processing*, 1(26):1–10, 2013. 2

[44] Y. Yi, H. Wang, B. Zhang, and J. Yu. Mic-tju in mediaeval 2015 affective impact of movies task. In *Working Notes Proceedings of the MediaEval Workshop*, 2015. 7

[45] A. L. Yuille. The convergence of contrastive divergences. *Department of Statistics, UCLA*, 2006. 5

[46] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian. Affective visualization and retrieval for music video. *IEEE Transactions on Multimedia*, 12(6):510–522, 2010. 2