Capturing Dependencies among Labels and Features for Multiple Emotion Tagging of Multimedia Data

Shan Wu,[1] **Shangfei Wang,**[*1] **and Qiang Ji**[2]
[1]Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, China
Email: SA14WS@mail.ustc.edu.cn, sfwang@ustc.edu.cn
[2]Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA
Email: qji@ecse.rpi.edu

Abstract

In this paper, we tackle the problem of emotion tagging of multimedia data by modeling the dependencies among multiple emotions in both the feature and label spaces. These dependencies, which carry crucial top-down and bottom-up evidence for improving multimedia affective content analysis, have not been thoroughly exploited yet. To this end, we propose two hierarchical models that independently and dependently learn the shared features and global semantic relationships among emotion labels to jointly tag multiple emotion labels of multimedia data. Efficient learning and inference algorithms of the proposed models are also developed. Experiments on three benchmark emotion databases demonstrate the superior performance of our methods to existing methods.

Introduction

We are surrounded by digital multimedia collections due to the popularity of the Internet and the proliferation of user-friendly equipment, such as smart phones. Such multimedia collections, including music, images, and videos, have gradually become the most common means of communication, entertainment, and knowledge and information sharing. Naturally, emotion tagging for multimedia data has attracted increasing attention in recent years, since emotion is a key factor in human communication and entertainment.

Automatic emotion annotation for multimedia data is a challenging task due to the complexity and subjectivity of human emotions, and the rich variety of multimedia content. Current research into multimedia emotion tagging mainly focuses on developing discriminative features and classifiers. Various visual and audio features have been adopted. Both static and dynamic classifiers are used for emotion annotation. An extensive review of emotion tagging of music, videos and images can be found in (Yang and Chen 2012; Joshi et al. 2011; Kim et al. 2010; Wang and Ji 2015; Wang and Wang 2005; Dorai and Venkatesh 2001; Wang and He 2008).

Most present research on multimedia emotion tagging assumes there is only one emotional tag for a medium. However, several emotional tags can be assigned to the same multimedia data. For examples, Figure 1(a) lists two shots of a video clip from the FilmStim database, both conveying disgust and fear. Figure 1(b) shows two images that convey happiness, peace, and tenderness. Some emotions often appear together, while others do not. For instance, Figure 1(a) are still shots from a video about war, which may make people fearful and disgusted, but rarely happy. On the contrary, Figure 1(b) describes two beautiful scenes which convey happiness, peace, and tenderness, but not fear and disgust. Therefore, emotion tagging should be formulated as a multi-label classification problem, and successfully exploring the dependency inherent in multiple emotions is the key to improve emotion tagging.

Figure 1: Samples which induce mixed emotions. (a) Two image frames from the FilmStim database induce fear and disgust, but not happiness; (b) Two images from website induce happiness, peace, and tenderness, but not fear and disgust.

However, it is only recently that a few researchers realized that the disjointedness of emotional labels is not valid in emotion detection from music, images and videos. Li and Ogihara (Li and Ogihara 2003) may be the first to formulate emotion detection from music as a multi-label classification problem. They decompose the problem into a set of binary classification problems, and adopt a support vector machine as the classifier for each binary classification. Their method can be regarded as binary relevance (BR) multi-label algorithm, which ignores the dependencies among labels. Later, Trohidis et al. (Trohidis et al. 2011) compare seven multi-label classification algorithms, i.e., binary relevance, label powerset (LP), random k-label sets (RAKEL), multi-label k-nearest neighbor (MLkNN), ranking by pairwise comparison (RPC), calibrated label ranking (CLR), and multi-label back-propagation (BPMLL), for emotion detection from music. LP, RAKEL, RPC, and CLR explore label dependencies from target labels through detecting label combinations, such as a pairwise or subset label combinations in the training data. Thus, it is only feasible for a few combinations, and it is hard to detect thousands of

[*]this is the corresponding author.

possible combinations. In addition, such emotion combinations only capture coexistent emotions. They cannot capture emotions that are mutually exclusive of each other. BPMLL and MLkNN explore label dependencies indirectly with the help of features and hypotheses. They respectively extend back-propagation and k-nearest neighbor to handle multi-label data. With the modified hypothesis, they can model the flexible label dependencies to some extent. However, they do not explore the relations from target labels. All these works demonstrate the potential of multi-label modeling for emotion detection from music. More recently, Wang et al. (Wang, Wang, and Ji 2013) propose a framework of multi-label multimedia emotion tagging for music as well as images and videos. Specifically, they propose a Bayesian network (BN) to automatically capture the label dependencies directly from the target emotion labels, and combine the captured emotion dependencies with their measurements to achieve accurate multi-emotion tagging of multimedia data. However, due to the first-order Markov assumption of BNs, this model can only capture the pairwise dependencies among multiple emotions. Wang et al. (Wang et al. 2015) further propose a three-layer restricted Boltzmann machine model to capture the higher-order relationships among emotion labels.

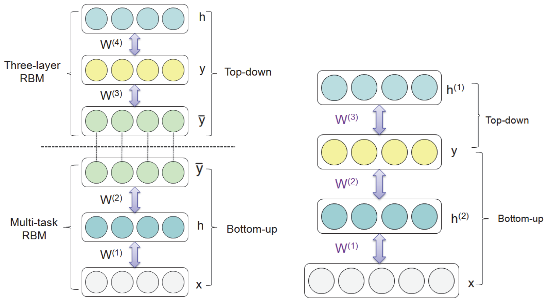
The above studies either address the dependencies among multiple emotions from emotion labels directly, or address them indirectly with the help of features and hypotheses. To the best of our knowledge, no work models multiple emotions' dependencies in both feature and label spaces. Since several emotions can be present in the same medium, the dependencies inherent in target labels and in the shared features among multiple emotions carry crucial top-down and bottom-up evidence (respectively) for improving multimedia emotion tagging, and they have not been thoroughly exploited yet.

To mitigate the limitations of the above methods, we propose two methods to tackle the problem of multimedia emotion tagging by exploiting the relationships of emotions from both shared features and target labels. Two methods are proposed to learn such relationships: independently and dependently. For independent learning, a multi-task Restricted Boltzmann Machine (RBM) classifier (Larochelle and Bengio 2008) is adopted to detect multiple emotions simultaneously by exploiting the relationships embedded in features. A three-layer RBM (Wang et al. 2015) is used to model the high-order dependencies among emotions by parameter learning. Finally, the outputs of the multi-task learning algorithm are used as the inputs to the three-layer RBM to obtain improved multiple emotion tagging. For dependent learning, we propose a new four-layer RBM model to simultaneously model relationships among feature and label spaces. Specifically, the bottom three layers capture the feature relationships, and the top two layers model high-order label dependencies. Experimental results on three multimedia databases demonstrate that multi-task learning outperforms single-task learning, and the relationship model from emotion labels further improves the performance of emotion tagging. Furthermore, it is more effective to learn relationships in both feature and label spaces dependently rather than independently.

To the best of our knowledge, this paper is the first work to assign multiple emotions to multimedia data by exploring the emotional relationships at both feature and label levels. By learning the shared features with a multi-task RBM classifier and modeling the dependencies among emotion labels with a hierarchy RBM model, the proposed approaches can exploit both top-down and bottom-up relations among emotions independently and dependently to improve multiple emotions tagging for multimedia. Furthermore, the proposed approaches also shed a light on the research of multi-label classification, since little work of multi-label classification performs shared feature learning and semantic label dependency capturing simultaneously (Cherman, Monard, and Metz 2011; Huang, Yu, and Zhou 2012; Wang et al. 2014).

Methods

We propose two hierarchical RBM models to capture dependencies in both feature and label spaces. The first method is shown in Figure 2(a), consisting of a multi-task RBM to learn the shared features among emotion labels and obtain label measurements, as well as a three-layer RBM to capture the high-order dependencies among multiple labels and combine measurements with label dependencies. Thus, this model learns the two kind of dependencies separately. Figure 2(b) shows the second model, which captures the dependencies in feature and label spaces simultaneously. The bottom three layer RBM can learn the shared features in a bottom-up manner, and the top two-layers RBM captures global relationships among labels from a top-down direction. By learning the weights between the four layers together, the proposed four-layer RBM captures the label dependencies as well as feature dependencies simultaneously.



(a) (b)

Figure 2: Two proposed methods. (a) Combining a multi-task RBM with a three-layer RBM to capture dependencies among features and labels independently. (b) Capturing dependencies among features and labels dependently.

Capturing dependencies among features and labels independently

Multi-Task RBM The multi-task RBM (Larochelle and Bengio 2008), as shown in the bottom part of Figure 2(a), is adopted to learn the shared features among different labels and obtain measurements of multiple emotions simultaneously. **h** represents the hidden nodes, $\bar{\mathbf{y}}$ is predicted labels, and **x** is the features which are continuous variables in our work. $\mathbf{W}^{(1)}$ connects the features to hidden nodes modeling the relationships among features and captures the commonalities among different labels. $\mathbf{W}^{(2)}$ connects the hidden nodes to labels capturing the variations among different labels. The hidden layer **h** captures not only the dependencies among features, more importantly, it represents the salient information for the input features **x**. In other words, **h** learns a feature representation of **x**, which serves as the input to a multi-task classifier to estimate image labels, producing $\bar{\mathbf{y}}$.

The total energy of this model is defined as follows,

$$E(\bar{\mathbf{y}}, \mathbf{x}, \mathbf{h}; \Theta) = -\sum_i b_i h_i - \sum_j c_j \bar{y}_j - \sum_i \sum_j h_i W_{ij}^{(2)} \bar{y}_j - \sum_i \sum_k h_i W_{ik}^{(1)} \frac{x_k}{\sigma_k} + \frac{1}{2} \sum_k \frac{(x_k - a_k)^2}{\sigma_k^2} \tag{1}$$

where $\Theta = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{a}, \mathbf{b}, \mathbf{c}\}$ represents the parameters, $\{a_k\}$, $\{b_i\}$ and $\{c_j\}$ are the biases of the inputs, hidden nodes and target classes respectively, and σ_k is the standard deviation of the Gaussian noise for x_k.

The joint distribution of the inputs and target labels of RBM is calculated by marginalizing over all the hidden units as shown in Equation 2, where $Z(\Theta) = \sum_{\mathbf{h},\mathbf{x},\bar{\mathbf{y}}} exp(-E(\bar{\mathbf{y}}, \mathbf{x}, \mathbf{h}; \Theta))$ is the partition function.

$$p(\bar{\mathbf{y}}, \mathbf{x}; \Theta) = \frac{\sum_h exp(-E(\bar{\mathbf{y}}, \mathbf{x}, \mathbf{h}; \Theta))}{Z(\Theta)} \tag{2}$$

Given the training set $D = \{(\mathbf{x}^{(i)}, \bar{\mathbf{y}}^{(i)})\}_{i=1}^N$, the parameters Θ are obtained by maximizing the log likelihood L which is defined by Equation 3, where N represents the total number of training samples.

$$\Theta^* = \underset{\Theta}{argmax}\, L(\Theta) = \underset{\Theta}{argmax}\, \frac{1}{N} \sum_{i=1}^{N} \log p(\bar{\mathbf{y}}^{(i)}, \mathbf{x}^{(i)}; \Theta) \tag{3}$$

The gradient of L with respect to any parameters $\theta \in \Theta$ can be calculated using Equation 4,

$$\frac{\partial L(\Theta)}{\partial \theta} = \langle \frac{\partial E}{\partial \theta} \rangle_{p(\bar{\mathbf{y}}, \mathbf{x}, \mathbf{h}; \Theta)} - \langle \frac{\partial E}{\partial \theta} \rangle_{p(\bar{\mathbf{y}} | \mathbf{x}, \mathbf{h}; \Theta)} \tag{4}$$

where $\langle \cdot \rangle_p$ represents the expectation over distribution p. The contrastive divergence algorithm (Larochelle and Bengio 2008) is adopted to learn the parameters.

After parameter learning, we predict the label by choosing the most probable state of $\bar{\mathbf{y}}^*$ under $p(\bar{\mathbf{y}}|\mathbf{x}; \Theta^*)$ as shown in Equation 5.

$$\bar{\mathbf{y}}^* = arg \max_{\bar{\mathbf{y}}} p(\bar{\mathbf{y}}|\mathbf{x}; \Theta) \tag{5}$$

where the posterior probability is defined as follows:

$$p(\bar{\mathbf{y}}|\mathbf{x}; \Theta) = \frac{e^{\mathbf{c}^T \bar{\mathbf{y}}} \prod_i (1 + e^{b_i + \sum_k W_{ik}^{(1)} x_k + \sum_j W_{ij}^{(2)} \bar{y}_j})}{\sum_{\bar{y}^\star} e^{\mathbf{c}^T \bar{\mathbf{y}}^\star} \prod_i (1 + e^{b_i + \sum_k W_{ik}^{(1)} x_k + \sum_j W_{ij}^{(2)} \bar{y}_j^\star})} \tag{6}$$

Since the computation cost of Equation 6 is exponential to the number of emotion labels, Gibbs sampling is adopted.

Three-Layer RBM We adopt a three-layer RBM model (Wang et al. 2015) (shown in the top part of Figure 2(a)) to capture the high-order semantic relationships among labels, and then infer the emotion labels by combining label dependencies with label measurements. The bottom layer is the measurements $\bar{\mathbf{y}}$ obtained from the multi-task RBM. The middle layer is the ground-truth **y** of emotion labels. The hidden layer **h** captures the high-order dependencies among the ground-truth labels **y**. Using the estimated $\bar{\mathbf{y}}$ from the multi-task RBM model, this three-layer RBM estimates the truth (**y**), subject to the label dependencies encoded in **h**.

The energy function of this model is defined in Equation 7,

$$E(\mathbf{y}, \bar{\mathbf{y}}, \mathbf{h}; \Theta) = -\mathbf{b}^T \mathbf{y} - \mathbf{c}^T \mathbf{h} - \mathbf{h}^T \mathbf{W}^{(4)} \mathbf{y} - \mathbf{y}^T \mathbf{W}^{(3)} \bar{\mathbf{y}} \tag{7}$$

where **b**, **c** represents the biases of target labels **y** and hidden units **h**, $\mathbf{W}^{(3)}$ measures the compatibility between measurements and target classes, and $\mathbf{W}^{(4)}$ captures the high-order dependencies among labels.

The parameters of the three-layer RBM model can be learned by maximizing the log conditional likelihood defined in Equation 8.

$$\Theta^* = \underset{\Theta}{argmax}\, L(\Theta) = \underset{\Theta}{argmax} \log p(\mathbf{y}|\bar{\mathbf{y}}; \Theta) \tag{8}$$

The gradient of the log likelihood function L is described in Equation 9.

$$\frac{\partial L(\Theta)}{\partial \theta} = \langle \frac{\partial E}{\partial \theta} \rangle_{p(\mathbf{h}, \mathbf{y} | \bar{\mathbf{y}}; \Theta)} - \langle \frac{\partial E}{\partial \theta} \rangle_{p(\mathbf{h} | \mathbf{y}, \bar{\mathbf{y}}; \Theta)} \tag{9}$$

A revised contrastive divergence algorithm is adopted to learn the parameters (Wang et al. 2013). After parameter learning, the emotion labels can be inferred by maximizing the posterior probability $p(\mathbf{y}|\bar{\mathbf{y}}; \Theta)$ according to

$$\mathbf{y}^* = arg \max_{\mathbf{y}} p(\mathbf{y}|\bar{\mathbf{y}}; \Theta) \tag{10}$$

The Gibbs sampling method is adopted for inference. Detailed learning and inference algorithms can be found in (Wang et al. 2015; 2013).

Capturing dependencies among features and labels dependently

We propose a novel four-layer RBM to capture the label and feature dependencies jointly as shown in Figure 2(b). It consists of two hidden layers. The first hidden layer $\mathbf{h}^{(1)}$ captures the high-order dependencies among emotion labels **y**. The second hidden layer $\mathbf{h}^{(2)}$ models the relationships between input features **x** and different emotion labels **y**. $\mathbf{W}^{(1)}$ connects the features to the hidden layer $\mathbf{h}^{(2)}$, modeling the feature commonalities among multiple labels. $\mathbf{W}^{(2)}$ connects the hidden layer $\mathbf{h}^{(2)}$ to labels, capturing the variations among labels. $\mathbf{W}^{(3)}$ connects labels to the hidden layer $\mathbf{h}^{(1)}$, modeling the high-order dependency among the target labels.

The total energy function of this model is defined in Equation 11,

$$
\begin{aligned}
&E(\mathbf{y},\mathbf{x},\mathbf{h}^{(1)},\mathbf{h}^{(2)};\Theta) = \\
&-\sum_i h_i^{(2)} b_i - \sum_j y_j c_j - \sum_t h_t^{(1)} d_t \\
&+\frac{1}{2}\sum_k \frac{(x_k - a_k)^2}{\sigma_k^2} - \sum_i \sum_k h_i^{(2)} W_{ik}^{(1)} \frac{x_k}{\sigma_k} \\
&-\sum_i \sum_j h_i^{(2)} W_{ij}^{(2)} y_j - \sum_t \sum_j h_t^{(1)} W_{tj}^{(3)} y_j
\end{aligned}
\tag{11}
$$

where $\{a_k\}, \{b_i\}, \{c_j\}$ and $\{d_t\}$ are the biases of the inputs, the hidden layer $\mathbf{h}^{(2)}$, target labels and the hidden layer $\mathbf{h}^{(1)}$ respectively, and σ_k is the standard deviation of the Gaussian noise for x_k. The joint distribution of the inputs and target classes is shown in Equation 12, where $Z(\Theta)$ is the partition function.

$$
p(\mathbf{y},\mathbf{x};\Theta) = \frac{\sum_{\mathbf{h}^{(1)}} \sum_{\mathbf{h}^{(2)}} exp(-E(\mathbf{y},\mathbf{x},\mathbf{h}^{(1)},\mathbf{h}^{(2)};\Theta))}{Z(\Theta)}
\tag{12}
$$

The proposed four-layer RBM model can be decomposed into two parts. The top two-layer RBM captures the high-order dependencies among emotion labels through the hidden layer $\mathbf{h}^{(1)}$. The bottom three-layer RBM constructs a better feature representation for emotion recognition through the latent units $\mathbf{h}^{(2)}$, which connect features to multiple labels. As a whole, the four-layer RBM model captures the global semantic relationships among emotion labels and the shared features simultaneously.

To learn parameters of the four-layer RBM model, we attempt to maximize the log likelihood as shown in Equation 13.

$$
\Theta^* = \underset{\Theta}{argmax}\, L(\Theta) = \underset{\Theta}{argmax} \log p(\mathbf{y},\mathbf{x};\Theta)
\tag{13}
$$

To update the parameters, a stochastic gradient descent method is applied. The gradient can be calculated using Equation 14. We employ the contrastive divergence algorithm to obtain the gradient, as shown in Algorithm 1.

$$
\frac{\partial L(\Theta)}{\partial \theta} = \langle \frac{\partial E}{\partial \theta} \rangle_{p(\mathbf{h}^{(1)},\mathbf{h}^{(2)},\mathbf{y},\mathbf{x};\Theta)} - \langle \frac{\partial E}{\partial \theta} \rangle_{p(\mathbf{h}^{(1)},\mathbf{h}^{(2)}|\mathbf{y},\mathbf{x};\Theta)}
\tag{14}
$$

After parameter learning, the emotion labels $\mathbf{y}^*$ of a test sample can be inferred according to Equation 15.

$$
\mathbf{y}^* = \underset{\mathbf{y}}{argmax}\, p(\mathbf{y}|\mathbf{x};\Theta)
\tag{15}
$$

where the posterior probability is defined as follows:

$$
p(\mathbf{y}|\mathbf{x};\Theta) = \frac{exp(\mathbf{c}^T\mathbf{y}) \prod_t (1 + exp(\alpha_{yt})) \prod_i (1 + exp(\beta_{yi}))}{\sum_{\mathbf{y}^\star exp(\mathbf{c}^T \mathbf{y}^\star) \prod_t (1+exp(\alpha_{y^\star t})) \prod_i (1+exp(\beta_{y^\star i}))}}
\tag{16}
$$

Where $\alpha_{yt} = d_t + \sum_j h_t^{(1)} W_{tj}^{(3)} y_j$, $\beta_{yi} = b_i + \sum_j h_i^{(2)} W_{ij}^{(2)} y_j + \sum_k h_i^{(2)} W_{ik}^{(1)} x_k$.

Algorithm 1 Four-Layer RBM Learning by Contrastive Divergence algorithm.

```
Input: Training data($\mathbf{x}^{(i)}, \mathbf{y}^{(i)}$),learning rate $\lambda$.
Output: Model parameters $\Theta = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{W}^{(3)}, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}\}$.
  repeat
    for each training instance $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ do
      % Positive Phase
      $\mathbf{y}^0 \leftarrow \mathbf{y}^{(i)}, \mathbf{x}^0 \leftarrow \mathbf{x}^{(i)}$,
      $\hat{\mathbf{h}}^{(1)0} \leftarrow sigmoid(d_t + \sum_j W_{tj}^{(3)} y_j^0)$
      $\hat{\mathbf{h}}^{(2)0} \leftarrow sigmoid(b_i + \sum_j W_{ij}^{(2)} y_j^0 + \sum_k W_{ik}^{(1)} x_k^0)$
      % Negative Phase
      $\mathbf{h}^{(1)0} \sim sigmoid(d_t + \sum_j W_{tj}^{(3)} y_j^0)$
      $\mathbf{h}^{(2)0} \sim sigmoid(b_i + \sum_j W_{ij}^{(2)} y_j^0 + \sum_k W_{ik}^{(1)} x_k^0)$
      $\mathbf{y}^1 \sim sigmoid(c_j + \sum_i W_{ji}^{(2)} h_i^{(2)0} + \sum_t W_{jt}^{(3)} h_t^{(1)0})$
      $\mathbf{x}^1 \sim p(\mathbf{x}|\mathbf{h}^{(2)0})$
      $\hat{\mathbf{h}}^{(1)1} \leftarrow sigmoid(d_t + \sum_j W_{tj}^{(3)} y_j^1)$
      $\hat{\mathbf{h}}^{(2)1} \leftarrow sigmoid(b_i + \sum_j W_{ij}^{(2)} y_j^1 + \sum_k W_{ik}^{(1)} x_k^1)$
      %Update
      for $\theta \in \Theta$ do do
        $\theta \leftarrow \theta - \lambda(\frac{\partial}{\partial\theta} E(\mathbf{y}^0, \mathbf{x}^0, \hat{\mathbf{h}}^{(1)0}, \hat{\mathbf{h}}^{(2)0}) - \frac{\partial}{\partial\theta} E(\mathbf{y}^1, \mathbf{x}^1, \hat{\mathbf{h}}^{(1)1}, \hat{\mathbf{h}}^{(2)1}))$
      end for
    end for
  until Converges
```

The computational complexity of calculating the posterior probability directly is exponential to the number of target labels. Therefore, for a small number of emotion labels, we may directly calculate the posterior probability to obtain the final class according to Equation 16. Otherwise, for a large number of emotion labels, we use the Gibbs sampling method. The detailed steps are presented in Algorithm 2.

Algorithm 2 Four-Layer RBM Inference by Gibbs Sampling

```
Input:
  test sample $\mathbf{x}$,model parameters $\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{W}^{(3)}, \mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$.
Output: $\mathbf{y}^*$
  method 1: Apply Equation 15 to find $\mathbf{y}^*$ by maximizing the posterior probability.
  method 2: Use Gibbs Sampling to find $\mathbf{y}^*$
  repeat
    for $chain = 1 \rightarrow C$ do do
      randomly initialize $\mathbf{y}^0$
      for $r = 0 \rightarrow M$ do do
        $\mathbf{h}^{(1)r} \sim sigmoid(d_t + \sum_j W_{tj}^{(3)} y_j^r)$
        $\mathbf{h}^{(2)r} \sim sigmoid(b_i + \sum_j W_{ij}^{(2)} y_j^r + \sum_k W_{ik}^{(1)} x_k)$
        $\mathbf{y}^{r+1} \sim sigmoid(c_j + \sum_i W_{ji}^{(2)} h_i^{(2)r} + \sum_t W_{jt}^{(3)} h_t^{(1)r})$
      end for
    end for
    for $j = 1 \rightarrow n$ do do
      collect the last K samples of $y_j$ from each chain
      calculate $p(y_j|\mathbf{x})$ based on the collected samples
    end for
  until Converges
```

Experiment

Experimental Conditions

As of now, there are a few multimedia databases which contain multiple emotion labels. In our work, we employ three data sets: the Music database (Trohidis et al. 2008), the NVIE database (Wang et al. 2010) and the FilmStim database (Schaefer et al. 2010).

The Music Emotion database consists of 593 songs which can be classified into six emotion labels, i.e., amazement, happiness, relaxation, quietness, sadness and anger. The number of samples for each emotion are 173, 166, 264, 148, 168 and 189 respectively. Due to copyright issues, the music clips of this database are not available. However, the database provides 8 rhythmic features and 64 timbre features of each sample. Detailed information can be find in (Trohidis et al. 2008). In our work, all of these 72 features are employed.

The NVIE video database contains 72 videos with 7 emotion labels, i.e., happiness, anger, sadness, fear, disgust, surprise and valance. These 72 videos were used as emotion induced videos during the construction of the NVIE expression database (Wang et al. 2010). The number of samples for these seven emotions are 28, 12, 17, 34, 29, 27 and 29 respectively. The constructors of this database do not provide features. We extract three visual features, i.e., lighting key, energy color and visual excitement, and 31 audio features, including average energy, average energy intensity, spectrum flux, Zero Crossing Rate (ZCR), standard deviation of ZCR, 13 MFCCs, log energy of MFCCs and the standard deviations of the above 13 MFCCs.

The FilmStim database includes 64 videos with six emotions (tenderness, fear, anger, joy, sadness and disgust). There are 25, 36, 21, 25, 27 and 37 samples for these six emotions, respectively. Since the database does not include the features, we extracted the same visual and audio features as those of the NVIE video database.

To validate the effectiveness of our proposed models in capturing the relationships in both feature and label spaces, four methods are used to recognize emotion labels from multimedia data: the single-task RBM, which is similar to the multi-task RBM described in the methods Section except that it contains only one label for each model; the multi-task RBM; our first proposed model; and our second proposed model. Ten-fold cross validation is adopted. Both example-based (i.e. accuracy, F1-measures and subset accuracy) and label-based (i.e. micro and macro F1-measures) multi-label evaluation measures are used. Detailed definitions of these evaluation measures can be found in (Sorower 2010).

Experimental Results of Emotion Tagging

The experimental results on three multimedia databases are summarized in Table 1. From this table, we can obtain the following observations:

First, multi-task RBMs significantly outperform single-task RBMs, since the results of multi-task are consistently higher than single-task's on three databases. Both Example-based and label-based metrics are improved by using a multi-task RBM to learn the shared features. This demonstrates that a multi-task RBM can make recognition results not only more accurate but also more balanced than a single-task RBM does.

Secondly, on all databases, the experimental results of our proposed two methods are better than those of multi-task RBMs in all the cases. This demonstrates the importance of the semantic relationships among emotions for multiple emotion tagging, and the power of our proposed two models in capturing shared features and high-order label dependencies.

Table 1: Results of our approach on three databases.

Database	Method	example-based		label-based	
		Acc.	F1.	MicF1.	MacF1.
Music	single-task(BR)	0.546	0.629	0.670	0.659
	multi-task RBM	0.577	0.662	0.690	0.684
	Our method 1	0.584	0.666	0.693	0.687
	Our method 2	**0.585**	**0.668**	**0.695**	**0.688**
	SVM(BR)	0.514	0.593	0.650	0.624
	BN	0.552	0.629	0.660	-
	Wang et al.	0.554	0.645	0.675	0.676
	Andreas et al.	0.510	0.580	0.650	0.630
NVIE	single-task(BR)	0.457	0.520	0.563	0.479
	multi-task RBM	0.495	0.562	0.581	0.501
	Our method 1	0.502	0.583	0.608	0.536
	Our method 2	**0.562**	**0.621**	**0.642**	**0.607**
	SVM(BR)	0.335	0.413	0.476	0.381
	BN	0.427	0.488	0.487	-
	Wang et al.	0.318	0.443	0.493	0.511
	Andreas et al.	0.480	0.570	0.580	0.550
FilmStim	single-task(BR)	0.353	0.465	0.514	0.476
	multi-task RBM	0.394	0.486	0.519	0.495
	Our method 1	0.422	0.521	0.548	0.535
	Our method 2	**0.437**	**0.528**	**0.568**	**0.551**
	SVM(BR)	0.286	0.370	0.429	0.294
	BN	0.329	0.413	0.457	0.263
	Wang et al.	0.382	0.526	0.555	0.528
	Andreas et al.	0.440	0.530	0.570	0.540

"Acc." refers to " accuracy," "F1." refers to "F1 score," "MicF1." refers to "micro F1 score," "MacF1." refers to "macro F1 score."

Table 2: Examples of tagging results.

Database	FilmStim		Music	
ground-truth	fear, anger, sadness, disgust	fear,anger, joy,disgust	quietness, sadness	happiness, relaxation
multi-task RBM	tenderness, joy	**fear**, sadness, **disgust**	relaxation, **quietness**	**relaxation**, sadness
method1	tenderness, joy,**sadness**	**fear**, **disgust**	relaxation, **quietness**, **sadness**	**relaxation**
method2	**fear, anger**, **sadness**, **disgust**	**fear, anger**, sadness, **disgust**	**quietness**, **sadness**	**happiness**, **relaxation**

Finally, comparing the first proposed method and the second proposed method, we find our second method outperforms the first method in most cases. Since four-layer RBM models the relationships in feature and label spaces dependently, it may capture a more comprehensive breadth of relationships than the first method.

The proposed two methods outperform not only single-task RBMs, but also multi-task RBMs, since both methods successfully capture the dependencies among features and emotion labels. Table 2 lists several samples from video database and audio database to further demonstrate the superiority of our methods. For example, when a video evokes complicated emotions, i.e., fear, sadness, anger, disgust, our methods are superior to the multi-task RBM, since the results of our methods are on target while the multi-task RBM is a poor match. Similarly, for a music evoking quietness and sadness, the tagging results of our methods are almost correct, especially method 2. This demonstrates the effectiveness of our methods in capturing dependencies among features and emotions. What's more, we find that method 2 always performs better than method 1. The hidden layer $\mathbf{h}^{(2)}$ in the multi-task part in Figure 2(b) plays the same role as the hidden layer $\mathbf{h}$ in the multi-task RBM in Figure 2(a). However, their values are different after training, since training for Figure 2(b) is done jointly for all layers with interactions among layers, while training for the top and bottom parts of the model in Figure 2(a) are done separately without any interaction between layers. Hence, method 2 can model more comprehensive relationships than method 1.

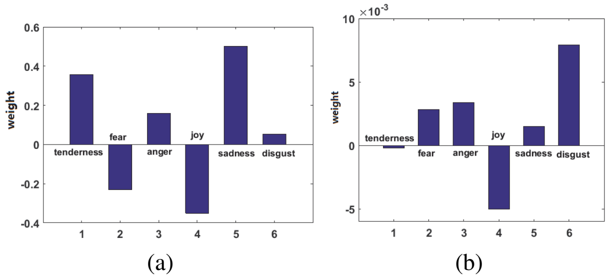
Semantic Relationship Analysis

In this section, we further analyze the semantic relationships among multiple emotions captured by the top hidden units of our two methods, using the FilmStim database as an example. As discussed in previous section, the semantic relationships captured by the two methods are different, since training for Figure 2(b) is done jointly for all layers with interactions among layers, while training for the top and bottom parts of the model in Figure 2(a) is done separately without any interaction between layers. The semantic pattern is measured by the weight between latent unit and emotion label. Larger weights represent a higher probability of presence, while smaller weights denote a higher probability of absence. For example, Figure 3(a) and Figure 3(b) show two patterns captured by two methods in the FilmStim database. Figure 3(a) demonstrates that tenderness, anger, sadness and disgust may appear together, but not fear and joy. Figure 3(b) shows that fear, anger, sadness and disgust may appear simultaneously, but not joy. Through further analyzing the global emotion relationships presenting in the ground-truth labels on the FilmStim database with correlation coefficients, we find that fear, anger, sadness and disgust tend to co-occur, and these four emotions and joy are mutually exclusive, which is successfully captured by method2.

Comparison with Related Work

There are only a few related works which address multiple emotion tagging for multimedia data. Wang et al. (Wang et al. 2015) adopted the same databases as ours. Therefore we directly compare our work with theirs as shown in Table 1, from which we find the following:

First, our approach using a BR(single-task RBM) outperforms Wang et al.'s SVM(feature-driven method), since all four example-based and label-based evaluation metrics of BR are much higher than SVM. Both the SVM in Wang et al.'s work and the BR used in our work recognize each emotion label independently, without considering relationships among emotions. The single-task RBM recognizes emotions from a new feature representation introduced by its hidden units, while the feature-driven method recognizes emotions from the visual-audio features directly. Therefore, the good performance of the single-task RBM may demonstrate the effectiveness of the hidden nodes of RBM for effective feature representation.



(a) (b)

Figure 3: Semantic relationships on FilmStim database captured by a top hidden-layer unit. (a) method1 (b) method2

Secondly, the performance of the multi-task classifier is better than that of Bayesian Network (BN). BN captures pair-wise dependencies among emotion labels, while multi-task RBM learns the commonalities and variations among emotions based on features. These results may demonstrate that feature dependencies are more powerful than pair-wise label dependencies for multiple emotion tagging.

Thirdly, our two proposed methods are superior to Wang et al.s (Wang et al. 2015) with both better example-based and label-based evaluation measures. Since Wang et al. already demonstrate the superiority of their proposed methods to current multi-label classifiers, including BPMLL and MLkNN. in their emotion tagging experiments, our proposed methods outperform current multi-label classifiers for multimedia emotion tagging.

In addition, we compare our methods to current multi-task methods. We use multi-task SVM proposed by Andreas et al. (Evgeniou and Pontil 2007) as examples. From Table 1, we find that our multi-task RBM achieves comparable results to multi-task SVM, and the proposed two methods outperform multi-task SVM in most cases. It further demonstrates the importance to capture the dependencies among both labels and features for emotion tagging.

Conclusion and further work

Although both feature dependencies and label dependencies are crucial for emotion tagging of multimedia data, little work addresses them simultaneously until now. In this paper, we propose two hierarchical models to systematically integrate the commonalities and variations across multiple emotions from shared features and the high-order dependencies from emotion labels for multimedia emotion tagging. Experimental results on three benchmark databases demonstrate the superiority of our proposed approaches to state-of-the-art methods due to their power in capturing emotion relationships from both features and labels. Furthermore, our proposed models can easily be adapted to other applications that involve multiple related outputs, such as facial action unit recognition, semantic scene segmentation and image annotation.

Acknowledgment

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

References

Cherman, E. A.; Monard, M. C.; and Metz, J. 2011. Multi-label problem transformation methods: a case study. *CLEI Electronic Journal* 14(1):4–4.

Dorai, C., and Venkatesh, S. 2001. Computational media aesthetics: Finding meaning beautiful. *IEEE multimedia* 8(4):10–12.

Evgeniou, A., and Pontil, M. 2007. Multi-task feature learning. *Advances in neural information processing systems* 19:41.

Huang, S.-J.; Yu, Y.; and Zhou, Z.-H. 2012. Multi-label hypothesis reuse. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, 525–533. New York, NY, USA: ACM.

Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q.-T.; Wang, J. Z.; Li, J.; and Luo, J. 2011. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE* 28(5):94–115.

Kim, Y. E.; Schmidt, E. M.; Migneco, R.; Morton, B. G.; Richardson, P.; Scott, J.; Speck, J. A.; and Turnbull, D. 2010. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, 255–266. Citeseer.

Larochelle, H., and Bengio, Y. 2008. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th international conference on Machine learning*, 536–543. ACM.

Li, T., and Ogihara, M. 2003. Detecting emotion in music. In *Proceedings of the International Symposium on Music Information Retrieval*, 239C240.

Schaefer, A.; Nils, F.; Sanchez, X.; and Philippot, P. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition & Emotion* 24(7):1153–1172.

Sorower, M. S. 2010. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*.

Trohidis, K.; Tsoumakas, G.; Kalliris, G.; and Vlahavas, I. P. 2008. Multi-label classification of music into emotions. In *ISMIR*, volume 8, 325–330.

Trohidis, K.; Tsoumakas, G.; Kalliris, G.; and Vlahavas, I. 2011. Multi-label classification of music by emotion. *EURASIP Journal on Audio, Speech, and Music Processing* 4(doi:10.1186/1687-4722-2011-426793).

Wang, W., and He, Q. 2008. A survey on emotional semantic image retrieval. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, 117–120. IEEE.

Wang, S., and Ji, Q. 2015. Video affective content analysis: a survey of state of the art methods. *Affective Computing, IEEE Transactions on* PP(99):1–1.

Wang, S., and Wang, X. 2005. Emotion semantics image retrieval: An brief overview. In *ACII 2005*, 490–497.

Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Chen, F.; and Wang, X. 2010. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia* 12(7):682–691.

Wang, Z.; Li, Y.; Wang, S.; and Ji, Q. 2013. Capturing global semantic relationships for facial action unit recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 3304–3311. IEEE.

Wang, S.; Wang, J.; Wang, Z.; and Ji, Q. 2014. Enhancing multi-label classification by modeling dependencies among labels. *Pattern Recognition* 47(10):3405 – 3413.

Wang, S.; Wang, J.; Wang, Z.; and Ji, Q. 2015. Multiple emotion tagging for multimedia data by exploiting high-order dependencies among emotions. *Multimedia, IEEE Transactions on* PP(99):1–1.

Wang, S.; Wang, Z.; and Ji, Q. 2013. Multiple emotional tagging of multimedia data by exploiting dependencies among emotions. *Multimedia Tools and Applications* 74(6):1863–1883.

Yang, Y.-H., and Chen, H. H. 2012. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3(3):40.