Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 267478, 12 pages
http://dx.doi.org/10.1155/2015/267478

Research Article

Topologically Ordered Feature Extraction Based on Sparse Group Restricted Boltzmann Machines

Zhong Chen,[1] Shengwu Xiong,[1] Zhixiang Fang,[2,3] Ruiling Zhang,[1,4] Xiangzhen Kong,[1] and Yi Rong[1]

[1] *School of Computer Science and Technology, Wuhan University of Technology, 122 Luoshi Road, Wuhan 430070, China*
[2] *State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China*
[3] *Engineering Research Center for Spatio-Temporal Data Smart Acquisition and Application, Ministry of Education of China, Wuhan University, 129 Luoyu Road, Wuhan 430079, China*
[4] *Institute of Information Technology, Luoyang Normal University, 71 Luolong Road, Luoyang 471022, China*

Correspondence should be addressed to Zhixiang Fang; zxfang@whu.edu.cn

Received 20 March 2015; Revised 28 July 2015; Accepted 9 September 2015

Academic Editor: Panos Liatsis

How to extract topologically ordered features efficiently from high-dimensional data is an important problem of unsupervised feature learning domains for deep learning. To address this problem, we propose a new type of regularization for Restricted Boltzmann Machines (RBMs). Adding two extra terms in the log-likelihood function to penalize the group weights and topologically ordered factors, this type of regularization extracts topologically ordered features based on sparse group Restricted Boltzmann Machines (SGRBMs). Therefore, it encourages an RBM to learn a much smoother probability distribution because its formulations turn out to be a combination of the group weight-decay and topologically ordered factor regularizations. We apply this proposed regularization scheme to image datasets of natural images and Flying Apsara images in the Dunhuang Grotto Murals at four different historical periods. The experimental results demonstrate that the combination of these two extra terms in the log-likelihood function helps to extract more discriminative features with much sparser and more aggregative hidden activation probabilities.

1. Introduction

Restricted Boltzmann Machines (RBMs) [1] are a type of product of experts model [2] based on Boltzmann Machines [3] but with a complete bipartite interaction graph. In general, RBMs, which are used as generative models to simulate input distributions of binary data [4], are viewed as an effective feature-representation approach for extracting structured information from input data. They have received much attention recently and have been successfully applied in various application domains, such as dimensionality reduction [5], object recognition [6], topic modeling [7], and feature learning [8]. In addition, RBMs have attracted much attention as building blocks for the multilayer learning systems (e.g., Deep Belief Networks (DBNs), Deep Boltzmann Machines (DBMs)), and variants and extensions of RBMs have a great many applications in a wide range of feature learning and pattern recognition tasks.

Due to the arbitrary connectivity of Boltzmann machines, they are too slow to be practical, and in order to obtain efficient and exact results, RBMs have the restrictions that there are no visible-visible or hidden-hidden connections, which leads to the obvious advantage that inferences in the RBMs are much easier than in Boltzmann Machines [9]. Therefore, the hidden units are conditionally independent and we may generate a more powerful learning model [10]. Lee et al. [11] proposed sparse RBMs (SRBMs) by pointing out that RBMs tend to learn distributed and nonsparse representations as the number of hidden units is increased; accordingly, they added a regularization term that penalized

a deviation of the expected activation with a low level to ensure that the hidden units would be sparsely activated. Moreover, in order to group similar activations of the hidden units and capture their local dependencies, Luo et al. [12] proposed sparse group RBMs (SGRBMs) using a novel regularization of the activation probabilities of the hidden units in RBMs. What SRBMs and SGRBMs have in common is that they have adopted sparsity to promote regularization, making them powerful enough to represent complicated distributions.

By introducing the $L_{1,2}$ regularizer into the activation probabilities of the hidden units, the SGRBMs have the following two properties: first, this model encourages few groups to be active when given observed data (this property yields sparsity at group level), and second, it results in only a few hidden units being active in a group (this property yields sparsity within the group). However, they did not consider overfitting problems, which lack corresponding strategies for controlling the reconstruction complexity of the weight matrix. In addition, they did not take into account the fact that all the extracted features in the hidden units are not topologically ordered (i.e., similar features are grouped together while they do not simultaneously discard group sparsity), and it is essential for a learning machine to obtain structured information from the input data. In 2002, Welling et al. [13] proposed a novel learning sparse topographic representation with products of Student t-distributions and found that if the Student t-distribution is used to model the combined outputs of sets of neutrally adjacent filters, then the orientation, spatial frequency, and location of the filters change smoothly across the topographic map. Later, Goh et al. [14] proposed a method for regularizing RBMs during training to obtain features that are sparse and topographically organized. The features learned are then Gabor-like and demonstrate a coding for orientation, spatial position, frequency, and color that vary smoothly with the topography of the feature map. For the purpose of efficiently extracting invariant features with group sparsity from high-dimensional data, in this paper, firstly we adopted a weight-decay strategy [15, 16] at group level based on SGRBMs, and secondly, by adding an extra term to penalize the topologically ordered factors in the log-likelihood function, the topologically ordered features at group level can be obtained.

The remaining sections of this paper are organized as follows. In Section 2, RBMs and Contrastive Divergence algorithms for RBM training are described in brief. In Section 3, a nontopologically ordered feature extraction approach is proposed to obtain sparse but not topologically ordered features between groups from the input data. In Section 4, a topologically ordered feature extraction approach is proposed to obtain structured information (i.e., sparse and topologically ordered features between the overlapping groups) from the input data. In Section 5, experimental results with two different datasets (namely, natural images and Flying Apsara images in the Dunhuang Grotto Murals) are shown to validate the proposed approach. Finally, the conclusions are in Section 6.

2. Restricted Boltzmann Machines and Contrastive Divergence

RBMs are a particular form of the Markov Random Field (MRF) model and are regarded as an undirected generative model which uses a layer of binary hidden units to model a distribution over binary visible units [17]. Suppose an RBM consists of n visible units $\mathbf{v} = (v_1, v_2, \ldots, v_n) \in \{0,1\}^n$ representing the input data and m hidden units $\mathbf{h} = (h_1, h_2, \ldots, h_m) \in \{0,1\}^m$ to capture the features of the input data. The joint probability distribution $P_{\boldsymbol{\theta}}(\mathbf{v}, \mathbf{h})$ is given by the Gibbs distribution with the energy function [17, 18]:

$$\begin{aligned} E_{\boldsymbol{\theta}}(\mathbf{v}, \mathbf{h}) &= -\mathbf{v}^T \mathbf{W}\mathbf{h} - \mathbf{b}^T \mathbf{v} - \mathbf{c}^T \mathbf{h}, \\ P_{\boldsymbol{\theta}}(\mathbf{v}, \mathbf{h}) &= \frac{1}{Z(\boldsymbol{\theta})} \exp\left(-E_{\boldsymbol{\theta}}(\mathbf{v}, \mathbf{h})\right), \end{aligned} \tag{1}$$

where $\mathbf{W} \in R^{n \times m}$ is the matrix of weights and $\mathbf{b} \in R^n$ and $\mathbf{c} \in R^m$ are vectors which represent the visible and hidden biases, respectively. All these are referred to as the RBM parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$, and $E_{\boldsymbol{\theta}}(\mathbf{v}, \mathbf{h})$ is the energy function and $Z(\boldsymbol{\theta}) = \sum_{\tilde{\mathbf{v}}, \tilde{\mathbf{h}}} \exp(-E_{\boldsymbol{\theta}}(\tilde{\mathbf{v}}, \tilde{\mathbf{h}}))$ is a corresponding normalized constant. Therefore, the marginal distribution of visible variables becomes

$$\begin{aligned} P_{\boldsymbol{\theta}}(\mathbf{v}) &= \sum_{\mathbf{h}} P_{\boldsymbol{\theta}}(\mathbf{v}, \mathbf{h}) \\ &= \frac{1}{Z(\boldsymbol{\theta})} \cdot \exp\left(\mathbf{b}^T \mathbf{v}\right) \\ &\quad \cdot \prod_{j=1}^{m} \left(1 + \exp\left(\mathbf{v}^T \mathbf{W}_{\cdot j} + c_j\right)\right). \end{aligned} \tag{2}$$

As the hidden units are independent given the states of the visible units and vice versa, when given the observed data, the conditional probabilities and conditional distributions of the hidden units are

$$\begin{aligned} P_{\boldsymbol{\theta}}\left(h_j = 1 \mid \mathbf{v}\right) &= \operatorname{sig}\left(\mathbf{v}^T \mathbf{W}_{\cdot j} + c_j\right), \\ P_{\boldsymbol{\theta}}\left(h_j \mid \mathbf{v}\right) &= \left(P_{\boldsymbol{\theta}}\left(h_j = 1 \mid \mathbf{v}\right)\right)^{h_j} \\ &\quad \cdot \left(1 - P_{\boldsymbol{\theta}}\left(h_j = 1 \mid \mathbf{v}\right)\right)^{1-h_j}, \\ P_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{v}) &= \prod_{j=1}^{m} P_{\boldsymbol{\theta}}\left(h_j \mid \mathbf{v}\right), \end{aligned} \tag{3}$$

where $\mathbf{W}_{\cdot j}$ is the jth column of $\mathbf{W}$, which is a vector that represents the connection weights between the jth hidden unit and all visible units, and $\operatorname{sig}(x) = 1/(1 + e^{-x})$ is the sigmoid activation function. Thus, the marginal distribution $P_{\boldsymbol{\theta}}(\mathbf{v})$ of the visible variables actually is a model of product of experts [2, 12]:

$$P_{\boldsymbol{\theta}}(\mathbf{v}) = \frac{1}{Z(\boldsymbol{\theta})} \cdot \exp\left(\mathbf{b}^T \mathbf{v}\right) \cdot \prod_{j=1}^{m} \frac{1}{\left(1 - P_{\boldsymbol{\theta}}\left(h_j = 1 \mid \mathbf{v}\right)\right)}. \tag{4}$$

Equation (4) deduces that all these hidden units for the individual components of the given data vector $\mathbf{v}$ are combined multiplicatively and will contribute probabilities according to the activation probabilities. If given a data sample, one specific hidden unit will be activated with a high probability, and the hidden unit is responsible for representing the data sample. If more data in the training data set activates a hidden unit with a higher probability, the hidden unit's feature will be less discriminative. Thus it is sometimes necessary to introduce sparsity at the hidden layer of an RBM [11, 19, 20].

For a training example $\mathbf{v}^0$, training an RBM is the same as modeling the marginal distribution $P_{\boldsymbol{\theta}}(\mathbf{v})$ of the visible units. A common practice is to adopt the log-likelihood gradient approach [16, 18] to maximizing the marginal distribution $P_{\boldsymbol{\theta}}(\mathbf{v})$, which aims to generate $\mathbf{v}^0$ with the largest probability. Using gradient descent approach can solve this problem:

$$\frac{\partial \ln P_{\boldsymbol{\theta}}\left(\mathbf{v}^0\right)}{\partial \boldsymbol{\theta}} = -\sum_{\mathbf{h}} P_{\boldsymbol{\theta}}\left(\mathbf{h} \mid \mathbf{v}^0\right) \frac{\partial E_{\boldsymbol{\theta}}\left(\mathbf{v}^0, \mathbf{h}\right)}{\partial \boldsymbol{\theta}} + \sum_{\mathbf{v},\mathbf{h}} P_{\boldsymbol{\theta}}\left(\mathbf{v}, \mathbf{h}\right) \frac{\partial E_{\boldsymbol{\theta}}\left(\mathbf{v}, \mathbf{h}\right)}{\partial \boldsymbol{\theta}}. \tag{5}$$

The second term of (5) is intractable because we cannot obtain any information about the marginal distribution $P_{\boldsymbol{\theta}}(\mathbf{v})$. In order to solve this problem, Hinton et al. [21] proposed Contrastive Divergence (CD) learning, which has become a standard way to train RBMs. The p-step contrastive divergence learning (CD-p) is in two steps: first, the Gibbs chain is initialized with the training example $\mathbf{v}^0$ of the training set. Second, the sample $\mathbf{v}^{(p)}$ is yielded after p steps, and each step q ($q = 1, 2, \ldots, p$) consists of sampling $\mathbf{h}^{(q-1)}$ from $P_{\boldsymbol{\theta}}(\mathbf{h} \mid \mathbf{v}^{(q-1)})$ and subsequently sampling $\mathbf{v}^{(q)}$ from $P_{\boldsymbol{\theta}}(\mathbf{v} \mid \mathbf{h}^{(q-1)})$. According to the general Markov Chain Monte Carlo (MCMC) theory, we know that when $p \to \infty$, the p-step contrastive divergence learning algorithm converges to the second term of (5) and becomes only visible in the proof of Bengio and Delalleau [22]. However, Hinton [16] pointed out that when initializing with the training example $\mathbf{v}^0$ of the training set, running one-step ($p = 1$) Gibbs sampling approximates this term in the log-likelihood gradient relatively well. Therefore, (5) can be approximated as

$$\frac{\partial \ln P_{\boldsymbol{\theta}}\left(\mathbf{v}^0\right)}{\partial \boldsymbol{\theta}} \approx -\sum_{\mathbf{h}} P_{\boldsymbol{\theta}}\left(\mathbf{h} \mid \mathbf{v}^0\right) \frac{\partial E_{\boldsymbol{\theta}}\left(\mathbf{v}^0, \mathbf{h}\right)}{\partial \boldsymbol{\theta}} + \sum_{\mathbf{h}} P_{\boldsymbol{\theta}}\left(\mathbf{h} \mid \mathbf{v}^{(1)}\right) \frac{\partial E_{\boldsymbol{\theta}}\left(\mathbf{v}^{(1)}, \mathbf{h}\right)}{\partial \boldsymbol{\theta}}. \tag{6}$$

Thus, the iterative update process of the jth column $\mathbf{W}_{\cdot j}$ from the weight matrix $\mathbf{W}$ for the training example $\mathbf{v}^0$ is represented by

$$\mathbf{W}_{\cdot j} \longleftarrow \mathbf{W}_{\cdot j} + \varepsilon\left(P_{\boldsymbol{\theta}}\left(h_j = 1 \mid \mathbf{v}^0\right) \cdot \mathbf{v}^0 - P_{\boldsymbol{\theta}}\left(h_j = 1 \mid \mathbf{v}^{(1)}\right) \cdot \mathbf{v}^{(1)}\right), \tag{7}$$

where ε is the learning rate. The first term of (7) decreases the energy of $\mathbf{v}^0$ [23]; at the same time, this term also guarantees that unit j is more likely to be activated when the hidden unit observes $\mathbf{v}^0$ again; this means that the hidden units are learning to represent $\mathbf{v}^0$ [12]. In the next section, we use a weight-decay strategy at group level for SGRBMs to capture features with group sparsity from the input data.

3. Nontopologically Ordered Feature Extraction Based on Sparse Group RBMs

In the unsupervised learning process, some of the hidden units may extract similar features if there is little difference between their corresponding weight vectors $\{\mathbf{W}_{\cdot j}\}_{j=1}^{m}$. This homogenization problem can be obvious and serious if the number of the hidden units is increased. To alleviate this problem, Lee et al. [11] introduced SRBMs and Luo et al. [12] introduced SGRBMs to remit statistical dependencies between all of the hidden units when adding a penalty term. SRBMs have been popular due to the fact that an RBM with a low average hidden activation probability is better at extracting discriminative features than nonregularized RBMs [8, 24]. This is especially the case in Luo et al. [12], who divided the hidden units equally into nonoverlapping groups to restrain the dependencies within these groups and penalized the overall activation level of a group. More discriminative features are learned when SGRBMs are applied to deep learning systems for classification tasks. However, Luo et al. [12] did not consider overfitting problems and did not propose any strategies for controlling the reconstruction complexity of the weight matrix. Thus, to equilibrate the reconstruction error (i.e., the learning accuracy of specific training samples) and reconstruction complexity (generalization ability), we have used a weight-decay strategy at group level based on SGRBMs to capture features with sparse grouping of the input data.

For an RBM with m hidden units, let $\mathbf{H} = \{1, 2, \ldots, m\}$ denote the set of all indices of the hidden units. The kth group is denoted by $\mathbf{G}_k$, where $\mathbf{G}_k \subset \mathbf{H}$, $k = 1, 2, \ldots, K$. Suppose all groups are nonoverlapping and of equal size [12] (see Figure 1). Given a grouping $\mathbf{G}$ and a sample $\mathbf{v}^{(l)}$, the kth group norm $N_{\mathbf{G}_k}(\mathbf{v}^{(l)})$ is given by

$$N_{\mathbf{G}_k}\left(\mathbf{v}^{(l)}\right) = \sqrt{\sum_{t \in \mathbf{G}_k} \left(P_{\boldsymbol{\theta}}\left(h_t^{(l)} = 1 \mid \mathbf{v}^{(l)}\right)\right)^2}, \tag{8}$$

where $N_{\mathbf{G}_k}(\mathbf{v}^{(l)})$ is the L_2 (Euclidean) norm of the vector comprising the activation probabilities $\{P_{\boldsymbol{\theta}}(h_t^{(l)} = 1 \mid \mathbf{v}^{(l)})\}_{t \in \mathbf{G}_k}$, which are considered as the overall activation level of the kth group. Given all the group norms $\{N_{\mathbf{G}_k}(\mathbf{v}^{(l)})\}_{k=1,2,\ldots,K}$, the mixed $L_{1,2}$ norm is

$$\sum_{k=1}^{K} \left|N_{\mathbf{G}_k}\left(\mathbf{v}^{(l)}\right)\right| = \sum_{k=1}^{K} \sqrt{\sum_{t \in \mathbf{G}_k} \left(P_{\boldsymbol{\theta}}\left(h_t^{(l)} = 1 \mid \mathbf{v}^{(l)}\right)\right)^2}. \tag{9}$$

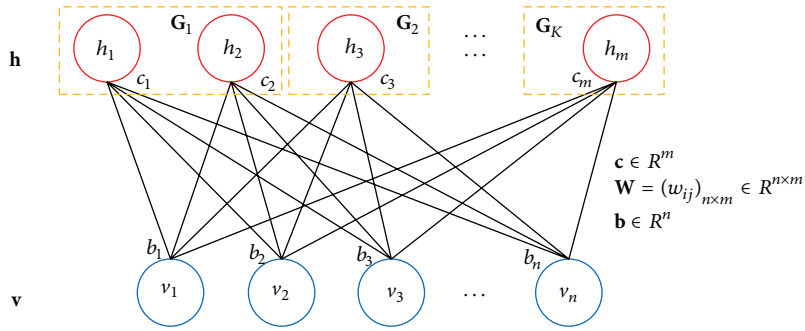

Figure 1: Grouping with nonoverlapping for nontopologically ordered SGRBMs.

The mixed $L_{1,2}$ norm of the grouping weight vectors $\{\mathbf{W}_{\cdot t}^{(l)}\}_{t\in\mathbf{G}_k}$ is shown by

$$M_{\mathbf{G}_k}\left(\mathbf{W}^{(l)}\right) = \sqrt{\sum_{t\in\mathbf{G}_k}\left\|\mathbf{W}_{\cdot t}^{(l)}\right\|_2^2} = \sqrt{\sum_{t\in\mathbf{G}_k}\sum_{i=1}^{n}\left(w_{it}^{(l)}\right)^2}, \quad (10)$$

$$\begin{aligned}\sum_{k=1}^{K}\left|M_{\mathbf{G}_k}\left(\mathbf{W}^{(l)}\right)\right| &= \sum_{k=1}^{K}\sqrt{\sum_{t\in\mathbf{G}_k}\left\|\mathbf{W}_{\cdot t}^{(l)}\right\|_2^2} \\ &= \sum_{k=1}^{K}\sqrt{\sum_{t\in\mathbf{G}_k}\sum_{i=1}^{n}\left(w_{it}^{(l)}\right)^2}.\end{aligned} \quad (11)$$

In fact, the mixed $L_{1,2}$ norm is considered as the overall weight strength level of the grouping.

We add these two $L_{1,2}$ regularizers ((9) and (11)) to the log-likelihood of the training examples. Thus, given the training set $\{\mathbf{v}^{(l)}\}_{l=1}^{L}$ comprising L examples, we need to solve the following optimization problem:

$$\begin{aligned}\max_{\boldsymbol{\theta}}\sum_{l=1}^{L}\Bigg[&\log P_{\boldsymbol{\theta}}\left(\mathbf{v}^{(l)}\right) - \lambda\sum_{k=1}^{K}N_{\mathbf{G}_k}\left(\mathbf{v}^{(l)}\right) \\ &- \eta\sum_{k=1}^{K}M_{\mathbf{G}_k}\left(\mathbf{W}^{(l)}\right)\Bigg],\end{aligned} \quad (12)$$

where λ and η are two regularization constants. The second term increases the sparsity of the hidden units at group level, and the third term decreases the reconstruction complexity of this model. We apply the contrastive divergence update rule (see (7)) to solve (12), and it is followed by one step of gradient ascent by using the gradient of the regularization terms.

By introducing these two regularizers, the iterative process to solve the optimal parameters (7) is updated as follows:

$$\begin{aligned}\mathbf{W}_{\cdot j}^{(l)} \longleftarrow \mathbf{W}_{\cdot j}^{(l)} + \varepsilon\Bigg[&P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)}\right)\cdot\mathbf{v}^{(l)} \\ &- P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)-}\right)\cdot\mathbf{v}^{(l)-} - \lambda\cdot\frac{1}{N_{\mathbf{G}_{k'}}\left(\mathbf{v}^{(l)}\right)} \\ &\cdot\left(P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)}\right)\right)^2\cdot\left(1 - P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)}\right)\right) \\ &\cdot\mathbf{v}^{(l)} - \eta\cdot\frac{1}{M_{\mathbf{G}_{k'}}\left(\mathbf{W}^{(l)}\right)}\cdot\mathbf{W}_{\cdot j}^{(l)}\Bigg],\end{aligned} \quad (13)$$

where $\mathbf{v}^{(l)-}$ is the CD-1 sampling from $\mathbf{v}^{(l)}$, and we assume the jth hidden unit belongs to the k'th group $\mathbf{G}_{k'}$. The last step in (13) is derived from $(d/dx)\mathrm{sig}(x) = \mathrm{sig}(x)(1 - \mathrm{sig}(x))$. Since the second and third terms in (12) control the sparsity of activation probabilities of the hidden units, then using (13), we can regard the hidden layer of RBMs as the nontopological feature extractor to capture features with group sparsity from the training data, abbreviated to NTOSGRBMs. In the next section, we obtain topologically ordered features from the input data at group level by adding an extra term to the mixed $L_{1,2}$ norm of the group weights based on SGRBMs and by adding an extra term to penalize the topologically ordered factors in the log-likelihood function.

4. Topologically Ordered Feature Extraction Based on Sparse Group RBMs

The approach of the sparsity-based feature extraction approaches is to employ regularizers to induce sparsity during discriminative feature representation [25, 26]. According to Luo et al. [12], the mixed $L_{1,2}$ norm encourages sparsity at group level; however, it does not contain any prior information about possible groups of covariates that we may want to select jointly [27, 28]. From the SGRBMs, we can learn a set of features with group sparsity that is useful for representing the input data; however, drawing inspiration from the human brain, we would like to learn a set of features that have both with group sparsity and topologically ordered in some manner. Here, *topologically ordered method* means that similar features are grouped together while retaining group sparsity. The aim of this constraint for hidden units is to group adjacent features together in the smoothed L_1 penalty. Instead of keeping all groups being nonoverlapping, therefore, we retain all of the overlapping groups. Then, suppose the $\widetilde{k}$th group is denoted by $\widetilde{\mathbf{G}}_{\widetilde{k}}$, where $\widetilde{\mathbf{G}}_{\widetilde{k}} \subset \mathbf{H}$, $\widetilde{k} = 1, 2, \ldots, \widetilde{K}$, and $\bigcup_{\widetilde{k}=1}^{\widetilde{K}}\widetilde{\mathbf{G}}_{\widetilde{k}} = \mathbf{H}$. Also, suppose that all

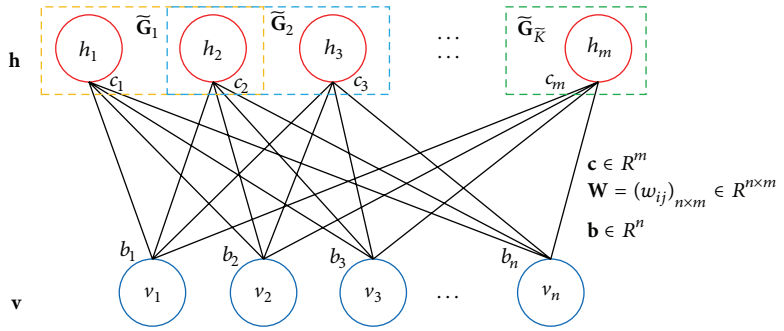

Figure 2: Grouping with overlapping for topologically ordered SGRBMs.

of these overlapping groups are of the same size; then all overlapping parts are of the same size, and each hidden unit belongs to two neighboring groups (see Figure 2).

Since $\sum_{\tilde{k}=1}^{\widetilde{K}} \sqrt{\sum_{\tilde{t}\in\widetilde{\mathbf{G}}_{\tilde{k}}} (P_{\boldsymbol{\theta}}(h_{\tilde{t}}^{(l)} = 1 \mid \mathbf{v}^{(l)}))^2}$ represents the overall activation level of hidden units for all groups, then (12) becomes

$$\max_{\boldsymbol{\theta}} \sum_{l=1}^{L} \Bigg[\log P_{\boldsymbol{\theta}}\left(\mathbf{v}^{(l)}\right) - \lambda \sum_{\tilde{k}=1}^{\widetilde{K}} \sqrt{\sum_{\tilde{t}\in\widetilde{\mathbf{G}}_{\tilde{k}}} \left(P_{\boldsymbol{\theta}}\left(h_{\tilde{t}}^{(l)} = 1 \mid \mathbf{v}^{(l)}\right)\right)^2} - \eta \sum_{\tilde{k}=1}^{\widetilde{K}} \sqrt{\sum_{\tilde{t}\in\widetilde{\mathbf{G}}_{\tilde{k}}} \sum_{i=1}^{n} \left(w_{i\tilde{t}}^{(l)}\right)^2} \Bigg]. \tag{14}$$

In fact, since we actually minimize the overall weight strength level of the grouping, the third term in (14) ensures that only a few hidden units can be activated in a group. From the perspective of information theory, the entropy of the distribution of the conditional probabilities for all hidden units in a group is relatively low. For the $\tilde{k}$th group $\widetilde{\mathbf{G}}_{\tilde{k}}$ in the hidden layer, the entropy of the conditional probabilities' distribution is defined as

$$T_{\widetilde{\mathbf{G}}_{\tilde{k}}}\left(\mathbf{v}^{(l)}\right) = -\sum_{\tilde{t}\in\widetilde{\mathbf{G}}_{\tilde{k}}} P_{\boldsymbol{\theta}}\left(h_{\tilde{t}}^{(l)} = 1 \mid \mathbf{v}^{(l)}\right) \log P_{\boldsymbol{\theta}}\left(h_{\tilde{t}}^{(l)} = 1 \mid \mathbf{v}^{(l)}\right). \tag{15}$$

For the neighboring two groups, the $\tilde{k}$th group $\widetilde{\mathbf{G}}_{\tilde{k}}$ and $(\tilde{k}+1)$th group $\widetilde{\mathbf{G}}_{\tilde{k}+1}$ at the hidden layer, we can define the topologically ordered factor (TOF) between these neighboring two groups as

$$F_{\widetilde{\mathbf{G}}_{\tilde{k}},\widetilde{\mathbf{G}}_{\tilde{k}+1}}\left(\mathbf{v}^{(l)}\right) = \left| \frac{T_{\widetilde{\mathbf{G}}_{\tilde{k}}\cap\widetilde{\mathbf{G}}_{\tilde{k}+1}}\left(\mathbf{v}^{(l)}\right)}{T_{\widetilde{\mathbf{G}}_{\tilde{k}}}\left(\mathbf{v}^{(l)}\right) + T_{\widetilde{\mathbf{G}}_{\tilde{k}+1}}\left(\mathbf{v}^{(l)}\right)} - \frac{C\left(\widetilde{\mathbf{G}}_{\tilde{k}}\cap\widetilde{\mathbf{G}}_{\tilde{k}+1}\right)}{C\left(\widetilde{\mathbf{G}}_{\tilde{k}}\right) + C\left(\widetilde{\mathbf{G}}_{\tilde{k}+1}\right)} \right|, \tag{16}$$

where $C(\cdot)$ is the count function to obtain the number of elements in a set. The structured features in the hidden units are topologically well-ordered if $F_{\widetilde{\mathbf{G}}_{\tilde{k}},\widetilde{\mathbf{G}}_{\tilde{k}+1}}(\mathbf{v}^{(l)})$ is close to zero. Since one important research direction in using the stochastic method (i.e., contrastive divergence [2] and approximate maximum-likelihood [29]) for RBMs is to design a regularization term [30], it is common to use weight-decay that regularizes the growth of parameters to avoid overfitting and stabilize learning. Therefore, in this paper, we have added an extra term to penalize the topologically ordered factors, so that the topologically ordered feature extraction based on SGRBMs (TOSGRBMs) can be extended by these two extra regularizers. Thus, given training data $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(L)}\}$, we need to solve the following optimization problem:

$$\max_{\boldsymbol{\theta}} \sum_{l=1}^{L} \Bigg[\log P_{\boldsymbol{\theta}}\left(\mathbf{v}^{(l)}\right) - \frac{\lambda}{2} \sum_{\tilde{k}=1}^{\widetilde{K}-1} F^2_{\widetilde{\mathbf{G}}_{\tilde{k}},\widetilde{\mathbf{G}}_{\tilde{k}+1}}\left(\mathbf{v}^{(l)}\right) - \eta \sum_{\tilde{k}=1}^{\widetilde{K}} \sqrt{\sum_{\tilde{t}\in\widetilde{\mathbf{G}}_{\tilde{k}}} \sum_{i=1}^{n} \left(w_{i\tilde{t}}^{(l)}\right)^2} \Bigg]. \tag{17}$$

This type of regularization can be seen as the combination of the group weight-decay and topologically ordered factors regularization. To address this problem, it is possible to use the alternating direction method of gradient ascent. The partial derivative of the first and third terms is shown in (13), and this problem may be turned into solving the partial derivative term $(\partial/\partial \mathbf{W}_{\cdot j}^{(l)}) \sum_{\tilde{k}=1}^{\widetilde{K}-1} F^2_{\widetilde{\mathbf{G}}_{\tilde{k}},\widetilde{\mathbf{G}}_{\tilde{k}+1}}(\mathbf{v}^{(l)})$. Suppose $j \in \widetilde{\mathbf{G}}_{\tilde{k}'} \cap \widetilde{\mathbf{G}}_{\tilde{k}'+1}$; thus $j \in \widetilde{\mathbf{G}}_{\tilde{k}'}$ and $j \in \widetilde{\mathbf{G}}_{\tilde{k}'+1}$; then

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{W}_{\cdot j}^{(l)}} \sum_{\tilde{k}=1}^{\widetilde{K}-1} F^2_{\widetilde{\mathbf{G}}_{\tilde{k}},\widetilde{\mathbf{G}}_{\tilde{k}+1}}\left(\mathbf{v}^{(l)}\right) \\ &= 2F_{\widetilde{\mathbf{G}}_{\tilde{k}'},\widetilde{\mathbf{G}}_{\tilde{k}'+1}}\left(\mathbf{v}^{(l)}\right) \cdot \left(-P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)}\right)\right) \\ &\quad \cdot \left(1 - P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)}\right)\right) \end{aligned}$$

$$\cdot \left(1 + \log P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)}\right)\right) \cdot \frac{\left(T_{\widetilde{\mathbf{G}}_{\widetilde{k}'}}\left(\mathbf{v}^{(l)}\right) + T_{\widetilde{\mathbf{G}}_{\widetilde{k}'+1}}\left(\mathbf{v}^{(l)}\right) - 2T_{\widetilde{\mathbf{G}}_{\widetilde{k}'} \cap \widetilde{\mathbf{G}}_{\widetilde{k}'+1}}\left(\mathbf{v}^{(l)}\right)\right)}{\left(T_{\widetilde{\mathbf{G}}_{\widetilde{k}'}}\left(\mathbf{v}^{(l)}\right) + T_{\widetilde{\mathbf{G}}_{\widetilde{k}'+1}}\left(\mathbf{v}^{(l)}\right)\right)^2}. \tag{18}$$

The objective function in (17) is optimized when using the iterated method described. In the above discussion, the form of (13) is not changed, but one significant difference is that we assume the jth hidden unit belongs to the $\widetilde{k}'$th group $\widetilde{\mathbf{G}}_{k'}$ and its neighboring group: the $(\widetilde{k}'+1)$th group $\widetilde{\mathbf{G}}_{k'+1}$. Thus, the iterative formula for the optimal solution of parameters of RBM is

$$\begin{aligned}\mathbf{W}_{\cdot j}^{(l)} \longleftarrow \mathbf{W}_{\cdot j}^{(l)} + \varepsilon \Bigg(& P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)}\right) \cdot \mathbf{v}^{(l)} \\ & - P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)-}\right) \cdot \mathbf{v}^{(l)-} + \lambda \cdot F_{\widetilde{\mathbf{G}}_{\widetilde{k}'}, \widetilde{\mathbf{G}}_{\widetilde{k}'+1}}\left(\mathbf{v}^{(l)}\right) \\ & \cdot P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)}\right) \cdot \left(1 - P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)}\right)\right) \\ & \cdot \left(1 + \log P_{\boldsymbol{\theta}}\left(h_j^{(l)} = 1 \mid \mathbf{v}^{(l)}\right)\right) \\ & \cdot \frac{\left(T_{\widetilde{\mathbf{G}}_{\widetilde{k}'}}\left(\mathbf{v}^{(l)}\right) + T_{\widetilde{\mathbf{G}}_{\widetilde{k}'+1}}\left(\mathbf{v}^{(l)}\right) - 2T_{\widetilde{\mathbf{G}}_{\widetilde{k}'} \cap \widetilde{\mathbf{G}}_{\widetilde{k}'+1}}\left(\mathbf{v}^{(l)}\right)\right)}{\left(T_{\widetilde{\mathbf{G}}_{\widetilde{k}'}}\left(\mathbf{v}^{(l)}\right) + T_{\widetilde{\mathbf{G}}_{\widetilde{k}'+1}}\left(\mathbf{v}^{(l)}\right)\right)^2} \\ & - \eta \cdot \frac{\mathbf{W}_{\cdot j}^{(l)}}{\widetilde{M}_{\widetilde{\mathbf{G}}_{\widetilde{k}'}}\left(\mathbf{W}^{(l)}\right) + \widetilde{M}_{\widetilde{\mathbf{G}}_{\widetilde{k}'+1}}\left(\mathbf{W}^{(l)}\right)} \Bigg),\end{aligned} \tag{19}$$

where $\widetilde{M}_{\widetilde{\mathbf{G}}_{\widehat{k}}}(\mathbf{W}^{(l)}) = \sqrt{\sum_{\widetilde{t} \in \widetilde{\mathbf{G}}_{\widehat{k}}} \sum_{i=1}^{n} (w_{i\widetilde{t}}^{(l)})^2}$ $(\widehat{k} = \widetilde{k}', \widetilde{k}'+1)$. The first term of (17) represents the true distribution of the input data, and maximizing it in fact minimizes the reconstruction error of RBMs. The third term of (17) represents the group sparsity of the hidden layer, and minimizing this penalty term is equivalent to minimizing the reconstruction complexity of RBMs. In addition, in order to obtain topologically ordered features from these features with group sparsity, we add the second term of (17) to penalize the topologically ordered factors.

The objective function in (17) is actually an optimization problem; however, it is not convex. In principle, we can apply gradient descent to solve this problem; however, computing the gradient of the log-likelihood term is expensive. Fortunately, the contrastive divergence learning algorithm gives an efficient approximation to the gradient of the log-likelihood [2]. Building upon this, in each iteration we can apply the contrastive divergence update rule, followed by one step of gradient descent using the gradient of the regularization terms [11].

5. Experimental Results

In this section, we compared the results of the proposed nontopologically ordered and topologically ordered feature extraction approaches based on SGRBMs. First, we applied the two approaches to model patches of natural images. Then, we applied them to analyze the structured features of Flying Apsara images from the Dunhuang Grotto Murals at four different historical periods.

5.1. Modeling Patches of Natural Images. The training data consists of 20,000 patches 8×8, randomly extracted from a standard set of ten 512×512 whitened images. All the patches were divided into minibatches, each containing 256 patches, and updated the weights of each minibatch (total batches = 2,000).

We trained a nontopologically ordered SGRBM with 20,000 real-valued visible units and 256 hidden units which were divided into 64 uniform nonoverlapping groups containing four hidden units each. The learning rate ε was set to 0.1 [16], and the regularization constants λ and η were empirically set to 0.1 and 0.5, respectively. The learned features are shown in Figure 3(a). For comparison, we also trained a topologically ordered SGRBM with 256 hidden units. The learned features are shown in Figure 3(b). Some features are extracted and localized Gabor-like edge detectors in different positions and orientations; these results are like those in [13, 14]. Since the hidden units within a group compete with each other in the modeling patches, each hidden unit in the nontopologically ordered SGRBMs focused on modeling more subtle patterns in the training data. As a result, the features learned with the topologically ordered SGRBMs are more aggregative at group level than those learned with the nontopologically ordered SGRBMs. Moreover, the learned features by TOSGRBMs shown in Figure 3 have an enforced a topological order, where the location, orientation, and frequency of the Gabor-like filters all change smoothly. In conclusion, from the perspective of the invariant feature learning, the topologically ordered feature extraction approach facilitate the training of the whole network to extract more discriminative features at the hidden layer.

We also compared our results with the standard SRBMs and SGRBMs. With the same parameter settings and the same number of iterations $(=10^4)$, Figure 4 shows that the SRBMs, SGRBMs, NTOSGRBMs, and TOSGRBMs extracted Gabor-like filter features; however, SRBMs and SGRBMs have many more redundant feature patches than NTOSGRBMs and TOSGRBMs. Moreover, since the grouped features are significant in SGRBMs, NTOSGRBMs, and TOSGRBMs, the features learned by the SGRBMs, NTOSGRBMs, and TOSGRBMs are more localized than those learned by the SRBMs.

In addition, we use Hoyer's sparseness measure [31] to determine the sparse representations learned by the SRBMs, SGRBMs, NTOSGRBMs, and TOSGRBMs. This measure is in the interval $[0, 1]$ and on a normalized scale. Figure 5 shows the activation probabilities of the hidden units that were computed using the regular SRBMs, SGRBMs,

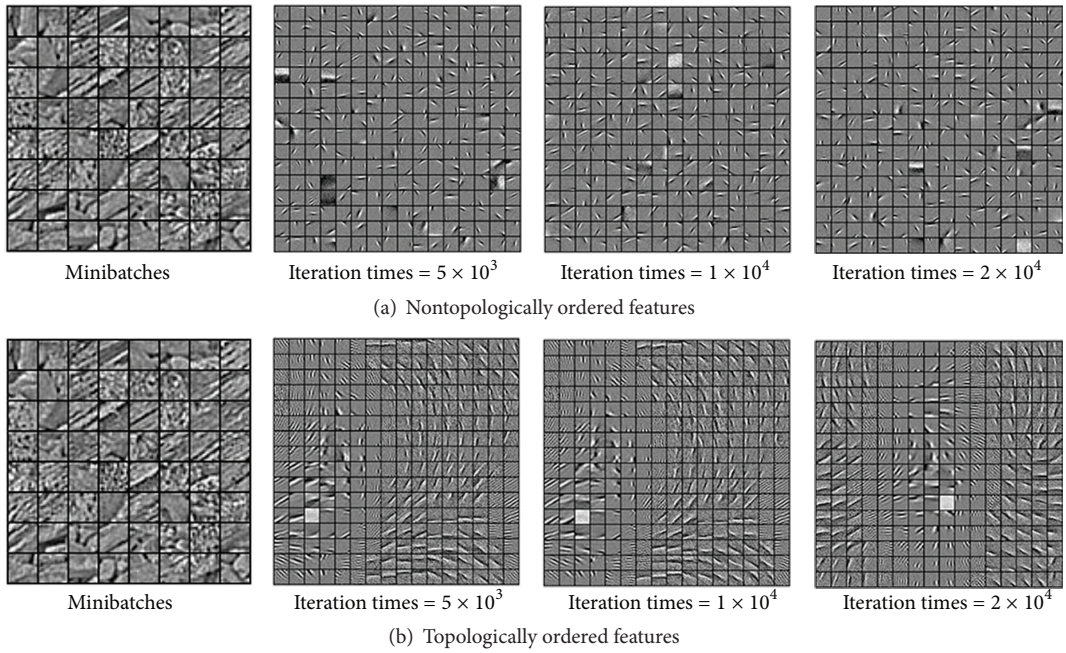

Figure 3: Learned features by the nontopologically ordered (a) and topologically ordered (b) feature extraction approach with 256 elements, learned on a dataset of 8×8 whitened natural image patches with 6×6 cyclic overlapping groups.

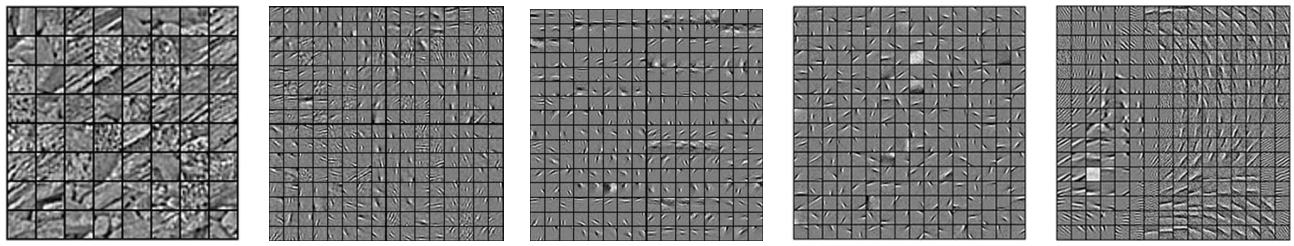

Figure 4: Comparison of learned features with 256 elements by SRBMs, SGRBMs, NTOSGRBMs, and TOSGRBMs.

NTOSGRBMs, and TOSGRBMs. It can be seen that the representations learned by the SGRBMs were much more sparse than for the other three models, although NTOSGRBMs and TOSGRBMs, with similar activation probabilities of hidden units, learned similar representations (close sparseness values) and learned much more sparse representations than the SRBMs, but much less sparse representations than the SGRBMs.

5.2. Modeling Patches of Flying Apsaras Images in the Dunhuang Grotto Murals. The image dataset of the Flying Apsaras in the Dunhuang Grotto Murals published in Zheng and Tai's book [32] contains 300 images. These images cover four historical periods: the early stage, the developing stage, the flourishing stage, and the terminal stage [33]. In the present study, as an example, the training data consisted of 20,000 8×8 randomly selected image patches of the Flying Apsara images. These patches were randomly extracted from a standard set of ten 512×512 fine-art paintings of the Flying Apsaras (Figure 6). Features from these images covering the four historical periods were exacted using both nontopologically ordered and topologically ordered SGRBMs. In addition, the parameters settings were the same as in the previous subsection.

In Figure 7, we see that both approaches show pronounced advantages in extracting discriminative features at group level, although the features learned by the topologically ordered SGRBMs were more aggregative than the nontopologically ordered SGRBMs. Moreover, since there is structural similarity between the features of the Flying Apsara images, their representations varied smoothly compared to the transformations and invariance achieved using TOSGRBMs. The features learned by TOSGRBMs (Figure 7) had enforced topological order, where the location, orientation, and frequency of the Gabor-like filters all change smoothly. It is concluded that, when using topologically ordered feature extraction based on SGRBMs, feature selection performed well because of the sparse and aggregative features at group level.

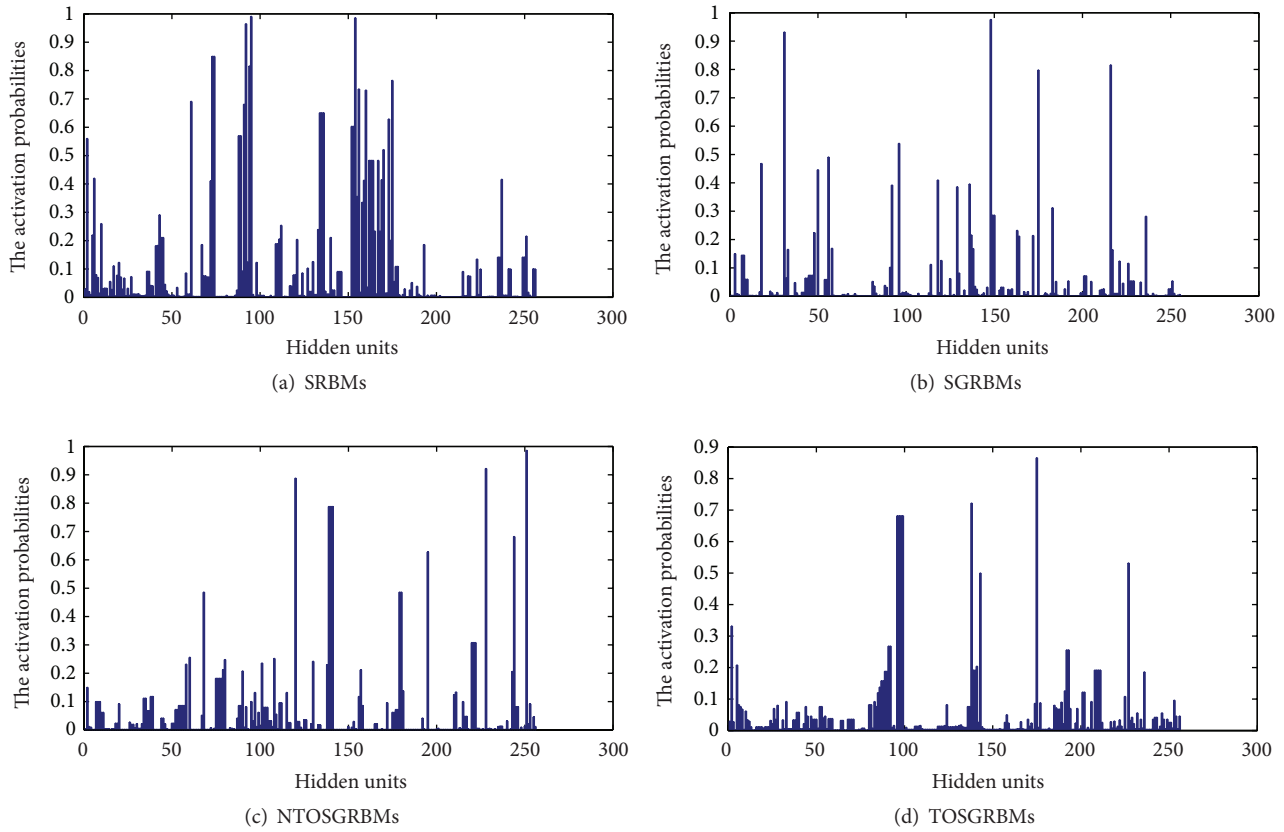

Figure 5: (a) Activation probabilities computed under the SRBMs, the sparseness of the vector is 0.68; (b) activation probabilities computed under the SGRBMs; the sparseness is 0.89; (c) activation probabilities computed under the NTOSGRBMs; the sparseness is 0.81; (d) activation probabilities computed under the TOSGRBMs; the sparseness is 0.78.

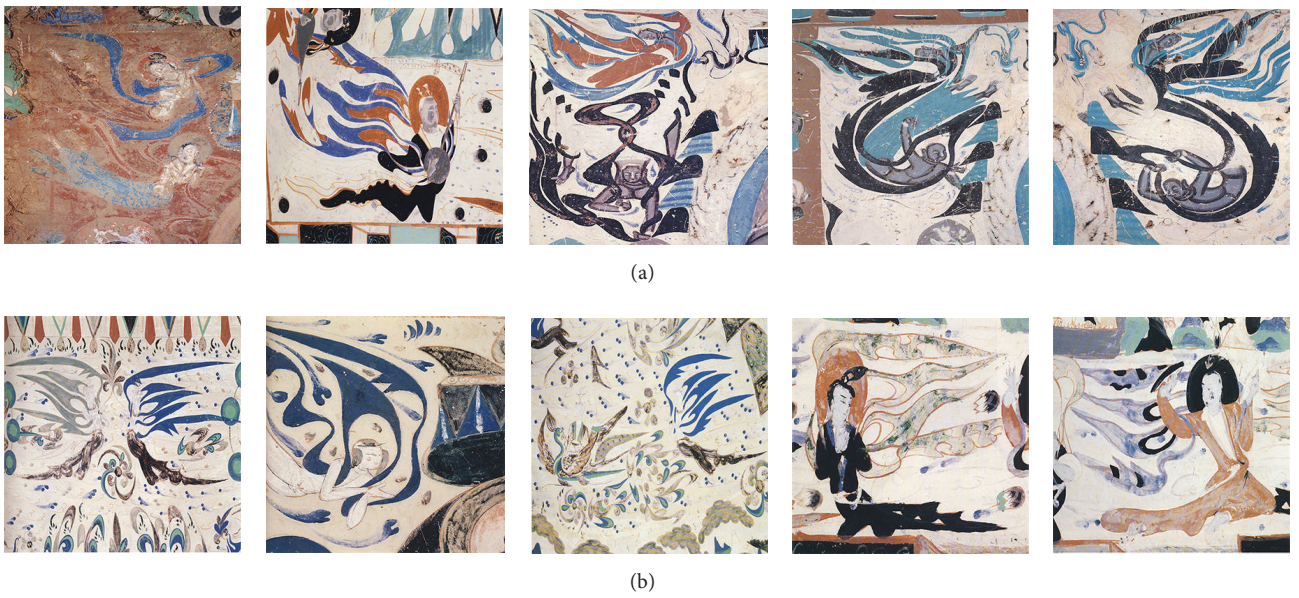

Figure 6: Ten 512×512 fine-art paintings with the same topic (the Flying Apsara art) but created at different periods. (a) Five fine-art paintings created at the period of Northern Wei (439–535 A.D.) and Western Wei (535–556 A.D.). (b) Five fine-art paintings created at the period of Western Wei (535–556 A.D.) and Northern Zhou (556–581 A.D.).

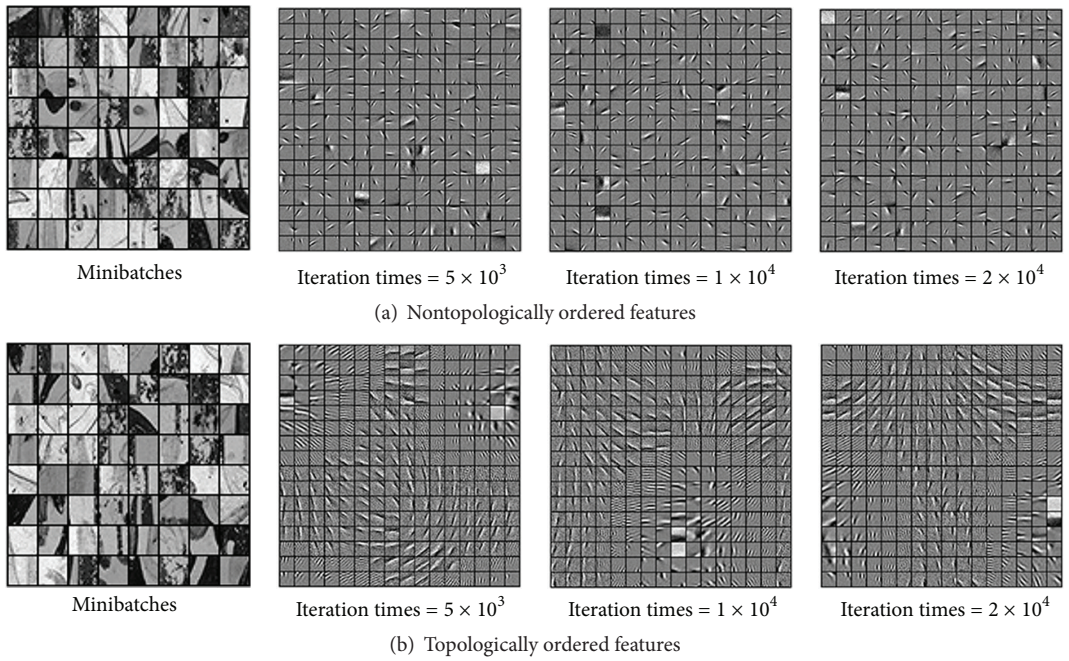

Figure 7: Learned features using the nontopologically ordered (a) and topologically ordered (b) feature extraction approaches (taking the early stage as an example).

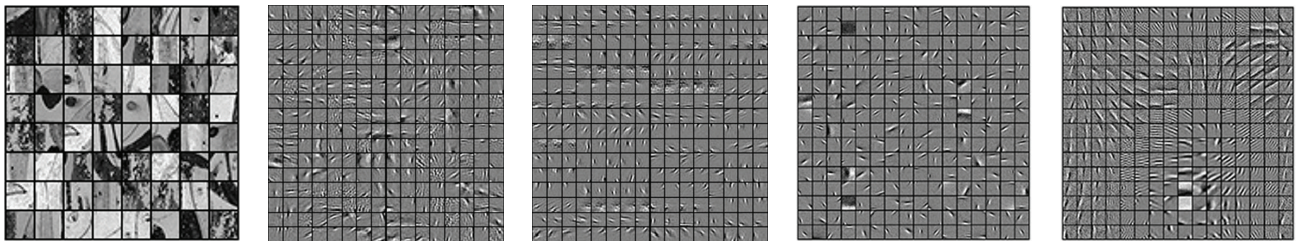

Figure 8: Comparison of learned features with 256 elements by SRBMs, SGRBMs, NTOSGRBMs, and TOSGRBMs.

Taking the early stage of the Flying Apsara images as an example, we also compared our learned features with those of the standard SRBMs and SGRBMs. With the same parameter settings and the same number of iterations ($=10^4$), Figure 8 shows that the grouped features are significant in SGRBMs, NTOSGRBMs, and TOSGRBMs. The features learned using the SGRBMs, NTOSGRBMs, and TOSGRBMs are more localized than those for the SRBMs. Moreover, the sparse features learned with SRBMs and NTOSGRBMs did not group similar feature patches as an aggregative representation. Both SGRBMs and TOSGRBMs learned to group similar features without discarding group sparsity; however, the sparse features learned by TOSGRBMs not only grouped similar feature patches as an aggregative representation but also retained similar feature patches in topological order at group level.

To measure the sparsity of the weight matrix $\mathbf{W}$ quantitatively, we used a sparsity indicator $S(\mathbf{W}) = (\sqrt{nm} - (\sum_{i=1}^{n}\sum_{j=1}^{m}|w_{ij}|)/(\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}w_{ij}^2}))/(\sqrt{nm} - 1)$ based on the relationship between the L_1 and L_2 norms. In accordance with Cauchy Inequality, $0 \le S(\mathbf{W}) \le 1$. The indicator takes a value of 1 if and only if $\mathbf{W}$ contains only one single nonzero component and 0 if and only if all elements are equal (but not equal to zero) and interpolates smoothly between the two extremes [31]. The Hoyer measure $S(\mathbf{W})$ is a normalized version of the L_2/L_1 measure. It approximates the L_0 measure but, because it is not as flat as the L_0 measure, it has a gradient that can be used in optimization problems [34].

Figures 9–11 list the different performances when applied to the Flying Apsara images for the four historical periods. Those show the average reconstruction error, the average topologically ordered factor, and the average sparsity of the weight matrix from five individual experiments with different random initializations for both nontopologically ordered and topologically ordered SGRBMs and different numbers of iterations (1×10^2, 2×10^2, 5×10^2, 1×10^3, 2×10^3, 5×10^3, 1×10^4, and 2×10^4). These results show that TOSGRBMs not only generate the topologically ordered features well

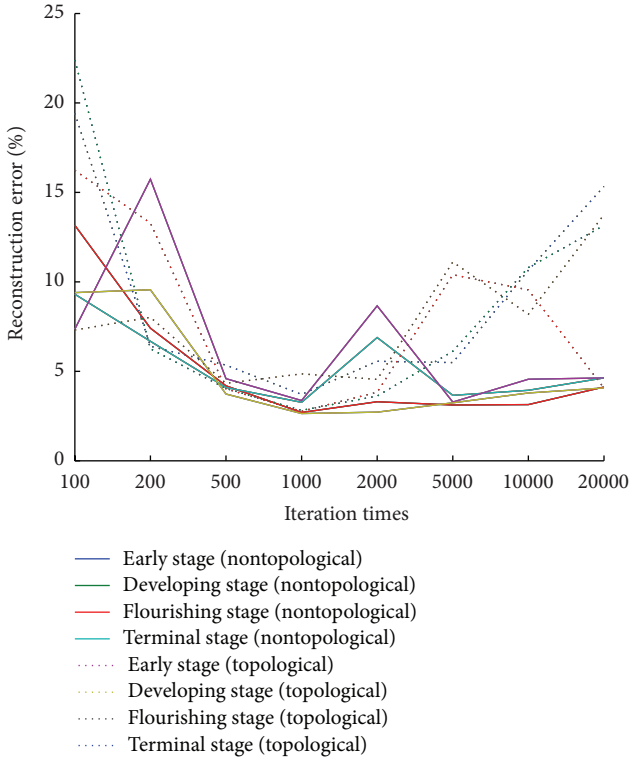

Figure 9: The different reconstruction errors of topologically ordered SGRBMs and nontopologically ordered SGRBMs when applied to image dataset of Flying Apsara at four different historical periods.

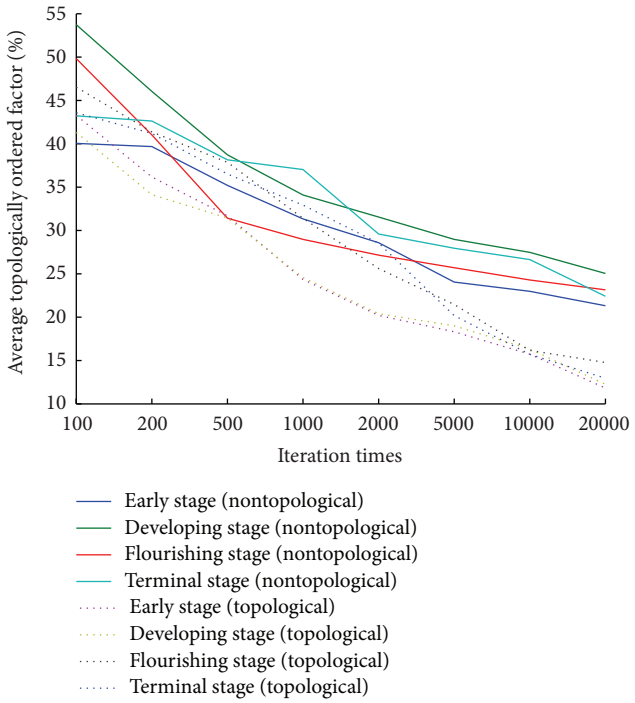

Figure 10: The different average topologically ordered factors of topologically ordered SGRBMs and nontopologically ordered SGRBMs when applied to image dataset of Flying Apsara at four different historical periods.

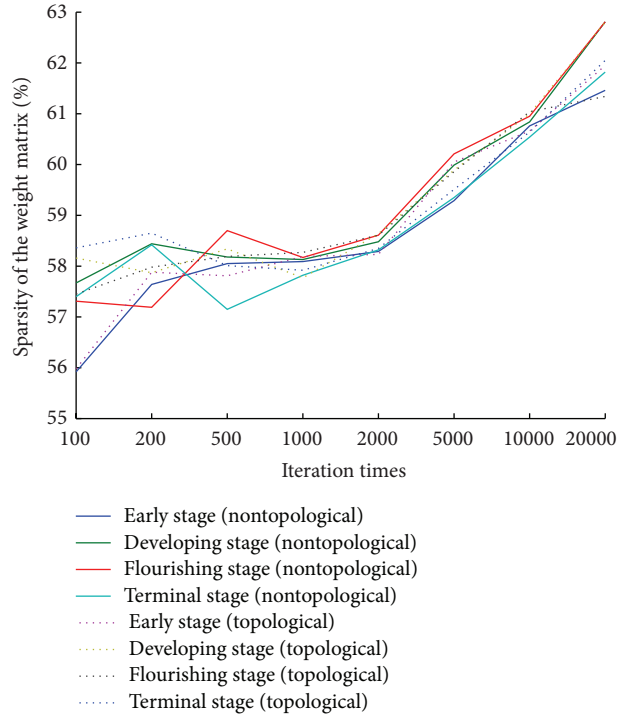

Figure 11: The different weight matrix sparseness of topologically ordered SGRBMs and nontopologically ordered SGRBMs when applied to image dataset of Flying Apsara at four historical periods.

(see Figures 7, 8, and 10) but also generate representative features well. Since the average sparsity of weight matrix is large (see Figure 11), TOSGRBMs give the hidden units low correlation when representing structured features of the input data, and according to the tendency of the average reconstruction errors (Figure 9), it can avoid over-fitting learning simultaneously. The topologically ordered SGRBMs approach achieved the best discriminative feature extraction and produced the best trade-off between reconstruction error and complexity. The topologically ordered SGRBMs have lower average topologically ordered factors, which indicate that the proposed topologically ordered SGRBMs decrease the similarities between extracted features and order them well topologically, because of the penalty on the topologically ordered factors of all groups.

6. Conclusions

For the purpose of extracting topologically ordered features efficiently from high-dimensional data, firstly we used a weight-decay strategy at group level based on SGRBMs to capture features with group sparsity from the input data. Secondly, by adding an extra term to penalize the topologically ordered factors in the log-likelihood function, we obtain topologically ordered features at group level. Experimental results on the image datasets of both natural images and the Flying Apsara images from the Dunhuang Grotto Murals at four different historical periods demonstrate that the combination of these two extra terms in the log-likelihood function

helps to extract better discriminative features with much sparser and more aggregative hidden activation probabilities. In conclusion, in our experiments the topologically ordered SGRBMs showed markedly prominent sparsity of weight matrix, discriminative features at the hidden layer, and sparse feature representations. Topologically ordered SGRBMs were therefore found to be superior to nontopologically ordered SGRBMs for those reasons.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Key Basic Research Program of China (no. 2012CB725303), the National Science Foundation of China (no. 61170202, no. 41371420, and no. 61203154), and the Fundamental Research Funds for the Central Universities (no. 2013-YB-003). The authors would like to thank the editor and the anonymous reviewers for their valuable comments.

References

[1] Y. Freund and D. Haussler, "Unsupervised learning of distributions on binary vectors using two layer networks," Tech. Rep., Computer Research Laboratory, University of California, Santa Cruz, Calif, USA, 1994.

[2] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[3] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines," *Cognitive Science*, vol. 9, no. 1, pp. 147–169, 1985.

[4] G. F. Montúfar, J. Rauh, and N. Ay, "Expressive power and approximation errors of restricted Boltzmann machines," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '11)*, pp. 415–423, Granada, Spain, December 2011.

[5] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[6] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th International Conference On Machine Learning (ICML '09)*, pp. 609–616, June 2009.

[7] G. E. Hinton and R. Salakhutdinov, "Replicated softmax: an undirected topic model," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '09)*, pp. 1607–1614, 2009.

[8] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS '11)*, pp. 215–223, Fort Lauderdale, Fla, USA, April 2011.

[9] M. Yasuda and K. Tanaka, "Approximate learning algorithm for restricted Boltzmann machines," in *Proceedings of the International Conference on Computational Intelligence for Modelling Control & Automation*, pp. 692–697, IEEE, Vienna, Austria, December 2008.

[10] P. Garrigues and B. Olshausen, "Learning horizontal connections in a sparse coding model of natural images," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '08)*, vol. 20, MIT Press, Cambridge, Mass, USA, 2008.

[11] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area v2," in *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS '07)*, pp. 873–880, December 2007.

[12] H. Luo, R. Shen, C. Niu, and C. Ullrich, "Sparse group restricted Boltzmann machines," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pp. 429–434, San Francisco, Calif, USA, August 2011.

[13] M. Welling, G. Hinton, and S. Osindero, "Learning sparse topographic representations with products of student-t distributions," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1359–1366, MIT Press, 2002.

[14] H. Goh, L. Kusmierz, J.-H. Lim, N. Thome, and M. Cord, "Learning invariant color features with sparse topographic restricted Boltzmann machines," in *Proceedings of the 18th IEEE International Conference on Image Processing (ICIP '11)*, pp. 1241–1244, IEEE, Brussels, Belgium, September 2011.

[15] S. Bengio, F. Pereira, Y. Singer, and D. Strelow, "Group sparse coding," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '09)*, pp. 82–89, 2009.

[16] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[17] J. Martens, A. Chattopadhya, T. Pitassi, and R. Zemel, "On the representational efficiency of restricted Boltzmann machines," in *Proceedings of the Advances in Neural Information Processing Systems (NIPS '13)*, pp. 2877–2885, December 2013.

[18] A. Fischer and C. Igel, "Training restricted Boltzmann machines: an introduction," *Pattern Recognition*, vol. 47, no. 1, pp. 25–39, 2014.

[19] H. Larochelle and Y. Bengio, "Classification using discriminative restricted Boltzmann machines," in *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*, pp. 536–543, ACM, Helsinki, Finland, July 2008.

[20] R. Salakhutdinov and H. Larochelle, "Efficient learning of deep Boltzmann machines," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, vol. 6, Sardinia, Italy, May 2010.

[21] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[22] Y. Bengio and O. Delalleau, "Justifying and generalizing contrastive divergence," *Neural Computation*, vol. 21, no. 6, pp. 1601–1621, 2009.

[23] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–27, 2009.

[24] 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)*, pp. 689–696, Bellevue, Wash, USA, June-July 2011.

[25] Q. Zhou, L. Zhang, and L. Ma, "Learning topographic sparse coding through similarity function," in *Proceedings of the 4th International Conference on Natural Computation (ICNC '08)*, pp. 241–245, IEEE, Jinan, China, October 2008.

[26] F. Wu, Y. Yuan, Y. Rui, S. Yan, and Y. Zhuang, "Annotating web images using NOVA: non-convex group spArsity," in *Proceedings of the 20th ACM International Conference on Multimedia (MM '12)*, pp. 509–518, ACM, Nara, Japan, November 2012.

[27] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proceedings of the 26th International Conference On Machine Learning (ICML '09)*, pp. 433–440, June 2009.

[28] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach, "Learning hierarchical and topographic dictionaries with structured sparsity," in *Wavelets and Sparsity XIV*, Proceedings of SPIE, 81381P, p. 13, International Society for Optics and Photonics, 2011.

[29] T. Tieleman, "Training restricted boltzmann machines using approximations to the likelihood gradient," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 1064–1071, July 2008.

[30] K. Cho, A. Ilin, and T. Raiko, "Tikhonov-type regularization for restricted Boltzmann machines," in *Artificial Neural Networks and Machine Learning—ICANN 2012*, vol. 7552 of *Lecture Notes in Computer Science*, pp. 81–88, Springer, Berlin, Germany, 2012.

[31] P. O. Hoyer, "Non-negative matrix factorization with sparsity constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[32] R. Zheng and J. Tai, *Collected Edition of Flying Apsaras in the Dunhuang Grotto Murals, China*, Commercial Press, Hong Kong, 2002 (Chinese).

[33] Z. Chen, S. Xiong, Z. Fang, Q. Li, B. Wang, and Q. Zou, "A kernel support vector machine-based feature selection approach for recognizing Flying Apsaras' streamers in the Dunhuang Grotto Murals, China," *Pattern Recognition Letters*, vol. 49, pp. 107–113, 2014.

[34] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.