Latent Semantic Representation Learning for Scene Classification

Xin Li XINLI@TEMPLE.EDU
Yuhong Guo YUHONG@TEMPLE.EDU
Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA

Abstract

The performance of machine learning methods is heavily dependent on the choice of data representation. In real world applications such as scene recognition problems, the widely used low-level input features can fail to explain the high-level semantic label concepts. In this work, we address this problem by proposing a novel patch-based latent variable model to integrate latent contextual representation learning and classification model training in one joint optimization framework. Within this framework, the latent layer of variables bridge the gap between inputs and outputs by providing discriminative explanations for the semantic output labels, while being predictable from the low-level input features. Experiments conducted on standard scene recognition tasks demonstrate the efficacy of the proposed approach, comparing to the state-of-the-art scene recognition methods.

1. Introduction

The success of machine learning algorithms generally depends on the choice of data representation, since a good representation can disentangle the underlying explanatory factors behind the observed data and facilitate classification model learning (Bengio et al., 2012). Though learning from low-level features extracted from raw inputs produces good performance in many classification tasks with simple label concepts, it is difficult to learn complex semantic label concepts, such as scene labels, directly from low-level features. A scene label typically expresses a semantic concept that can be described by presence patterns of a set of high-level objects. For example, as shown in Figure 1, a "*street*" scene may consist of objects such as *tree*, *road*, *building* and *sky*, and a "*coast*" scene may consist of objects such as *ship*, *sea*, *sky*, and *mountain*. Recognition of such high-level semantic concepts often requires human ingenuity and high-level statistical domain knowledge, neither of which can be captured in low-level features. Thus a critical challenge for automatic scene recognition lies in the semantic gap between the low-level image features, such as the local gradient-based SIFT and HOG features (Lowe, 2004; Dalal & Triggs, 2005), and the high-level semantic scene concepts.

In this paper, we address the challenge of semantic scene classification by learning a latent semantic representation of the input data. Specifically, we propose to use a patch-based latent layer of variables to model the intrinsic contextual structure of the semantic output concepts, while ensuring them to be predictable from the original low-level input features. The latent variables can be viewed as an intermediate representation between the low-level inputs and the high-level outputs. Moreover, we encode the spatial information of the images by using a Laplacian regularizer over the latent representation vectors of the patches within each image, which enforces a spatially smooth change of the semantic contents. We formulate the learning problem as a joint optimization problem over both the latent representation variables and the prediction model parameters, which simultaneously minimizes the regression losses from the inputs to the latent representation and the prediction losses from the latent representation to the output labels. We expect such a model can automatically capture intrinsic explanations of the output semantic concepts, and hence improve the overall prediction performance from the low-level inputs to the high-level outputs. Our experimental results on standard scene classification datasets show that the proposed approach outperforms a few baseline and state-of-the-art scene classification methods.

2. Related Work

For image analysis and scene classification, numerous data representations built upon low-level features have been introduced in the recent decade. The widely employed bag-of-word model (Sivic et al., 2005) uses a histogram representation which is efficient to compute from the low-level input features, but lacks contextual information for com-

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

Figure 1. Scene is a semantic concept that consists of different objects. Top row presents the scene images, *open country*, *street*, and *coast*. Bottom row presents the regions occupied by different object categories in each corresponding image.

plex semantic scene recognition. Contextual information can be interpreted as object interactions or co-occurrences. Many works exploit such interactions between objects to learn intermediate representations of the data for improving recognition performance (Li & Guo, 2012; Blaschko & Lampert, 2009; Choi et al., 2010; Divvala et al., 2009; Fidler et al., 2009; Sadeghi & Farhadi, 2011). Most of these works model pairwise object interactions. For example, Sadeghi & Farhadi (2011) proposed a semantic concept, *visual phases* (objects performing an action or a pair of objects interacting with each other), to assist recognition tasks. In addition, the probabilistic graphical model proposed in (Li & Guo, 2012) integrates a chain structure to capture the co-occurrence of objects. Moreover, the work in (Li et al., 2012) models the visual appearance of a group of objects to capture high-order contextual interactions. Kumar & Koller (2010) presented a two-layer model based on bottom-up over-segmentation algorithms, where the first layer assigns each pixel to a unique connected region and the second layer assigns each region to a unique label. Kwitt et al. (2012) proposed a spatial pyramid matching architecture to combine the mid-level theme representation with the spatial pyramid structure for scene recognition. This approach however relies on predefined meaningful semantic themes and requires weakly supervision such as the presence knowledge of the semantic themes. Another group of works explore contextual information by identifying intermediate representations using topic models. For example, Wang & Grimson (2007) proposed a spatial latent Dirichlet allocation model which clusters co-occurring and spatially neighboring visual words into the same topic. He & Zemel (2008) presented a hybrid framework for image labeling, which combines a generative topic model with discriminative prediction models. Most recently, a state-of-the-art work in (Niu et al., 2012) proposes a discriminative latent Dirichlet allocation model to capture two types of contextual information, global spatial layout and visual coherence in uniform local regions, for scene recognition. However, these models mostly have high requirement for object or scene component identification, and involve complicated training processes.

From the perspective of learning intermediate representations with latent variable layers, there are some related works on learning layer-wise models. Hinton et al. (2006) proposed a fast learning algorithm for multi-layer generative deep belief nets. Lee et al. (2009) presented a generative convolutional deep belief network, which learns useful high-level visual features, such as object parts, from unlabeled object and natural scene images. Jarrett et al. (2009) investigated a two-stage system with random nonlinear filters for feature extraction. These models however are generative models and are not optimized to capture latent semantic representations that are most discriminative for the target labels. In addition to these, Bergamo et al. (2011) proposed a compact code learning method for object categorization, which uses a set of latent binary indicator variables as the intermediate representation of images. However, they identify latent concepts from the whole image instead of local patches, without considering the spatial distribution of the latent concepts. Moreover, the latent variables are represented implicitly using indicator functions in their model, which eliminates the capacity of encoding prior knowledges over the latent representations.

Different from these methods, the proposed approach in this paper employs a patch-based latent layer of variables to model the contextual structure of the semantic output concepts. The proposed model has a larger modeling capacity than previous contextual information based methods since the high-level visual concepts in our model can be any useful visual entities such as objects, object parts, their composites and co-occurrences, and the spatial information between the patches can be reserved in the latent representation by enforcing spatial Laplacian regularizers. Patch-based learning has also been exploited in previous work (Ranzato et al., 2006) for unsupervised image feature

extraction based on an autoencoder model. Their method however is an unsupervised encoding and decoding process from patch to patch.

3. Proposed Approach

In this section, we present a patch-based latent semantic representation learning model for scene recognition. We first formulate the problem as a joint minimization problem over the latent variables and the prediction model parameters, while encoding the spatial information across patches within each image with a Laplacian regularizer. Then we develop an efficient alternating optimization procedure to solve it. Below we use $\mathbf{1}$ to denote any column vector with all 1 values assuming its size can be determined from context, and use I_s to denote an identity matrix with size s.

3.1. Latent Variable Model

Scene recognition is a multi-class prediction problem. Given t labeled images $\{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^t$, where $\mathbf{x}^i$ denotes the i-th image and $\mathbf{y}^i \in \{-1, +1\}^k$ is its scene label vector, we aim to learn a prediction model from $\mathbf{x}^i$ to $\mathbf{y}^i$. We first partition each input image into a bag of n non-overlapping patches, where each patch forms a low-level input feature vector with length d. The observed features from the n patches of the i-th image can be represented as a matrix $X^i \in \mathbb{R}^{n\times d}$, whose j-th row, X^i_j, contains the input vector from its j-th patch. The scene labels are very semantic and abstractive concepts, and they are difficult to be predicted directly from the low-level input features. To bridge this gap, we first learn a latent contextual representation vector $Z^i_j \in \{0, 1\}^{1\times m}$ for each j-th patch of the i-th image and assume each entry of Z^i_j indicates the existence of a latent high-level visual entity. We then learn the output label concept of an image based on the summary of the high-level latent visual entities inferred from its local patches. The m latent visual entities can be any individual or composite visual concepts from the set of images, but they need to be both directly predictable from the low-level input features and discriminative for the target semantic scene labels. Under this assumption, we formulate the scene label prediction problem with latent representation variables as the following unified optimization over two loss functions

$$\begin{aligned}\min_{\{Z^i\},\Theta,W} \quad & \sum_{i=1}^{t}\begin{pmatrix}\alpha_i\sum_{j=1}^{n}\mathcal{L}\big(Z^i_j, f(X^i_j;\Theta)\big)\\ +\mathcal{V}\big(\mathbf{y}^i, g(\sum_j Z^i_j; W)\big)\end{pmatrix} \\ & + \gamma_f R(\Theta) + \gamma_g R(W) + \gamma_z\textstyle\sum_i R_z(Z^i) \\ \text{subject to} \quad & Z^i \in \{0,1\}^{n\times m} \text{ for } i = 1,\dots,t;\end{aligned} \tag{1}$$

where $f(\cdot)$ is the function that predicts the latent visual entities from the input features of each patch and $g(\cdot)$ is the function that predicts the output labels from the latent visual entities contained in the whole image; Θ and W denote the model parameters of these two prediction functions respectively; $\mathcal{L}(\cdot)$ and $\mathcal{V}(\cdot)$ are loss functions; $R(\cdot)$ and $R_z(\cdot)$ are regularization functions. The prediction process encoded in this latent variable model is intuitively demonstrated in Figure 2.

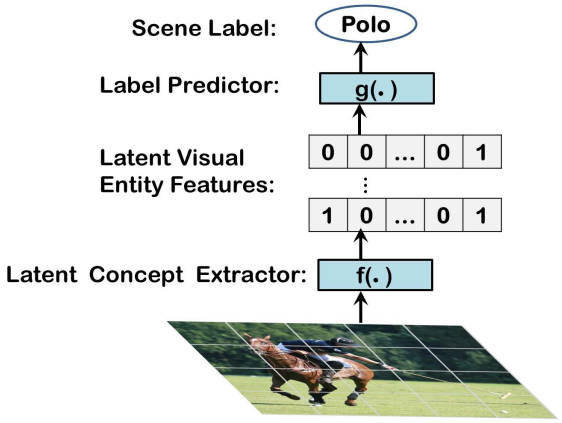


Figure 2. Illustration of the prediction process of the proposed latent variable model.

To produce a concrete optimization problem, we consider simple least square loss functions for $\mathcal{L}(\cdot)$ and $\mathcal{V}(\cdot)$, a Frobenius-norm regularization function $R(\cdot) = \|\cdot\|_F^2$, and the following linear prediction functions

$$f(X^i_j; \Theta, \mathbf{b}) = X^i_j\Theta^\top + \mathbf{b}^\top, \tag{2}$$

$$g(\textstyle\sum_j Z^i_j; W, \mathbf{q}) = \sum_j Z^i_j W + \mathbf{q}^\top. \tag{3}$$

Moreover, since the integer constraints induce hard optimization problems, we relax the integer constraints over Z^i into inequality constraints $Z^i \geq 0$ while enforcing a L1-norm regularization function R_z over Z^i to promote its sparsity. The optimization problem we obtained is

$$\begin{aligned}\min_{\{Z^i\},\Theta,\mathbf{b},W,\mathbf{q}} \quad & \sum_{i=1}^{t}\begin{pmatrix}\alpha_i\|Z^i - X^i\Theta^\top - \mathbf{1}\mathbf{b}^\top\|_F^2\\ +\|\mathbf{y}^{i\top} - \mathbf{1}^\top Z^i W - \mathbf{q}^\top\|_2^2\end{pmatrix} \\ & + \gamma_f\|\Theta\|_F^2 + \gamma_g\|W\|_F^2 + \gamma_z\textstyle\sum_i\|Z^i\|_1 \\ \text{subject to} \quad & Z^i \geq 0, \text{ for } i = 1,\dots,t;\end{aligned} \tag{4}$$

where $\Theta \in \mathbb{R}^{m\times d}$ and $\mathbf{b} \in \mathbb{R}^m$ are the model parameters of the linear function $f(\cdot)$; $W \in \mathbb{R}^{m\times k}$ and $\mathbf{q} \in \mathbb{R}^k$ are the model parameters of the linear function $g(\cdot)$; $\|\cdot\|_F, \|\cdot\|_2, \|\cdot\|_1$ denote Frobenius norm, Euclidean norm and entrywise L1-norm respectively. It is obvious that our proposed model can capture both *local* information, through the patches, and *global* information, through the summarization of the latent visual entity representations in the whole image, for target semantic label prediction.

3.2. Laplacian Regularization

The latent representation variables in our proposed model encode the high-level visual concepts that are directly predictable from the input features while being useful for identifying the target semantic scene labels. For each patch, its latent representation vector can be viewed as its mid-level prediction outputs. To better identify these mid-level outputs, we next propose to enforce a Laplacian regularization term over each set of latent vectors Z^i for each i-th image. Laplacian regularization has been typically used in semi-supervised learning scenarios to enforce the smoothness of the prediction values on unlabeled instances with respect to the intrinsic affinity structure of the input data (Belkin & Niyogi, 2002; Belkin et al., 2005). We propose to exploit this output regularization principle to improve the mid-level latent output learning by exploiting the affinity structures of the images.

In particular, we consider a natural affinity structure for each image, i.e., the spatial adjacency structure. For the i-th image X^i, we construct a spatial adjacency matrix $A^i \in \{0,1\}^{n\times n}$ over all the n patches, such that $A^i_{ab} = 1$ only if the a-th patch and the b-th patch are spatial neighbors in the i-th image. The Laplacian matrix based on this spatial adjacency matrix can be computed as $L^i = diag(A^i\mathbf{1}) - A^i$. With the spatial Laplacian matrices constructed for all images, the Laplacian regularized optimization problem is obtained as following

$$\begin{aligned}\min_{\{Z^i\},\Theta,\mathbf{b},W,\mathbf{q}} \quad & \sum_{i=1}^{t}\begin{pmatrix}\alpha_i\|Z^i - X^i\Theta^\top - \mathbf{1}\mathbf{b}^\top\|_F^2 \\ +\|\mathbf{y}^{i\top} - \mathbf{1}^\top Z^i W - \mathbf{q}^\top\|_2^2\end{pmatrix} \\ & + \gamma_f\|\Theta\|_F^2 + \gamma_g\|W\|_F^2 + \gamma_z\textstyle\sum_i\|Z^i\|_1 \\ & +\mu\textstyle\sum_i \mathrm{tr}(Z^{i\top}L^iZ^i) \\ \text{subject to} \quad & Z^i \geq 0, \text{ for } i = 1,\ldots,t\end{aligned} \tag{5}$$

3.3. Optimization Algorithm

The joint minimization problem in (5) is a non-convex optimization problem. We develop an iterative optimization algorithm to solve it by alternatively optimizing the model parameters and the latent variable values. In each iteration, given fixed latent variables $\{Z^i\}$, the model parameters for the prediction functions $f(\cdot)$ and $g(\cdot)$ can be trained independently with closed-form solutions.

Proposition 1 *Given fixed $\{Z^i\}_{i=1}^t$, the minimization problem over $\{\Theta, \mathbf{b}\}$ in (5) has the following closed-form solution:*

$$\Theta = \big(\textstyle\sum_i \alpha_i \hat{Z^i}^\top \hat{X^i}\big)\big(\gamma_f I_d + \textstyle\sum_i \alpha_i \hat{X^i}^\top \hat{X^i}\big)^{-1}, \tag{6}$$

$$\mathbf{b} = \bar{Z}^\top - \Theta\bar{X}^\top, \tag{7}$$

where

$$\bar{X} = \frac{1}{n\sum_i \alpha_i}(\textstyle\sum_i \alpha_i \mathbf{1}^\top X^i), \quad \hat{X^i} = X^i - \mathbf{1}\bar{X}, \tag{8}$$

$$\bar{Z} = \frac{1}{n\sum_i \alpha_i}(\textstyle\sum_i \alpha_i \mathbf{1}^\top Z^i), \quad \hat{Z^i} = Z^i - \mathbf{1}\bar{Z}. \tag{9}$$

Proposition 2 *Given fixed $\{Z^i\}_{i=1}^t$, the minimization problem over $\{W, \mathbf{q}\}$ in (5) has the following closed-form solution:*

$$W = (M^\top HM + \gamma_g I_m)^{-1}M^\top HY, \tag{10}$$

$$\mathbf{q} = \frac{1}{t}(Y - MW)^\top \mathbf{1}, \tag{11}$$

where $H = I_t - \frac{1}{t}\mathbf{1}\mathbf{1}^\top$, $Y = [\mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^t]^\top$, and $M = [\mathbf{1}^\top Z^1; \mathbf{1}^\top Z^2; \cdots; \mathbf{1}^\top Z^t]$.

Proposition 1 and Proposition 2 can be proved by simply setting the partial derivatives of the optimization objective function regarding each of the model parameters to zeros.

Given fixed prediction model parameters $\{\Theta, \mathbf{b}, W, \mathbf{q}\}$, the optimization problem over the latent variables $\{Z^i\}$ can be decomposed into a set of independent sub-problems, one for each Z^i matrix, which enables the capacity of exploiting parallel computation resources for large scale computation. Specifically, the optimization problem over each latent Z^i matrix is a quadratic minimization problem with non-negativity constraints:

$$\min_{Z^i} \quad \ell(Z^i) \quad \text{subject to} \quad Z^i \geq 0 \tag{12}$$

where

$$\begin{aligned}\ell(Z^i) = {} & \alpha_i\|Z^i - X^i\Theta^\top - \mathbf{1}\mathbf{b}^\top\|_F^2 \\ & + \|\mathbf{y}^{i\top} - \mathbf{1}^\top Z^i W - \mathbf{q}^\top\|_2^2 \\ & + \gamma_z \mathbf{1}^\top Z^i \mathbf{1} + \mu\mathrm{tr}(Z^{i\top}L^iZ^i).\end{aligned} \tag{13}$$

However standard second-order quadratic solvers are very inefficient for solving this minimization problem and have scalability problem for large images, since each Z^i is typically large and the Hessian matrix will be quadratically large. We propose to use an efficient and scalable first-order projected gradient descent algorithm to conduct minimization, as shown in Algorithm 1. In this algorithm, for each iteration, we first compute the gradient matrix from the objective function

$$\begin{aligned}\nabla\ell(Z^i) = {} & 2EZ^iWW^\top + 2\alpha_i Z^i - 2\alpha_i(X^i\Theta^\top + \mathbf{1}\mathbf{b}^\top) \\ & + 2\mathbf{1}(\mathbf{q}^\top - \mathbf{y}^{i\top})W^\top + \gamma_z + 2\mu L^iZ^i\end{aligned} \tag{14}$$

where E is a square matrix of size n with all 1 values. Then we take a gradient update over Z^i with stepsize $1/\rho$ and project it onto the non-negativity constraints. We use $\rho = 2\alpha_i + 2n\|WW^\top\|_F + 2\mu\sqrt{m}\|L^i\|_F$, which guarantees the convergence of the projected gradient descent procedure.

Algorithm 1 Projected gradient descent algorithm

Procedure:

while not converged

1. compute the gradient matrix $\nabla\ell(Z^i)$ at the current point Z^i using Eq. (14).
2. update $Z^i = \max(0, Z^i - \frac{1}{\rho}\nabla\ell(Z^i))$.

end while

Lemma 1 *Given the continuously differentiable function $\ell(\cdot)$ in (13), and $\rho = 2\alpha_i + 2n\|WW^\top\|_F + 2\mu\sqrt{m}\|L^i\|_F$. For any $L \geq \rho$, let*

$$Q_L(A, B) = \ell(B) + \langle A - B, \nabla\ell(B)\rangle + \frac{L}{2}\|A - B\|_F^2. \tag{15}$$

Then we have $\ell(A) \leq Q_L(A, B)$ for all $A, B \in \mathbb{R}^{n\times m}$.

Proof: It is easy to check that $\rho = 2\alpha_i + 2n\|WW^\top\|_F + 2\mu\sqrt{m}\|L^i\|_F$ is a Lipschitz constant of $\nabla\ell(\cdot)$, such that

$$\|\nabla\ell(A) - \nabla\ell(B)\|_F \leq \rho\|A - B\|_F,\ \forall\, A, B \in \mathbb{R}^{n\times m}$$

Then following (Beck & Teboulle, 2009, Lemma 2.1), we can draw the conclusion of this lemma. □

Let $p_\rho(Z^i) = \arg\min_{A\geq 0} Q_\rho(A, Z^i)$. It has a closed-form solution

$$p_\rho(Z^i) = \max(0, Z^i - \frac{1}{\rho}\nabla\ell(Z^i)). \tag{16}$$

Following Lemma 1, we have

$$\ell(p_\rho(Z^i)) \leq Q(p_\rho(Z^i), Z^i) \leq Q(Z^i, Z^i) = \ell(Z^i) \tag{17}$$

Thus the projected gradient descent steps in Algorithm 1 is guaranteed to continuously improve the convex objective function (13) and reach the optimal solution.

3.4. Testing

With the trained model, given a new test image, we first collect n_s patches from it and represent them as a $n_s \times d$ matrix X^s. Then, the trained prediction model can be applied by setting $Z^s = X^s\Theta^\top + \mathbf{1}\mathbf{b}^\top$ and $\mathbf{y}^s = \mathbf{1}^\top Z^s W + \mathbf{q}^\top$ sequentially. The final prediction of the scene-level category y for this test image is

$$y = \arg\max_{r\in\{1,\cdots,k\}} \mathbf{y}^s(r). \tag{18}$$

4. Experimental Results

We evaluated the proposed method on 3 standard scene datasets: MIT LabelMe Urban and Natural Scene (LabelMe) (Oliva & Torralba, 2001), 15 Natural Scene dataset (Lazebnik et al., 2006) (Scene 15) and UIUC Sports (Li & Fei-Fei, 2007). The LabelMe dataset contains 2694 color images across 8 scene categories. The Scene 15 dataset contains 15 scene classes, with $200 \sim 400$ images per class. The UIUC Sports dataset has 8 complex sports scene classes and each class has around $800 \sim 2000$ images. In all experiments, we randomly selected 80 images per category for training and used the rest for testing for all methods except the convolutional neural networks which need more training data. All results reported in this section are averages over 10 runs, with different random selections of training and testing images.

In each experiment, we compared the proposed spatial regularized latent semantic representation learning method, denoted as *SR-LSR*, with its variant *LSR* that drops the spatial Laplacian regularizers by setting $\mu = 0$, and eight other related methods for scene classification:

(1) *Bag-of-word based SVM (SVM)*, which is a baseline method that trains SVM classifiers with the dictionary-based bag-of-word model.
(2) *Neural Network with a single hidden layer (1-NN)*.
(3) *Neural Network with two hidden layers (2-NN)*.
(4) *Deep Belief Net (DBN)* (Hinton & Salakhutdinov, 2006) with 3 hidden layers.
(5) *Convolutional Neural Network with three feature stages (CNN-L3)* (LeCun et al., 1998).
(6) *Convolutional Neural Network with two feature stages (CNN-L2)*.
(7) *Chain Model*, which is the probabilistic graphical model with latent object chain structure for scene recognition (Li & Guo, 2012).
(8) *CA-TM*, which is the recent discriminative latent Dirichlet allocation model from (Niu et al., 2012).

For the proposed approach, we set the number of latent variables, i.e., the m value in our model, same as the number of latent units in each hidden layer of the 1-NN, 2-NN and DBN methods. Without special specification, the m value we used in the experiments is 20. For CNN-L3, we used the same model setting as the LeNet-5 in (LeCun et al., 1998). CNN-L2 has the same setting as CNN-L3 except we dropped the third feature stage (LeCun et al., 2010). Moreover, we used much more training data for CNN-L3 and CNN-L2 to get reasonable results. Specifically, 1600, 6400 and 3000 training images are used on LabelMe, UIUC Sports and Scene 15 respectively.

In each experiment, we used 5-fold cross-validation technique to select the trade-off parameters for all methods. For the proposed method, we conducted parameter selection for the trade-off parameters γ_g and γ_z from the set $[0.005, 0.05, 0.1, 0.5, 1, 5]$, and performed selection for μ from the set $[0.1, 0.5, 1, 5, 10]$, while setting $\gamma_f = 0.5$ and

Table 1. Classification results on the *LabelMe* dataset. Each column contains the average classification accuracies of a comparison method on all scene categories. The first eight rows contain results over individual categories and the last row contains their averages. The bold and italic numbers highlight the best and the second best results respectively on each category.

Methods	SVM	1-NN	2-NN	DBN	CNN-L3	CNN-L2	Chain Model	CA-TM	LSR	SR-LSR
coast	0.625	0.446	0.532	**0.971**	0.681	0.747	0.685	0.890	0.882	*0.916*
forest	0.844	0.766	0.786	**0.980**	0.610	0.875	0.880	*0.950*	0.912	*0.950*
highway	0.633	0.678	0.667	0.239	0.501	0.729	0.431	0.840	**0.911**	*0.910*
insidecity	0.720	0.623	0.746	0.925	0.915	0.908	0.662	0.920	*0.948*	**0.950**
mountain	0.572	0.429	0.473	0.446	0.497	0.503	0.499	0.810	*0.881*	**0.886**
opencountry	0.355	0.361	0.442	0.000	0.387	0.511	0.370	0.760	*0.860*	**0.889**
street	0.588	0.665	0.552	0.774	0.563	0.648	0.691	0.860	*0.897*	**0.905**
tallbuilding	*0.808*	0.544	0.511	0.431	0.373	0.211	0.579	**0.930**	0.708	0.797
Average	0.601	0.544	0.575	0.578	0.547	0.682	0.636	0.870	*0.884*	**0.898**

Table 2. Classification results on the *UIUC Sports* dataset. Each column contains the accuracy results of one comparison method across all class categories. The first eight rows contain the classification accuracies over eight scene categories and the bottom row contains their averages. The bold and italic numbers highlight the best and the second best results respectively on each category.

Methods	SVM	1-NN	2-NN	DBN	CNN-L3	CNN-L2	Chain Model	CA-TM	LSR	SR-LSR
badminton	0.947	0.728	0.474	**1.000**	0.677	0.810	*0.985*	0.940	0.939	0.938
bocce	0.822	0.585	0.678	**0.966**	0.411	0.746	0.880	0.490	*0.963*	0.885
croquet	0.649	0.597	0.035	0.000	0.689	0.641	0.634	0.740	*0.749*	**0.793**
polo	0.323	0.329	0.439	0.000	0.472	*0.713*	0.698	0.690	0.703	**0.746**
rockclimbing	0.337	0.446	0.446	0.010	0.230	0.484	0.441	**0.940**	0.429	*0.641*
rowing	0.776	0.688	0.618	0.229	0.307	0.538	0.779	0.750	*0.829*	**0.920**
sailing	0.734	0.642	0.633	0.716	0.770	0.505	0.891	0.830	**0.917**	*0.899*
snowboarding	0.564	0.427	0.482	0.018	0.356	0.243	0.679	*0.710*	0.682	**0.850**
Average	0.642	0.553	0.510	0.373	0.417	0.562	0.756	0.780	*0.794*	**0.839**

all $\{\alpha_i\}$ as 1. We treated each image as a bag of 16×16 patches and extracted a HOG feature vector with length 72 (Dalal & Triggs, 2005) from each patch. We further normalized each HOG vector to have unit L2-norm. For the baseline bag-of-word model, we used a dictionary with 500 visual words (HOG vectors). But for CNN-L3 and CNN-L2, we used raw image data as inputs (LeCun et al., 2010).

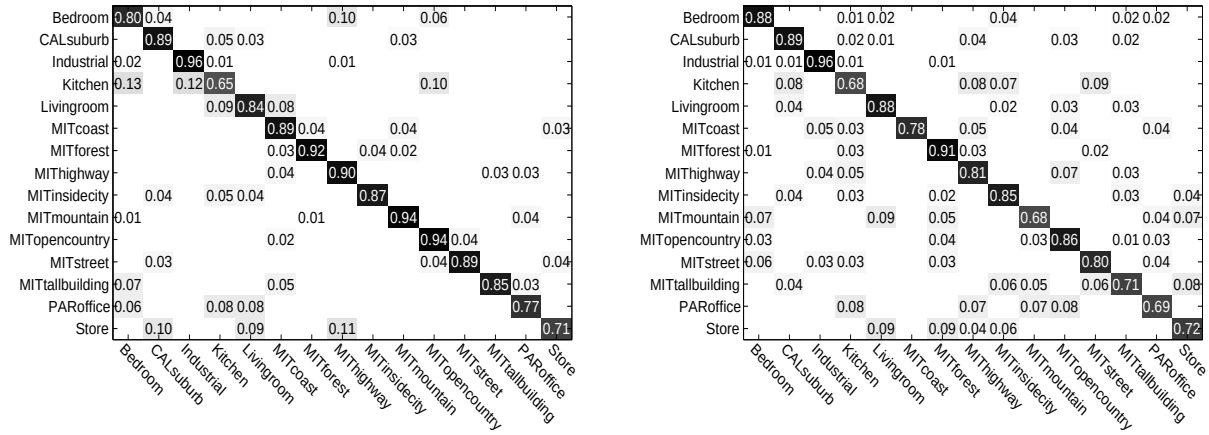
4.1. Scene Classification Results

We evaluated the performance of the proposed method and the other comparison methods in terms of test classification accuracy. The average results over the three scene datasets, *LabelMe*, *UIUC Sports* and *Scene 15*, are reported in Table 1, Table 2, Table 3 and Figure 3 respectively. From Table 1 we can see that our proposed *SR-LSR* method and its variant *LSR* have superior performance on the *LabelMe* dataset, comparing to the other eight methods. Among the eight comparison methods, the neural network methods, *1-NN* and *2-NN*, do not have advantages over the baseline SVM method. The *DBN* method which usually requires a large amount of data for robust deep learning (Ciresan et al., 2012), produces the best results on two categories, but has detection failures on another category since our training set is small. Though more training data has been used for *CNN-L3* and *CNN-L2*, their performance is mediocre among all the other comparison methods. Moreover, *CNN-L2* demonstrates better performance than *CNN-L3*. The *Chain Model* and the *CA-TM* are both based on probabilistic graphical models. *Chain Model* does not show any clear advantage over the baseline methods. But the state-of-the-art work *CA-TM* clearly outperforms the other seven comparison methods on most categories. Nevertheless, the proposed *SR-LSR* and its variant *LSR* outperform *CA-TM* and all the other comparison methods on five out of the total eight categories, and *SR-LSR* achieves the best overall accuracy result averaged over the eight categories. With the spatial Laplacian regularization, *SR-LSR* outperforms *LSR* on almost all categories, and the improvements are significant on many categories including *coast*, *forest*, *opencountry* and *tallbuilding*.

Similar comparison results are observed in Table 2 on the *UIUC Sport* dataset with complex sport scene classes. The

Table 3. Average classification results on the Scene 15 dataset.

Methods	SVM	1-NN	2-NN	DBN	CNN-L3	CNN-L2	Chain Model	CA-TM	LSR	SR-LSR
Accuracy	0.745	0.551	0.680	0.693	0.411	0.763	0.789	0.825	*0.847*	**0.857**



(a) SR-LSR (b) LSR

Figure 3. The confusion matrices of the prediction results produced by the proposed *SR-LSR* (with spatial regularization) method and its variant *LSR* (without spatial regularization) respectively on the Scene 15 dataset.

neural network methods *1-NN* and *2-NN* again have inferior performance than the SVM baseline. *DBN* though produces best results on two categories, it fails to detect *croquet* and *polo* and has very poor overall performance. With more training data, *CNN-L3* and *CNN-L2* produce reasonable results across categories. But they are outperformed by a few other comparison methods. These suggest the deep architecture learning models, which usually require a huge amount of training instances, are not appropriate options for the standard scene classification data we have here. The probabilistic graphical model based methods, *Chain Model* and *CA-TM*, demonstrate good performance on this dataset which suggests contextual information (encoded by intermediate representations) is quite helpful. The proposed *SR-LSR* and *LSR* again maintain their advantages by outperforming all the other comparison methods on five and three individual categories respectively. Moreover, *LSR* produces a better overall average result than the other eight methods, while *SR-LSR*, with additional spatial regularizers, further outperforms *LSR* by 0.045 in terms of the average accuracy over all categories.

Table 3 presents the average accuracy results over 15 categories of the *Scene 15* dataset for all methods. We can see that the proposed *SR-LSR* and *LSR* outperform all the other methods. Figure 3 presents the confusion matrices produced by the prediction results of the proposed methods. From the two confusion matrices, we observe that our proposed methods produce reasonable results even on the indoor categories (e.g. bedroom, kitchen, living room, office, store) which are more difficult to predict (Quattoni & Torralba, 2009). By comparing the two matrices, we can see that the confusion matrix of *SR-LSR* is more sparse. This suggests the latent representation learned with spatial regularization can effectively eliminate some irrelevant scene label categories. Moreover, we can see that the spatial regularization has more impact on outdoor scenes than indoor ones. For example, with the spatial regularization, *SR-LSR* outperforms *LSR* by 0.11 on *MITcoast*, by 0.26 on *MITmountain* and by 0.14 on *MITtallbuilding*. This might be due to the fact that there are more semantic content changes across space in indoor scenes than outdoor scenes.

In summary, the proposed method demonstrates effective performance and outperforms all the other comparison methods on all the three scene datasets.

4.2. Interpretation and Impact of the Latent Variables

In our experiments, we also investigated the meaning of learned latent representations. The latent variables in our proposed model are expected to capture a set of visual enti-

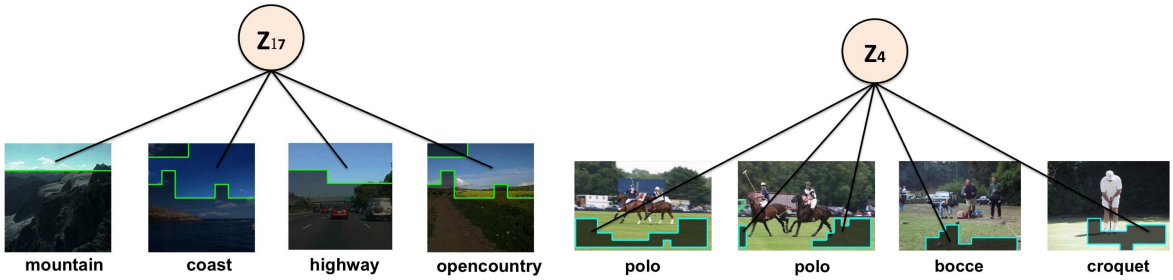


Figure 4. Examples of the latent concepts learned from low-level local features by *SR-LSR*. We used $m = 20$ (i.e. the latent vector has m entries, $Z = [Z_1, \ldots, Z_m]$) in our experiments, and here are the Z_{17} and Z_4 learned in *LabelMe* and *UIUC Sports* respectively.

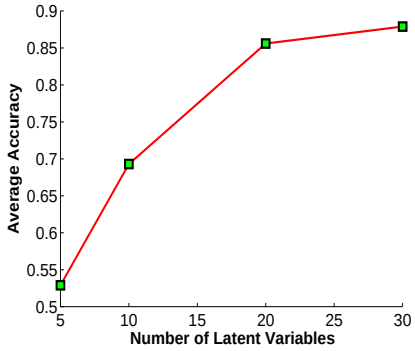


Figure 5. The impact of the number of latent variables, m, on the *Scene15* dataset with $m \in \{5, 10, 20, 30\}$.

ties, i.e., the high-level visual concepts, which can explain the target semantic scene labels. To verify this assumption, we performed visualization on the patches that are mapped to a specific latent concept. Recall that in our model, the j-th patch of the i-th image is mapped into a latent representation vector Z_j^i with length m, corresponding to m latent variables. The larger is an entry value of Z_j^i, the more related this patch is to the corresponding latent concept. The patch is considered to be mapped to the r-th latent concept if the r-th entry of Z_j^i has the largest value among the whole vector. The r-th latent concept can then be visualized by displaying the patches that are mapped to it. Figure 4 presents two examples of our learned latent concepts on the *LabelMe* dataset and the UIUC Sports dataset respectively. The concept Z_{17} has a close relationship with patches over *sky* regions, whereas Z_4 has a strong connection with patches over *grass* regions. This suggests these latent visual concepts are meaningful and are shared across different scene categories. Though it is not appropriate to conclude that Z_{17} is straightly *equivalent* to *sky* or Z_4 is straightly equivalent to *grass*, as these concepts are learned from the low-level gradient-based HOG features, in general our latent representation can capture visual entities that are useful for scene label prediction.

We also studied the impact of the number of latent variables, m, on the performance of our proposed method *SR-LSR*. We tested a range of m values from the set $\{5, 10, 20, 30\}$. The average classification results for different m values on the *Scene15* dataset are presented in Figure 5. We can see that with the increase of the m value from 5 to 20, the classification performance of the proposed approach improves dramatically. It suggests that small number of latent variables can restrain the model from learning useful latent representations for the target prediction task. Nevertheless, from $m = 20$ to $m = 30$, the performance change is very small. On the other hand, with the increase of m value, the computational cost increases dramatically, since the optimization needs to be conducted to learn more latent variable values for each patch in each image. This justifies the selection of $m = 20$ in our previous experiments since $m = 20$ provides a good trade-off between the classification performance and the computational cost.

5. Conclusion

In this paper, we proposed a patch-based latent variable model tailored for semantic scene classification tasks, where a latent layer of variables are used to model high-level latent contextual visual concepts that are both predictable from the low-level feature inputs and discriminative for the semantic output labels. The proposed model can capture both *local* information, through the patches, and *global* information, through the summarization of the latent representation vectors in the whole image and the spatial regularization across patches, for target semantic label prediction. We formulated the model as a joint minimization problem for latent representation learning and prediction model training, and developed an efficient alternating optimization algorithm to solve it, which has closed-form solutions for the model parameter learning step and an efficient projected gradient descent procedure for the latent variable learning step. Our empirical results on three standard scene datasets demonstrated that the proposed method can achieve promising scene classification results and outperform the state-of-the-art scene recognition methods.

References

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2, No. 1:183–202, 2009.

Belkin, M. and Niyogi, P. Using manifold structure for partially labeled classification. In *Proc. of NIPS*, 2002.

Belkin, M., Niyogi, P., and Sindhwani, V. On manifold regularization. In *Proceedings of AISTATS*, 2005.

Bengio, Y., Courville, A., and Vincent, P. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012. URL `http://arxiv.org/abs/1206.5538`.

Bergamo, A., Torresani, L., and Fitzgibbon, A. Picodes: Learning a compact code for novel-category recognition. In *Proceedings of NIPS*, 2011.

Blaschko, M. and Lampert, C. Object localization with global and local context kernels. In *Proceedings of BMVC*, 2009.

Choi, M., Lim, J., Torralba, A., and Willsky, A. Exploiting hierarchical context on a large database of object categories. In *Proceedings of CVPR*, 2010.

Ciresan, D., Meier, U., and Schmidhuber, J. Multi-column deep neural networks for image classification. In *Proceedings of CVPR*, 2012.

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *Proceedings of CVPR*, 2005.

Divvala, S., Hoiem, D., Hays, J., Efros, A., and Hebert, M. An empirical study of context in object detection. In *Proceedings of CVPR*, 2009.

Fidler, S., Boben, M., and Leonardis, A. Evaluating multi-class learning strategies in a generative hierarchical framework for object detection. In *Proceedings of NIPS*, 2009.

He, X. and Zemel, R. Learning hybrid models for image annotation with partially labeled data. In *Proceedings of NIPS*, 2008.

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

Hinton, G., Osindero, S., and Teh, Y. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, July 2006.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. What is the best multi-stage architecture for object recognition? In *Proceedings of ICCV*, 2009.

Kumar, M. and Koller, D. Efficiently selecting regions for scene understanding. In *Proceedings of CVPR*, 2010.

Kwitt, R., Vasconcelos, N., and Rasiwasia, N. Scene recognition on the semantic manifold. In *Proceedings of ECCV*, 2012.

Lazebnik, S., Schmid, C., and Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of CVPR*, 2006.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

LeCun, Y., Kavukcuoglu, K., and Farabet, C. Convolutional networks and applications in vision. In *Proceedings of ISCAS*, 2010.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of ICML*, 2009.

Li, C., Parikh, D., and Chen, T. Automatic discovery of groups of objects for scene understanding. In *Proceedings of CVPR*, 2012.

Li, L. and Fei-Fei, Li. What, where and who? classifying events by scene and object recognition. In *Proceedings of ICCV*, 2007.

Li, X. and Guo, Y. An object co-occurrence assisted hierarchical model for scene understanding. In *Proceedings of BMVC*, 2012.

Lowe, D. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, November 2004.

Niu, Z., Hua, G., Gao, X., and Tian, Q. Context aware topic model for scene recognition. In *Proc. of CVPR*, 2012.

Oliva, A. and Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, May 2001.

Quattoni, A. and Torralba, A. Recognizing indoor scenes. In *Proceedings of CVPR*, 2009.

Ranzato, M., Poultney, C., Chopra, S., and LeCun, Y. Efficent learning of sparse representations with an energy-based model. In *Proceedings of NIPS*, 2006.

Sadeghi, M. and Farhadi, A. Recognition using visual phrases. In *Proceedings of CVPR*, 2011.

Sivic, J., Russell, B., Efros, A., Zisserman, A., and Freeman, W. Discovering objects and their location in images. In *Proceedings of ICCV*, 2005.

Wang, X. and Grimson, E. Spatial latent dirichlet allocation. In *Proceedings of NIPS*, 2007.