JOINT DICTIONARY LEARNING AND TOPIC MODELING FOR IMAGE CLUSTERING

Lingbo Li, Mingyuan Zhou, Eric Wang and Lawrence Carin

Department of Electrical and Computer Engineering, Duke University, Durham NC 27708, USA

ABSTRACT

A new Bayesian model is proposed, integrating dictionary learning and topic modeling into a unified framework. The model is applied to cluster multiple images, and a subset of the images may be annotated. Example results are presented on the MNIST digit data and on the Microsoft MSRC multi-scene image data. These results reveal the working mechanisms of the model and demonstrate state-of-the-art performance.

Index Terms—Bayesian, dictionary learning, sparse coding, topic modeling, image clustering, annotating

1. INTRODUCTION

There has been much recent interest in developing statistical models for automatic clustering and annotation of images, based on local image features as well as available meta-data such as image annotations [1, 2, 3, 4, 5, 6, 7, 8]. Statistical topic models, such as probabilistic Latent Semantic Analysis (pLSA) [9] and Latent Dirichlet Allocation (LDA) [10], originally developed for text analysis, have been successfully applied for these image-analysis tasks by representing an image as a bag of visual words [2]. Local image descriptors, e.g., scale-invariant feature transform (SIFT) [11], are commonly used to extract features from local patches, segments, or super-pixels [7]. The extracted local features are used to design a discrete codebook (i.e., vocabulary) with vector quantization (VQ). When analyzing images, each local descriptor is subsequently assigned to one of the codewords [2, 6, 7], with these codes playing the role of discrete words in traditional documents. Although significant success has been achieved with this approach, there is no principled way to define the codebook size, and hence this parameter must be tuned and is in general a function of the dataset considered. Further, the feature extraction (e.g., via SIFT) is performed separately from the subsequent statistical analysis, making it unclear which features should be used and why one class of features should be preferred.

Recent research on dictionary learning and sparse coding has demonstrated superior performance in a number of challenging image processing applications, including image denoising, inpainting and sparse image modeling [12, 13, 14]. Recent advances in image classification show that substantially improved performance may be achieved by extracting features from local descriptors with dictionary learning and sparse coding, this replacing VQ [15, 16]. However, it is not clear how to integrate these tools with topic modeling, to constitute an overall statistical model.

In this paper we develop a novel Bayesian model that integrates dictionary learning, sparse coding and topic modeling, for joint analysis of multiple images and (when present) associated annotations. The model links topics to probabilities for use of particular dictionary elements, with the dictionary learned jointly while performing topic modeling. The learned model clusters all images into groups, based upon dictionary usage, and a statistical distribution is also provided for words that may be associated with previously non-annotated images (only a subset of the images are assumed annotated when learning the model). Below we develop the modeling framework and explain how inference is performed; the analysis is demonstrated on common databases, with comparisons to previous research that has addressed similar problems.

2. REVIEW OF BAYESIAN DICTIONARY LEARNING AND TOPIC MODELING

Let \( \mathbf{x}_i \in \mathbb{R}^P \) represent the \( i \)th data sample and \( \{ \mathbf{x}_i \}_{i=1}^N \) represents the complete data set under analysis. For the application considered here, each \( \mathbf{x}_i \) corresponds to a set of contiguous pixels (from a small image “patch” extracted from an overall image). The set \( \{ \mathbf{x}_i \}_{i=1}^N \) represents data extracted from \( N \) image patches, across all images of interest. Each \( \mathbf{x}_i \) is assumed to be represented as a linear combination of a sparse set of atoms from a dictionary \( \mathbf{D} \in \mathbb{R}^{P \times K} \), where the columns of \( \mathbf{D} \) represent dictionary atoms. A prior is placed on \( \mathbf{D} \), and a posterior density function on \( \mathbf{D} \) is learned based on \( \{ \mathbf{x}_i \}_{i=1}^N \). Further, the size of the dictionary (total number of active atoms across all \( \mathbf{x}_i \)) is unknown, and to be inferred; i.e., it is anticipated that only a subset of the \( K \) candidate dictionary elements are used. Specifically, for each \( i \), \( \mathbf{x}_i = \mathbf{D} \alpha_i + \epsilon_i \), where \( \alpha_i \in \mathbb{R}^K \) is sparse and \( \| \epsilon_i \|_2 \| \mathbf{x}_i \|_2 \ll 1 \). Additionally, a prior is placed on \( \{ \alpha_i \}_{i=1}^N \), and the statistics of the residual are also to be inferred.

In recent research [14], it has been demonstrated that the beta process (BP) and Bernoulli process (BeP) may be coupled to constitute a prior on \( \{ \alpha_i \}_{i=1}^N \) and \( \mathbf{D} \), to impose the desired sparseness and to infer the dictionary composition and size; this construction also imposes that many of the \( \mathbf{x}_i \) will use a similar subset of columns of \( \mathbf{D} \).

In the model developed below, we will consider analysis of multiple images simultaneously. Each image is assumed drawn from a distribution over topics, and therefore each image is associated with one topic. Each topic is characterized by a distribution over object types that may occur in the image, and in the absence of annotations the number of object types is inferred via the images alone. When annotations are available, the number of objects is linked to the total number of unique words across all annotations. To link the topic model to dictionary learning, each object type will have an associated probability of using columns of \( \mathbf{D} \), and therefore each object type places a prior on the sparseness of the coefficients \( \alpha_i \). In this manner topic modeling and dictionary/feature learning may be performed jointly.

3. THE BAYESIAN HIERARCHICAL MODEL

Given a set of \( M \) images, we represent each image as a set of local patches. The \( m \)th image is represented as \( \{ \mathbf{x}_{mi} \}_{i=1}^{N_m} \), where \( N_m \) represents the total number of patches in this image, and \( \mathbf{x}_{mi} \) is the data from the \( i \)th patch. We use Bayesian dictionary learning on the
data \(\{x_{mi}\}_{m=1,M;i=1,N_m}\) to infer a dictionary \(\mathbf{D}\) under which each \(x_{mi}\) is sparsely represented. Specifically, each \(x_{mi}\) is represented as

\[
x_{mi} = \mathbf{D}(z_{mi} \odot s_{mi}) + \epsilon_{mi}
\]

where \(\odot\) represents the pointwise/Hadamard vector product, \(K\) is the truncation level on the possible number of dictionary atoms, \(z_{mi} = [z_{mi1}, \ldots, z_{miK}]^T\), \(s_{mi} = [s_{mi1}, \ldots, s_{miK}]^T\), \(z_{mik} \in \{0, 1\}\) indicates whether the \(k\)th atom is active within patch \(i\) in image \(m\), \(s_{mi} \in \mathbb{R}\), and \(\epsilon_{mi}\) is the residual error. Note that under an appropriate dictionary \(\mathbf{D}\), \(z_{mi}\) represents the specific sparseness pattern of dictionary usage for \(x_{mi}\). This part of the model is as in previous Bayesian dictionary learning [14], and the unique component of the model is to link the sparse binary vector \(z_{mi}\) to a topic model.

We assume that each image is associated with a topic (scene class). Each topic is in turn characterized by a distribution over objects. Finally, each object is characterized by a distribution on the usage of particular dictionary elements.

Let \(r_m \in \{1, \ldots, T\}\) indicate the topic (scene type) the \(m\)th image is associated with; this random variable is assumed drawn from a multinomial distribution \(\mu = (\mu_1, \ldots, \mu_T)^T\) with a uniform Dirichlet prior as

\[
r_m \sim \sum_{t=1}^{T} \mu_t \delta_t, \quad \mu \sim \text{Dir}(\alpha_\mu / T, \ldots, \alpha_\mu / T),
\]

where \(\delta_t\) is a unit measure at the point \(t\).

Each topic is characterized by a distribution over object types, with a maximum of \(J\) object types assumed. The probability vector

\[
\nu_t \sim \text{Dir}(\alpha_\nu / J, \ldots, \alpha_\nu / J)
\]

defines the probability that each of the \(J\) objects is observed in topic \(t \in \{1, \ldots, T\}\). Hence, if topic \(r_m \in \{1, \ldots, T\}\) is associated with image \(m \in \{1, \ldots, M\}\), then the objects associated with image \(m\) are drawn from \(\nu_{r_m}\). Let

\[
h_{mi} \sim \sum_{j=1}^{J} \nu_{r_m j} \delta_j
\]

represent an indicator variable defining which of the \(J\) objects is associated with patch \(i\) in image \(m\).

We now wish to place a probability distribution on use of dictionary elements (columns of \(\mathbf{D}\)) that is linked to which object a given patch is associated with. Hence, for each object type, we define a probability over usage of the \(K\) potential dictionary elements (columns of \(\mathbf{D}\)). Specifically, the vector \(\pi_j\) defines the probability that each of the \(K\) columns of \(\mathbf{D}\) is employed to represent object type \(j \in \{1, \ldots, J\}\), where the \(k\)th component of \(\pi_j\) is a probability satisfying \(\pi_{jk} \in (0, 1), k \in \{1, \ldots, K\}\). This \(K\)-dimensional vector of probabilities is defined as

\[
\pi_j \sim \prod_{k=1}^{K} \text{Beta}(\eta_{j0}, \eta_{j}(1 - \eta_{j0})).
\]

Then as in conventional dictionary learning [14], the binary vector

\[
z_{mi} \sim \prod_{k=1}^{K} \text{Bernoulli}(\pi_{h_{mi} k})
\]

defines which dictionary elements are used for representation of \(x_{mi}\).

Fig. 1. Graphical representation of the model.

Summarizing, for the \(m\)th image, we first draw a topic \(r_m\). Then, for each patch \(i\) in image \(m\) we draw an object type \(h_{mi} \sim \text{Mult}\{\nu_{r_m}\}\). Finally, for this object type there is an associated probability vector of Bernoulli inputs \(\pi_{h_{mi}}\), from which the binary vector \(z_{mi}\) is drawn, defining which columns of \(\mathbf{D}\) are used for representation of the data in patch \(i\) of image \(m\), \(x_{mi}\).

If annotations are available for at least a subset of the \(M\) images, it is desirable to leverage the information they provide. When available, the words associated with image \(m\) are represented as \(y_m = (y_{m1}, \ldots, y_{mJ})\), where \(y_{mj}\) denotes the number of times word \(j\) is present in the annotation to image \(m\). Typically, \(y_{mj}\) will be either one or zero. Since the number of words in the annotation \(|y_m|\) may be very different than the number of patches \(N_m\), we scale \(y_{m}\) such that the words and image features contribute comparably within the likelihood function. Specifically, we perform the scaling \(y_m' = (N_m / |y_m|) y_m\), where in each component of \(y_m'\) we take the nearest non-negative integer. This scaled annotation count is assumed drawn as

\[
y_m' \sim \text{Mult}(\nu_{r_m}, N_m)
\]

such that the topic-dependent draw of words in the annotation is consistent with the associated draw of patch-dependent objects within the image.

Figure 1 shows a graphical representation of the proposed model, where shaded and unshadowed nodes indicate observed and latent variables, respectively. An array indicates dependence between variables. The boxes are plates that denote repetition, with the number of repetitions indicated by the variables in the corner of boxes.

4. MODEL INFERENCE

Because each consecutive layer in the hierarchical model is in the conjugate-exponential family, efficient Gibbs sampling inference can be used. The inference equations for the dictionary \(\mathbf{D}\), the binary sparse codes \(z\) and the real sparse codes \(s\) are similar to that in [14], and are omitted for brevity. Below we briefly summarize update equations for unique aspects of the proposed model:

- **Sampling \(\pi_j\):** the dictionary usage for object \(j\) is sampled from a beta distribution as:

\[
p(\pi_j | -) \sim \text{Beta}(a_j, b_j)
\]
where \( a_j = a_0 + \sum_{m=1}^{M} \sum_{i=1}^{N_m} \delta(h_{mi} = j)z_{mi} \), and \( b_j = b_0 + \sum_{m=1}^{M} \sum_{i=1}^{N_m} \delta(h_{mi} = j)(1 - z_{mi}) \).

**Sampling \( r_m \):** the scene category topic indicator \( r_m \) is sampled from a \( T \)-dimensional multinomial distribution as:

\[
p(r_m = t|\cdot) \propto \mu_t \prod_{j=1}^{T} t_{j}^{y_{mj} + \sum_{i=1}^{N_m} \delta(h_{mi} = j)}.
\]

**Sampling \( h_{mi} \):** the object indicator \( h_{mi} \) is sampled from a \( J \)-dimensional multinomial distribution as:

\[
p(h_{mi} = j|\cdot) \propto \nu_{mj} \prod_{k=1}^{K} \pi_{jk}^{x_{mik}}(1 - \pi_{jk})^{1-x_{mik}}.
\]

**Sampling \( \nu_{tj} \) and \( \mu_t \):**

\[
p(\nu_{tj}|\cdot) \sim \text{Dir}(\nu_{tj}^*), \quad p(\mu_t|\cdot) \sim \text{Dir}(\mu_t^*, ..., \mu_T^*)
\]

where \( \nu_{tj}^* = \frac{y_{tj}}{T} + \sum_{m=1}^{M} \left[ y_{mj} + \sum_{i=1}^{N_m} \delta(h_{mi} = j) \right] \delta(r_m = t) \) and \( \mu_t^* = \frac{z_t}{T} + \sum_{m=1}^{M} \delta(r_m = t) \).

5. IMPLEMENTATIONS AND EXPERIMENTAL RESULTS

5.1. MNIST Handwritten Digits

We first test the model using the MNIST handwritten digit database, considering 50 samples per digit (digits 0 through 9), and therefore \( N = 500 \) samples are considered in total. In this experiment annotations are not considered. We randomly select 50 partially overlapping patches per digit, and each patch is of size 15 \( \times \) 15 (the original digit images are of size 28 \( \times \) 28). All the patches are used to constitute the data matrix \( X \in \mathbb{R}^{P \times N} \), where \( P = 225 \) and \( N = 25,000 \).

The matrix \( X \) is pre-whitened with principal component analysis (PCA) and the first \( L = 100 \) principle components are preserved as features (\( L = 100 \) keeps about 95\% energy of the original data, achieves a good balance between accuracy and complexity).

We set truncation levels as \( K = 200, J = 50 \) and \( T = 20 \) (similar results were found for larger truncations). Note that these truncation levels are upper bounds on the associated parameter, while the model infers the number of components needed. For example, while we truncate \( T = 20 \), we expect roughly ten topics to be inferred, associated with the ten digits. The inferred dictionary atoms are shown in Fig. 2 in order of importance (probability to be used). For some runs, the proposed model infers more than 10 non-zero topic weights, i.e., some digits such as 4 and 5 tend to occupy more than one topic and there may be a total of 12 topics inferred (there is more than one way some digits may be expressed, and the different type rendering may constitute a unique topic). In order to draw a confusion matrix, multiple topics of the same digit are combined according to the ground truth. The average confusion matrix is calculated in Fig. 3 with the average performance 80.4\%. Note that this performance is achieved with an unsupervised model.

5.2. Microsoft Image Data

For the MSRC data (from Microsoft Research), we choose 320 images from 10 categories of images with manual annotations available. The categories are “tree,” “building,” “cow,” “face,” “car,” “sheep,” “flower,” “sign,” “book” and “chair.” The numbers of images are 45 and 35 in the “cow” and “sheep” classes, respectively and 30 in all the other classes (here the category is expected to be associated with a topic in our model). Each image has size \( 213 \times 320 \) or \( 320 \times 213 \). We evenly divide each (color) image into \( 32 \times 32 \times 3 \) non-overlapping patches. Similarly to the experiment setting for the MNIST digit data set, we choose \( L = 100, K = 200 \) and \( T = 20 \). No parameter optimization has been performed.

For annotations, we remove all annotation-words that occur less than 8 times (approximately 1\% of them). There are 15 unique annotation-words: “building”, “grass”, “tree”, “cow”, “sheep”, “sky”, “water”, “face”, “car”, “flower”, “sign”, “book”, “chair”, “road” and “people”. For each category, we randomly choose 10 images, and remove their annotations, treating them as non-annotated images within the analysis (to allow quantification of inferred-annotation quality). We assume that each annotation word corresponds to a visual object in the image, thus \( J = 15 \).

As shown in Fig. 3, the inferred dictionary includes both color and texture features. With these data we typically infer 14 topics (there are actually 10 scene types from which the images are constituted). We use the same method as in the MNIST experiment to integrate multiple inferred topics/scenes, to compute a confusion matrix; the confusion matrix is shown in Fig. 4. The average performance is 86.8\%, outperforming the results in [7] by 3.9\% under the same test settings.

Based on the learned posterior word distribution \( \nu_{tj} \) for the \( t \)th scene class, we can further infer which objects are most probable for each scene class (topic). Figure 5 shows the \( \nu_{tj} \) for 9 classes, with the largest five probabilities displayed, a good connection is manifested between the words and image types.
6. CONCLUSION

A novel Bayesian statistical model is proposed for joint dictionary learning and topic modeling. The binary sparse codes of each patch under the learned dictionary are linked to the topic modeling by associating the objects in the image with a dictionary usage probability vector (and each topic has a distribution over objects). Feature selection is not required since the learned dictionary is able to discover the important color and texture features from the data, and link them with the topic model. The model is applied to automatic image clustering, with state-of-the-art performance realized on typical image databases.

7. REFERENCES