Enhancing LDA with Segmentation

Bryan Russell

Created: October 13th, 2004
Modified: October 19th, 2004

Motivation

When trying to learn clusters of “words” (SIFT features [3]) for images using generative aspect models, such as pLSA and LDA, we treat each image as a “document” and consider all words belonging to that document. Preliminary results suggest that these models discriminate between object classes reasonably well when there is only one object in the image. However, this approach can be problematic if we consider images with multiple objects in the scene. The task of classifying the single image containing two target objects cannot be done easily using the existing model.

Moreover, for images with single objects it may be desirable to not consider the words belonging to the background and only use the words contained within the object. Figure 1 shows an image with the “words” superimposed and labelled with their class. While the background (context [7]) contains important information about the target object, the clustering task may become easier if we can separate the foreground and background words.

To cope with the above problems, we will incorporate a model for segmentation into LDA. We will show that this model can be solved efficiently using the same stochastic descent techniques as presented in [2].

The Model

We will augment the LDA graphical model in Figure 2 \(^1\) with a Markov Random Field (MRF) that models segmentation. We consider the following Gibbs energy for segmentation:

\[
U(\{g^{(d)}\}_{d=1}^D, \{z^{(d)}\}_{d=1}^D, \{\psi^{(t)}\}_{t=1}^T) = -\sum_{d=1}^D \sum_{i=1}^{M_d} \log \frac{\psi_{z_i}^{(d)} g^{(d)}_{i}}{g^{(d)}_{i}} \\
+ \delta \sum_{(i,j) \in E} [z_i \neq z_j] \frac{\exp(-\epsilon(g_i - g_j)^2)}{\text{dist}(i,j)}
\]

where \(\{g^{(1)}, \ldots, g^{(D)}\}\) are the grayscale pixels of the image corpus containing \(D\) images with each document \(d\) containing \(M_d\) pixels, \(\{z^{(1)}, \ldots, z^{(D)}\}\) are the topic assignments for all the pixels in the corpus, \([\cdot]\) is the indicator function, \(E\) is the set of edges connecting neighboring pixels, \(\text{dist}(\cdot,\cdot)\)

\(^1\)For more information on the generative aspect model (LDA), refer to [2, 4, 1].
Figure 1: When LDA is applied to clustered SIFT features extracted from a corpus of images containing motorbikes and airplanes [6], we get sampled topic labels for each feature. In (b), we have indicated the class assignment for each SIFT feature extracted from (a) by coloring the ellipse corresponding to the SIFT region. Here, we learned two classes and colored the ellipses green and magenta. We discriminated between the two classes by counting the number of topic assignments in a given image. While this scheme was successful for most images in the corpus, this particular image was misclassified since there are more features labeled magenta than green, which corresponded to the airplane and motorbike classes respectively.

Figure 2: The graphical model that describes the generative aspect model (LDA). The portion of the graph inside the boxes should be interpreted as being duplicated as many times as indicated in the top left corner of the box.
Figure 3: The graphical model that describes the generative aspect model (LDA) enhanced with a model for segmentation. In (a), we see the two scenarios for a pixel location: it may or may not be the center of a SIFT feature. If the pixel location does not have a SIFT center, then only the segmentation model influences the topic assignment. Otherwise, the LDA model influences the topic assignment in addition to the segmentation model. Although the graphical model in (a) explicitly shows how LDA and the segmentation model interact, it does not fully capture the pairwise constraints in the segmentation model. (b) explicitly shows these constraints.
is the Euclidean distance between pixel positions (in most cases, this is one), $\psi_q^t \triangleq p(g_i^{(d)} = q | z_i^{(d)} = t)$ is a multinomial over the pixel values at location $i$ in document $d$ conditioned on the topic setting for that pixel, and $\delta$ and $\epsilon$ are user-specified parameters. These parameters are normally set to $\delta = 50$ and:

$$\epsilon = \left(2((z_i - z_j)^2)^{-1} \right.$$

for all pixel locations $i$ and $j$. Local minima solutions to this energy function has recently been explored using an interactive interface known as Grab-Cuts [5].

We will convert this energy function into a Gibbs distribution and insert it into the LDA likelihood:

$$p(\{w^{(1)}_d\}_{d=1}^D, \{g^{(d)}_i\}_{d=1}^D, \{z^{(d)}_i\}_{d=1}^D, \{\theta^{(d)}_t\}_{d=1}^D, \{\phi^{(t)}_j\}_{t=1}^T, \{\psi^{(t)}_i\}_{t=1}^T | \alpha, \beta, \gamma)$$

$$\propto \prod_{d=1}^D p(\theta^{(d)} | \alpha) \prod_{t=1}^T p(\phi^{(t)} | \beta) p(\psi^{(t)} | \gamma) \prod_{i \in S_d} \phi^{(x^{(d)}_i)}_{w^{(d)}_i} \theta^{(d)}_{z^{(d)}_i}$$

$$\exp(-U(\{g^{(d)}_i\}_{d=1}^D, \{z^{(d)}_i\}_{d=1}^D, \{\psi^{(t)}_i\}_{t=1}^T))$$

(3)

where $\{w^{(1)}, \ldots, w^{(D)}\}$ are the extracted SIFT feature labels (i.e. the “words” of the image corpora), $\theta^{(d)}$ are multinomial parameters over the topics for document $d$, $\phi^{(t)}$ are multinomial parameters over the vocabulary words given the topic setting $t$, $S_d$ is the set of pixel indices where a SIFT center occurs, and $\theta^{(d)}_t, \phi^{(t)}_j,$ and $\psi^{(t)}_i$ are sampled from Dirichlet distributions:

$$p(\theta^{(d)} | \alpha) = \frac{\Gamma \left( \sum_{t=1}^T \alpha_i \right)}{\Gamma(1)} \prod_{t=1}^T \theta_i^{(d)}^{(\alpha_i - 1)} = \frac{\Gamma(T\alpha)}{\Gamma(T)} \prod_{t=1}^T \theta_i^{(d)}^{(\alpha_i - 1)}$$

(4)

$$p(\phi^{(t)} | \beta) = \frac{\Gamma \left( \sum_{i=1}^V \beta_i \right)}{\Gamma(1)} \prod_{j=1}^V \phi_i^{(t)}^{(\beta_j - 1)} = \frac{\Gamma(V\beta)}{\Gamma(V)} \prod_{i=1}^V \phi_i^{(t)}^{(\beta_j - 1)}$$

(5)

$$p(\psi^{(t)} | \gamma) = \frac{\Gamma \left( \sum_{i=1}^P \gamma_i \right)}{\Gamma(1)} \prod_{i=1}^P \psi_i^{(t)}^{(\gamma_i - 1)} = \frac{\Gamma(P\gamma)}{\Gamma(P)} \prod_{i=1}^P \psi_i^{(t)}^{(\gamma_i - 1)}$$

(6)

where $\Gamma$ is the Gamma function, $T$ are the number of topics to learn, $V$ are the number of vocabulary words (SIFT cluster centers), and $P$ are the number of pixel values (usually 256). Note that we assume a simpler model and use scalar values for $\alpha, \beta,$ and $\gamma$ instead of vectors. Figure 3 shows the graphical model for this new model.

Notice that we consider location for each “word” (SIFT descriptor) in the image. Intuitively, LDA is affecting the segmentation model at the points where there is a SIFT center. On a high level, we can think of a flat sheet with sparse points containing signals from another process. These signals will be propagated throughout the sheet. Moreover, the process that generates these signals will be affected by the state of the sheet.

The intuition here is that the LDA model will assist the segmentation by clustering topics that (hopefully) correspond to objects. Moreover, the segmentation will assist LDA in learning the topics by segmenting salient regions and forcing all words within the segmentation to have the same class label.
Solution to the Model

The idea here is that we will use the Gibbs sampling framework where we iterate over all pixels in the dataset and sample over the posterior of a topic assignment for a single pixel given all other topic assignments. We initialize this process by randomly assigning topics to each pixel. We have two scenarios: a pixel without and with a SIFT feature centered at that pixel. For the former case, we have:

\[ p(z_i^{(d)} = t | \{ \mathbf{w}^{(d)} \}_{d=1}^{D}, \{ \mathbf{z}_{-i} \}, \{ \mathbf{g}^{(d)} \}_{d=1}^{D}, \alpha, \beta, \gamma) \]

\[ \propto \left( \frac{n_i^{t} + \gamma}{P\gamma + Mt} \right) \prod_{j \in \{(i,e) \in E\}} \exp \left( -\delta[t \neq z_j^{(d)}] \exp(\epsilon (g_{i}^{(d)} - g_{j}^{(d)})^2) \right) \]

where \( \{ \mathbf{z}_{-i} \} \) indicates all the topic variables except \( z_i^{(d)} \) and \( n_i^t \) is the number of pixels with grayscale \( q \) assigned to topic \( t \).

For the latter case, we have:

\[ p(z_i^{(d)} = t | \{ \mathbf{w}^{(d)} \}_{d=1}^{D}, \{ \mathbf{z}_{-i} \}, \{ \mathbf{g}^{(d)} \}_{d=1}^{D}, \alpha, \beta, \gamma) \]

\[ \propto \left( \frac{n_i^d + \alpha}{T\alpha + N^d} \right) \left( \frac{n_i^{t(0)} + \beta}{V\beta + N^t} \right) \left( \frac{n_i^{t(q)} + \gamma}{P\gamma + M^t} \right) \]

\[ \prod_{j \in \{(i,e) \in E\}} \exp \left( -\delta[t \neq z_j^{(d)}] \exp(\epsilon (g_{i}^{(d)} - g_{j}^{(d)})^2) \right) \]

where \( n_i^d \) is the number of times a word (SIFT center) is assigned to topic \( t \) in document \( d \), \( n_i^t \) is the number of times word \( v \) is assigned to topic \( t \), \( N^d \) is the number of words in document \( d \), and \( N^t \) are the number of words assigned to topic \( t \).

Similar to [2], we sample a topic label for each pixel for all of the images in the dataset based on the above posterior formulas. With these topic assignments, we can recover the following model parameters:

\[ \hat{\phi}_{v}^{(t)} = n_i^t + \beta \]

\[ \hat{\phi}_{t}^{(d)} = n_i^d + \alpha \]

\[ \hat{\psi}_{q}^{(t)} = n_i^t + \gamma \]

Results

To test the idea, we generated various one-dimensional signals. The results are shown in Figures 4, 5, and 6.
Figure 4: Simple 1-D example. We generate two different signals with three examples of each (six signals altogether in the training set, each separated by the dashed lines). The height of the signals represent image intensities. Intuitively, each signal contains one object. SIFT feature centers are located at the positions where the signal is filled in, with the cluster center label (word) given above the signal. (a) The ground truth, where the color indicates the correct topic label. (b) The recovered topic labels using LDA with segmentation. At location $i$, the topic is determined by taking the topic with highest probability according to Equation 7 or Equation 8. (c) The probability of a topic assignment at each location (Equation 7 or Equation 8), color coded according to topic label. (d) Recovered $\phi$ parameters. The parameter settings for this example were $\bar{T} = 3$, $\bar{\alpha} = 0.9$, $\bar{\beta} = 0.1$, $\bar{\gamma} = 1000$, $\delta = 4$, and $\bar{\epsilon} = 2$. Notice that we recover the topics almost perfectly.
Figure 5: The same set of signals as in Figure 4, but with an additional signal containing two different words. This signal is supposed to simulate two objects in the same image. Notice that the topics were recovered almost perfectly. The parameter settings for this example were $\hat{T} = 3$, $\hat{\alpha} = 0.9$, $\hat{\beta} = 0.1$, $\hat{\gamma} = 1000$, $\delta = 4$, and $\hat{\epsilon} = 2$. 
Figure 6: For this test, each signal simulates the case of two distinct objects in an image. Again, the topics were recovered almost perfectly. The parameter settings for this example were $\hat{T} = 3$, $\hat{\alpha} = 0.9$, $\hat{\beta} = 0.1$, $\hat{\gamma} = 1000$, $\hat{\delta} = 2$, and $\hat{\epsilon} = 2$. 
Problems to be Resolved

Several issues need to be resolved to make this model better.

- Move away from modelling at the pixel level and towards the SIFT level. Currently, the model is tractable for a few images; we need a model that will scale to thousands of images. Perhaps we can have an edge set where the K-nearest neighbors of each SIFT feature are connected. The segmentation energy formula can then be modified to model this new representation.

- Compare and contrast with the hierarchy of SIFT features (SIFT of SIFTs) idea.

- Often, a SIFT feature useful for describing an object may have its center lie outside of the object shape. This model should account for this and perhaps include these outliers when learning topics.

References


