Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Graphical Model-Based Learning in High Dimensional Feature Spaces

Zhao Song and **Yuke Zhu**
{*zhaos, yukez*}*@sfu.ca*
School of Computing Science
Simon Fraser University
Burnaby, Canada

Abstract

Digital media tend to combine text and images to express richer information, especially on image hosting and online shopping websites. This trend presents a challenge in understanding the contents from different forms of information. Features representing visual information are usually sparse in high dimensional space, which makes the learning process intractable. In order to understand text and its related visual information, we present a new graphical model-based approach to discover more meaningful information in rich media. We extend the standard Latent Dirichlet Allocation (LDA) framework to learn in high dimensional feature spaces.

Introduction

The significance of combining textual and visual information to achieve better understanding has been addressed for forty years (Winograd 1973). A consolidation of text and images often provides additional information to resolve ambiguity. For instance, sample pictures of products accompanied by descriptive text on online shopping websites give customers a clearer perception of the products. Various features are developed for representing visual information; most, if not all, are sparse in high dimensional space, which makes the learning process intractable. In this paper, we limit our discussion on pictorial information; however, our model can be easily extended to other media forms.

Latent Dirichlet Allocation is a generative probabilistic topic model for collections of discrete data (Blei, Ng, and Jordan 2003). In recent years, there has been an increasing interest in developing new topic models based on standard LDA. In addition, LDA has been successfully applied to tackle computer vision problems (Wang and Grimson 2007).

Much work has been reported for integrating text and images using LDA frameworks (Barnard et al. 2003; Feng and Lapata 2010). However, previous work lacks the effectiveness for general feature vectors. In this paper, we propose a novel approach to information retrieval from both textual and visual sources. We extend the standard Latent Dirichlet Allocation (LDA) framework to effectively learn in high dimensional feature spaces.

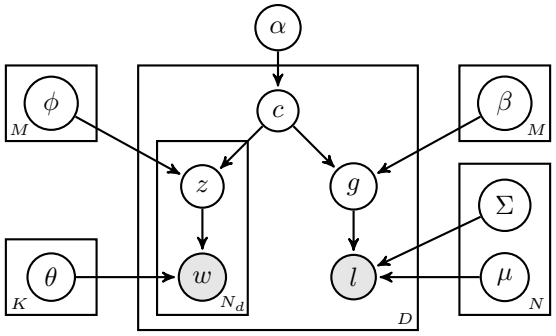

Figure 1: A graphical representation of our model.

Proposed Approach

The goal of our approach is to discover K text-image topics with M clusters and N Gaussian distributions. Let θ denote the word distribution set of all the topics. Let N_d denote the word count in document d. We assume that the distribution of each cluster is a set of Gaussians. Each Gaussian distribution g can be represented by mean vector $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Let $p(c|\alpha)$ denote the weight of cluster c, $\sum_{c\in C} p(c|\alpha) = 1$, where α is a weight distribution over all the clusters. Since Gaussians are generated from clusters, we use $\beta = \{\beta_c\}_{c\in C}$ to represent Gaussian distributions over all clusters, $\beta_c = \{p(g|c)\}_{c\in C}$ where $p(g|c)$ is the probability of Gaussian distribution g given cluster c and $\sum_{g\in G} p(g|c) = 1$ for each c. The topics are also generated from clusters. We use $\phi = \{\phi_c\}_{c\in C}$ to indicate the topic distributions from all the clusters, $\phi_c = \{p(z|c)\}_{z\in Z}$ where $p(z|c)$ is the probability of topic z given cluster c and $\sum_{z\in Z} p(z|c) = 1$ for each c. To generate a text-image document d in collection D:

1. Draw a cluster c from the discrete distribution of cluster importance α, $c \sim Discrete(\alpha)$.
2. Draw a Gaussian g from the multinomial β_c, $g \sim \beta_c$.
3. Draw an image feature vector v_d from Gaussian distribution of $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$.
4. To generate each word in document d:
   (a) Draw a topic $z \sim$ multinomial(ϕ_c)
   (b) Draw a word $w \sim$ multinomial(θ_z)

Let us denote all parameters by $\Psi = \{\theta, \alpha, \beta, \phi, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. Given the data collection $\{(w_d, v_d)\}_{d \in D}$, where w_d is the text and v_d is the feature vector of the image, the log-likelihood of the collection given Ψ can be defined as follows:

$$L(\Psi; D) = \log p(D|\Psi) = \log \prod_{d \in D} p(w_d, v_d|\Psi)$$

To compare the topics in these clusters, we compute $p(z|v)$ for topic $z \in Z$ given vector v. Given the estimated Ψ, we first estimate the probability of vector v given topic z.

$$\begin{aligned} p(v|z, \Psi) &= \sum_{c \in C} \sum_{g \in G} p(v|g, \Psi) p(g|c, \Psi) p(c|z, \Psi) \\ &= \sum_{c \in C} \sum_{g \in G} p(v|\mu_g, \sigma_g) p(g|c, \Psi) \frac{p(z|c) p(c|\alpha)}{p(z|\Psi)} \end{aligned}$$

where $p(z|\Psi) = \sum_{c \in C} p(z|c) p(c|\alpha)$ and $p(v|\mu_g, \sigma_g)$ is based on step 3 in generative process. After we get $p(v|z, \Psi)$, $p(z|v, \Psi)$ can be computed by Bayes' law.

$$\begin{aligned} p(z|v, \Psi) &\propto p(v|z, \Psi) p(z|\Psi) \\ &\propto \sum_{c \in C} p(z|c) p(c|\alpha) \sum_{g \in G} p(v|g) p(g|c) \end{aligned}$$

We omit the EM steps due to space constraints. The parameter estimation algorithm is discussed in the Appendix[1].

Our model is based on the assumption that correlation exists between text and its associated visual information: correlated text in the documents tends to accompany with similar visual features. Figure 1 is a graphical representation of the proposed model.

Dimension Reduction

High dimensionality in the representations of rich media often brings difficulty in the learning process. We need dimension reduction techniques to map high dimensional representations into a lower dimensional feature space. In addition, we want to maintain a certain order of the components in the lower dimensional vector for our subsampling procedure. We use Principal Component Analysis (PCA), which uses linear transformation for the dimension reduction. The transformation of PCA is well defined such that the first component in the low dimensional representation has the greatest variance, the second component has the second greatest variance, and so forth. Thus, PCA allows us to subsample the first k components with the greatest variance. Since the optimal k that truncates noisy components is latent, it is optimized by perplexity-based gradient descent in the learning process.

Image Representation

We use a bag-of-words approach (Li and Perona 2005) for image representation. Each image is treated as a visual document that can be integrated into our graphical models. We use a difference-of-Gaussian (DoG) detector to identify interesting local points, and compute the Scale-Invariant Feature Transform (SIFT) descriptors for these points. In order to generate the codebook, we quantize the feature vectors using k-means clustering. The cluster centroids constitute a visual word codebook $\mathcal{V}$ of size k. Each feature vector is affiliated to the closest centroid in $L2$ norm. Thus, we represent an image as a sparse histogram $\mathcal{H}$ over $\mathcal{V}$, where the i-th component of $\mathcal{H}$ is occurrence frequency of the i-th visual word over the total word counts of the image.

[1]Please see Appendix for further explanation: http://www.sfu.ca/~zhaos/pub/aaai2013_appendix.pdf

Model Evaluations

Experimental Datasets

We evaluate our approach on two real-life datasets containing textual and visual information. **Social20** is an image collection of 20 visual concepts (Li, Snoek, and Worring 2010), which is obtained by randomly selecting 1,000 images for each concept from the image hosting website *Flikr.com*. We select all the images with complete metadata (user tags, titles, descriptions and geographic information) from five broad concepts for evaluation. **Attribute Discovery Dataset** is an image dataset with four shopping categories (Berg, Berg, and Shih 2010). This dataset is collected from the shopping comparison website *like.com*. Textual descriptions of product sample pictures are provided.

Experimental Setup

We compare our proposed model against two baseline methods. The first baseline is the standard text-based LDA using the textual information from metadata. In the second baseline, we regularize the LDA with image feature vectors. We perform 5-fold cross-validation, and use perplexity to evaluate the performance of these models. In the experiment, we report three types of perplexity: textual perplexity, visual perplexity and combined perplexity (see Appendix).

References

Barnard, K.; Duygulu, P.; Forsyth, D.; de Freitas, N.; Blei, D. M.; and Jordan, M. I. 2003. Matching words and pictures. *Journal of Machine Learning Research* 3:1107–1135.

Berg, T. L.; Berg, A. C.; and Shih, J. 2010. Automatic attribute discovery and characterization from noisy web images. *European Conference on Computer Vision* 663–676.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Feng, Y., and Lapata, M. 2010. Visual information in semantic representation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL* 91–99.

Li, F.-F., and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. *IEEE Computer Vision and Pattern Recognition* 524–531.

Li, X.; Snoek, C. G. M.; and Worring, M. 2010. Unsupervised multi-feature tag relevance learning for social image retrieval. *ACM International Conference on Image and Video Retrieval*.

Wang, X., and Grimson, E. 2007. Spatial latent dirichlet allocation. *Neural Information Processing Systems* 1–8.

Winograd, T. 1973. A procedural model of language understanding. *Computer Models of Thought and Language*.