Inter-Battery Topic Representation Learning

Cheng Zhang[1], Hedvig Kjellström[1], Carl Henrik Ek[2]

[1]Robotics, Perception and Learning (RPL)
KTH Royal Institute of Technology, Sweden
{chengz,hedvig}@kth.se

[2]Department of Computer Science
University of Bristol,UK
carlhenrik.ek@bristol.ac.uk

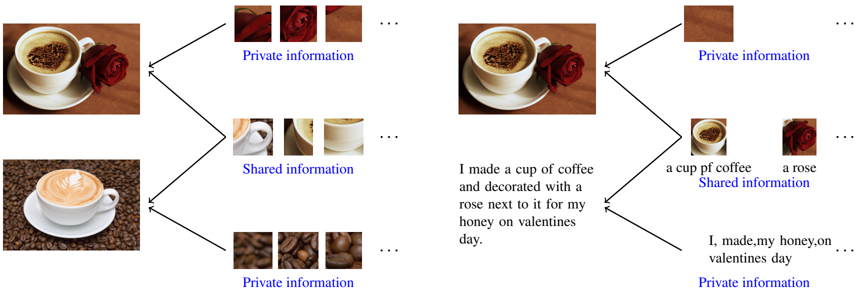
Abstract. In this paper, we present the Inter-Battery Topic Model (IBTM). Our approach extends traditional topic models by learning a factorized latent variable representation. The structured representation leads to a model that marries benefits traditionally associated with a discriminative approach, such as feature selection, with those of a generative model, such as principled regularization and ability to handle missing data. The factorization is provided by representing data in terms of aligned pairs of observations as different views. This provides means for selecting a representation that separately models topics that exist in both views from the topics that are unique to a single view. This structured consolidation allows for efficient and robust inference and provides a compact and efficient representation. Learning is performed in a Bayesian fashion by maximizing a rigorous bound on the log-likelihood. Firstly, we illustrate the benefits of the model on a synthetic dataset,. The model is then evaluated in both uni- and multi-modality settings on two different classification tasks with off-the-shelf convolutional neural network (CNN) features which generate state-of-the-art results with extremely compact representations.

Keywords: Factorized Representation, Topic Model, Multi-View Model, CNN Feature, Image Classification

1 Introduction

The representation of an image has a large impact on the ease and efficiency with which prediction can be performed. This has generated a huge interest in directly learning representation from data [1]. Generative models for representation learning treat the desired representation as an unobserved latent variable [2–4]. Topic models, which are generally a group of generative models based on Latent Dirichlet Allocation (LDA) [3], have successfully been applied for learning representations that are suitable for computer vision tasks [5–7]. A topic model learns a set of topics, which are distributions over words and represents each document as a distribution over topics. In computer vision applications, a topic is a distribution over visual words, while a document is usually an image or a video. Due to its generative nature, the learned representation

This research has been supported by the Swedish Research Council (VR) and Stiftelsen Promobilia.



(a) Uni-modal example (b) Multi-modal example

Fig. 1. Examples of using factorized representations in different scenarios. (a) gives an example of modeling ”a cup of coffee” images. Different images with a cup of coffee all share certain patterns, such as cup handles, cup brims, etc. Moreover, each image also contains patterns that are not immediately related to the ”cup of coffee” label, such as the rose or the coffee beans. These can be considered as private or instance-specific for each image. (b) gives an example of modeling the image and its caption. Different modalities describe the same content as ”a cup of coffee” and ”a rose”. However, the wooden table pattern is not described in the caption and words such as ”I made”, ”my honey” etc. do not correspond to the content of the image. This information can be considered as private or modality-specific.

will provide rich information about the structure of the data with high interpretability. It offers a highly compact representation and can handle incomplete data, to a high degree, in comparison to other types of representation methodologies. Topic models have been demonstrated with successful performance in many applications. Similar to other latent space probabilistic models, the topic distributions can easily be adapted with different distributions with respect to the types of the input data. In this paper, we will use a LDA model as our basic framework and apply an effective factorized representation learning scheme.

Modeling the essence of the information among all sources of information for a particular task has been shown to offer high interpretability and better performance [6, 8–12]. For example, for object classification, separating the key features of the object from the intra-class variations and background information is key to the performance. The idea of factorized representation can be traced back to the early work of Tucker, ’An Inter-Battery Method of Factory Analysis’ [8], hence, we name the model presented in this paper Inter-Battery Topic Model (IBTM).

Imagine a scenario in which we want to visually represent ”a cup of coffee”, illustrated in Figure 1 (a). Apart from a cup of coffee, such images commonly contain additional information that is not correlated to this labeling, e.g., the rose and the table in the upper image and the coffee beans in the lower image. One can think of the information that is common among all images of this class and thus correlated with the label, as the *shared* information. Images with a cup of coffee will share a set of ”cup of coffee” topics between them. In addition, each image does also contain information that can be found only in a small share of the other images. This information can be thought of as *private*. Since the shared, but not the private, information should be employed in the estimation task (e.g., classification), it is highly beneficial to use a factorized model which represents the information needed for the tasks (shared topics) separately from the information that is not task related (private topics).

A similar idea can be applied in the case when two different modalities of the data are available. A common case is images as one modality and the captions of the images as another, as shown in Figure 1 (b). In this scenario, commonly not all of the content in the image has its corresponding caption words; and not every word in the caption has its corresponding image patches. However, the important aspects of the scene or object depicted in the image are also described in the caption, and vice versa, the central aspects of the caption are those that correlate with what is seen in the image. Based on this idea, an ideal multi-modal representation should factorize out information that is present in both modalities (words describing central concepts, and image patches from the corresponding image areas) and represent it separately from information that is only present in one of the modalities (words not correlated with the image, and image patches in the background). Other modality examples include video and audio data captured at the same event, or optical flow and depth measurements extracted from a video stream.

To summarize, there is a strong need of modeling information in a factorized manner such that shared information and private information are represented separately. In our model, the shared part of the representation will capture the aspects of the data that are essential for the prediction (e.g., classification) task, leading to better performance. Additionally, inspecting the factorized latent representation gives a better understanding of the structure of the data, which is helpful in the design of domain-specific modeling and data collection.

The main contribution of this paper is *a generative model, IBTM, for factorized representation learning, which efficiently factorizes essential information for an estimation task from information that is not task related* (Section 3). This results in a very effective latent representation that can be used for predication tasks, such as classifications. IBTM is a general framework, which is applicable to both single- and multi-modal data, and can easily be adapted to data with different noise levels. To infer the latent variables of the model, we derive an efficient variational inference algorithm for IBTMs.

We evaluate our model in different experimental scenarios (Section 4). Firstly, we test IBTM with a synthetic dataset to illustrate how the learning is performed. Then we apply IBTM to state-of-the-art datasets in different scenarios to illustrate how different computer vision tasks benefit from IBTM. In a multi-modal setting, modality-specific information is factorized from cross-modality information (Section 4.2.1.2 and 4.2.2.2). In a uni-modal setting, instance-specific information is factorized from class-specific information (Section 4.2.1.1 and 4.2.2.1).

2 Related Work

With respect to the scope of this paper, we will summarize the related work mainly from two aspects: Topic Modeling and Factorized Models.

Topic Modeling. Latent Dirichlet Allocation (LDA) [3] is the corner stone of topic modeling. In computer vision tasks [5–7], topic modeling assumes that each visual document is generated by selecting different themes while the themes are distributions over visual words. In correspondence with other works in representation learning, the themes can be interpreted as factors, components or dictionaries. The topic distribution for each document can be interpreted as factor weights or as a sparse and low-dimensional representation of the visual document. This has achieved promising results in different tasks and provided an intuitive understanding of the data structure. For computer vision tasks, topic modeling has been used for classification, either with supervision in the model [13–17] or by learning the topic representation in an unsupervised manner and applying standard classifiers such as softmax regression on the latent topic representation [12]. Another interesting direction using topic modeling in computer vision is the multi-modal extension of topic models; it has been applied to tasks such as image annotation [11, 18–20], contextual action/object recognition [7] and video tagging [6]. Being a generative model, it represents all information found in the data. However, for a specific task, only a portion of this information might be relevant. Extracting this information is essential for a good representation of the data. Hence a model that describes key information for the current task is beneficial.

Factorized Models. The benefit of modeling the between-view variance separately from the within-view variance was first pointed out by Tucker [8]. It was rediscovered in machine learning in recent years by Ek et.al. [21]. Recent research in latent structure models has also shown that modeling information in a factorized manner is advantageous for both uni-modal scenarios [10, 12, 22], in which only one type of data is available and multi-modal scenarios [6, 9, 21], in which different views correspond to different modalities. For uni-modal scenarios, a special words topic model with a background distribution (SWB) [22] is one of the first studies on factorized representation using topic model for information retrieval tasks. In addition to topics, SWB uses a words distribution for each document to represent document specific information and a global word distribution for background information. As shown in the experiments, this text-specific factorization model is less suitable for computer vision tasks than IBTM. Works that apply such a factorized scheme on multi-modal topic modeling [6, 11] include the multi-modal factorized topic model [11] and Video Tags and Topics Model (VTT) [6]. The multi-modal factorized topic model which is based on correlated topic models [23] only provides an implicit link between different modalities with hierarchical Dirichlet priors since the factorization is enforced on the logistic normal prior, while VTT is only designed for the specific application.

In this paper, we present a general framework IBTM which models the topic structure in a factorized manner and can be applied to both uni- and multi-modal scenarios.

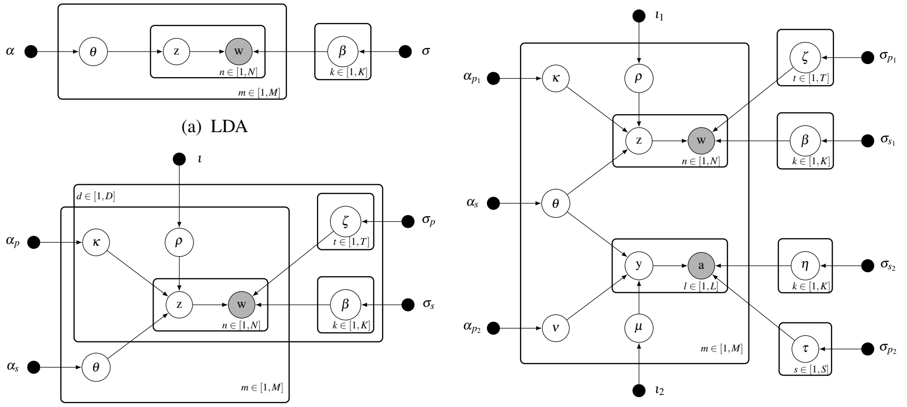
3 Model

In this section, firstly, we will shortly review LDA [3] which IBTM is based on and then present the modeling details and inference of IBTM. Finally, we will describe how the latent representation can be used for classification tasks with which we evaluate our approach.

3.1 Latent Dirichlet Allocation

LDA is a classical generative model which is able to model the latent structure of discrete data, for example, a bag of words representation of documents. Figure 2 (a) shows the graphic representation of LDA [3]. In LDA, the words (visual words) w are assumed to be generated by sampling from a per document topic distribution $\theta \sim Dir(\alpha)$ and a per topic words distribution $\beta \sim Dir(\sigma)$. The Dirichlet distribution is a natural choice as it is conjugate to multinomial distribution.

3.2 Inter-Battery Topic Model



(a) LDA

(b) Generalized IBTM

(c) Two view IBTM

Fig. 2. Graphical representations

We propose the IBTM which models latent variables in a factorized manner for multi-view scenarios. Firstly, we will explain how to apply IBTM to a two view scenario such that it easily can be compared to other models [7, 8, 18, 20]. In the following, we present the more generalized IBTM, which can encode any number of views.

Two View IBTM. The two view version of IBTM, shown in Figure 2 (c), is an LDA-based model, in which each document contains two views and the words w and a from the two views are observed respectively. The two views can represent different types of data, such as two modalities, for example, image and caption as in Figure 1 (b); or two different descriptors for the same data, for example, SIFT and SURF features of the same image. They can also be two instances of the same class, for example, the two cups of coffee as in Figure 1 (a).

The key of IBTM is that we assume that topics are factorized. We do not force topics from two views to be matched completely since commonly each view has its view-specific information. Hence, in our model, a shared topic distribution between two views for each document is separated from a private topic distribution for each view. As in Figure 2 (c), $\theta \sim Dir(\alpha_s)$ is the *shared* per topic distribution for each document, and correspondingly $\beta \sim Dir(\sigma_{s1})$ and $\eta \sim Dir(\sigma_{s2})$ are the per shared topic words distributions for each view. $\kappa \sim Dir(\alpha_{p1})$ and $\nu \sim Dir(\alpha_{p2})$ are the *private* per document topic distributions for each view respectively, and correspondingly $\zeta \sim Dir(\sigma_{p1})$ and $\tau \sim Dir(\sigma_{p2})$ are the private per topic word distributions for each view. To determine how much information is shared and how much information is private, partition parameters $\rho \sim Beta(\iota_1)$ and $\mu \sim Beta(\iota_2)$ are used for each view. In this case, to generate topic assignments for each word in each view, z and y are sampled as

$$z \sim Mult([\rho * \theta; (1-\rho) * \kappa])^1, \quad y \sim Mult([\mu * \theta; (1-\mu) * \nu]). \tag{1}$$

In the extreme cases, if $\rho = 0$ and $\mu = 0$, no information is shared between the two views and IBTM becomes two separated LDA. Otherwise, if $\rho = 1$ and $\mu = 1$, IBTM becomes a regular multi-modal topic model [7, 20].

The whole IBTM is represented as:

$$\begin{aligned}
&p(\kappa, \theta, \nu, \rho, z, w, \mu, y, a, \zeta, \beta, \eta, \tau|\Theta) \\
=&\left(\prod_{t=1}^{T} p(\zeta_t|\sigma_{p1})\right)\left(\prod_{k=1}^{K} p(\beta_k|\sigma_{s1})\right)\left(\prod_{k=1}^{K} p(\eta_k|\sigma_{s2})\right)\left(\prod_{s=1}^{S} p(\tau_s|\sigma_{p2})\right)\prod_{m=1}^{M}\Bigg(p(\kappa_m|\alpha_{p_1})p(\theta_m|\alpha_s) \\
&p(\nu_m|\alpha_{p_2})p(\rho_m|\iota)p(\mu_m|\iota_2)\left(\prod_{n=1}^{N} p(z_{mn}|\kappa_m, \theta_m, \rho_m)p(w_{mn}|z_{mn}, \beta, \zeta)\right)\left(\prod_{l=1}^{L} p(y_{ml}|\nu_m, \theta_m, \mu_m)p(a_{ml}|y_{ml}, \eta, \tau)\right)\Bigg)
\end{aligned}$$

where $\Theta = \{\alpha_{p_1}, \alpha_s, \alpha_{p_2}, \sigma_{p_1}, \sigma_{p_2}, \sigma_{s_1}, \sigma_{s_2}, \iota_1, \iota_2\}$, and as in the graphic representation of IBTM in Figure 2 (b), the total number of documents is M; the number of words for each document is N and L for the first view and the second view respectively; the number of shared topics for both views is K; the number of private topics is T and S and the vocabulary size is V and W for the first view and the second view respectively.

Mean Field Variational Inference. Exact inference on this model is intractable due to the coupling between latent variables. Variational inference and sampling based methods are the two main groups of methods to perform approximate inference. Variational inference is known for its fast convergence and theoretical attractiveness. It can also be easily adapted to online requirements when facing big data or streaming data. Hence, in this paper, we use mean field variational inference for IBTM. The fully factorized variational distribution is assumed following the mean field manner:

$$q(\kappa, \theta, \nu, \rho, z, \mu, y, \zeta, \beta, \eta, \tau) = q(\kappa)q(\theta)q(\nu)q(\rho)q(z)q(\mu)q(y)q(\zeta)q(\beta)q(\eta)q(\tau)\ .$$

For each term above, the per document topic distributions are: $q(\kappa) = \prod_{m=1}^{M} q(\kappa_m|\delta_m)$ where $\delta_m \in \mathbb{R}^T$; $q(\theta) = \prod_{m=1}^{M} q(\theta_m|\gamma_m)$ where $\gamma_m \in \mathbb{R}^K$; $q(\nu) = \prod_{m=1}^{M} q(\nu_m|\varepsilon_m)$ where $\varepsilon_m \in \mathbb{R}^S$. The per word topic assignments are: $q(z) = \prod_{m=1}^{M}\prod_{n=1}^{N} q(z_{mn}|\phi_{mn})$ where $\phi_{mn} \in \mathbb{R}^{K+T}$ such that the first K topics correspond to the shared topics and the last T topics correspond to the private topics; $q(y) = \prod_{m=1}^{M}\prod_{l=1}^{L} q(y_{mn}|\chi_{mn})$ where $\chi_{mn} \in \mathbb{R}^{K+S}$ such that the first K topics correspond to the shared topics and the last S topics correspond to the private topics. The per document beta parameters are: $q(\rho) = \prod_{m=1}^{M} q(\rho_m|r_m)$ and $q(\mu) = \prod_{m=1}^{M} q(\mu_m|u_m)$. Finally, the per topic words distributions are: $q(\zeta) = \prod_{t=1}^{T} q(\zeta_t|\xi_t)$, $q(\beta) = \prod_{k=1}^{K} q(\beta_k|\lambda_k)$, $q(\eta) = \prod_{k=1}^{K} q(\eta_k|\upsilon_k)$, $q(\tau) = \prod_{s=1}^{S} q(\tau_s|o_s)$. All the variational distributions follow the same family of distributions under the model assumption.

[1] We use $[A;B]$ to indicate matrix and vector concatenation

Applying Jensen's inequality on the log likelihood of the model, we get the evidence lower bound (ELBO) $\mathscr{L}$:

$$\log p(w,a,\mathbb{Z}|\Theta) = \log\int \frac{p(w,a,\mathbb{Z}|\Theta)q(\mathbb{Z})}{q(\mathbb{Z})}d\mathbb{Z} \geq \mathop{\mathbb{E}}_{q}[\log p(w,a,\mathbb{Z}|\Theta)] - \mathop{\mathbb{E}}_{q}[\log q(\mathbb{Z})] = \mathscr{L}$$

where $\mathbb{Z} = \{\kappa,\theta,\nu,\rho,z,\mu,y,\zeta,\beta,\eta,\tau\}$.

By maximizing the ELBO, we get the update equations for the variational parameters. Only the ones that differ from LDA are presented here and derivation details are presented in the supplementary material. The update equations for the per document topic variational distribution are:

$$\delta_{mt} = \alpha_{p1} + \sum_{n=1}^{N}\phi_{mn(K+t)}, \quad \gamma_{mk} = \alpha_s + \sum_{n=1}^{N}\phi_{mnk} + \sum_{l=1}^{L}\chi_{mlk}, \quad \varepsilon_{ms} = \alpha_{p2} + \sum_{l=1}^{L}\chi_{ml(K+s)}.$$

The update equation for the topic assignment in the first view is, when $i \leq K$:

$$\phi_{mni} = \exp\left(\left(\Psi(\gamma_{mk}) - \Psi(\sum_{i=1}^{K}\gamma_{mi})\right) + \left(\Psi(r_{m1}) - \Psi(r_{m1}+r_{m2})\right) + \sum_{v=1}^{V}[w_{mn}=v]\left(\Psi(\lambda_{iv}) - \Psi(\sum_{p=1}^{V}\lambda_{ip}))\right) - 1\right);$$ [2]

and when $i > K$ (as $i = K+t$):

$$\phi_{mni} = \exp\left(\left(\left(\Psi(\delta_{m(i-K)}) - \Psi(\sum_{p=1}^{T}\delta_{mp})\right) + \left(\Psi(r_{m2}) - \Psi(r_{m1}+r_{m2})\right)\right) + \sum_{v=1}^{V}[w_{mn}=v]\left(\Psi(\xi_{iv}) - \Psi(\sum_{p=1}^{V}\xi_{ip}))\right) - 1\right).$$

The update equations for the partition parameters are:

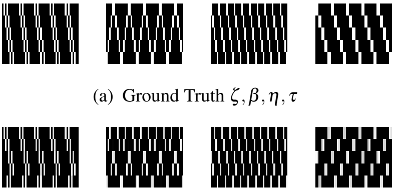
$$r_{m1} = \iota_{11} + \sum_{n=1}^{N}\sum_{i=1}^{K}\phi_{mni}, \quad r_{m2} = \iota_{12} + \sum_{n=1}^{N}\sum_{i=K}^{K+T}\phi_{mni}$$

The update for the second view follows equivalently.

In the implementation, all global latent variables are initialized randomly except for the shared per topic word distribution for the second modality, which is initialized uniformly. Due to the exchangeability of Dirichlet distribution which leads to rotational symmetry in the inference, initializing only one of the shared per topic word distribution randomly will increase the robustness of the model performance.

Generalized IBTM. It is straight-forward to generalize the two view IBTM to more views. The graphical representation of the generalized IBTM is shown in Figure 2 (b), where D is the total number of views. When $D = 2$, the models in Figure 2 (b) and 2 (c) are identical. The inference procedure can be adapted easily, since the updates of both topic assignments and partition parameters for each view follow the same form. The only difference is the per document shared topic variational distribution $\gamma_{mk} = \alpha_s + \sum_{d=1}^{D}\sum_{n=1}^{N^{(d)}}\phi_{mnk}^{(d)}$, where $\phi_{mnk}^{(d)}$ is the variational distribution of the topic assignment for the d-th view.

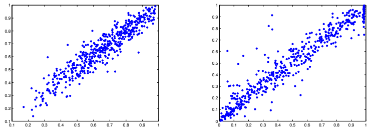
[2] $\Psi(x)$ is the digamma function.



(a) Ground Truth ζ, β, η, τ

(b) Estimated ζ, β, η, τ

Fig. 3. (a) shows the ground truth of the four per topic word distributions in the two-view IBTM model; (b) shows the inferred distributions. Each row in the distributions represents a topic and each column presents a word. There are 5 topics in each distribution ($K = T = S = 5$), and the vocabulary size is 50 for both views ($V = W = 50$).



(a) Estimation of ρ (b) Estimation of μ

Fig. 4. Visualization of synthetic data experiment on inference of partition parameters ρ and μ. The x-axis is the ground truth and the y-axis is the estimation. Each dot in the plot presents the ρ and μ for a document.

3.3 Classification

Topic models provide a compact representation of the data. Both LDA and IBTM are unsupervised models and can be used for representation learning. The topic representation can be applied to different tasks, for example, image classification and image retrieval. Commonly, the whole topic representation will be employed for these tasks using LDA. Using IBTM, we will only rely on the shared topic space which represents the information essence. For image classification, we can simply apply a Support Vector Machine or softmax regression, taking the shared topic representation as the input. In our experimental evaluation, softmax regression is used. Although there are different types of supervised topic models [13, 16] where class label is encoded as part of the model, the work in [12] shows that the performance on computer vision classification tasks using supervised model and unsupervised model with an additional classifier is similar. The minor improvement on the performance commonly comes with significant improvement of computation cost. Hence, we keep IBTM as a general framework for representation learning in an unsupervised manner.

4 Experiments

In the experiments, firstly, we will evaluate the inference scheme and demonstrate the model behavior in a controlled manner in Section 4.1. Then we will use two benchmark datasets to evaluate the model behavior in real world scenarios in Section 4.2. For this purpose, we use the LabelMe natural scene data for natural scene classification [18, 24, 25] and the Leeds butterfly dataset [26] for fine-grained categorization.

4.1 Inference Evaluation using Synthetic Data

To test the inference performance, we generate a set of synthetic data using the model given different topic distributions ζ, β, η, τ and hyper-parameters for $\mu, \rho, \kappa, \theta, \nu$. We generate 500 documents and each document has 100 words for each view. Given the generated data, a correct inference algorithm will be able to recover all the latent parameters. Figure 3 (a) shows the ground truth that we used for the per topic words distribution and the estimation of these latent variables using variational inference as described in Section 3.2. All the topics are correctly recovered. Due to the exchangeability of Dirichlet distribution, the estimation gives different order of the topics which is shown as row-wise exchanges in Figure 3(b). Figure 4 shows the parameter recovery for the partition parameters ρ and μ which are generated from beta distribution. In the example, we use $\iota_1 = (4,2)$ and $\iota_2 = (1,1)$ as hyper-parameters for the beta distributions. In this setting, the first view is comparably clean; the second view is more noisy with big variations on the noise level among the data. As Figure 4 shows, almost all the partition parameters are correctly recovered.

4.2 Performance Evaluation using Real-World Data

In this section, model performance is evaluated on real-world data. We present two experimental groups. The first one is using the LabelMe natural scene dataset [18, 24] and the second one is using the Leeds butterfly dataset [26] for fine-grained classification. We focus on the model performance where we investigate the distribution of topics and partition parameters. This will provide us with insight into the data structure and model behavior. Thereafter, we will present the classification performance. In these experiments, the classification results are obtained by applying softmax regression on the topic representation. In all experimental settings, the hyper-parameters for the per document topic distributions are set to $\alpha_* = 0.8$, the hyper-parameters for the per topic word distributions are set to $\sigma_* = 0.6$ and the hyper-parameters for the partition variables are set to $\iota_* = (5,5)$[3]. We also perform experiments with different features, including off-the-shelf CNN-features from different layers and traditional SIFT features. Here, we only present the results using off-the-shelf CNN conv5_1 features as an example. We use the pre-trained Oxford VGG 16-layer CNN [27] for feature extraction. We create sliding windows in 3 scales with a 32 pixels step size to extract features, in the same manner as [28], and use K-means clustering to create a codebook and represent each image using a bag-of-visual-words. The vocabulary size is 1024. In general, the performance is robust when higher layers are used and when the vocabulary size is sufficient. More results using different features and different parameter settings are enclosed in the supplementary material.

4.2.1 LabelMe Dataset. We use the LabelMe Dataset as in [18, 25] for this group of the experiments. The LabelMe dataset contains 8 classes of 256×256 images: highway, inside city, coast, forest, tall buildings, street, open country and mountain. For each class, 200 images are randomly selected, half of which are used for training, and half of which are used for testing. This results in 800 training and 800 testing images. We perform the experiment in two different scenarios: Image and Image, where only images are available; and Image and Annotation, where different modalities are available.

[3] α_* includes α_{p_1},α_s and α_{p_2}. σ_* includes σ_{p_1}, σ_{p_2}, σ_{s_1} and σ_{s_2}. ι_* includes ι_1 and ι_2.

4.2.1.1 Image and Image. In this experiment, we explore the scenario in which only one modality is available. We want to model essential information that captures the within class variations and explains away the instance specific variations. Both views are bag-of-CNN Conv5_1 feature representations of the image data. For each document, two training images from the same class are randomly paired. This represents the scenario as shown in the introductory Figure 1 (a). For the experimental results presented below, the numbers of topics are set to $K = 15$, $T = 15$, $S = 15$.[4]

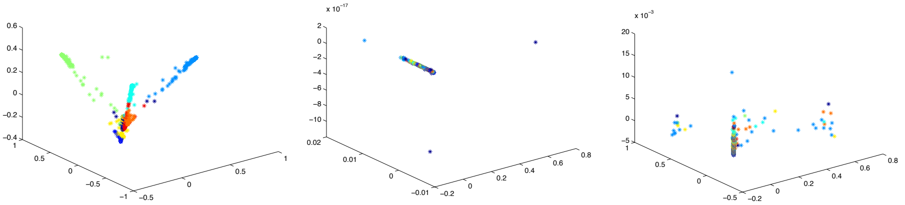
Figure 6 shows the histograms of the partition parameters in this case. Figure 6 (a) and (b) appear to be similar. This is according to intuition; since both views are images and they are randomly paired within the same classes, the statistical features are expected to be the same for both views. Most partition parameters are larger than 0.8, which means that large parts of information can be shared between images from the same class and that the CNN Conv5_1 features provide a good raw representation of the images. For image pairs with more variation that does not correlate with the image class, the partition parameters will be smaller. The essential information ratio varies among images which causes the partition parameters to vary among different images.

Figure 5 visualizes the document distribution in different topic representation spaces. Figure 5 (a) shows that documents from different classes are well separated in the space defined by the shared topic representation. Figure 5 (b) and (c) show that documents from different classes are more mixed in the private topic spaces. Thus, the private information is used to explain instance specific features of a data point, but not class-specific features – these have been pushed into the shared space, according to the intention of the model. The variations in the private spaces are small due to the low noise ratio in the dataset. For the classification performance where only images are available, using IBTM with classification using only the shared representation leads to a classification rate of 89.75%. The classification results are summarized in Table 1. A standard LDA obtains better performance than PCA with the same number of dimensions. IBTM outperforms LDA with the same number of topics and can even obtain better results than using the full dimension (1024) of bag-of-Conv5_1 features together with linear SVM. While using SWB [22] [5], the performance is unsatisfactory for such computer vision tasks due to the noisy properties of images. The results show that IBTM is able to learn a factorized latent representation, which separates task-relevant variation in the data from variation that is less relevant for the task at hand, here classification.

4.2.1.2 Image and Annotation. In this experiment, we explore the scenario when two different modalities are available for different views. We use the bag-of-Conv5_1 representation of images as the first view and the image annotations as the second view. The word counts for the annotations are scaled with the annotated region. For each document, 79 Conv5_1 features are extracted from the image view, and the sum over the word histogram for each view is normalized to 100. The number of topics is set to $K = 15$, $T = 15$, $S = 15$ in the experimental results presented here. Figure 7 shows the

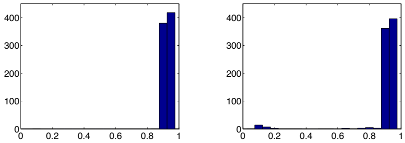
[4] The performance is robust with a sufficient amount of topics, 15 or higher. More results with different numbers of topics are presented in the supplement.

[5] We implemented SWB using Gibbs Sampling following the description in the paper [22]. The parameter settings are the same as in [22]. Linear SVM is used for classification using the topic representation from SWB. More analysis using SWB is presented in the supplementary material of this paper.



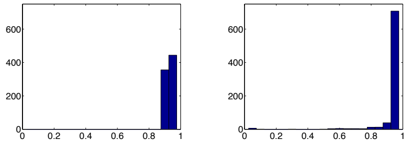
(a) θ, Shared (b) κ, Private (c) ν, Private

Fig. 5. Visualization of the shared topic representation (θ) and private topic representations (κ and ν) for LabelMe experiments using randomly paired images from the same class. The documents of different classes are colored differently and the plots show the first three principal components after applying PCA on the per document topic distributions for all the training data.



(a) Hist of ρ (Img) (b) Hist of μ (Img)

Fig. 6. The histogram over partition parameters of the LabelMe image-image experiment. Img indicates that this modality uses natural images.

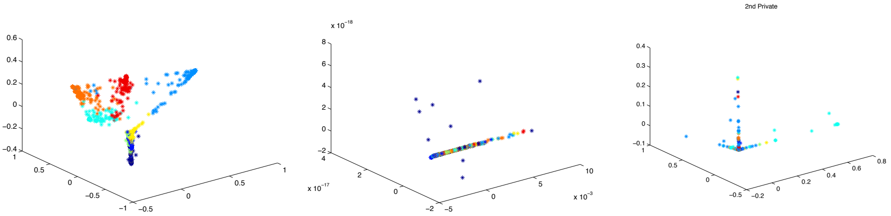


(a) Hist of ρ (Img) (b) Hist of μ (Ann)

Fig. 7. The histogram over partition parameters of the LabelMe image-annotation experiment. Img indicates that this modality uses natural images. Ann indicates that this modality uses image annotations.

DocNADE [25]	SupDocNADE[25]	Full SVM	PCA15 SVM	LDA15	SWB15 [22]	IBTM15
81.97%	83.43%	87%	80.88%	85.25%	59.88%	**89.75%**

Table 1. The performance comparison for Image and Image experiment with the LabelMe dataset. Full SVM shows the performance using SVM on the bag of Con5_1 features, while PCA 15 SVM shows the performance after applying PCA and using the top 15 principal components. LDA 15 shows the result using LDA with 15 topics and classification by softmax regression. IBTM 15 shows the result using IBTM with 15 shared topics and classification by softmax regression only on the shared topics.



(a) θ, Shared (b) κ, Private (c) ν, Private

Fig. 8. Visualization of the shared topic representation (θ) and private topic representations (κ and ν) for LabelMe experiments using image features for the 1st view and annotation for the 2nd view.The documents of different classes are colored differently and the plots show the first three principal components after applying PCA on the per document topic distributions for all the training data.

histograms of the partition parameters ρ and μ for the two views respectively. Figure 7 (b) shows that the partition parameters are more concentrated around large values compared to Figure 7 (a), which indicate that most annotation information is more essential. This is consistent with the intuition of the relative noise levels in image vs annotation data.

Full SVM	PCA 15	LDA15	SWB 2V [22]	IBTM15 1V	IBTM15 2V
87.63%	84.88%	85.38%	61%	**89.38%**	**95%**

Table 2. The performance comparison for the image-annotation experiment for the LabelMe dataset. "IBTM15 1V" shows the prediction performance with only images available (1 view testing) and "IBTM15 2V" shows the prediction performance with both images and annotation available (2 view testing). For "SWB 2V", we concatenate words from images and captions for each document for both training and testing to use SWB since it is a single-view model

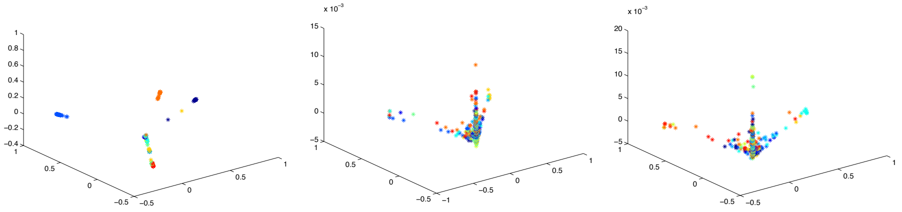
Figure 8 shows the distribution of documents using different topic representations. As in the previous experiment, documents from different classes are well separated in the shared topic representation and are more mixed in the private topic representations. Table 2 summarizes the classification performance.[6] IBTM is able to outperform other methods with a performance of 89.38% even when only images are available for testing. When both modalities are available, the performance goes up to 95%, while ideal classification by humans for this dataset is reported to be 90% in [24].

4.2.2 Leeds Butterfly Dataset. In this section, the Leeds butterfly dataset [26] is used to evaluate the IBTM model on a fine-grained classification task. This dataset contains 10 classes of butterfly images collected from Google image search, both the original images with cluttered background and segmentation masks for the butterflies are provided in the dataset. For each class, 55 to 100 images have been collected and there are 832 images in total. In this experiment, 30 images are randomly selected from each class for training and the remaining 532 images are used for testing. Similarly to above, we perform the experiment in two different scenarios: Image and Image, where only the natural images with cluttered backgrounds are available; and Image and Segmentation, where one modality is the natural image and the other modality is the segmented image.

4.2.2.1 Image and Image. In this experiment, we use only the natural images to evaluate the model performance in the uni-modal scenario. The experimental setting is similar to Section 4.2.1.1, where two images from the same class are paired randomly. $K = 15$, $T = 3$ and $S = 3$ are used for the results presented here. The histograms in Figure 10 are to the previous dataset, however, with smaller values. As natural images of butterflies have more background information that is not related to the class of the butterfly, while for the LabelMe dataset, almost the whole image has information contributing to the natural scene class.

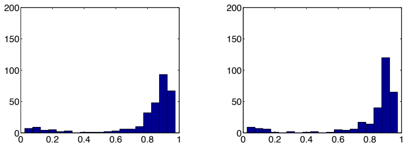
Figure 9 visualizes the image distribution in the different topic representations, where the shared topic representation separates images from different classes better

[6] The 0.65% difference of Full SVM performance in Table 1 and Table 2 were due to different random data partitions.



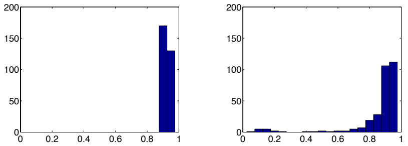
(a) θ, Shared (b) κ, Private (c) ν, Private

Fig. 9. Visualization of the shared topic representation (θ) and private topic representations (κ and ν) for experiments on the Leeds Butterfly dataset using randomly paired images from the same class. The documents of different classes are colored differently and the plots show the first three principal components after applying PCA on the per document topic distributions for all the training data.



(a) Hist of ρ (Img) (b) Hist of μ (Img)

Fig. 10. The histogram over partition parameters of the Leeds Butterfly image-image experiment. Img indicates that this modality uses natural images.



(a) Hist of ρ (Seg) (b) Hist of μ (Img)

Fig. 11. The histogram over partition parameters of the Leeds Butterfly image-segmentation experiment. Img indicates that this modality uses natural images. Seg indicates that this modality uses segmented images

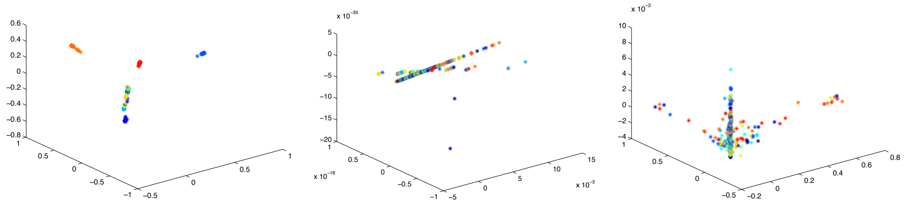
than the private ones. Table 3 summarizes the classification performance for this dataset. There ”II IBTM 15” shows the result of IBTM using only natural images, which obtains the highest performance 95.86% in this uni-modality setting with only 15 topics.

NLD[26] [7]	Full SVM	PCA 15	II SWB15 [22]	IS SWB15 [22]	LDA15	II IBTM15	IS IBTM 1V	IS IBTM 2V
56.3%	95.49%	88.35%	80.26%	94.55%	91.92%	**95.86%**	**96.05%**	**99.06%**

Table 3. The performance comparison with the Leeds Butterfly dataset. ”II” shows the prediction performance for the paired image setting (Image-Image) for IBTM and only images for SWB. ”IS” shows the prediction performance for the image and its segmentation image setting. In this setting, ”1V” means that only images are available (1 view testing) and ”2V” means that both images and segmentations are available (2 view testing). ”IS SWB” shows the performance of using SWB with concatenated words from images and segmentations.

4.2.2.2 Image and Segmented Image. In this experimental setting, natural images and segmented images are used as two different views for training to demonstrate the multi-modality scenario. The segmented images are used as the first view and the natural images are used as the second view. Since the model is symmetric, the order of the views

[7] Learning Models for Object Recognition from Natural Language Descriptions (NLD) trained a classification model based on text descriptors. All images are tested to use visual information to extract attributes to fit the text template for testing. The experiment setting is different from our experiments. However, we include the result from the original paper for completeness.



(a) θ, Shared (b) κ, Private (c) ν, Private

Fig. 12. Visualization of the shared topic representation (θ) and private topic representations (κ and ν) for experiments on the Leeds Butterfly dataset using images paired with their segmentation masks. The documents of different classes are colored differently and the plots show the first three principal components after applying PCA on the per document topic distributions for all the training data.

has no impact on the model. Figure 11 shows the histogram of the partition parameter. It is apparent that the partition parameters of the segmented images are more concentrated around the large values. Thus, the model has learned that the segmented images contain more relevant information. This is consistent with human intuition. Figure 12 shows the topic distribution using shared and private latent representations where the shared topic representations for different classes are naturally separated. Classification performance is summarized in Table 3. SWB performs better with this dataset than with the LabelMe dataset. The reason for this is probably that the visual words here are less noisy than in LabelMe. ”IS IBTM15” denotes the performance of testing with only natural images and ”IS IBTM15” shows the performance of testing with both natural images and their segmentation. We can see that IBTM performs better than other methods even if only natural images are available for testing. With the segmentation, the performance is almost ideal.

5 Conclusion

In this paper, we proposed a different variant of the topic model IBTM with a factored latent representation. It is able to model shared information and private information using different views which has been proven to be beneficial for different computer vision tasks. Experimental results show that IBTM can effectively encode the task-relevant information. Using this representation, the state-of-the-art results are achieved in different experimental scenarios.

In this paper, the focus lay on exploring the concept of factorized representations and the experiments were centered around two view scenarios. In future work, we plan to evaluate the performance of IBTM by using any number of views and in different scenarios such as cue-integration. In the end, efficient inference algorithms are the key for probabilistic graphic models in general. In this paper, we used variational inference in a batch manner. In the future, more efficient and robust inference algorithms [29, 30] can be explored.

References

1. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. PAMI **35**(8) (2013) 1798–1828
2. Tipping, M.E., Bishop, C.M.: Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society **61**(3) (1999) 611–622
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. JMLR **3** (2003) 993–1022
4. Lawrence, N.D.: Gaussian Process Latent Variable Models for visualisation of high dimensional data. In: NIPS. (2004) 329–336
5. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR. Volume 2., IEEE (2005) 524–531
6. Hospedales, T.M., Gong, S.G., Xiang, T.: Learning tags from unsegemented videos of multiple human actions. In: ICDM. (2011)
7. Zhang, C., Song, D., Kjellstrom, H.: Contextual Modeling with Labeled Multi-LDA. In: IROS, Tokyo, Japan (2013)
8. Tucker, L.R.: An Inter-Battery Method of Factory Analysis. Psychometrika **23** (June 1958)
9. Damianou, A., Ek, H.C., Titsias, M., Lawrence, N.D.: Manifold Relevance Determination. In: ICML. (2012) 145–152
10. Zhang, C., Ek, C.H., Damianou, A., Kjellstrom, H.: Factorized Topic Models. In: ICLR. (2013)
11. Virtanen, S., Jia, Y., Klami, A., Darrell, T.: Factorized multi-modal topic model. arXiv preprint arXiv:1210.4920 (2012)
12. Zhang, C., Kjellstrom, H.: How to Supervise Topic Models. In: ECCV workshop on Graphical Models in Computer Vision. (2014)
13. Blei, D.M., McAuliffe, J.D.: Supervised Topic Models. arXiv preprint arXiv:1003.0783 (2010)
14. Zhang, C., Ek, C.H., Gratal, X., Pokorny, F.T., Kjellström, H.: Supervised hierarchical Dirichlet processes with variational inference. In: ICCV workshop on Inference for probabilistic graphical models. (2013)
15. Lacoste-Julien, S., Sha, F., Jordan, M.I.: DiscLDA: Discriminative learning for dimensionality reduction and classification. In: NIPS. (2008)
16. Zhu, J., Ahmed, A., Xing, E.P.: MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification. In: ICML. (2009)
17. Zhu, J., Chen, N., Perkins, H., Zhang, B.: Gibbs max-margin supervised topic models with fast sampling algorithms. In: ICML. (2013)
18. Wang, C., Blei, D., Fei-Fei, L.: Simultaneous image classification and annotation. In: CVPR, IEEE (2009) 1903–1910
19. Wang, Y., Mori, G.: Max-margin Latent Dirichlet Allocation for Image Classification and Annotations. In: BMVC. (2011) 7
20. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: International Conference on Research and Development in Information Retrieval, ACM (2003) 127–134
21. Ek, C.H., Rihan, J., Torr, P., Rogez, G., Lawrence, N.: Ambiguity modeling in latent spaces. In: Machine learning for multimodal interaction. Springer (2008) 62–73
22. Chemudugunta, C., Smyth, P., Steyvers, M.: Modeling general and specific aspects of documents with a probabilistic topic model. In: NIPS. Volume 19. (2006) 241–248
23. Blei, D., Lafferty, J.: Correlated topic models. In: NIPS. Volume 18. (2006) 147
24. Li, L.J., Su, H., Lim, Y., Fei-Fei, L.: Objects as attributes for scene classification. In: Trends and Topics in Computer Vision. Springer (2012) 57–69
25. Zheng, Y., Zhang, Y.J., Larochelle, H.: Topic Modeling of Multimodal Data: An Autoregressive Approach. In: CVPR. (June 2014) 1370–1377

26. Wang, J., Markert, K., Everingham, M.: Learning models for object recognition from natural language descriptions. In: BMVC. (2009)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
28. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: ECCV, Springer (2014) 392–407
29. Minka, T.P.: Divergence measures and message passing. In: Microsoft Research Technical Report. (2005)
30. Hoffman, M.D., Blei, D.M.: Structured stochastic variational inference. In: AISTATS. (2015)