OBJECT CATEGORY RECOGNITION THROUGH VISUAL-SEMANTIC CONTEXT NETWORKS

BY ISHANI CHAKRABORTY

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computer Science

Written under the direction of

Ahmed Elgammal

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2014

ABSTRACT OF THE DISSERTATION

OBJECT CATEGORY RECOGNITION THROUGH VISUAL-SEMANTIC CONTEXT NETWORKS

by ISHANI CHAKRABORTY

Dissertation Director: Ahmed Elgammal

Understanding and interacting with one's environment requires parsing the image of the environment and recognizing a wide range of objects within it. Despite wide variations, such as viewpoint, occlusion and background clutter, humans can achieve this task effortlessly and almost instantaneously. In this thesis, we explore computational algorithms that teach computers to recognize objects in natural scenes. Inspired by the findings in human cognition, our algorithms are based on the notion that visual inference involves not only recognizing individual objects in isolation but also exploiting rich *visual and semantic associations* between object categories that form complex scenes. We view artificial object recognition as a fusion of information from two interconnected representations. The first is the *inter-image* representation in which an image location is *visually associated* with previously learnt object categories based on appearance models to find the most likely interpretations. The second is the *intra-image* representation in which the objects in an image are *semantically associated* with each other to find the most meaningful spatial and structural arrangements. The two representations are

interconnected in that the visual process proposes object candidates to the semantic process, while the semantic process verifies and corrects the visual processs hypotheses.

The primary goal of this thesis is to develop computational models for visual recognition that characterize the visual and semantic associations and their inter-dependencies to resolve object identities. In order to do so, we model object associations in contextual spaces. Unlike traditional approaches for object recognition that use context as a postprocessing filter to discard inconsistent object labels, we stratify scene generation into a Bayesian hierarchy and simultaneously learn semantic and visual context models for objects in scenes. The semantic-visual contexts among objects are represented through latent variables in this hierarchy. The intra-image associations within a scene are modeled as semantic context while the inter-image relations due to appearance similarities between object categories are modeled as visual context. To combine the complementary information derived from the two spaces, object labels are inferred by *context switching*; labels activated by appearance matches constrain semantic search while semantic coherence, in turn, constrains object identities. We demonstrate how this novel context network for modeling associations between objects leads to highly accurate object detection and scene understanding in natural images, especially when training data is impoverished and negative exemplars are not easily available.

Dedication

For my parents, Sheela and Gangadhar

Table of Contents

List of Figures

Chapter 1

Introduction

"An object is not so attached to its name that one cannot find for it another one which is more suitable." - René Magritte

Understanding a scene requires parsing an image of the world and recognizing a wide range of objects within it, along with their properties. Humans perform this task seamlessly and effortlessly, which belies the computational magnitude required to perform this task. We can detect and categorize objects from among tens of thousands of possibilities within a fraction of a second [22]. How are we able to achieve this remarkable feat? Although we are far from reaching a conclusive answer, researchers in the fields of neuroscience and psychophysics have made great strides in this direction. It is now known that more human brain is devoted to vision than to any other cognitive task.

The field of computer vision was conceived about fifty years ago with the ambition of constructing artificial systems that mimic human vision. Vision was studied not only to extract various aspects about the world from images but also as "an inquiry into the nature of internal representations" by which this information is captured [71]. Since much of human visual perception remained unclear, computer vision has since then independently developed a strong theoretical and algorithmic basis by fusing concepts from mathematics, signal processing, statistics and insights into human cognition. The convergence of statistical methods with biologically plausible formulations has led to increasingly more accurate and detailed models for vision in recent years.

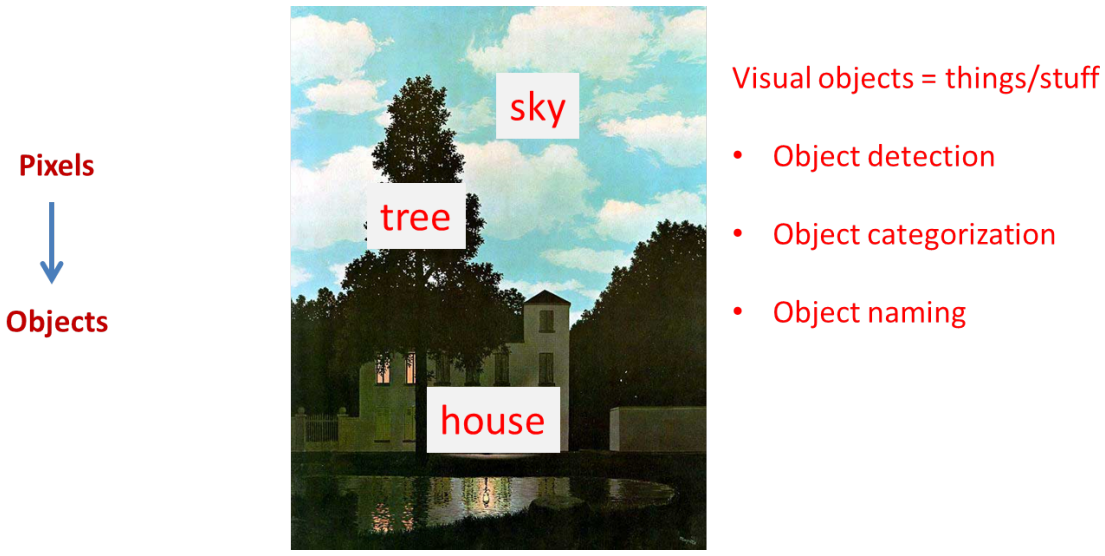


Figure 1.1: Object Recognition is a classical problem in the field of Computer Vision that aims to find familiar objects in an image or video. An object recognition algorithm is expected to process pixels in images to parse objects out of them. This is one of the fundamental problems of Computer Vision and is variously referred to as object detection, recognition, localization and naming. This is the core problem that we investigate in this thesis.

In this thesis, we are interested in teaching computers to recognize objects in natural scenes. This problem supercedes several related tasks variously known as visual object detection, localization, categorization and naming and is essentially the ability to automatically assign labels to familiar objects (objects that have been "seen" previously by the computer). The semantic aspect of this problem is to pick the correct noun from the vocabulary of English language. Doing this requires understanding the contents of the scene itself. For example, a chair-like object should be named armchair if it is seen in the sitting room but would be named lounge chair if it appears outdoors. The visual counterpart to this problem, i.e., associating an image region with an object label is implicitly harder. This is because "object" in itself is an impoverished concept - it presupposes a well-defined spatial presence and extent in images. Even

Figure 1.2: Some reasons why visual object recognition is a challenging problem. (a) Objects are often defined by their functions; buildings can have very different appearances. (b) Objects, such as cars vary tremendously across viewpoints. (c) Some objects can be learnt from many easily available examples, e.g., coffee mugs; other objects are harder to obtain,e.g., anteaters.

for well-defined objects, the crux of the problem is to learn to identify them over a large range of viewing conditions. See Figure 1.2. Often, objects are defined by their functions, not their appearance (intra-class variability). Objects can occur at any location (position variability), any size (scale variability), at any three dimensional configuration (pose variability), at any lighting condition, or in different environments (background clutter). Another factor that influences recognition is the familiarity with the object itself, i.e., how many instances of the object are available for training (frequency variability). For example, images of different types of coffee mugs are easily available, but it is much harder to collect images of different types of ant-eaters. An ideal object detector would be able to identify thousands of such diverse object categories despite all types of variations.

The traditional approach to object detection is to construct image-based appearance classifiers trained on a large database of examples (positive objects and negative, other objects or background scene). Given a new image, every classifier is applied to the image locally within sliding windows. Object is detected when a classifier "fires" at a location. The limitation of such independently trained classifiers is that they are rarely able to model all the sources of variabilities arising in real world images. Moreover, they do not exploit the rich inter-relationships between objects in scenes. For example, Figure 1.3 (left) shows the output from one of the leading object detectors [27] for a complex scene. The objects labels are shown at the detected

Figure 1.3: Object labels identified in the image. Left: Context-agnostic object detection, Middle: Context as post-processing filter, Right: Our proposed joint context model

window locations. While some of the objects are correctly detected (e.g., bed, curtain, flowers, window, etc.), several other detections are not just incorrect (e.g., countertop, showcase etc.) but also out-of-place in the image. This is a major limitation of sliding-window based object detectors, that we seek to address in our thesis.

Now, consider the object labels in the center frame of Figure 1.3. At the start of the detection process, the image regions are visually ambiguous and can be perceived as many different objects. However, when some key objects e.g., bed, curtains etc. are detected with high confidence, the computer should use this information to rapidly modify the confidence levels of other objects. This "completion" of information is caused by the object associations within the bedroom scene that guide understanding of the complete image and the objects contained in it. Consequently, all the image regions are better recognized due to the contextual influence of global features, scene layout and the associated objects. This influence of the surrounding environment that is encoded in a vision algorithm for object determination is defined as *context*. Using contextual knowledge about objects has been shown to improve upon context-agnostic object detection. However, the role of contextual cues in most state-of-the-art methods is rather limited. It is usually applied as a post-processing step to filter inconsistent object labels from scenes. Figure 1.3 (middle) shows the output labels obtained from a leading context based

object detector [18]. This is the result after context is used as a second step to arbitrate object labels on the detections obtained by the context-agnostic object detector, shown in the left image. Observe that although all the out-of-place detections are removed, the overall output is very sparse. Only generic, high level objects (e.g., bed and floor) are retained. Moreover, contextually correct detections (e.g., flowers) are eliminated from the final output. In general, context models that are applied as a secondary step to improve label agreement are often reductionist by nature and do not provide rich explanation of the scenes. This is because there is essentially a unidirectional flow of information from objects to context. Once the context of the scene is established, only the contextually correct detections are retained, but there is no mechanism to retrieve or relabel the incorrect and missing detections.

In complex scenes, the combined effect of appearance variations and naming conventions leads to ambiguity in scene interpretation. We attempt to mitigate the effects of these variations by identifying and modeling the sources of ambiguity in images. We hypothesize that ambiguity can be captured primarily along two contexts: semantic context and visual context. Semantic context captures the prior knowledge about what objects to expect in a scene [34]. This is based on learning from real world images in which scenes are usually composed of specific objects. Based on the correlation of objects in the semantic space, one can find the likelihood of an object to be found in some scenes but not others [34]. In contrast, visual context captures similarities in visual appearance of objects. Locally, appearance features of an image region might not inform about the exact identity of the object in that region (i.e., the answer to the question "What is this?" [29]). The low-level features extracted from the image region can be attributed to multiple object representations simultaneously. So instead, it provides a set of initial guesses of what the object might be (the answer to "What is this like?"). In this sense, by mapping an image region onto the visual space, we find the objects that are more likely to be found at that specific image region but not others.

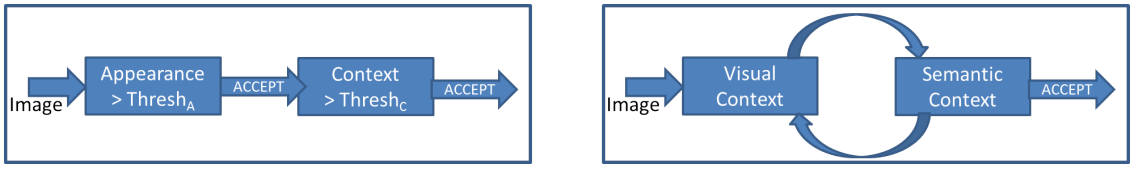


Figure 1.4: The role of contextual cues in most state-of-the-art methods is rather limited. It is usually applied as a post-processing step to filter inconsistent object labels from scenes. We propose joint context model in which objects are inferred through interactive context switching.

We represent the ambiguity of naming objects in semantic-visual space in terms of the linguistic notions of polysemy and synonymy (Figure 1.5). In the visual space, we define visual polysemy as the ambiguity that arises when an object displays diverse appearances, e.g., the term "building" in Figure 1.2. This makes it difficult to develop a single appearance classifier for the object, since that classifier needs to capture different modes of appearance features. In contrast, if different names are associated with similar looking objects, that is referred to as visual synonymy. For example, a "car", "van" and "truck" mostly have similar appearances. A generic car classifier can possibly fire at image locations containing vans and trucks. Similar ambiguities are also observed in the semantic space of object names. Semantic polysemy refers to the situation when same objects appear in different scenes. E.g., a "cup" might appear in a kitchen or on an office table and essentially cohabit with different sets of objects. A classifier that predicts objects based on expected object co-occurrences would be confounded if it uses the cup as a cue for grounding the semantic context of the scene. Semantic synonymy, on the other hand, arises when different combination of objects appear in similar scenes. This affects the performance of classifiers that use rigid contextual rules between objects to disambiguate label agreement.

We model ambiguity of object names based on the multiple hypotheses that arise from the semantic-visual spaces and infer the *most probable* and the *best conforming* hypothesis. To

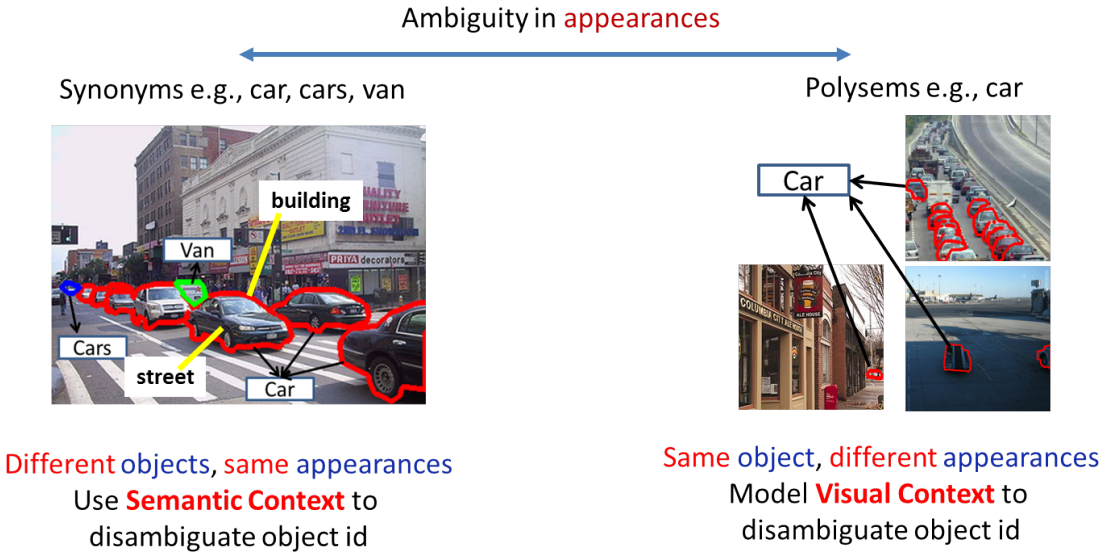


Figure 1.5: We represent the ambiguity of naming objects in semantic-visual space in terms of the linguistic notions of polysemy and synonymy. In this thesis we mitigate the effect of ambiguity through contextual knowledge in semantic and visual spaces.

do this, we hypothesize an object to be a node in a large network in which it is associated to other objects by two types of associations. Intra-image associations capture spatial correlations within the same image, i.e., the object label at an image location is correlated semantically with the object labels at all the other locations within the image. This is modeled by learning the semantic context between objects. The second type are the inter-image associations, in which an object label is associated with labels in other images based on the visual appearance similarities between their image locations. This is modeled by learning the visual context of label appearance. Due to this network of associations, object labels are implicitly related to one another in semantic and visual context spaces through many-to-many relations. Some of these associations are strong and repeatable, while others are weak and rare. We hypothesize that by modeling the statistics of their inter-relationships we can discover strongly associated groups

of objects in a context network.

Based on this hypothesis, we construct Bayesian models in which object associations are represented as probability distributions in a latent, low dimensional, joint context space. Specifically, our model is based on the Latent Dirichlet Allocation that has been successfully implemented in Natural Language Processing for resolving ambiguity in text. In NLP, LDA is applied to collections of documents, in which every document is represented by a bag-of-words vector. The model projects each document into a low dimensional space represented by a set of bases that capture the grouping tendencies of words, also known as topics of the collection [117]. For example, if the documents are scientific abstracts, the model is able to group words under common topics such as biology or astronomy [83]. In statistical terms, an LDA models each document as an admixture of latent topics, in which each document is a mixture distribution of topics and each topic is a mixture distribution of words.

LDA provides a formal approach for expressing many-to-many relations between documents and words through topics. By translating the definitions of documents, words and topics in the image domain, one can construct similar relationships between image, local image regions and object labels within an LDA framework. The relational graph over object labels, as explained above, is analogous to the word-document relation if we assume object labels to be words and inter-image or intra-image associations to be documents. The intra-image associations constitute semantic documents while the inter-image associations constitute visual documents. By learning the models in both spaces, one can discover latent topics of label groupings.

1.1 Thesis Contributions

In this thesis, I propose Bayesian models that stratify scene generation into a hierarchy of semantic and visual context models of objects. The semantic-visual contexts are represented

through latent object variables in a Bayesian hierarchy. The intra-image associations within a scene are modeled as semantic context while the inter-image relations due to appearance similarities between object categories are modeled as visual context. I propose two Bayesian models that combine these two types of contexts in a causal hierarchy. Finally, I demonstrate how this novel generative framework for modeling associations between objects leads to highly accurate object detection and scene understanding in natural images, especially when training data is impoverished and negative exemplars are not easily available.

The proposed models are based on the Bayesian framework of Latent Dirichlet Allocation (LDA). In the following chapter, I provide the technical background of LDA and describe the details about the terminologies, image representation, model definitions, and parameter estimation and inference. This provides the foundation for the technical description of the two computational models explained thereafter.

In the third chapter, I present ViewLDA, a Bayesian model that learns view-specific object categories with minimal supervision and without negative exemplars. It is a completely unsupervised approach for jointly solving object localization and viewpoint estimation in a unified model. Our approach is to model the unified visual and semantic context of objects and background in images. First, a generic object detector is learnt by modeling the visual context of objects across various viewpoints. Then, a view-specific object detector is learnt by modeling the spatial context of objects and corresponding backgrounds in images. In a new image, this unified model is applied to localize objects by probabilistically classifying regions into object and background. We show results on a challenging dataset of multiview car images where our results surpass the state of the art classification when the domain of training and testing images are varied.

In the fourth chapter, I present Visual Semantic Integration model (VSIM) that connects

the semantic and visual contexts through their shared object names. VSIM models scene interpretation as a top-down approach where semantically contextual labels are first created to represent a coherent scene composition. These labels are then re-interpreted with their visually contextual counterparts in the appearance space. Specifically, VSIM is a probabilistic, hierarchical model of latent context and observed features. In the first level, the image is modeled as a distribution over latent semantic contexts which determines the semantic labels that compose the scene. In the next level, each semantic labels visual context determines the appearance features that are finally the observed variables in the model. Inference in VSIM is initiated in a bottom-up manner, where observed image regions are the only cues used to infer the semantic and visual object labels in the image. I.e., the goal of VSIM is to infer the semantic object labels in an image, given its appearance features. We derive an iterative Data Augmentation algorithm that pools the label probabilities and maximizes the joint label posterior of an image. Our model surpasses the performance of state-of-art methods in several visual tasks on the challenging SUN09 dataset. Finally, I conclude my thesis by summarizing the main contributions and suggesting some possible directions for future work.

Chapter 2

Literature Review

Visual object recognition has a long and multi-branched history. In this section, we first provide an overview of the basic frameworks and then follow with a more focused view on a particular branch that uses contextual information of objects for visual recognition. We discuss different types of semantic contextual representations and learning techniques that incorporate semantic knowledge about the world through spatial, scale and multimodal relationships. Next, we looked at the role played by context in Natural Language Processing algorithms. In contrast to the "prior" or "arbitrator" role in vision, context in NLP is actively modeled within the inference machine jointly with other appearance features. We review the work on neural network based connectionist networks that model multiple hypotheses in which context based detections are continuously updated during inference. We compare and contrast these methods to the hierarchical networks proposed in Computer Vision. Finally, we explore latent semantic model based networks used for representation and learning of contextual information, both in vision and language problems.

2.1 Evolution of visual object detector

In the English dictionary, there are 10,000 to 30,000 words that qualify as names of objects [6]. An image usually contains multiple and diverse variety of these object categories. The role of an object recognition is to detect and localize all the objects within the image. Keeping the

enormity of the problem in view, researchers have adopted a "divide and conquer approach" and stratified the problem into several sub-tasks that ease application development and analysis [19]. Since visual object detector has a long and multi-branched history of innovation, we concentrate only on some significant developments from the last decade.

One interesting aspect about the development of new algorithms is their implicit relationship to the scope of the problem that manifests in the form of the datasets available at that time. Appropriate datasets are essential for various stages of object recognition research, e.g., for learning visual models of objects, detecting and localizing instances of these objects in images, and evaluating the performance of algorithms [84]. It is not surprising then that research is, to a great extent, driven by competition to outperform the results on the most current data at hand. When a substantial number of research groups have tested and understood the behavior of algorithms on a particular data, that problem is assumed to be closed and a newer, harder challenge is developed for contention. Hence, it is instructive to observe jointly, the evolution of vision algorithms and dataset challenges to get an overview of the problem.

- **MPEG7 and shape alignment**: The MPEG7 shape dataset was released around 2000 [56]. It consists of simple binary silhouettes of objects such as apple, bat, bird etc. The main mode of variation in these images is the deformation of the shape along the contours. So, the goal for the researchers trying to optimize performance on this dataset was to develop shape descriptors for matching non-rigid closed contours. The main contributions in this direction were the shape context algorithm from Belongie et al. [79] who developed an edge samples based descriptor, optimal correspondence of visual parts by Latecki et al. [56] and curve alignment algorithm from Sebastian et al. [93], which were all based on path-following along contours. These methods together produced competitive performance of 80% accuracy on MPEG data. However, the scope of the problem was limited since the objects were on clean background (no clutter) and contained no texture.

- **Caltech101 and parts based object detection**: Beiderman proposed the Recognition-by-components theory in 1987 to explain object recognition [6]. According to the RBC theory, humans are able to recognize objects by separating them into geons. Geons are a small set of simple, canonical shapes found in nature and are the main component parts for all objects. When real world images of 101 object categories were released by Caltech in 2003, the researchers put this theory into practice by developing parts-based object detectors. It was a natural progression from whole body detection because shape variations in many natural objects can be viewed as deformations along individual parts, where each individual part is, in itself, more or less rigid. For example, if we consider a human body to be an object, its deformations are restricted only along the bone joints. The Caltech 101 objects are of various colors and textures, as well as with a bit of background clutter, which makes them more challenging than MPEG shapes. To reliably detect and classify the objects, the idea of deformable parts models was devised, in that the confidence of individual parts was accumulated based on geometric constraints between the parts to identify the entire object. The forerunner to this framework was the Constellation model proposed by Burl et al [12], later refined by Fergus et al. which proposes a Bayesian generative model for characterizing individual parts and their in-between relations [31]. A probabilistic representation is used for the object's shape, appearance, occlusion and scale. The typical objects in Caltech 101 were airplane, motorbike, and faces. The model parameters are learnt using Maximum Likelihood. Recognition is performed on a query image by first detecting local parts and their scales, and then evaluating the parts in a Bayesian manner. Another significant contribution in similar direction was Pictorial Structures for Object recognition proposed by Felzenszwalb et al [28]. The basic idea is to represent an object by a collection of parts arranged in a deformable configuration. The appearance of each part is modeled using local features, and the pairwise deformable configuration is represented by spring-like connections between parts. This was also a generative, model-based approach and aimed at recognizing faces and human bodies in images. These models are conceptually

appealing but have not shown a significant improvement in practice. As noted in [27], on difficult datasets deformable models are often outperformed by "conceptually weaker" models such as simple bag-of-words. This performance gap of part based systems was overcome by Felzenszwalb et al.'s seminal work, Discriminatively Trained, Multiscale, Deformable Part Models [27]. Their model includes both coarse global template covering an entire object and part templates at finer resolution. The part templates and the full model parameters are learned discriminatively. This method by far gives the best performance on most of the state-of-the-art object datasets, e.g., Caltech 101 and VOC Pascal.

- **VOC Pascal, SUN09 datasets and context based object detection**: The VOC Pascal Challenge [25] was introduced to meet the need for data with more realistic and less restrictive image conditions, multiple object class instances within a single image, to include partial occlusion with size and orientation variations, etc [84]. It consists of 20 object classes, each with significant intra-class variability (that Caltech 101 lacked). More recently, SUN09 dataset by MITs Torralba et al. have introduced over 500 object categories with high inter as well as intra class variations [18]. It consists of objects in various spaces (indoors and outdoors), of diverse classes (things like car and stuff like sky), from different viewpoints and of different frequencies (3000 sky versus 30 microwave instances). The most significant aspect of these datasets is the presence of several different objects within a single image. Traditional methods require applying 500 object detectors individually across the entire image, which make the process very costly. Moreover, the false detections from each object detector reduce precision as the number of detectors increase, making image understanding messy. Hence, these datasets instigated a new paradigm to object recognition that can not only discriminate between different locations within the same image but can also *exploit* their simultaneous appearance within the same frame of reference. This is based on the idea that in the real world, objects rarely occur in isolation; they co-occur with other objects and within particular environments. This assumption of coherent composition of objects in the real world is defined as context. Several different

object recognition models have used statistical methods that can exploit semantic context in real world scenes. Below, we explore different types of contexts used in vision.

2.2 Context in Object Recognition

Context has been studied in psychophysics and more recently, in computer vision as a facilitator of visual recognition. In psychophysics, one of the pioneering contributions was from the Gestalt school which proposed the Principles of Perceptual Organization [113]. Gestalt psychologists studying object perception suggested that "people acquire meaning from the totality of a group of stimuli rather than from any one individual stimulus". Max Wertheimer, a Gestaltist, introduced principles of grouping, which were empirical measures of associations between elementary parts. The main principles were similarity (physically similar things belong together), proximity (things in close proximity belong together), continuity (things form close, continuous forms), closure (things are perceived as being complete), and simplicity (preference to simple forms). The treatise also included principles of context which considered how perception of an object is influenced by its surroundings. Two important principles of context are figure and ground (people interpret a stimulus in the context of its background) and contrast (people notice stimulus that stands out from its surroundings).

Contextual knowledge in artificial vision algorithms is typically modeled as a prior that influences the likelihood of an object to be found in a scene. One of the first algorithms that proposed context based recognition was the CONDOR system by Strat and Fischler, who estimated the scale of objects based on camera information to generate scene hypothesis [99]. Since then, researchers have increasingly adopted contextual priors into their computational frameworks. Contextual mid-level cues have shown to not only improve the quality of object estimates but also reduce the amount of information processing needed for low-level features [88].

There are diverse representation schemes and algorithms to integrate context into recognition algorithms. Several approaches have been proposed that employ context at different levels of compositionality, e.g., object-based context [18] or scene-based context [104]. Context can be integrated at parts level to detect an object [88], or at object level to detect a scene [54]. A detailed survey can be found in the article by Galleguillos et al. [35]. We explore some typical probabilistic systems that use the context as a prior at scene level to inter-connect objects. These works can be classified into the following categories.

- **Spatial context**: The most commonly used spatial contextual constraint is the assumption of Markovian smoothness, i.e., neighboring regions share the same color, texture, and/or object labels e.g., in this work by He et al. [41]. Object co-occurrence in scenes was used by Rabinovich et al. by querying the Google Sets web application which generates a list of possibly related items, e.g., body, face, eyes, lips [86]. This information is captured by a binary adjacency matrix. Spatial geometric relations between objects is captured in several works, both at the pixel level, e.g., in Textonboost method proposed by Shotton et al. [97] and at image regions e.g., by Kumar et al. [54] and Choi et al. [18]. Both these methods employed a hierarchical model to capture context at different levels in the scene.
- **Scale context**: Prior information about the relative sizes of objects that appear together in a scene reduces the need for multiscale search and thus computation can be focused towards more likely scales. For example, pedestrian detection in street images is difficult problem due to the variability of human form and pose. However, the size of detected vehicles have been used as a prior for estimating pedestrian sizes, which also improves overall detection, as shown in Hoiem et al. [47]. Scale context in 2D images assume objects to be at relatively similar depth, which is a rather limiting assumption. However, if camera meta-data is available or can be easily computed then scale context becomes an useful prior for 3D analysis of scenes. For example, a combination of 3D analysis and scale context improves people layout estimation in [15]. Since, understanding the 3D layout of scenes is usually a difficult task, another method

of computing the scale information is by capturing the "gist" of the scene, which is a low-dimensional representation of the complete image [104]. Because it captures the textures in the image, it is useful for predicting what types of objects may appear in the image, and at which location/scale.

- **Multimodal context**: Information from image blobs and labels are combined in the work by Carbonetto et al. [14], where they learnt models of objects and their spatial relationships, given captioned images. More recently, using captioned images allowed learning of richer relationships between object names in terms of prepositions and comparative adjectives in this work by Gupta et al. [39]. With the advent of big-data, language+vision integration has become a rapidly developing field, and several research efforts are now exploring the possibilities of combining text and visual features for various multimodal tasks. Combined text and vision was shown to perform effective image annotation in Jamieson et al. [51]. Blei et al. [9] proposed correspondence latent Dirichlet allocation (corrLDA) to model the joint distribution of images and text, as well as the conditional distribution of text given the image. Quattoni et al. [85] have devised a semi-supervised learning algorithm that exploits text captions to constrain the visual representation to one that predicts the presence or absence of individual words. An image clustering algorithm was proposed by Bekkerman et al. based on combinatorial Markov random fields [5]. The related semantics of words and images have also been considered. For example, in Leong et al. [57], semantic relatedness between words and images is used to better exploit multimodal content. Supervised word sense disambiguation improved when pictures were also used in Barnard et al. [4]. In contrast, Saenko et al. [91] combined visual and textual information to solve the reverse problem of image sense disambiguation.

2.3 Context in NLP: Word Sense Disambiguation

Context is used in Natural language processing to resolve Word Sense Disambiguation (WSD), which is one of the fundamental problems in Artificial Intelligence. WSD is a process that governs identification of the correct sense (or meaning) of a word in a sentence. This problem arises because in most languages there are words that have multiple meanings (*polysems*). For example, bank could refer to the financial institution or to a bank of a river. Additionally, different words can have the similar meanings (*synonyms*). For example, bank and shore both refer to the land bordering a body of water. This relatedness of words and meanings makes it difficult to understand the sense in which a word is used in a sentence.

The sense of the word being used is determined by its semantic context. Statistical approaches for understanding word meanings are required to model a representation for semantic context of words. Also, in complex discourses, the challenge is to understand how and when contexts come into the comprehension process, and which context gets activated during a particular discourse [24]. This is known as the *lexical access* problem and it looks at how contexts are used for word sense disambiguation. One hypothetical system of lexical access in NLP is the *connectionist network*, in which words are connected by soft constraints in a network [101, 112]. Learning involves a multiple hypothesis approach where the strength of the connections change continuously, thus changing the contextual influence they have upon each other. We use the example provided in [23] to illustrate this approach. Consider the sentence: "The pot, which Mary bought from John, made her cough". Initially, the cooking-pot context will have the most activation, but "cough" will cause the marijuana context to become more active. If the subsequent sentence is "Mary was allergic to the flower in it" then the mention of being allergic will send more activation to contexts representing a different reason for coughing and, along with "flower", cause the flower-pot interpretation now to be most preferred. Thus, activation spread is implemented via a parallel, interconnected and efficient mechanism.

Connectionist networks have been typically implemented as neural networks, where an activation function determines the activation of a context. Words contribute to a activation using weights; positive weights contribute activation and negative weights lead to a reduction in the net input. Activation spreads in parallel from individual words. The most activated context represents the current interpretation. When new inputs are received, the most active contextual nodes may drop in activation, leading to a reinterpretation of prior inputs [23]. Connectionist networks have been combined with machine learning algorithms to solve multitude of different discriminatory tasks in [19].

2.4 Context in Vision: Image Sense Disambiguation

An image can be compared to a complex sentence in which connections between image regions, object likelihoods and semantic relations between objects together lend explanation to the image. In fact, this analogy to the connectionist network has been extensively studied in visual psychophysics by Bar et al. [29], where the authors found evidence of an interactive context network in the brain that facilitates object prediction. They propose that the contextual associations analyzed during the recognition of an object or scene are represented and activated in corresponding *context frames of reference*. Experiments on the type of information represented in context frames suggest that it includes both spatial and visual aspects of contextual relations. The associations mediated by an activated context frame can lead to sensitizing, or priming, specific representations of other contextually related objects [29]. Spreading activation of contextually related objects can thereby help in facilitating subsequent recognition of objects as a function of their contextual associations. *The concurrence of evidence from linguistics and cognitive vision is a significant event and an underlying motivation behind our proposed joint context network paradigm for automated vision*.

In the last few years, disambiguation of visual objects in images has been studied using

multimodal data. Much of this work has focused on grouping tendencies of images and text. Loeff et al. [65] introduced the problem of image sense discrimination (ISD) for images retrieved by an internet search engine for an ambiguous keyword. They used a simple k-means clustering based method on ambiguous images returned by keywords (e.g., crane could mean the bird or the instrument to lift objects). In their study they concluded that senses are "fuzzier" for images than for words, making it more complicated to disambiguate images. Saenko et al. [91] devised a method to filter abstract senses from image search results. They propose an unsupervised method that, given a word, selects non-abstract senses of that word from an online ontology and generates images depicting these disambiguated entities. Their method uses both image features and the text associated with the images to relate latent topics to particular senses. The basic assumption in the sum rule is that the features of one modality are independent of sense given the other modality. Lucchi et al. [67] propose to learn a ranking function that optimizes a ranking cost and simultaneously discovers disambiguated senses of the query that fit the supervised task.

In contrast to clustering, a deeper understanding of relationships between objects can be represented through object hierarchies and networks. The hierarchical context model by Choi et al. [18] is a context tree that represent relationships between object categories. Their tree provides a rich representation of spatial object dependencies between 107 object catgories (SUN09 dataset) and enables efficient inference and learning algorithm. A similar tree structured model for large scale visual learning was proposed by Gao et al. [36], but based on visual relations between objects. In their work, they organize a set of binary classifiers in a DAG structure where the root contains coarse classifiers (e.g., indoors versus outdoors scenes) and the leaves are classifiers that perform fine grain classification e.g., between living room and dining room. Deselaers et al. [21] investigate insightful questions on Imagenet such as "Do semantic categories form clusters in visual space?", "Are there visual prototypes for a semantic category?", "Is visual similarity correlated to semantic similarity?" and "Are semantic categories visually

separable?". They concentrate on 1) How the visual variability within a category changes with depth in the hierarchy. 2) Determination of a visual prototype for every category 3) Measurement of the relation between semantic and visual similarity, and 4) Analysis of how within-class and between-class visual similarities change as a function of how broadly classes are semantically defined. They use these computations to devise a distance between image pairs that estimates whether they are from the same basic-level category (e.g., between car and dog). Although these works have focused on important issues of image and text ambiguities, none of them compare or combine the semantics across the two domains. The Visual Memex model, proposed by Malisiewicz et al. [70] reasons about object identities and their contextual relationships. It assumes that each object instance is a concept in itself. Visual Memex is a vast graph of these concepts, with nodes representing all the object instances in the dataset, and connecting edges representing the different types of associations between them. There are two types of edges in visual memex: 1) visual similarity edges, e.g. this car looks like that car), and 2) contextual associations, e.g. this car is next to this building. This work is closest in spirit to our context network. However, they do not reason about the many-to-many relations between object connections.

2.5 Latent Semantic Models for context networks

Latent semantic variable models (LSM), also known as topic models in text literature is a method for modeling collections of discrete data. The theory of topic modeling evolved in the field of Information Retrieval, where the problem is to identify documents that contain words from the query. Instead of direct word to word matching, topic modeling allows better information representation through the basis of a conceptual topic or meaning of a document. Formally, topic modeling is a probabilistic dimensionality reduction technique. The main idea is to map high-dimensional count vectors, such as vector space representations of text documents, to a

lower dimensional representation in a latent semantic space [45]. The actual technique has evolved from Singular Value Decomposition based Latent Semantic Indexing (LSI) developed by Deerwester et al. [20] to probabilistic LSI (pLSI) [45] that assumes a mixture decomposition of documents derived from a latent class model. Finally, the pLSI has been extended with a fully Bayesian treatment in Latent Dirichlet Allocation [11].

Although LSMs are primarily used for modeling text data, it has been enthusiastically adopted by the Computer Vision community due to the success of bag-of-words model in various vision tasks. The basic assumption underlying the use of LSM is considering an image as a document and its local features as discrete units that are conditionally independent of each other, given a latent topic. Based on this assumption, LSM was initially applied to object categorization in [98] and [100], and for scene classification in [58]. The basic model underlying Latent Dirichlet Allocation, a specific type of LSM, is an unsupervised, generative, parametric model with user-defined number of topics and hyperparameters.

Due to LDA's flexibility in accommodating more information by adding random variables, changing the type of distributions at the nodes, and modifying dependencies, the basic model has been modified to suit the problems in computer vision. For example, in addition to feature (word) correlations captured through topics, topic correlations is captured through a logistic normal prior in Correlated Topic Model [10] and it is shown to improve recognition rates in multi-action sports video sequences in [110]. To allow data-driven estimation of topic count, an LDA is generalized to a non-parametric, hierarchical Dirichlet process [103] that utilizes a Dirichlet process prior to estimate the number of topics during training. This model is used in Optimol [59] to perform incremental, online learning on Internet images. A nested Chinese restaurant process (nCRP) prior accommodates multi-level hierarchy in LDA structure [8], as exemplified in the semantivisual hierarchy of images in [62].

Modeling multimodal data, in form of images and texts has been jointly represented using

CorrLDA and its variants [9, 3, 114] for image annotation and is an active area for further research for diverse textual domains such as advertizing [17] and closed captions in video streams [77].

Different types of regularization have been proposed to constrain topic learning in models. Sparse Topic Model [109] associates each topic with a subset of the vocabulary while a bridge between dictionary learning for sparse coding and topic models is established in [52]. Sparsity is applied to a real world problem in [107], where temporal activity mining is performed by applying a sparsity constraint on the activity topic discovery. Methods for manifold regularization based topic discovery assume that data from texts or images lie on a low-rank, non-linear manifold within the high-dimensional space of the original data [13]. This constraint has been used to model images that share similar visual features or same annotations using similar topics in [96].

To allow supervision, topic models have been extended to include side information by upstream and downstream conditioning. In downstream conditioning, the supervised response variable is generated from topic assignment variables. An example of downstream conditioning is the max-margin latent Dirichlet allocation used for image classification and annotation in [111], which extends MedLDA [117] for multi-label problems. In contrast, if the supervised response variables directly or indirectly generate latent topic variables, the process is known as upstream conditioning. A joint max-margin and max-likelihood learning method for upstream scene understanding models is presented for scene classification in [118] and part based object and scene classification in [82]. For more refined supervision, such as in form of domain knowledge of a specific scenario knowledge base and LDA is combined by defining a Markov Random Field over topic assignments and parameters in Logic-LDA [2].

The flexibility in defining different components of images and videos using random variables and through directed dependencies makes topic modeling techiques amenable to a large

variety of computer vision problems. Some examples are as follows. Common object estimation, resulting in background subtraction has been investigated in [32] and in [40], which includes a predefined shape model as a top-down cue. Identity and locations of visual words in buildings specific to topics is learnt in a geometrically consistent way using geometric LDA [83]. Random Field Topic Model for semantic region analysis in crowded scenes is proposed in [116]. Dependencies between between static and dynamic features are modeled in 3D urban scenes in [37]. Abnormal patterns in crowd behavior is studied in [72]. Differences between genders have been mapped in [63].

Overall, LSM modeling provides a means to describe rich, multimodal information using simple, intuitive models. However, some of its simplistic assumptions can be viewed as drawbacks based on the problem at hand. For example, although bag-of-words model assumption is simple and robust for many applications, the loss of spatial information can be a major limitation for object or scene detection problems. To include spatial information, several papers model location nodes using Gaussian distribution [64, 100]. Another limitation of LSMs is their purely generative representation, which reduces their discriminative power. To allow discriminative information, hybrid models have been proposed [111, 82, 118]. Topics are groups of features that are assumed to be coherent and semantically relevant. However, due to variations in dataset and stochasticity of learning algorithms, the topics discovered might be meaningless and erroneous. Machine learning techniques such as Iterative Topic Modeling [48] aim to correct and update the topics using human in the loop. Finally, the learning process in LSMs is driven by the hyperparameters that are either set apriori or estimated from the data. The effects of either method is still a matter of research study [108].

Chapter 3

Latent Dirichlet Allocation

3.1 Introduction to LDA

The LDA model forms the basis for the latent variable models described in the next two chapters. By giving a general outline of the formulation, modeling and inference technique of LDA here, we can concentrate on the novel aspects of the models and inference algorithms in the subsequent chapters.

3.2 Latent Dirichlet Allocation: Formulation and Inference

LDA is based on probabilistic Latent Semantic Indexing (pLSI) [45] which is an admixture model. It is formulated as a weighted sum of component distributions whose weighting proportions sum to one, as shown below.

$$P(d, w_i) = p(d) \sum_z P(w_i|z)P(z|d), \tag{3.1}$$

where d is a document, w is a word in the document and z is the latent topic. A document may contain multiple topics. $p(z|d)$ serves as the mixture weights of the topics for a particular document d. In pLSI, d is a multinomial random variable whose values are determined by the training documents and the model learns the topic mixtures $p(z|d)$ only for those documents on which it is trained. Hence, pLSI is not truly a generative model of documents since there is no direct way of using it to assign probabilities to a previously unseen document [11]. When

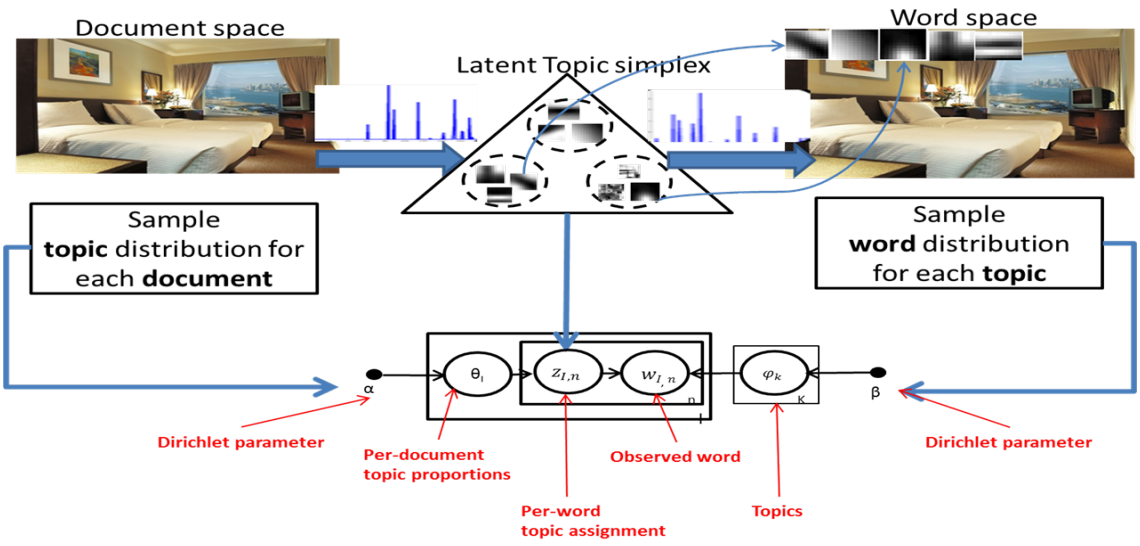


Figure 3.1: Generative graphical model of Latent Dirichlet Allocation illustrating image generation.

pLSA is treated as a Bayesian model, we get the following LDA formulation.

$$P(\theta, w, \phi|\alpha, \beta) = p(\theta|\alpha) \sum_{z} p(w_i|z, \beta)p(z|\theta), \tag{3.2}$$

Unlike pLSA, documents are also sampled as part of the generative process. The document generation is controlled by the random variable θ, which follows a Dirichlet distribution.

3.2.1 LDA Model Formulation

We formulate the LDA model in terms of visual words and images. This is analogous to words and documents in textual LDA models. The generative graphical model for generating visual words from topics is shown in Figure 3.1. Let the total number of images be D. Assume each image has N visual words. We assume a constant number of words per document for notational simplicity. In practice, there can be variable number of words in each image.

In the graphical model, each word $w_{d,n}$ from a image d is modeled as a sample drawn

from a mixture of hidden nodes called topics. The topic samples $z_{d,n}$ are, in turn, generated according to a vector of parameters θ_d, specific to the image. θ for each image is sampled from a Dirichlet distribution α. The number of topics (elements in the θ vector) is represented using T and is a parameter of the model. Conceptually, θ represents the different proportions of topics in an image. Next, topic-specific words are sampled from this distribution as follows. First, for each word $w_{d,n}$, a topic indicator $z_{d,n}$ is drawn from θ_d. Then the corresponding topic-specific word distribution $\phi_{z_{d,n}}$ is used to sample the word. The topic-specific word distributions ϕ are sampled once for the entire corpus of images. Formally, the generative model is as follows.

$$\begin{aligned} \theta_d &\sim Dir(\theta|\alpha) \qquad & \phi_k &\sim Dir(\phi|\beta) \\ z_{d,n} &\sim Mult(z|\theta) \qquad & w_{d,n} &\sim Mult(w|\phi_{z_{d,n}}) \end{aligned} \tag{3.3}$$

The generative model includes hidden variables that need to be estimated by conditioning on observed variables. In this model, the observed variables are the words in each image. In addition, the parameters of the Dirichlet distribution, also known as hyperparameters, as well as the number of topics T are specified by the user. Thus, the hidden variables are the image-specific topic proportion matrix θ of size DxT, the topic-specific word distribution matrix ϕ of size TxV, and DxN topic indicators z for all words. The joint distribution of all the observed and hidden variables, given the hyperparameters is

$$p(\theta, z, w, \phi|\alpha, \beta) = \left\{\prod_D p(\theta_d|\alpha)\right\}\left\{\prod_D\prod_N p(z_{d,n}|\theta_d)\right\}\left\{\prod_D\prod_N p(w_{d,n}|\phi_{z_{d,n}})\right\}\left\{\prod_T p(\phi_k|\beta)\right\} \tag{3.4}$$

$$= \left\{\prod_D \frac{\Gamma(\sum_T(\alpha_k))}{\prod_T(\Gamma(\alpha_k))}\prod_T \theta_k^{\alpha_k}\right\}\left\{\prod_D\prod_T \theta_k^{\#_k^d}\right\}\left\{\prod_T\prod_V \phi_{k,v}^{\#_v^k}\right\}\left\{\prod_T \frac{\Gamma(\sum_V(\beta_v))}{\prod_V(\Gamma(\beta_v))}\prod_V \phi_{k,v}^{\beta_v}\right\} \tag{3.5}$$

Each bracketed term in Equation 3.5 can be derived from the corresponding term in Equation 3.4 by replacing the probabilities with the respective density functions as modeled in Equation 3.3.

$$J.P.D. = \left\{ \prod_{D} \frac{(\Gamma(\sum_{T} \alpha_k))}{\prod_{T} \Gamma(\alpha_k)} \prod_{T} \theta_k^{\alpha_k + \#_k^d} \right\} \left\{ \prod_{T} \frac{\Gamma(\sum_{V} (\beta_v))}{\prod_{V} (\Gamma(\beta_v))} \prod_{V} \phi_{k,v}^{\beta_v + \#_v^k} \right\} \tag{3.6}$$

3.2.2 Inference

Inference in generative models can be interpreted as: 1) Inference of parameters, a.k.a parameter estimation and 2) Inference of posterior distribution over parameters, given new observations. As we shall explain later, these two problems are inherently similar for generative models, hence the validity of the common nomenclature. In this section, we explain the parameter estimation formulation for LDA.

The main goal of inference is to estimate the parameters of the LDA model ϕ, θ and the hidden variables z, given the observed visual words in images and hyperparameters.

$$p(\theta, \phi, z|w, \alpha, \beta) = \frac{p(\theta, \phi, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \tag{3.7}$$

This function does not have a closed form solution since computing the denominator (a.k.a. partition function) requires integration over the entire space of variables θ and ϕ, which makes it intractable. To solve it, researchers have proposed approximate solutions using Variational Bayes [81], Expectation Propagation [76] and Gibbs Sampling and it variants [38]. In our approaches, we choose Gibbs Sampling (GS) as the inference framework since it is a simple and a globally optimal framework. It can also be extended to handle complicated distributions that arise from the new models that proposed in the subsequent chapters.

Gibbs Sampling (GS) is a Markov Chain Monte Carlo (MCMC) algorithm that approximates a high-dimensional probability distribution by iteratively sampling from another related

distribution, called proposal distribution and constructing a Markov chain of its samples. A sample is a set of values assigned to each variable in the distribution. It has been proven that after sufficient number of iterations, the samples generated are from the desired distribution, if certain conditions are satisfied. GS is a special case of MCMC where each variable of the distribution (x_i) is sampled one at a time, conditioned on the values of all other variables ($x_{\sim i}$). The algorithm can be enumerated as -

- For iterations j from 1 to N
  - Choose variable x_i randomly.
  - Sample x_i from proposal distribution $p(x_i|x_{\sim i)}$

Calculating the proposal distribution

In Gibbs sampling, the main problem is to find a proposal distribution that is easy to sample from. In this section, we describe a method to compute the proposal distribution for LDA model, as explained in [42]. This method is based on the work of Griffiths et al. [38] who proposed a collapsed version of Gibbs Sampling for LDA. Collapsing is the process of integrating out a subset of the hidden variables to make the inference problem simpler. It also leads to faster convergence. In the LDA model the parameters θ and ϕ can be interpreted as statistics of the associations between the observed $w_{d,n}$ and the corresponding $z_{d,n}$, the state variables of the Markov chain of MCMC simulations [42]. This strategy of integrating out some of the parameters for model inference is often referred to as collapsed or Rao-Blackwellised approach, which is often used in Gibbs sampling.

$$\int_{\theta}\int_{\phi} p(\theta, z, w, \phi|\alpha, \beta)d\theta d\phi = p(z, w|\alpha, \beta) \tag{3.8}$$

Starting with the probabilities over image regions (left term), we integrate the term over θ_d, as shown below.

$$\prod_D \int_{\theta_d} \frac{(\Gamma(\sum_T \alpha_k))}{\prod_T \Gamma(\alpha_k)} \left\{ \frac{(\Gamma(\sum_T (\alpha_k + \#_k^d))}{\prod_T \Gamma(\alpha_k + \#_k^d)} \prod_T \theta_{d,k}^{\alpha_k + \#_k^d} \right\} \frac{\prod_T \Gamma(\alpha_k + \#_k^d)}{(\Gamma(\sum_T \alpha_k + \#_k^d))} d\theta_d \tag{3.9}$$

$$= \prod_D \frac{(\Gamma(\sum_T \alpha_k))}{\prod_T \Gamma(\alpha_k)} \left\{ \int_{\theta_d} p(\theta_{d,k} | \alpha_k + \#_k^d) d\theta_d \right\} \frac{\prod_T \Gamma(\alpha_k + \#_k^d)}{(\Gamma(\sum_T \alpha_k + \#_k^d))} \tag{3.10}$$

$\#_k^d$ denotes counts, the number of times topic k appears in document d. To complete the multinomial distribution $p(\theta_{d,k}; \alpha_k + \#_k^d)$ we multiply the function with the normalization coefficients and then divide by the same coefficients to retain balance. The parenthesized term in the above function is a probability density function that integrates to one. Representing the normalizing constants as $\Delta = \prod(\Gamma(.))/\Gamma(\sum(.))$, the above equation simplifies to

$$p(z, w | \alpha, \beta) = \prod_D \frac{\Delta(\alpha_k + \#_k^d)}{\Delta(\alpha_k)} \prod_T \frac{\Delta(\beta_v + \#_v^k)}{\Delta(\beta_v)}, \tag{3.11}$$

where $\#_v^k$ denotes the number of times term v is assigned to topic k in the whole corpus. From the joint distribution above, we can derive the proposal distribution for the Gibbs samples as follows.

$$p(z_i | z_{\sim i}, w, \alpha, \beta) = \frac{p(z, w | \alpha, \beta)}{p(z_{\sim i}, w | \alpha, \beta)} = \frac{p(w | z, \beta)}{p(w | z_{\sim i}, \beta)} . \frac{p(z | \alpha)}{p(z_{\sim i} | \alpha)} \tag{3.12}$$

$$= \frac{\Delta(\alpha_k + \#_k^d)}{\Delta(\alpha_k)} \prod_T \frac{\Delta(\beta_v + n_v^k)}{\Delta(\beta_v)} \tag{3.13}$$

$$= \frac{\Gamma(\#_v^k + \beta_v)\Gamma(\sum_V \#_{v,\sim i}^k + \beta_v)}{\Gamma(\#_{v,\sim i}^k + \beta_v)\Gamma(\sum_V \#_v^k + \beta_v)} . \frac{\Gamma(\#_k^d + \alpha_k)\Gamma(\sum_T \#_{k,\sim i}^d + \alpha_k)}{\Gamma(\#_{k,\sim i}^d + \alpha_k)\Gamma(\sum_T \#_k^d + \alpha_k)} \tag{3.14}$$

The Gamma function for real positive integer is the factorial function of the number reduced by 1, i.e., $\Gamma(x) = (x-1)$. Expanding the above functions as factorials and noting that the count matrices $\#_k^d$ and $\#_{k,\sim i}^d$ differ only by the value of one specific entry at i, we are able to cancel most of the terms in the numerator with terms in the denominator. Finally, we get a simple form of the proposal distribution, as follows.

$$p(z_i|z_{\sim i}, w, \alpha, \beta) = \frac{\#_k^d + \alpha_k - 1}{\sum_T(\#_k^d + \alpha_k) - 1} \frac{\#_v^k + \beta_v - 1}{\sum_V(\#_v^k + \beta_v) - 1} \tag{3.15}$$

3.2.3 Parameter Estimation

In the J.P.D. (Equation 3.4), the first term captures the posterior probability of parameter θ after observing several topic assignments on words. This is a posterior Dirichlet distribution. Using the expectation of Dirichlet distribution $Dir(a) = \frac{a_i}{\sum_i a_i}$, we estimate the parameter as follows.

$$p(\theta|\mathcal{M}, \alpha) = \frac{1}{Z_\theta} \prod_T \theta_k^{\alpha_k + \#_k^d} \tag{3.16}$$

$$= Dir(\theta|\#^d + \alpha) \tag{3.17}$$

Applying this concept to both the parameters yields the following formulations.

$$\theta_k^d = \frac{\#_k^d + \alpha_k}{\sum_T(\#_k^d + \alpha_k)} \phi_v^k = \frac{\#_v^k + \beta_v}{\sum_V(\#_v^k + \beta_v)} \tag{3.18}$$

3.2.4 Burn-in and Convergence of Gibbs Samples

For successful implementation of Gibbs sampling, the important user-defined parameters are (1) choosing an appropriate starting distribution, (2) a substantial burn-in period, and (3) after the Gibbs iterations have converged, picking uncorrelated samples for parameter estimation and inference. We assume an uniform distribution as starting distribution, since that agrees with our prior assumptions about image and scene content. For adequete burn-in period, we

used rejected the first 200 samples, and then retained every other 100th sample to minimize correlation between samples. These samples were used for parameter estimation, as explained below. We tested for convergence by (1)performing inference from multiple starting points and (2)checking if the label to topic assignments converge after burn-in iterations. In both cases, convergence was achieved.

3.2.5 Inference on new data

LDA is a generative model in which the main purpose is to model the data in such a manner as to estimate soft associations between latent topics and observed entities. How to use this topic structure when new data arrives is dependent on the task we need to solve. In the following chapters we will model different scenarios and explain inference techniques over different components. Hence, discussion on inference is postponed to the specific chapters.

Chapter 4

View-specific Object Recognition by including Context in a Topic Model Cascade.

4.1 Introduction

We inhabit in a three dimensional world, whereas images are 2D and capture only single views of the objects. This loss of information makes object category detection and localization in images a very challenging task, specially when the objects to be detected show large variations in appearance across different viewpoints. The problem of identifying objects across viewpoints is defined as *multi-view object detection*. From a machine learning standpoint, the learning objective of a multi-view object detector requires appearance features representing different views to regress onto a single classifier. Since the features do not share common appearances, they are spatially fragmented in feature space. Hence, learning a single classifier quickly becomes an intractable problem. This makes learning a multi-view object detector a hard problem to solve.

Traditional approaches decompose the problem of multi-view detection by annotating and partitioning the training images into separate views, either by manually labeling or by an initial unsupervised clustering of images [94, 55, 106]. This requires an extensive annotation and preprocessing stage. The first stage is needed to define *what is the object* in the image. Annotations spatially localize the object regions, either as bounding boxes or through image masks. The second stage is needed to define *what is the viewpoint* of these objects bounded within

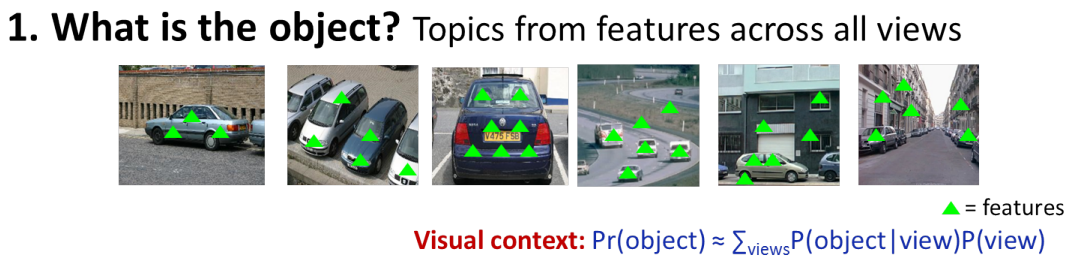


Figure 4.1: Multiview object detection involves two problems. 1) What is the object? We identify object features through latent visual topics learnt across all views. 2) What is the viewpoint? We identify joint distributions of object and correlated background through latent semantic topics learnt across all scene compositions.

these regions. This is done either manually or by feature clustering. Finally, an independent set of classifiers are trained for each view.

Our approach is based on two observations. First, objects share appearance features across viewpoints and across different instances. By showing the machine a few examples of the specific object, we expect it construct a weak, general object classifier. Second, it is generally observed that there is a strong correlation between object attributes such as size, location, viewpoint etc. and its context attributes such as perspective, spatial layout of other objects, etc. in

a scene. Due to such regularities in appearance across large number of natural images, context can help in disambiguating the object's fine grained identity. Our algorithm is based on this coarse to fine search principle.

Based on these two observations, we present a completely unsupervised approach for jointly solving *object localization* and *viewpoint estimation* in a *unified model*. Our approach is to model the unified visual and semantic context of objects and background in images. First, a generic object detector is learnt by modeling the visual context of objects across various viewpoints. Then, a view-specific object detector is learnt by modeling the spatial context of objects and corresponding backgrounds in images. In a new image, this unified model is applied to localize objects by probabilistically classifying regions into object and background.

The novelty of our approach lies in modeling the object as a function of the background. This is in contrast to most object detection systems in which large number of positive and negative examples of an object are presented to a classifier and a decision boundary is learnt to separate object (signal) from background (noise). In contrast, in our formulation instead of eliminating the background as generic noise, we model the background/scene features as a factor that influences the object signal. The appearance and the location of the object in a scene is modulated by the background factor.

Following this notion, we develop a model that localizes an object category by including its background context to restrict and refine the object's feature space. In particular, we represent a viewpoint as a joint distribution of the object appearance and its background context appearance. For example, in a highway scene, a car usually appears in its back or side-back view while the surrounding context has a linear perspective and includes trees, billboards etc. Based on such regularities, we aim at modeling object attributes specifically for each correlated context across different scenes.

Figure 4.2: Cars testset contains images from various viewpoints

We propose a *topic model cascade* to model view-specific object detectors. First, to indicate what features constitutes of an object, a Latent Dirichlet Allocation model is used on the image features to identify objects in scenes. Then, the object hypotheses generated by the LDA are further refined by jointly re-learning the feature distributions with the background regions in a View-LDA model. Intuitively, while object model needs atleast some kind of supervision (which we provide weakly through positive exemplars), the background, by virtue of its numerosity, can be learned in an unsupervised manner. By exploiting the object appearances and scene regularities, View-LDA simultaneously clusters features and images into topics (groups

of features from a single object) and scenes (groups of images with similar viewpoints). In the process, the object topic in each scene cluster adapts to its correlated scene context and learns to accurately classify object and background in that scene. In summary, we describe an algorithm that automatically learns an object model and adapts it according to view specific context. We present topic model cascade, a generalizable formulation that is used to capture visual and semantic context of objects jointly. Finally, we present a novel Gibbs sampling inference algorithm to learn the paramters and perform posterior inference on topic model cascade.

4.2 Related Work

Topic models were developed for text analysis, but have been adopted successfully for object detection and scene classification tasks [80, 40, 60]. The original, unsupervised LDA model has been adapted to include document [40] and topic level supervision [68, 33]. The purpose of supervisory information is to tune the model towards discovery of semantically coherent topics, such as object-centric topics for detection tasks. In [80], completely annotated images are used whereas in [40], a "shape-aware" model is applied on weakly supervised data while assuming automatic discovery of object-topic. In real images though, objects may not be salient and their appearance varies due to view, intra-class variability etc., which makes automatic object-topic discovery less likely. Our solution is to include weak supervision by seeding the learning process with a small number of object-only images across all variabilities. Without any manual labeling, this balances out the numerosity of the background and leads to discovery of object-centric topics. is usually resolved by manually or pre-learning separate views of the object as a pre-training stage, which is usually cumbersome.

In general, context is viewed as a scene attribute that helps to disambiguate certain object attributes such as the probability of its presence, possible locations etc. [105]. In effect,

it is often modeled independently of the object. Recently however, there has been a trend towards adaptive learning in which context plays a more active role in refining the object model. Specifically, jointly modeling of scene cues has shown to improve object description in the spatio-featural space [90, 50]. In topic models, one way to capture this dependence is by modeling correlations among topics. For example, in correlated topic model [10], topics are drawn from a single Gaussian mean and varying covariances. Instead of controlling the correlation using a single parameter, we propose a more flexible, data-driven approach. Via our proposed View-LDA model, we not only capture the correlation among features in a topic space, but also the correlation among images in the scene space. The topics are based on low-level image features and are shared by all images. The image clusters are based on topic distributions and separate different (object, context) configurations. A similar idea of co-clustering was proposed in the text analysis literature in [95]. View-LDA model can also be interpreted as a variant of the author-topic model [89], where the authors, that manifest as scenes in our formulation, are unknown and are discovered by document clustering.

Globally, an image captures a specific viewpoint of a scene. Locally, each region within the image captures either a part of the object or background. Given cues about objects and background parts in the image, we want to classify the image into one of the predefined views. We also want to solve the reverse problem i.e., given the viewpoint of the image, we want to associate regions with object or background parts. We develop a solution which tackles these two problems simultaneously.

4.3 Proposed Overall Approach

We use a hierarchical latent generative model (or topic model) to characterize visual and semantic contexts of objects in images. The latent variables in the model (also known as topics)

represent correlated features specified by the contextual cues. We construct an extended hierarchy of variables, which we call a topic model cascade. The first level of the hierarchy models a generic object, where the visual context of the object category is modeled through weakly supervised learning. The weak supervision is provided through a object cues; a dataset of images containing objects only, which are mixed with natural images (images containing objects in context) for object discovery. Features from the object cues and the natural images are correlated based on spatial proximity and similarity. This correlation is captured through latent variables, whose distributions are learnt from the corpus of images. The learnt distributions are used to weakly label object features from background. The weakly labeled image regions are cascaded to the second level of the hierarchy which defines a view specific model. In this level, images along with the region labels are used together to learn distinct views and view-specific object models. Joint distribution of features from the object hypotheses and background regions in images are probabilistically partitioned into view clusters. Object features from similar view images are pooled to update the object hypotheses. This iterative pooling of features for each view and for each object leads to an expert and highly fine-grained view specific object detection.

We now describe our overall approach in detail using Figure 4.3 as reference. Our training dataset consists of a small number of object cues (images containing objects only) and natural images (images with object and background context). Each image is segmented into superpixels and local features are extracted and quantized into visual words. Thus, each image segment becomes a document and the features get captured as bag of words corresponding to each document. A Latent Dirichlet Allocation is applied to this word-document corpus to discover coherent clusters of visual words. Each topic softly clusters frequently cooccurring visual words. Some topics consistently occur within object regions; we call them ”object topics”. To automatically identify object topics, we adopt an information theoretic approach. We measure entropy of topics across object cues. We select object topics based on entropy of topics. A high

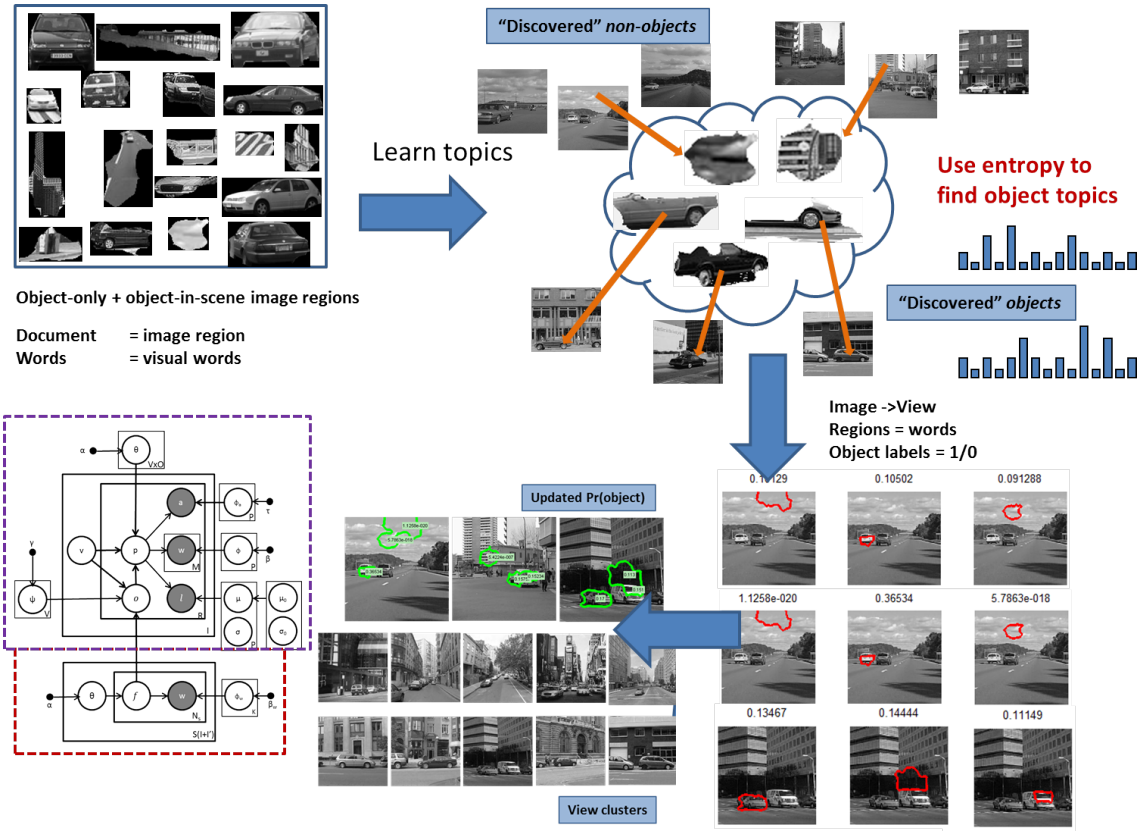


Figure 4.3: Overall approach

entropy of a topic means maximal presence of that topic in object cues. The image segments in natural images with high density of object topics are cued as objects. Hence, all image segments are weakly classified into object or background. Next, the cued locations and image features are supplied to a View-LDA model that simultaneously clusters images and regions into *views* and *parts*. Specifically, in the View-LDA, images are documents and regions are words. A word is associated with a side information, which is the prior object label obtained from the previous step. The latent variables are views and parts. Each view is a distribution over parts of the image that corresponds uniquely to that view's (object, background) configuration. By learning the detailed view and parts configuration of the image, we update the object scores

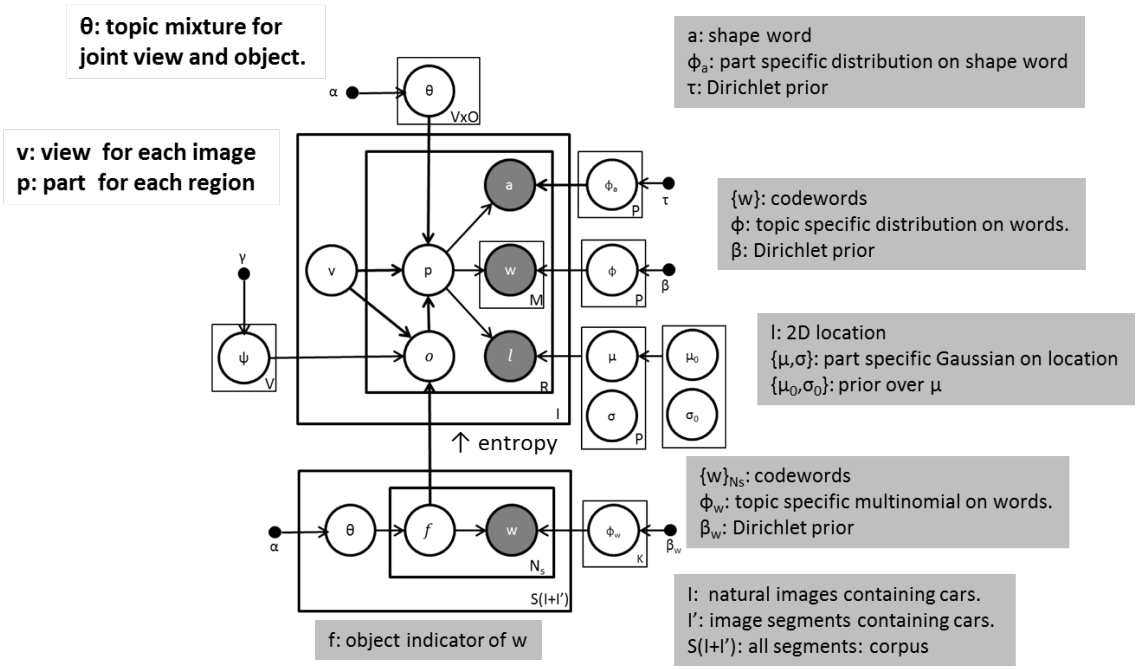


Figure 4.4: View-LDA graphical model

at each image region. The assumption is that the post-View-LDA object probability at correct regions will increase while the probabiities at the background decreases. Intuitively, context is exploited at two levels. First, LDA captures visual context of an object among local features within image segments. Then, View-LDA captures spatial and co-occurrence relations between regions within an image which is jointly abstracted out as the view parameter of the image.

4.4 Mathematical Formulation

4.4.1 Visual context across object views

In the first stage of our topic-model cascade, we use positive exemplars to cue object locations in natural images. Let the number of images be I and the number of positive exemplars be I'. We segment all the images and represent local features as visual words. The input to the

model consists of a corpus of image segments S, where each segment consists of a variable number of visual words $\vec{w_{N_1}}, \vec{w_{N_2}}, \cdots, \vec{w_{N_S}}$. Let the number of segments extracted from the images be represented as $S(I + I')$. We treat each image region as a bag of visual words and fit a Latent Dirichlet Allocation model on the segments corpus. The LDA model discovers latent "topics" f which are coherent clusters of visual words that capture frequently occurring patterns in the segments. We model the topic proportions in a segment as a multinomial and denote it by parameter θ. The topic-specific word distribution is denoted by the multinomial ϕ. The uniform Dirichlet hyperparameters are denoted by α and β, respectively. The N_s visual words in each segment are sampled according to the hidden topic indicator f. The number of topics is denoted by K and is set by the user. The total number of visual words is the size of the vocabulary of the codebook. We denote it by V. Based on collapsed Gibbs sampling formulation derived in Section x, the proposal distribution for the topic of the i^{th} visual word v is:

$$p(f_i = k | f_{\sim i}, \vec{w}, \alpha, \beta) = \frac{\#_k^s + \alpha_k}{\sum_K (\#_k^s + \alpha_k)} \frac{\#_v^k + \beta_v}{\sum_V (\#_v^k + \beta_v)}, \tag{4.1}$$

where $\#_k^s$ is the number of times topic k appears in image segment s and $\#_v^k$ is the number of times visual word v is assigned to topic k in the entire sample set. After convergence of the Gibbs sampling iterations, each local feature in each segment is assigned to a topic.

4.4.2 Inferring object hypothesis from topics

The LDA captures spatially localized features that are typically found in the image corpus. Since all the images in the corpus contain the object of interest (OOI), a subset of these topics should capture patterns associated with the object. For example, if "car" is the OOI, a topic could be associated with a subpart of a car, e.g., its tires or a side door. To determine which of the topics capture "object-like" parts we look at the entropy of the topic distributions within

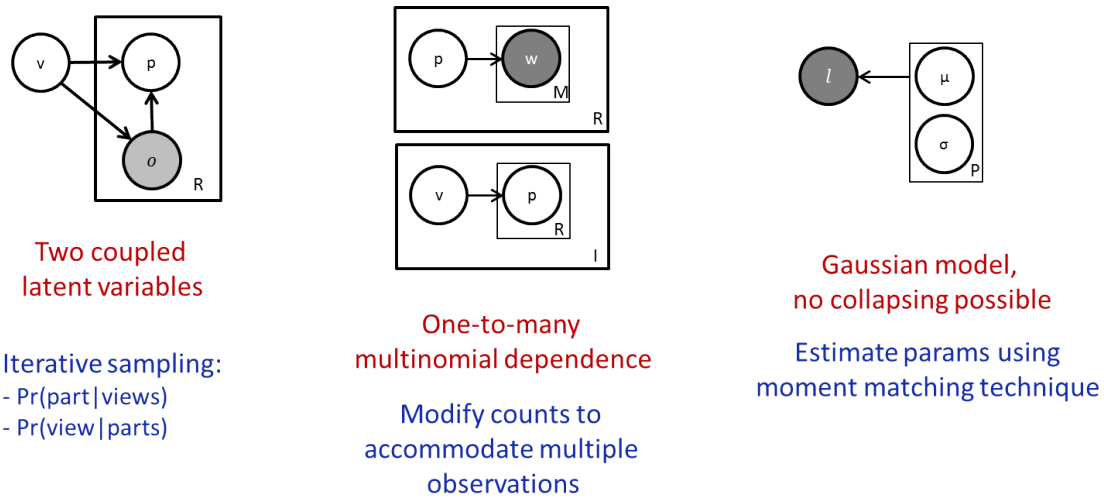


Figure 4.5: This figure illustrates the main modeling differences between LDA and View-LDA that lead to learning challenges.

the positive exemplar set. The intuition is that a topic that captures object parts will appear in many different object images, leading to high entropy of the topic across images. However, if it is a background topic, it should have low entropy. This use of entropy has been included earlier in [89] to identify diversity of authors in documents. The region-specific topic distribution is transposed θ' and the entropy of each topic is computed.

$$H(\theta_i{}') = -\sum_{S(I')} \theta_s{}' log_2 \theta_s{}' \tag{4.2}$$

In each image, we rank the regions based on the average entropy content of their constituent topics. The top 10 regions in each image are hypothesized as objects. The images and the object hypotheses are cascaded onto the next step.

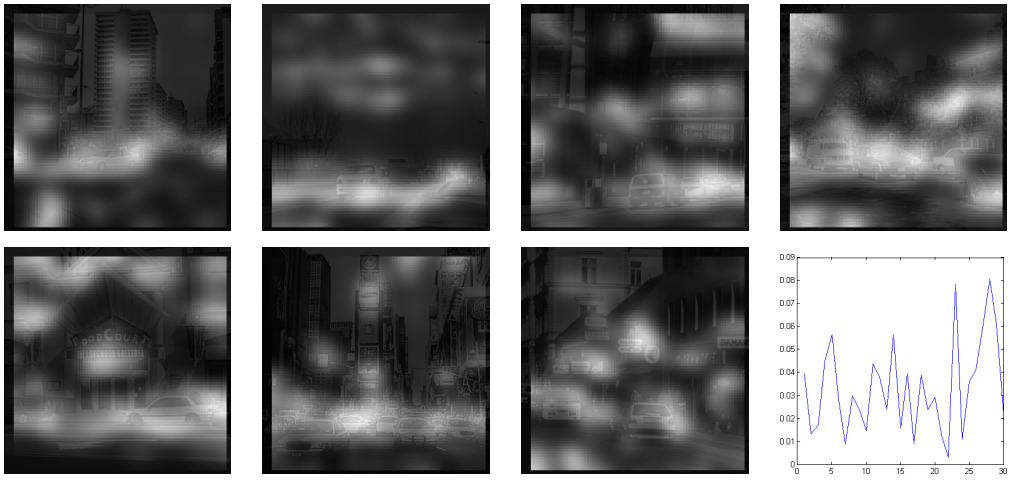


Figure 4.6: Generic object confidence maps in scene images and the average document-topic distribution over object images. Zoom in for better view.

4.4.3 Semantic and spatial context across scene compositions

The View-LDA models images and regions into *views* and *parts*. The model is learnt by using the appearance features, locations and object hypotheses of the image regions as observations. Each image is associated with a latent view variable v which determines what parts of the object and of the background are visible in the image. Each image region is associated with two latent variables, an identity variable o and a part variable p. The object identity variable o is the object hypothesis in the learning phase. It is conditioned on the view v, which determines the number of times o is sampled in this view; capturing the extent of the object in the image. The multinomial variable ψ models this relation between the view and the object identity. The part variable is conditioned on both the view and the identity variables and is modeled by the multinomial θ. It captures the semantics of the appearance of parts from the object and background. For example, assume that the view is a "street" and the object of interest is a "car". Then, if the identity is "object", one part variable may model the tires of the car. If

Figure 4.7: Test images clustered by view posterior and the corresponding object-part distribution of that view.

however, the identity is "background" for the same view, the part may model the windows of a building (a typical background object in street scenes). The parts are dependent on the visual appearance (visual words w and region properties a) and location l of the regions they model. The appearance relations are captured by multinomials ϕ and ϕ_a and β and τ are Dirichlet hyperparameters to a and w, respectively. For location, we use two dimensional Gaussian priors (μ, σ), with parameters mean, μ and covariance, σ. The conjugate prior of a multivariate Gaussian mean is also a Gaussian. We represent it by (μ_0, σ_0). The conjugate prior over covariance is an Inverse-Wishart distribution [7]. Since in our experiments, estimation of covariances between x and y coordinates often lead to singularities, we fixed the values of covariance matrix to 1000 in each direction.

4.4.4 Parameter Estimation and Inference

The joint distribution of all the variables, except the hyperparameters in the model is given by:

$$J.P.D. = Pr(\overbrace{v, o, p}^{\text{hidden variables}}, \underbrace{a, w, l}_{\text{observables}}, \overbrace{\psi, \theta, \phi_a, \phi, \mu, \sigma}^{\text{parameters}}) \tag{4.3}$$

$$= \prod_{I} Pr(v) \prod_{R} Pr(O|\psi_v) Pr(p|\theta_{v,o}) Pr(a|\phi_{a_p}) Pr(l|\mu_p, \sigma_p) \prod_{M} Pr(w|\phi_p) \tag{4.4}$$

We learn the parameters of the model by Gibbs sampling the proposal distributions which are computed by collapsing (integrating out) parameters θ, ψ, ϕ_a, ϕ (as explained in Chapter 3). The conjugate priors μ and σ are difficult to be integrated out, hence they are optimized within each iteration. In the learning phase, the identity variables o are known. Therefore, there are two unknown variables, v and p. The proposal distribution of p being assigned to value k is:

$$\begin{aligned} Pr(p_i = k|\vec{p_{\sim i}}, V, O, W, A, L, \Omega) \quad \propto \quad & Pr(p_i|p_{\sim i}, o_i, v_i, \Omega) \times Pr(a_i|p_i, A_{\sim i}, \Omega) \cdots \\ & \times Pr(\vec{w_i}|p_i, W_{\sim i}, \Omega) \times Pr(l_i|p_i, L_{\sim i}, \Omega), \end{aligned} \tag{4.5}$$

where Ω is the set of all hyperparameters in the model. The first two terms are transformed into count ratios by utilizing the Dirichlet-Multinomial solution for single entry difference in count matrices (Chapter 3). The third term relates each part to multiple visual words. This implies that excluding the current part (p_i) affects multiple word counts in the count matrix. The solution is derived as follows.

Deriving Gibbs proposal for one-to-many multinomial relation

The conditional likelihood ratio of words, given a topic z:

$$\frac{p(\vec{w}|z=k)}{p(\vec{w_{\sim i}}|z=k)} = \frac{\Delta(n_z+\beta)}{\Delta(n_{z,\sim i}+\beta)} \tag{4.6}$$

$$= \frac{\prod_V \Gamma(n_v^k+\beta_v)}{\prod_{V,\sim i}\Gamma(n_v^k+\beta_v)} . \frac{\Gamma(\sum_{V,\sim i}(n_v^k+\beta_v))}{\Gamma(\sum_V(n_v^k+\beta_v))}, \tag{4.7}$$

where V is the vocabulary size for words. For solving the first term, we see that all the terms cancel out in the expanded form except for the terms that appear with the i^{th} entity. Also, the difference between the complete set in the numerator and the i^{th} removed set in the denominator for each of those terms is the number of times (count) the term appears in the i^{th} entity (c_v). Replacing the Gamma with its factorial equivalent yields the following form:

$$= \prod_{v,v\in i} \frac{(n_v^k+\beta_v-1)!}{(n_v^k-c_v+\beta_v-1)!} = \prod_{v,v\in i}\prod_{c=0}^{c_v-1}(n_v^k+\beta_v-c) \tag{4.8}$$

For solving the second term, we see that when the sums are expanded within the Gamma function, the overall difference between the numerator and the denominator is due to the reduced counts of all the terms not appearing in the i^{th} entity.

$$= \frac{(\sum_V n_v^k + \sum_V \beta_v - 1 - \sum_{v,v\in i} c_v)!}{(\sum_V n_v^k + \sum_V \beta_v - 1)!} = \frac{1}{\prod_{c=1}^{c=C_w-1}(\sum_V n_v^k + \sum_V \beta_v - c)}, \tag{4.9}$$

where $C_w = \sum_{v,v\in i} c_v$.

Estimating Gaussian parameters

The fourth term of Equation 4.4 involves estimating the location of the i^{th} region, given its part assignment. The posterior distribution over l_i is a Gaussian whose parameters can be estimated using moment matching technique. The sample mean and the covariance of all the regions

associated with a given part p is:

$$\mu^k_{\sim i} = \frac{1}{n^k} \sum_{I,R} \sum_{j:p_j=k \cap j \neq i} x_j \quad \sigma^k_{\sim i} = \Delta_k + \sum_{I,R} \sum_{j|p_j=k \cap j \neq i} (x_j - \mu^k_i)(x_j - \mu^k_i)^T. \tag{4.10}$$

$$Pr(l_i|p_i, l_{\sim i}, \Omega) = \mathcal{N}(l_i; \mu_{p_i,\sim i}, \sigma_{p_i,\sim i}) \tag{4.11}$$

The final form for the proposal distribution of parts is:

$$\left(\frac{n^{vo}_p + \alpha}{\sum_P n^{vo}_p + P\alpha}\right)\left(\frac{n^p_a + \tau}{\sum_A n^p_a + A\tau}\right)\left(\frac{\prod_{w,w\in i}\prod_{c=0}^{c_w-1}(n^k_w + \beta_w - c)}{\prod_{c=0}^{c=C_w}(\sum_W n^k_w + \sum_W \beta_w - c)}\right)\left(\mathcal{N}(l_i; \mu_{p_i,\sim i}, \sigma_{p_i,\sim i})\right), \tag{4.12}$$

Above, we derived a formula to sample parts for image regions. Parts proposal is conditioned on the view variable (term 1 of Equation 4.12), which we assume as known. However, since it is also hidden, we derive a sampling formulation, in which parts are assumed to be known. Accordingly, this can be set up as an iterative sampling solution, where parts are sampled based on views and vice versa. For view sampling we observed that conditioned on z and o, the view samples are independent of other variables. Removing the parameters, the equation becomes:

$$Pr(v_i = v|v_{\sim i}, p, o, w, a, l, \Omega) \propto Pr(v_i|v_{\sim i}, p, o) \propto P(v_{\sim i}, p, o|v_i) \tag{4.13}$$

$$= Pr(p_i|v_i, v_{\sim i}, o)Pr(o_i, v_{\sim i}|v_i) \tag{4.14}$$

The above equations are derived by Bayes rule and by noticing that the values of the part and identity of the i^{th} entity is independent of the $\sim i^{th}$ views, given the i^{th} view. Hence, the sampling distribution is:

$$Pr(v_i|v_{\sim i}, p, o) \propto Pr(o_i|v_i, o_{\sim i}) \times Pr(p_i|v_i, o_i, p_{\sim i}), \tag{4.15}$$

The relations $(view- > label)$ and $(view, label- > part)$ are similar to the relation between part and visual words; i.e., they follow a one-to-many multinomial dependence as derived in Equation 4.7. We directly write the view proposal:

$$Pr(v_i = v|v_{\sim i}, p, o) \propto \left(\frac{\prod_{o \in i} \prod_{c=0}^{c_o - 1} (n_o^v + \gamma - c)}{\prod_{c=0}^{c=C_o} (\sum_2 n_o^v + 2\gamma - c)} \right) \left(\frac{\prod_{p \in i} \prod_{o \in i} \prod_{c=0}^{c_{p,o} - 1} (n_{p,o}^v + \alpha - c)}{\prod_{c=0}^{c=C_{p,o}} (\sum_{p,2} n_{p,o}^v + 2P\alpha - c)} \right), \tag{4.16}$$

where c_o is the number of times the identity o appears in the i^{th} image and C_o is the sum over all os. Similarly, $c_{p,o}$ and $C_{p,o}$ capture similar statistics but by including part assignment. The parts and view proposals are coupled, i.e., they are conditioned on one another. The view posterior is dependent on the parts assignments and vice versa. We use iterative sampling to decouple the variables [40, 95].

After a number of Gibbs iterations, the multinomial parameters are estimated from the samples as follows. The Gaussian parameters are also updated in every iteration given by Equation x.

$$\gamma_v^o = \frac{n_o^v + \alpha}{\sum_O n_o^v + O\alpha} \quad \theta_{vo}^k = \frac{n_k^{vo} + \alpha}{\sum_K n_k^{vo} + K\alpha} \tag{4.17}$$

$$\tau_k^a = \frac{n_a^k + \beta}{\sum_A n_a^k + A\beta} \quad \phi_k^w = \frac{n_w^k + \beta}{\sum_W n_w^k + W\beta} \tag{4.18}$$

Inference and Evaluation

Given a new image, our goal is to find the locations of the objects. To achieve this using our contextLDA model we need the posterior label probabilities $(Pr(o|observables, \mathcal{M})$ for each image region. This is explained below.

We follow the method of query sampling proposed by [45] and run the inference algorithm on the new image exclusively. The difference in the posterior Gibbs sampling is that the states of Gibbs sampling are augmented with the observations of the new image. It could be assumed to be a model update step in that a new image is inserted into the already existing corpus and the counts of the model need to be updated because of the new observations.

We first initialize the algorithm by randomly assigning parts and labels to regions and a view to the image. Then a number of loops through Gibbs posterior sampling are performed. In this, only the counts of tokens associated with the new image are changed in each iteration. Specifically, for each sample, we first iterate over all regions and update their parts and labels, respectively. Next, after updating the region assignments, we update the view assignment for the complete image. This gives us a single sample of $\{view, parts, labels\}$.

At the end of the sampling routine, we compute the parts distribution, marginalized by the probabilities over view samples.

$$\tilde{\theta}^o = \sum_V \frac{\#_v}{\sum_V \#_v} \frac{\tilde{\#}_p^{v,o} + \alpha_p}{\sum_P \tilde{\#}_p^{v,o} + \alpha_p} \tag{4.19}$$

This gives us the parts distribution for the object and the background label. In our images the objects occupy a much smaller part of the image as compared to the background. Hence, we rely on the background parts distribution and use that to assign a view to the image. Specifically, using symmetric Kullback-Leibler Divergence we compare the parts distribution of the new image with that obtained from the learned corpus. This measure has been earlier used for comparing discrete random variables in [43].

$$D_{KL}(\tilde{P}_{o=2}||P^v_{i,o=2}) = \sum_P Pr(\tilde{P}_{o=2} = p)(log_2 Pr(\tilde{P}_{o=2} = p) - log_2 Pr(P^v_{i,o=2} = p)) \tag{4.20}$$

$$D_{KLSymm}(\tilde{P}_{o=2}||P^v_{i,o=2}) = \frac{1}{2}(D_{KL}(\tilde{P}_{o=2}||P^v_{i,o=2}) + D_{KL}(P^v_{i,o=2}||\tilde{P}_{o=2})) \tag{4.21}$$

The image is assigned to the view whose parts distribution gives the minimum KL Divergence.

$$v^* = argmin_v D_{KLSymm}(\tilde{P}_{o=2}||P^v_{i,o=2}) \tag{4.22}$$

Finally, the label probabilities at each image region is calculated by

$$Pr(o|v^*, \mathcal{M}) = \sum_P \theta^o_p \sum_o Pr(p|o, v^*) \tag{4.23}$$

$$\tag{4.24}$$

where $c^j_{K,O} = \sum_{K,O} c^j_{k,o}$ and $c^j_{k,o}$ denotes frequency of the $(k, o)^{th}$ pair observed in the j^{th} image.

4.5 Implementation details and results

In this thesis, we focus on a specific object: cars. Cars demonstrate all the properties that makes its detection hard (as shown in Figure 4.2), namely, large intra class variations (different makes and models), dramatic change of appearance between views (frontal versus side profile) and usually appears within a large and complex background (street, highway etc.). Car detection is

Figure 4.8: Posterior object probabilities predicted by View LDA. For each image, these were the top 5 object regions hypothesized by the initial LDA model. We observe that View-LDA is successfully able to increase the object posterior of true positives regions and decrease it at false alarm regions.

also an essential in many different applications e.g., tracking, surveillance and 3D reconstruction of scenes. Because of these reasons, detecting cars makes an interesting case study for understanding multi-view detection.

For training, we use images from the multi-view car training set[1] as object-only images. The first 50 images from each of the 7 viewpoints were considered. For natural images with cars, we collected images from MIT LabelMe scene dataset[2], Google street view[3] screenshots and ETHZ Multiview car dataset. There were 167 images in total with 576 instances of cars. There was an average of 2.13 cars per image. A subset of our test images are shown in figure x.

[1] http://www.vision.ee.ethz.ch/ bleibe/data/datasets.html#cars-all

[2] http://labelme.csail.mit.edu/

[3] http://maps.google.com/

Our test set contain images of moving cars on streets and highways and parked cars on streets.

In the LDA step, features in object images are SIFT descriptors[4] on Kadir-Brady interest points[5] extracted within given segmentation masks. In scene images, each image region is a document. To find regions, each image is segmented into superpixels by constrained normalized cut followed by grouping[6]. There are 40 regions per image. This segmentation method is ideally suited over others because it gives convex regions that bound the object regions well. Each region is then represented using dense PHOW features, a variant of dense SIFT descriptors extracted at multiple scales[4]. Features are vector quantized into codewords using k-means clustering algorithm. Each region also includes the location, and region properties. The location is the (x, y) coordinate of the region center and the region properties are the aspect ratio and orientation. During model learning, the parameters for the LDA model were set to $\alpha = 0.1$ and $\beta = 0.01$ to get sparse topic distributions. In the View-LDA model, α was changed to $= \frac{50}{T}$. The number of topics in both models was set to $T = 30$. The number of scene clusters is set to $V = 6$. The number of iterations in Gibbs sampling procedure was set to 200 for training and 50 for inference.

After LDA, 5 highest entropy topics are used to label regions as objects in scene images. In Figure 4.6, the plot shows the average document-topic distribution over object images. It shows that the generic object description is an incoherent mixture over multiple topics. In contrast, in Figure 4.7, we show object-topic distributions for three scene clusters. The topic concentration is sparse and peaky which points to a more restricted and decisive object space.

To illustrate how the object model adapts to its correlated context, we show some results in Figure 4.9. Each row displays the confidence maps and detected regions based on view-specific and scene-independent model respectively. For example, if we ignore the scene in the

[4]http://www.vlfeat.org/overview/

[5]http://www.robots.ox.ac.uk/ timork/salscale.html

[6]http://www.cs.sfu.ca/ mori/research/superpixels/code/

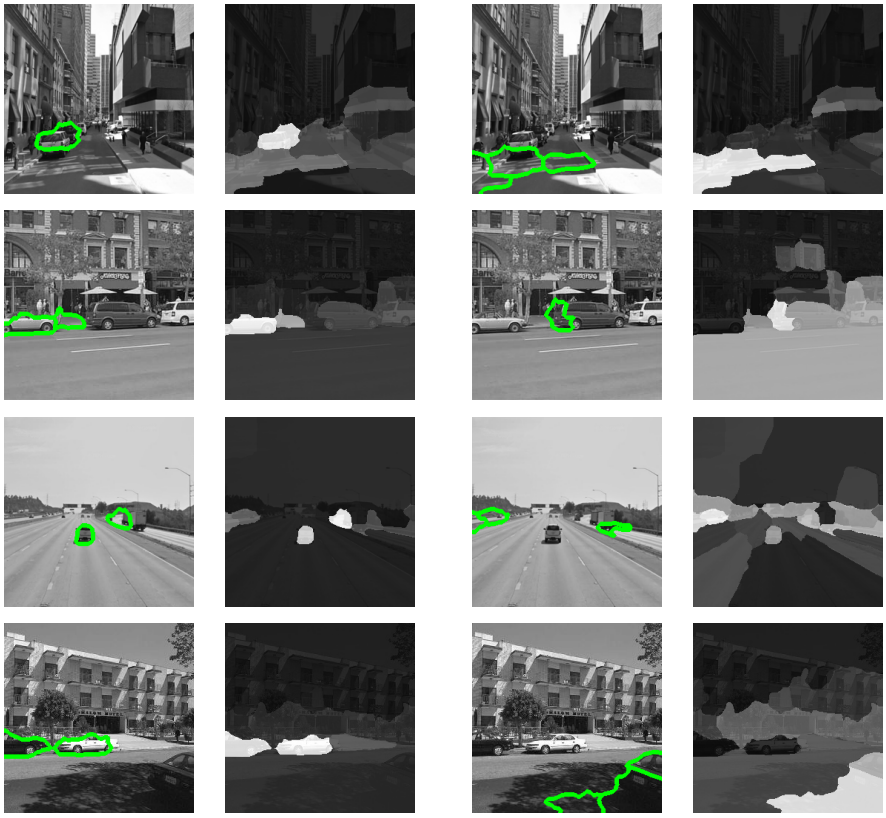


Figure 4.9: Detected regions and probability maps based on view-specific (Left) and scene-independent (Right) object model.

first image, the shadows on the road can be erroneously detected as object. However, when the context is mapped (street scene, buildings at an angle), the actual object regions (back-view of cars) are automatically selected.

To quantitatively evaluate our results on the car dataset, we consider two metrics(a) number of images with correct detections and since each image can contain multiple car regions,(b) number of correct regions. Each region can contain a single complete car, a part of the car or multiple cars. The results are shown in Table 4.1.

4.5.1 Comparison with semi-supervised LDA

We compare our results with the semi-supervised LDA model (SS-LDA) [68]. The main difference between our model and SS-LDA is in the dependencies between topics and labels. View-LDA model is an upstream model in which topics are conditioned on labels. On the other hand, SS-LDA is a downstream model in which labels are generated from topics. Except for the additional features (labels, locations and region properties) associated with the topics, the SS-LDA is the similar to LDA and so is the learning procedure. Object topics in SS-LDA are identified as the topics that have high correlation with object label. The object probability calculation in a new image is identical to View-LDA, except that the scene dependence is ignored.

We notice that our model outperforms the SS-LDA model by a huge margin. In fact, the upstream, scene-independent model also performs better that SS-LDA which can be considered as a scene-independent, downstream model. Intuitively, this is mainly because of the data-driven, noisy labels that we use to model the topics. In upstream models, topics are discovered around high frequency (hence,high density) regions in the feature space. This narrows the confidence region of the generic model to a sharp peak in feature space. On the other hand, in the downstream model, the topics are first generated and then labeled, due to which more overlaps may exist between object and context topics. The first row in Table 4.1 shows the post-learning object probabilities for the two models. Higher the difference between object and background averages, better is the separability at runtime. In general, the upstream modeling with view specific detection due to View-LDA outperforms both the other object-background separation based approaches.

	View-LDA	SS-LDA	Scene-indep LDA
Post training P_{av} (Object/Backgnd)	0.143/0.068	0.113/0.093	-
Detections/image	67/76	50/76	55/76
Detections/region	162/206	112/206	134/206
FPPI	0.3460	0.4653	0.4472

Table 4.1: Results across 576 test images.

4.6 Conclusions

In this paper, we represent a scene as a joint configuration of object and its context. We describe an algorithm that automatically learns an object model and adapts it according to local context cues. In effect, a cascaded topic model is proposed that jointly models scene features to learn object detectors that accurately classify object and view-specific context. To demonstrate the efficacy of this method we perform multi-view car detection on a challenging dataset. We further discuss and justify our design choices empirically. Specifically, we show that the view-specific, upstream model for object detection outperforms object-background separating, downstream models.

Figure 4.10: Object detection results. Top, View-LDA based detections. Bottom, SS-LDA based detections.

Chapter 5

Recognizing Object Labels through Visual and Semantic Contexts

5.1 Introduction

In the recent years there has been a tremendous interest in image understanding through multi object category detection. A significant corpus of research on this topic is focused on semantic context modeling in which a single scene context is imposed on an image to filter erroneous or incompatible object detections [61, 100]. Typically, these models represent scene semantics by the count statistics of visual words which are quantized indices of high dimensional image features. These simplifying assumptions about the scene content and the local image regions limit the diversity of detected objects. Consequently, these systems show high performance for simple, natural scenes e.g., animal in grassland, airplane on ground etc., but there is little evidence of robust detection of large number of diverse objects that appear in complex, man-made spaces e.g., studios, offices, etc.

In contrast to prior work, our goal is to construct a context model that can *label multiple objects* in *complex scenes* with high precision. Since object labeling is essentially a man-made construct, we are inspired by the human brain's ability to fluently *parse layered scene semantics, discern multiple contexts and uniquely identify objects*. Consider the famous Nighthawks painting in Figure 5.1. In the first glance, we recognize the scene as a bar-room with people. On closer observation, other less dominant aspects of the scene such as the shopfronts, buildings etc., are enhanced and the overall scene context of the painting is revealed as a roadside

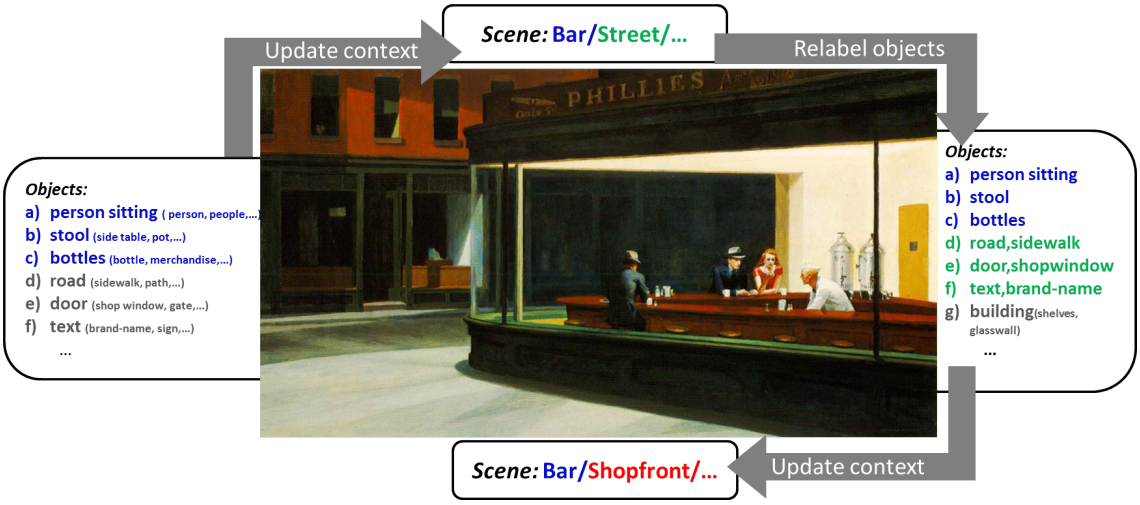


Figure 5.1: An illustration of label interpretation using visual and semantic context.

bar at nighttime. In the light of this overall context, we recognize generic object categories in more specific terms, for example “road” as “sidewalk”, and “text” as “brandname”. We repeat this process of linking names with concepts several times depending on the complexity of the scene.

There is evidence that top-down facilitation of object recognition in human brain is triggered by semantic associations, and more interestingly, by visual associations among objects. In Moshe-Bar et al. [30], cognitive experiments proved that in recognizing objects, “contextual analysis of initial guesses primes recognition of contextually related objects.” While the initial guesses of objects are driven by a rapid projection of coarse visual input comprising mainly of low spatial frequencies, the primed representations are driven by the semantic context of the nearby objects. The experiment informs that ”most likely” interpretation of an object is a function of two key mechanisms - appearance information from the to-be-recognized object itself and the semantic information from the context in which the object appears. Inspired by this theory, we attempt to model this biological phenomenon through a probabilistic model that explains the ”most-likely” interpretation of object through the concept of a ”visualsemantic

label". A visualsemantic label of an object is modeled such that object label associations are captured through (a) *visual context* and (b) *semantic context*. The visual context groups objects with appearance similarities. The semantic context groups objects that appear together in a scene. The two context spaces interact through the *visualsemantic model* such that the information from the labels in one space is used to update their identities in the other space. Thus, visually ambiguous labels are separated through semantic context e.g., trees generally occur in many scenes; however palm-trees specifically occur in sea beaches. Semantically ambiguous labels (labels that do not co-occur frequently) persist through visual context, e.g., a tree outside the window of a bedroom (a tree doesn't fit the bedroom context), or objects in complex spaces e.g., a kitchen with a dining area. By reasoning about object labels jointly in the two contexts, we are able to outperform current state-of-the-art context models, as shown in our evaluations. Our main contribution is that we propose a model that challenges the dominant view of context by broadening its definition to include visual and semantic relations within a unified framework.

- **Unified Context Model**: We present *Composite Scene Detection* (CSD) that models semantic and visual contexts of object labels. This is the first work that *explicitly models the ambiguities of textual labels as well as of the image features* and proposes a solution for resolving them in an unified manner.

- **From Visual words to Visual topics with nnLDA**: We *overcome the lossy "visual words"* representation prevalent in topic model framework by proposing a novel nnLDA to model *visual topic manifolds*.

- **Data driven semantic topic correlations with PAM**: Unlike prior methods that are parametric, we employ a Pachinko Allocation Model that allows *data-driven, graph-based, flexible topic correlations*.

- **Novel optimization technique**: We propose Data Augmentation algorithm using collapsed Gibbs sampling for inference in our linked topic-model.
- **Outperforms the state-of-the-art**: Our method is novel and improves object detection scores in the highly challenging SUN09 dataset.

5.2 Related Work

The separate tasks of object labeling and scene classification have recently been linked together in scene understanding models, mainly through the use of semantic context [87, 61, 100, 34]. Such systems improve recognition rates by filtering appearance information through pairwise [87] or grouped relationships [100] between labels that constrain object classification. More recently, researchers have sought to include other kinds of label relations based on language use. Visual synset is applied for object categorization in [115]. Spatial relations between object classes is modeled in [39] by examining "beyond nouns" to higher language constructs. In [18, 73], a semantic-spatial object class hierarchy has been proposed. In [49], the relative order of words used in captions is used as a cue for detecting object prominence in images. Sentence generation through objects, adjectives, and spatial relations is proposed in [53]. In [26], the authors evaluate similarity between an image and a sentence through an intermediate "meaning space" based on fitting semantic triplets of object, action, and scene. Visual hierarchy of object classes is proposed in [92, 36]. However, all these works consider a *unidirectional flow of information* either from language space [18, 49], or from images [26, 92]. In effect, none of these works consider interaction between language ambiguities and appearance ambiguities to solve scene understanding.

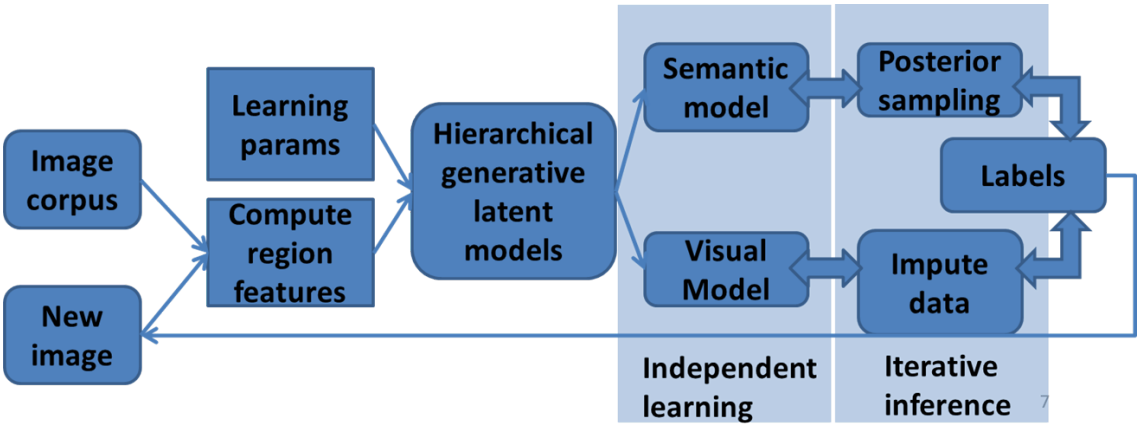


Figure 5.2: Flowchart of the overall approach.

5.3 Modeling Context

Recognizing objects in images requires associating labels to image regions. An object label is a textual word (noun) used to describe the features in an image region. It is hard to glean any information from a single textual word or a single local feature in isolation. To delve deeper, we search for the context in which it occurs. A textual word occurs in conjunction with other words in an image of a scene. The features of an image region lies in some neighborhood as other features in a feature space. Hence, based on the representation space, one can group entities into contextual clusters. A new image can be disambiguated by searching for a set of object labels that best explain the contextual clusters its image regions fall into. This is the underlying assumption of our modeling approach.

We model object labels in the semantic space based on the cooccurrences within images and in the visual space based on the appearance similarities between the corresponding image regions. We use the concept of generative, hierarchical latent modeling (e.g., topic models) to discover coherent clusters of labels (e.g., topics) in each of these spaces. We introduce the idea of cascaded generative models, where the probabilities from one model conditions the beliefs

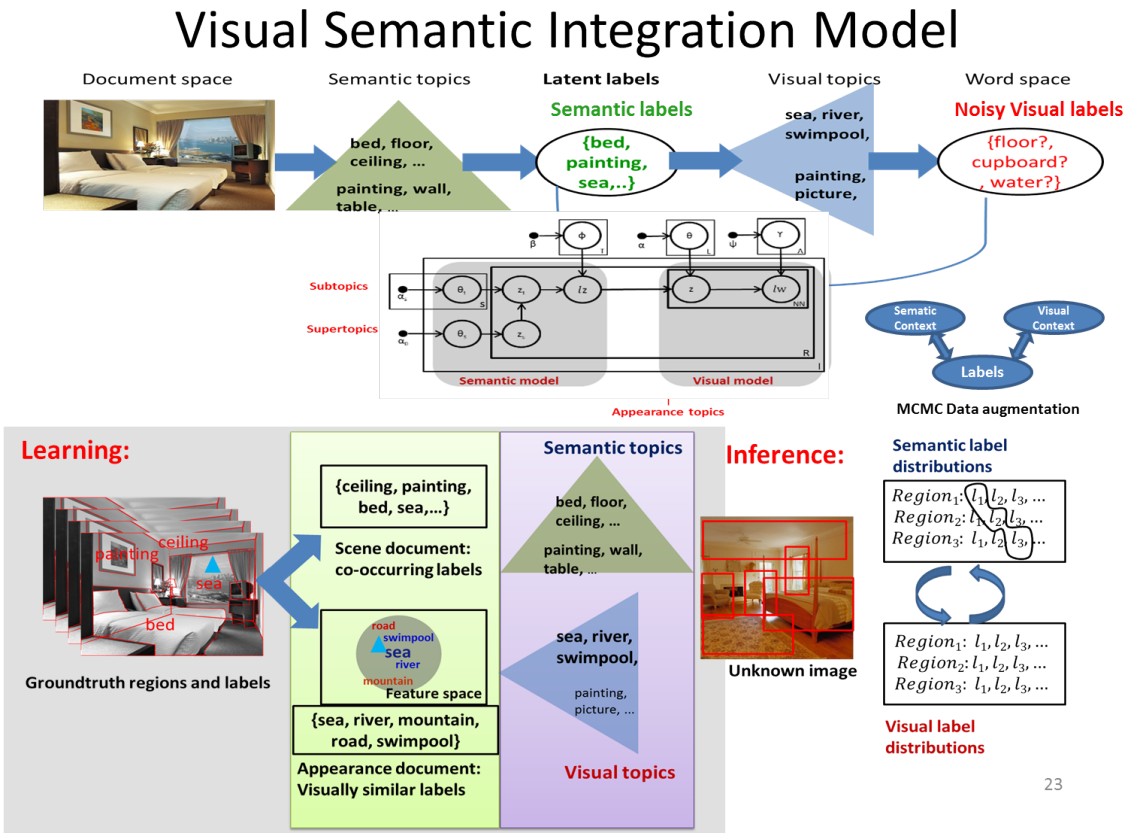


Figure 5.3: Visualization of VSIM

in the next model. We learn the parameters of the model from annotated images. During inference, the generative process is reversed to explain the image formation. We introduce an iterative sampling framework for inferring the posterior distributions in cascaded models, which is similar to Expectation Maximization approach in the Maximum Likelihood setting.

Specifically, we divide an image into regions and compute their features. These features are observations in the inference engine from which the scene parameters and the local label probabilities are inferred. Each region is thus projected as a discrete probability distribution

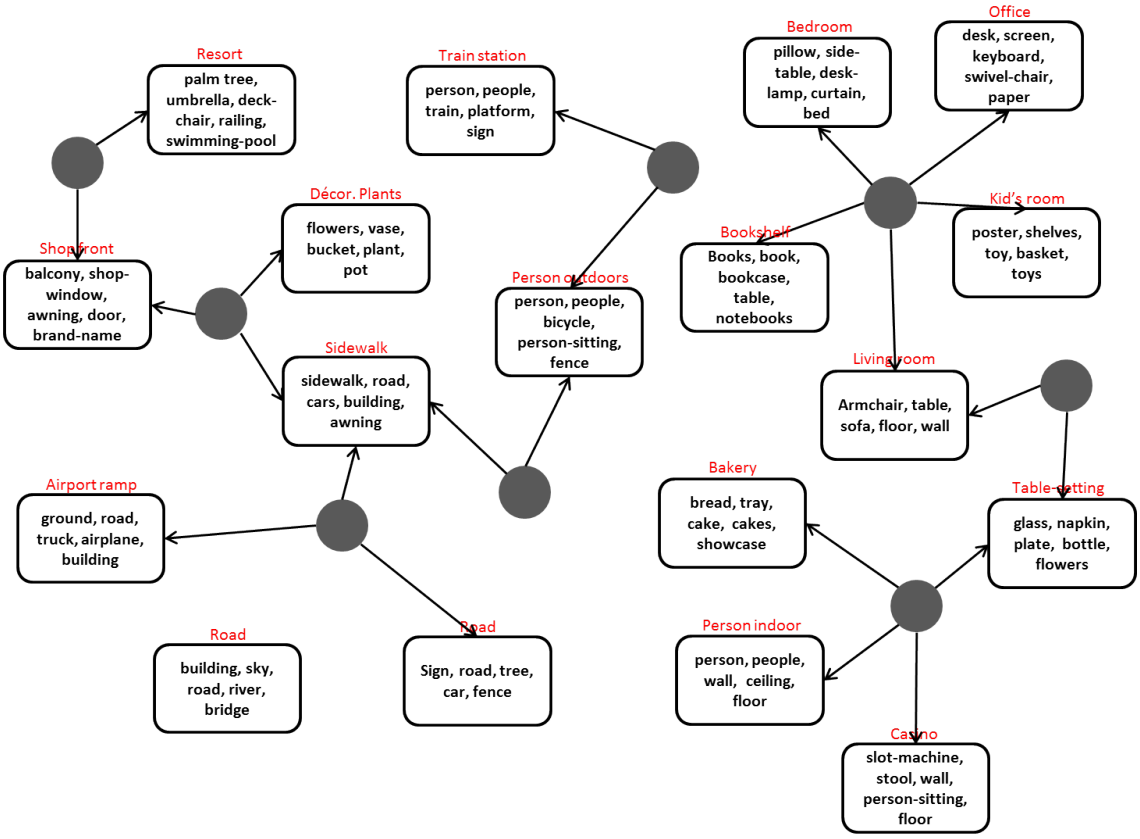


Figure 5.4: Subset of the PAM context graph. Supertopics (gray nodes) and subtopics (most frequent labels). Our interpretation of each subtopic is denoted in red. Zoom in for better view.

over all the labels. Finally, maximum aposteriori values are used to decide on the label assignments. The pipeline of our approach is illustrated in Figure 5.3. In the following, we provide a conceptual overview of the method and defer the detailed mathematical formulation in Section 4.

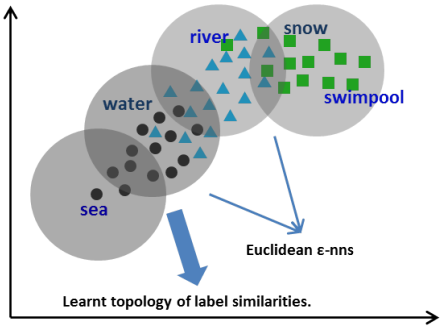


Figure 5.5: Illustration of topic manifold of {sea, river, snow, water, swimpool} in nnLDA.

shopwindow, doubledoor, gate, door, entrance
car, boat, van, truck, bus
sea, water, river, swimming pool, snow
desk, table, worktop, countertop, stove
stairs, steps, handrail, fence, bench
painting, picture, poster, bulletin board
person, person sitting, people
bottle, jar, vase, candle, person
building, wall, mirror, glasswall
bottles, building, merchandise, boxes, railing
book, sign, pack, text
plate, keyboard, book, paper

Figure 5.6: Top labels from some nnLDA topics.

5.3.1 Scene context in semantic space

Label cooccurrences through topics, but also a higher level semantic relationship between topics. Hence, we model the scene context through a probabilistic, directed acyclic graph (DAG) of topics known as Pachinko Allocation model(PAM) [74], which learns arbitrary-arity, nested, and possibly sparse topic correlations through *high level supertopics and low-level subtopics*.

Figure 5.4 depicts a subset of the topic graph inferred from groundtruth labels in SUN09 dataset [18]. The top labels in each subtopic are coherent and easy to interpret. In addition, natural images can consist of multiple subscenes, as captured by the supertopics. For example, a bookshelf can be found in isolation or in a living room (Fig. 5.4). The hierarchy in the topic structure of PAM is able to capture such complex interactions.

5.3.2 Appearance Context in visual space

Most state-of-the-art context models use visual words to represent context. Approximating a region's high dimensional feature vector using a scalar visual word leads to considerable information loss about the region's identity. In this paper, we harness the availability of labeled data to represent a region as a noisy realization of a union of object labels. Through this process, we also discover the visual contexts of object labels as topic manifolds in the visual space.

This idea is illustrated in Fig. 5.5 for the "sea" class. Image features depicting "sea" are physically close to instances of "water". They are also indirectly close to "swimming pool" instances, since both "sea" and "swimming-pool are close to "water". In general, the instances from "sea", "water" and "swimming-pool" all lie along the same manifold in the color space and hence should be captured in a single visual topic. Our proposed locality constrained, nearest neighbor Latent Dirichlet Allocation (nnLDA) model enables such visual topic discovery. Fig. 5.6 shows top labels from a few other topics found by nnLDA. Each element in the chain defines a distribution over words, and acts as the mean of the distribution over the subsequent element in the chain. Thus, each element in the chain can be thought of as introducing some additional corruption. All words are drawn from the final distribution in the chain.

5.3.3 Inferring context

So far, we have explained the motivation behind grouping object labels in the semantic and visual spaces. Our goal is to infer the most probable labels in a given image by analyzing the groupings in each space. Since labels are shared by both the semantic and visual contexts, a subset of highly probable labels in one space can inform the probability of label assignments in another space and vice versa. Based on this hypothesis, we adopt an iterative solution which alternates between the two spaces to *maximize the joint posterior probabilities of labels and contexts*. Specifically, we propose a solution based on Data Augmentation using collapsed

Gibbs sampling that allows us to solve an E-M like iterative optimization. The complete model and its mathematical formulation is described in the next section. Intuition of data imputation/posterior sampling...independently generate label distributions for each region using visual context. Use scene context to find compatible scenes and weight label probabilities. Use semantic label probabilities to modify

5.4 The Visual-Semantic Model

Fig. 5.8 summarizes our Visual-Semantic model. Its semantic component is a multi-level network of latent topics of a Pachinko Allocation Model (PAM) while its visual component is a locality constrained, nearest neighbor Latent Dirichlet Allocation (nnLDA). The labels generated from scene topics sample topic manifolds in the feature space to generate observed labels. During inference, the joint posterior over latent object labels and scene parameters are maximized using Data Augmentation algorithm.

5.4.1 Semantic Context: Generating object labels.

We model the co-occurrence context of object labels using a three-level Pachinko Allocation Model (PAM) [74]. Given an image corpus of size I, labels l in each image d are generated by topics at two levels. The per-image supertopic multinomial vector, θ_s is sampled from a symmetric Dirichlet hyperparameter α_0, while subtopic multinomial vectors θ_t are sampled for each supertopic from an asymmetric Dirichlet hyperparameter vector α_s. The role of α_s is crucial since it establishes sparse dependencies between super and subtopics. While subtopic correlations through supertopics lead to "sharing of statistical strength", the directed acyclic graph (DAG) structure allows each subtopic to be shared among multiple parent supertopics. The label mixing multinomials ϕ per subtopic are sampled corpus-wide from a symmetric Dirichlet hyperparameter β. Finally, each label l in the image is sampled from a topic path

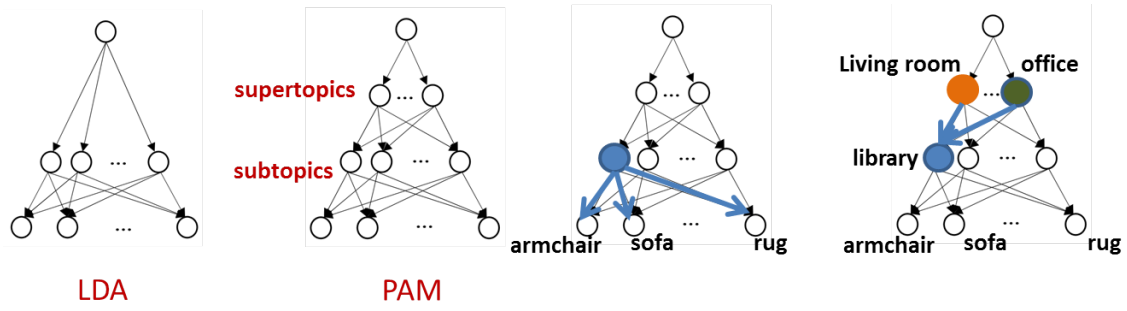


Figure 5.7: Illustration of PAM context graph. The directed acyclic graph allows a subtopic to spawn multiple objects. Each subtopic, on the other hand, can be dependent on multiple supertopics. This allows a flexible representation of inheritence and dependencies of scene properties, which is not possible in a tree structure (e.g., [18].

(z_s, z_t). We call these labels as the *semantic labels* because they are *generated* as opposed to the *observed*. The generative process is summarized as follows. The semantic model is a Pachinko Allocation Model in which the hierarchy of supertopics and subtopics captures various scene compositions using object labels.

- For each image $d = \{1 \cdots I\}$, sample a distribution θ_s over supertopics and a distribution θ_t over subtopics for each supertopic. $\theta_s \sim Dir(\alpha_0), \theta_t \sim Dir(\alpha_s)$.
- For each subtopic $k = \{1 \cdots T\}$, sample a distribution ϕ_k over labels. $\phi_k \sim Dir(\beta)$.
- For each label $lz = \{1 \cdots L\}$ -
  - Sample a topic path; a supertopic $z_s = \{1 \cdots S\}$ and a subtopic $z_t = \{1 \cdots T\}$. $z_s \sim Mult(\theta_s), z_t \sim Mult(\theta_{t,z_s})$.
  - Sample a label from subtopic, $lz = Mult(\phi_{z_t})$.

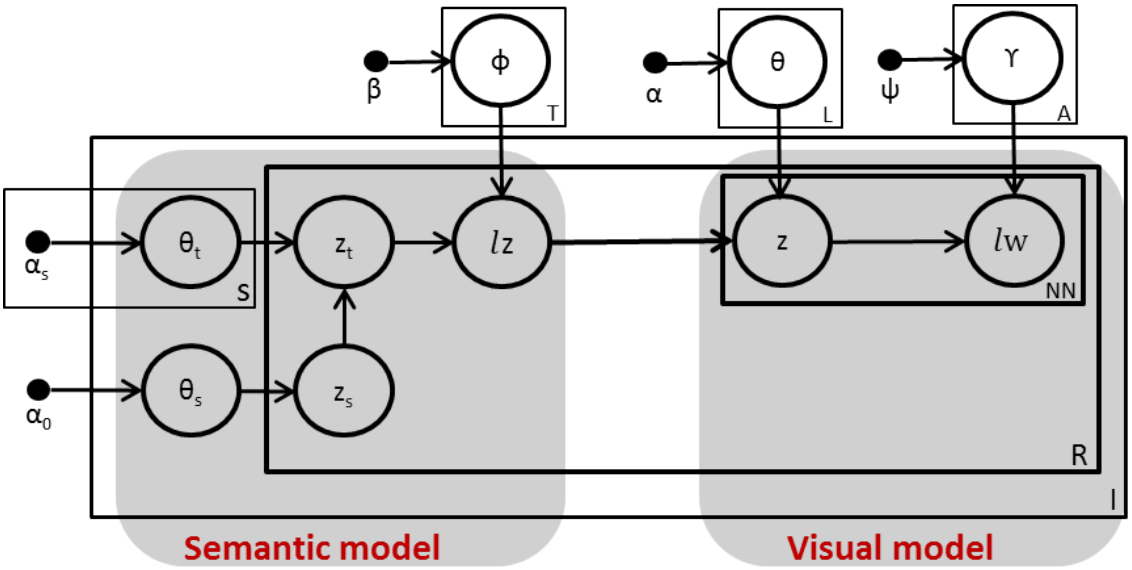


Figure 5.8: Visual-Semantic graphical model

5.4.2 Visual Context: Generating image regions

We define visual context as clusters of visually similar object-class labels and learn these clusters from instances of labeled image regions. Let $\{\vec{f_1}, \vec{f_2}, \cdots \vec{f_r}\}$ be the features of image regions r in some feature space (e.g., SIFT) with the corresponding semantic labels $\{lz_1, lz_2, \cdots lz_r\}$. Since the actual appearance interaction among object labels is unknown, we approximate it through local proximities among *label instances*. Specifically, we compute a ϵ-nearest neighbors graph on the scatter of label instances in each feature space and select the nearest labels of each instance, as follows.

$$\vec{lw_i} = \{l' \mid \forall j; \parallel f_i(lz_i) - f_j(l') \parallel \leq \epsilon\}, \tag{5.1}$$

where, $\parallel \ . \ \parallel$ is a distance norm in feature space s. Each label instance lz is associated with a set of observed labels $lw = \{l'\}$. This induces a many-to-many bipartite relation between the label instances and the semantic labels, similar to a document-term matrix, except that both

documents and terms are represented by labels. By imposing an LDA-based admixture, we consider each visual label to be a sample from a mixture model where the mixture components are multinomial random variables θ. In effect, we discover topics z that are low-rank, non-linear manifolds which capture maximally correlated (proximate) label distributions across multiple feature spaces. Specifically, the generative model is as follows

- For each label $lz = \{1 \cdots L\}$, sample a distribution θ_l over labels. $\theta_l \sim Dir(\alpha)$.
- For each visual topic $z = \{1 \cdots A\}$, sample a distribution γ_z over labels. $\gamma_z \sim Dir(\psi)$
- For each of the NN labels
  - Sample a visual topic z, conditioned on the multinomial parameter θ_l. $z \sim Mult(\theta_l)$
  - Sample a label $lw = Mult(\gamma_z)$.

5.4.3 Parameter Learning and Inference

The joint probability distribution of all the variables in the model is given by:

$$J.P.D. = Pr(\overbrace{zs, zt, lz, z}^{\text{hidden variables}}, \underbrace{lw}_{\text{observables}}, \overbrace{\theta s, \theta t, \phi, \theta, \gamma}^{\text{parameters}}) \tag{5.2}$$

$$= \overbrace{\{\prod_{I} p(\theta s|\alpha_0)(\prod_{S} p(\theta t|\alpha_s)) \prod_{R} p(zs|\theta s)p(zt|\theta s_{zs})p(lz|\phi_{zt}) \prod_{NN} p(z|\theta_{lz})p(lw|\gamma_z)\}}^{\text{image dependent samples}} \tag{5.3}$$

$$\cdot \overbrace{\{\prod_{T} p(\phi|\beta)\}\{\prod_{L} p(\theta|\alpha)\}\{\prod_{A} p(\gamma|\psi)\}}^{\text{corpus-wide samples}}$$

In this formulation, despite the seemingly daunting set of unknowns, the linear dependencies are exploited to simplify the inference into independent subproblems.

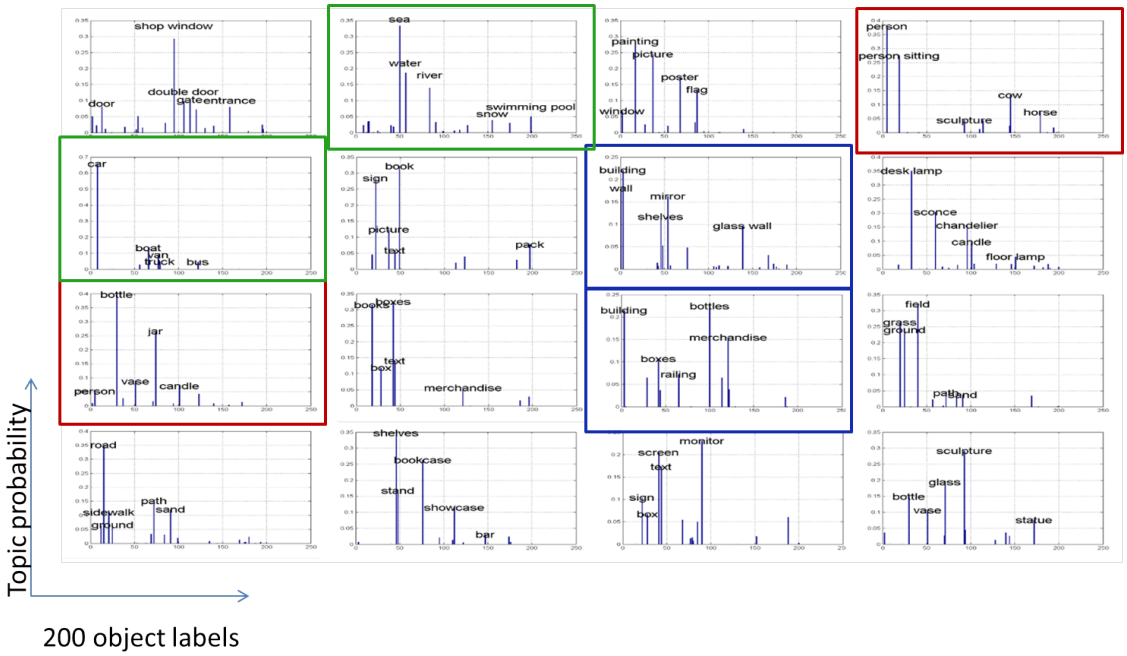


Figure 5.9: Visual polysemy and synonymy captured in topical clusters of nnLDA. As expected, "person" is correlated with "person-sitting", but is also correlated with "bottle" and "vase" (think of person in a wide angle view). Similarly, "building" can be modernistic e.g., similar to "glass" surfaces, or more traditional, similar to bookcases. Visual synonymy is captured in "car" and "truck" and "bus" and also in the cluster "sea", "water", "river".

For learning the parameters in the model, we use annotated images; images in which the all the objects are localized and labeled by human annotators. This gives us two pieces of information: 1) the collection of object labels in each image. This is used to learn the semantic model, and 2) the object locations in each image. These are used to extract local features related to the object appearances which informs the visual model. By knowing all the labels in the annotated image, lz becomes observed and the semantic model and visual model become conditionally independent, conditioned on lz. Therefore, parameter learning proceeds independently in the two models.

Parameter Learning in Semantic Model

Extracting the joint distribution of semantic model from the above J.P.D. formulation yields:

$$JPD_{semantic} = \{\prod_{I} p(\theta s|\alpha_0)(\prod_{S} p(\theta t|\alpha_s)) \prod_{R} p(zs|\theta s)p(zt|\theta s_{zs})p(lz|\phi_{zt}) \prod_{T} p(\phi|\beta)\} \tag{5.4}$$

The proposal distribution of supertopics and subtopics for the i^{th} entity is derived to be:

$$P(zs_i = s, zt_i = t|lz = l, zs_{\sim i}, zt_{\sim i}, \alpha_0, \alpha s_t, \beta) \propto \tag{5.5}$$
$$\left(\frac{n_s^d + \alpha_0}{\sum_S n_s^d + S\alpha_0}\right)\left(\frac{n_{st}^d + \alpha s_{st}}{\sum_T n_{st}^d + \sum_T \alpha s_{st}}\right)\left(\frac{n_l^t + \beta}{\sum_L n_l^t + L\beta}\right),$$

where zs_i and zt_i are the supertopic and subtopic assignments for lz_i, and $zs_{\sim i}$ and $zt_{\sim i}$ are the topic assignments of all the remaining regions in the image. Excluding the current token, n_s^d is the number of occurrences of topic s in image d and n_{st}^d is the number of times subtopic t is sampled from supertopic s within image d. n_l^t denotes the number of times the label l is assigned to subtopic t in the entire corpus.

Estimating scene Dirichlet hyperparameters. In addition to the topics per label, we also estimate αs within each Gibbs iteration of the PAM. This is because these hyperparameters capture different correlations among subtopics and can be thought to be the structural links that connect the supertopics and subtopics. Hence, the strength of the these connections need to be estimated in a data-driven manner. Several efficient techniques for Dirichlet hyperparameter optimization are described in [75]. We use the moment matching technique to estimate the approximate MLE of αs. In this technique, the model mean and variance of each αs_{st} is computed by matching them to the sample mean and variance, which are estimated from the super and subtopic samples.

$$E[\alpha s_{st}] = \frac{\alpha s_{st}}{\sum_{T} \alpha s_{st}} = \frac{\alpha s_{st}}{exp(log \sum_{T} \alpha s_{st})} \tag{5.6}$$
$$= \frac{1}{N} \sum_{I} \frac{n^d_{st}}{\sum_{T} n^d_{st}}$$
$$var[\alpha s_{st}] = \frac{1}{N} \sum_{I} (\frac{n^d_{st}}{\sum_{T} n^d_{st}} - E[\alpha s_{st}])^2$$
$$log \sum_{T} \alpha s_k = \frac{1}{T-1} \sum_{T-1} log \left(\frac{E[\alpha s_{st}](1 - E[\alpha s_{st}])}{var[\alpha s_{st}]} - 1 \right)$$

Parameter Learning in Visual Model

Starting with the joint distribution of visual model,

$$JPD_{visual} = \prod_{NN} p(z|\theta_{lz}) p(lw|\gamma_z) \{ \prod_{L} p(\theta_l|\alpha) \} \{ \prod_{A} p(\gamma_a|\psi) \} \tag{5.7}$$

The proposal distribution for visual topic of the i^{th} entity is derived to be:

$$P(z_i = a|lw, lz, z_{\sim i}, \alpha, \psi) \propto \tag{5.8}$$
$$\left(\frac{n^{lz}_a + \alpha}{\sum_{A} n^{lz}_a + A\alpha} \right) \left(\frac{n^a_{lw} + \beta}{\sum_{L} n^a_{lw} + L\beta} \right),$$

z_i corresponds to the visual topic of the i^{th} semantic label, n^{lz}_a is the number of times topic a in sampled for semantic label lz and n^a_{lw} denotes the number of times a visual label lw is assigned to topic a across the entire corpus. It is interesting to understand the role of the counts that relate a topic to the labels. For a pair of labels (lz, lw) and a topic a, 1) Both, semantic

count n_a^{lz} and visual count n_{lw}^a are low. Sampled topics would be randomly generated for this label pair. 2) Semantic count n_a^{lz} is high but visual count is low n_{lw}^a. The visual label is an outlier, i.e., whose appearance randomly matches a source feature. 3) Semantic count n_a^{lz} is low but visual count is high n_{lw}^a. The visual label has a generic appearance that is loosely associated with many different objects, e.g., wall, sky etc. 4) Semantic count and visual count are both high. There is strong correlation between the appearance features of these two labels. It is easy to see that topic sampling will favor the last scenario and consistently generate topic a which will strongly link a pair of visually correlated labels (lz, lw).

We use symmetric values of hyperparameters in the visual space which are typically set as $\alpha = 1$ and $\psi = 0.01$ for sparse topic discovery. The multinomials $\tilde{\theta}$ and $\tilde{\gamma}$ are calculated as the expectation of their corresponding Dirichlet distributions.

Inference

Given an image, the central challenge in using VSL model is to compute posterior probabilities over the semantic labels lz for each image region, conditioned on the observed visual labels lw.

$$P(lz|lw) = \sum_{zs}\sum_{zt}\sum_{z} P(lz, zs, zt, z|lw) = \sum_{zs}\sum_{zt}\sum_{z} P(lz|zs, zt, z, lw)P(zs, zt, z|lw) \tag{5.9}$$

$P(zs, zt, z|lw)$ gives the predictive likelihood of latent data given observations and $P(lz|zs, zt, z, lw)$ denotes the conditional probability of lz given augmented data (zs, zt, z, lw). Using conditional independence, $P(lz|lw)$ is simplified, as above. The second term consists of predictive distribution over semantic topics, which is computed by marginalizing over lz.

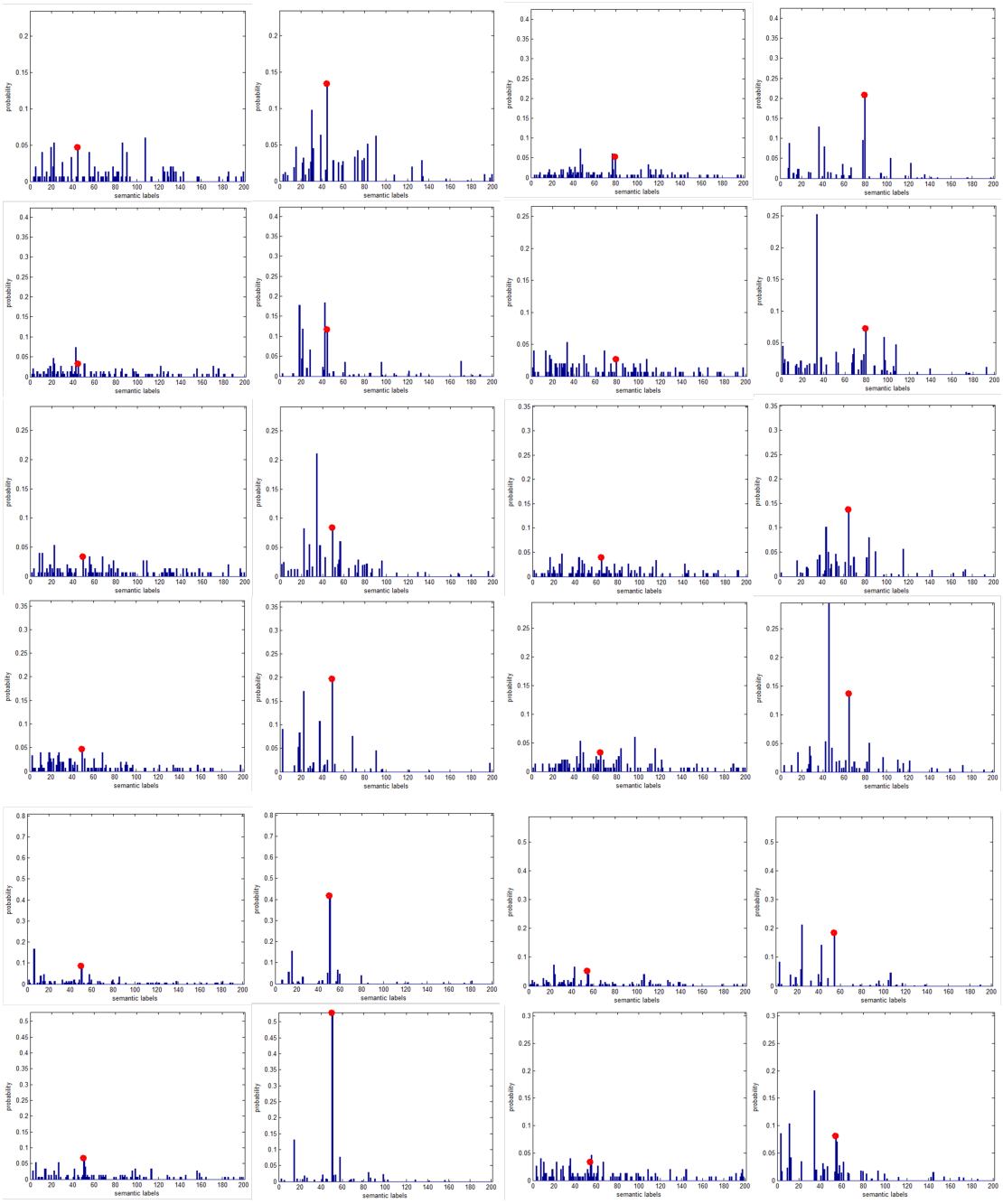


Figure 5.10: k-nearest neighbor labels(left) versus final semantic label distributions (right)

$$P(lz|lw) = \sum_{zs} \sum_{zt} \sum_{z} P(lz|zs, zt, z) P(zs, zt, z|lw) \tag{5.10}$$

$$P(zs, zt, z|lw) = \sum_{lz} P(zs, zt|lz) P(z|lw) P(lz|lw) \tag{5.11}$$

The above formulation leads to a couple inference problem. For solving $P(lz|lw)$, we need the predictive topic probabilities $P(zs, zt, z|lw)$. However, the link between the semantic topics and the observed lw passes through lz, which needs to be marginalized out. This creates dependence on $P(lz|lw)$, which we seek to solve in the first place.

We propose Data Augmentation algorithm to solve this inference problem. The idea of DA refers to a scheme of augmenting the observed data so as to make it more easy to analyze. It is equivalent to the Expectation Maximization approach, but applies to posterior sampling. DA was introduced in the statistics literature by Tanner and Wong [102]. The general framework consists of an iterative sampling framework with two steps 1) Data imputation step, in which the current guess of the posterior distribution $p(lz|lw)$ is used to generate multiple samples of the hidden variables (zs, zt, z) from the predictive distribution in Eq. x2, and 2) Posterior sampling step, in which the posterior is updated to be the mixture of the m augmented posteriors and approximated to be the average of $p(lz|lw, zs, zt, z)$ over the imputed z's. Thus, by successive substitution a stationary distribution over the posterior is achievable. Formally, the two steps can be represented as:

Data Imputation

$$\{zs^{(t+1)}, zt^{(t+1)}, z^{(t+1)}\} \sim \sum_{lz} P(zs, zt|lz) P(z|lw, lz) P^{(t)}(lz|lw) \tag{5.12}$$

We initialize the algorithm by performing nnLDA inference on the visual labels obtained from each image region. Based on the parameters of the nnLDA, visual topics are sampled and then predictive likelihood of lz is computed for each region.

$$p(\tilde{a}|lw, \tilde{a}_{\sim i}; \mathcal{M}) = \left(\frac{\tilde{n}^d_{a,\sim i} + \alpha}{\sum_A \tilde{n}^d_a + A\alpha} \right) \left(\frac{n^a_{lw} + \tilde{n}^a_{lw,\sim i} + \beta}{\sum_L n^a_{lw} + + \tilde{n}^a_{lw,\sim i} + L\beta} \right), \tag{5.13}$$

Next, the topic distributions for the region are computed as:

$$\tilde{\theta}^r_a = \frac{\tilde{n}^r_a + \alpha_a}{\sum_A \tilde{n}^r_a + \alpha_a} \tag{5.14}$$

Using the estimated topic proportions vector, we compute the likelihood of lz:

$$\begin{aligned} P(lz|lw) &= \sum_A P(lz = l|z = a)P(z = a|\vec{lw}) = \sum_A \frac{P(z = a|lz = l)P(lz = l)}{P(z = a)}(z = a|\vec{lw}) \\ &= \theta^l_a \frac{n_l}{n_a} \tilde{\theta}^r_a, \end{aligned} \tag{5.15}$$

where n_l is the prior count of semantic label $lz = l$ and n_a is the count of corpus wide count of topic a. Using this multinomial distribution, we generate samples of lz, which are used as observations for the semantic model. Specifically, using each $lz = \{lz_{R_1}, lz_{R_2}, ...\}$, we sample (zs, zt) according to Eq. 6.5. Thus, for each lz, we generate a complete set of (zs, zt, z). We can also compute the semantic topic proportions $\{\tilde{\theta}s, \tilde{\theta}t\}$ from the supertopic and subtopic samples.

Posterior Sampling

$$P^{(t+1)}(lz|lw) \sim \frac{1}{nsamples} \sum_{nsamples} P(lz|zs^{(t+1)}, zt^{(t+1)}, \mathcal{M}) \cdot P^(t)(lw) \tag{5.16}$$

Based on the estimated scene parameters $\{\tilde{\theta}s, \tilde{\theta}t\}$ and the model estimates of ϕ, we compute the semantic label distribution for the image. This can be computed simply by following the generative process of the model. This signal is used to modulate the visual label distribution at each image region. The average over multiple samples is used to update the $P(lz|lw)$ distribution.

5.5 Experiments

5.5.1 Dataset and Experiment Settings

We evaluate our proposed Visual-Semantic approach on the SUN09 database [18]. This dataset is specifically designed to leverage contextual information as it exists in the real world. It is a large collection of complex images capturing both indoor and outdoor environments. Each image consists of an average of 7 different annotated objects and the average occupancy of each object in the image is 5%. As in the real-world, the frequencies of observation of each object also varies. Some objects occur very frequently in images, such as wall, sky, building, etc. while other objects such as microwave, armchair, palm-tree etc. are much less frequent. The definition of an "object" encompasses *things* and *stuff*, as described in [44]. *Things* are well-defined and well-localized objects such as computer screen and showcase, while *stuff* are extended regions such as sky or text.

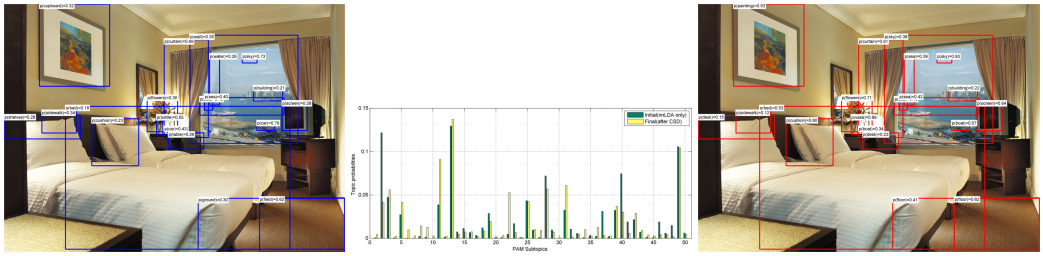
The dataset consists of 576 annotated object categories. We test our algorithm on 200 objects with 30 or more instances in the training set. There are 12,000 annotated images out of which 4367 images are marked as training images. We use these for learning our model. The

remaining 4317 images are in the testing set. Test images also come with annotations, hence our inferred scores are compared to the ground-truth for evaluations.

For learning the model, we use the annotated ground-truth locations and labels provided with the dataset. In test images, we use all the bounding boxes provided by the baseline detector [18] as image regions (a 256×256 image has > 400 regions). We hypothesize these blocks as object-like regions that should be analyzed for potential objects. Other options for object-like regions include objectness measure [1] or other image segmentation e.g., uniform grid.

Feature Representation

Each image region is represented using three types of features, similar to the work in [70, 46]. To capture color, we separate the R, G, B channels and bin their values uniformly between $[0, 255]$ into 11 bins. We generate a histogram of each channel and concatenate the R, G, B histograms. We also append the corresponding means and variances of each color channel, to generate a 36 length vector. To capture texture, we use the filter bank method proposed in [69]. The filter bank is a set of 40 filters. Each image is convolved with each filter to generate filter responses. Through k-means clustering, exemplar filter responses are precomputed into a texton codebook. We use a codebook of 100 textons. Each filter response is assigned to the closest texton in the codebook to generate a texton-word. A histogram of 100 texton-words is generated for each image region. Additionally, to capture transformation-invariant edge features, we densely sample the image and compute SIFT [66] descriptors for each local patch. These descriptors are vector quantized using a precomputed SIFT codebook. The features within each region are histogrammed based on a precomputed SIFT codebook of 400 SIFT words.



(a) Initial nnLDA detections. (b) Semantic subtopic distribution. (c) Final detections and associated probabilities.

Figure 5.11: A room with a view: Detecting *both* the room and the outside view. Zoom in for detailed view of object probabilities and topic distributions in initial and final rounds of VSL inference. Find details in text below.

Model representation

The semantic model which is represented by a Pachinko Allocation Model is learnt with 20 supertopics and 50 subtopics. The supertopic Dirichlet α_0 is set to a uniform value of 1. The subtopic Dirichlet αs is learnt by the model estimator. The visual model is represented by a nnLDA model. We choose a neighborhood of $NN = 50$ samples to generated each set of visual labels. 50 topics are learnt in the model. During parameter estimation, the Gibbs sampling is run for 1000 iterations in each model. For posterior inference of topics, we use 100 iterations. From multinomial distributions we generate 500 samples which are imputed and averaged to compute the expected value of a distribution. The data augmentation framework, consisting of data imputation and posterior sampling phases is run for 6 iterations.

Qualitative explanation

Before we proceed with the quantitative evaluation, we briefly provide here an intuitive illustration of our system on the **room with a view** example (Figure 5.11). The image shows a

”complex” scene, in that two separate contexts are coexisting in the same scene - the indoor bedroom scene and the outdoor view. To present an insight into our algorithm, we focus on a subset of regions extracted from the image. In the left image, we show the initial label inference with visual context model only. On the right, we show the final label inference on the same regions after applying our VSL model.

On initialization using the visual model, we obtain a *coarse* inference of the image in which a set of generic objects and the related scene categories are recognized. On the left, we overlay the image with the maximum aposteriori semantic labels and the corresponding probabilities of some of the objects in the scene. In the absence of a semantic context, most of the detected objects are unrelated. Next, using multiple samples of semantic labels from the scene we infer the semantic model. The subtopic distribution of the semantic model is plotted in green in the middle panel. The plot heights indicate the dominance of the certain subtopics, for example the the 13^{th} and 49^{th} topics are the top two semantic subtopics. Next, the inferred semantic parameters are used to sample the semantic label distribution of the scene. These are used as priors for re-adjust the visual label distribution for each region. Post normalization, we obtain a posterior distributions over visual labels that are modulated by the semantic context. This completes one iteration of the data augmentation procedure. After 6 iterations, the maximum aposteriori semantic labels for the regions are shown in the right panel.

Some MAP labels remain unchanged and interestingly, show an increase in beliefs, such as ”bed”, ”cushion”, and ”curtain”. Some MAP labels are driven to more *specific* labels e.g., ”water” is relabeled as ”sea”. Also, “car” is changed to “boat”, which could be because “boat” is visually similar to “car” but fits better with “sea”. These alternate, more specific labels are available from local appearances but are emphasized by the overall boat-sea semantic context. The center plot shows the change in the semantic pattern, where yellow plots show the final

subtopic distribution. For example, subtopic 11 and 22 both of which are related to sea semantics show significant increase. This also leads to correction of labels related to the *dominant* scene, e.g., "cupboard" is relabeled as "painting". Finally, incorrectly detected regions without semantic support end up with very low probabilities in final inference (e.g.,sidewalk) as do the weights of the initially detected, incorrect semantic (e.g., subtopic 2, related to sky, mountain, snow).

Experiment design

To evaluate our algorithm, we test the performance on the following tasks. These tasks reveal the performance at different stages of the algorithm as compared to the overall system as well as to the baseline models.

- Semantic label prediction with visual context model.
- Semantic scene prediction of semantic model.
- Prediction of most confident objects in an image.
- Object recognition.

As baseline, we consider the following state-of-the-art models:

- Felzenswalb's object detector. This detector achieves state of the art results on the several datasets. The detection results of this detector are available as part of the SUN09 dataset.
- hcontext object detector: This detector models object co-occurrences and spatial relationships using a tree graphical model. It is used on top of Felzenswalb's object detector to prune away non-contextual detections. The detection results are available as part of the SUN09 dataset.

- Other visual-textual models: We compare our approach to other generative approaches that model images and text jointly. We have implemented and run these models using the same features as used for our approach.

  – Correspondence Latent Dirichlet Allocation [9]: This model is an extension of the LDA model to include textual words. According to the model, the visual words, w and textual words, l can be determined by the same topic indicators, z generated from the parameter that captures the semantics of the scene θ.

  – Total Scene understanding model [61]: The original model captures words and captions in an image. Since our images do not come with captions, we only consider the visual part of the model for comparison. According to the model, the scene category c determines the textual labels l which, in turn, determines the visual words in the image.

5.5.2 Quantitative Evaluations

Semantic label prediction with visual context model: We evaluate the improvement in using nnLDA compared to the ϵ-nearest neighbors approach. We use average precision gain (AP gain) in label retrieval as the metric for comparison (Table 5.1). The mean AP gain across 200 object categories is approx. 4.0%, and 168 classes show an overall positive gain. It implies that sharing appearance information through visual context improves object recognition performance in general. It is interesting to note that the categories with maximum gain are all *rare categories* (categories with few training examples in the dataset) that are visually similar to frequent categories e.g., pillow, text. In contrast, the objects with maximum detection loss are categories with distinct appearances e.g., shoes, ingots. This is an expected behavior since the discriminative appearance of these objects is diluted by being fitted into topic groups, as also observed in [92].

Object classes	*Highest AP gain*	*Least AP gain*
	pillow ($+29.47$)	shoes (-40.95)
	text ($+15.37$)	ingots (-13.19)
	desk ($+14.83$)	fish (-8.80)
	armchair ($+12.55$)	chandelier (-7.30)
	flowers ($+12.45$)	monitor (-6.30)
	cabinet ($+12.01$)	glass (-6.27)
	fence ($+11.19$)	faucet (-6.06)
Mean Average Precision improvement = $+4.19\%$		

Table 5.1: Average precision gains with nnLDA compared to ϵ-NN.

TSU	CorrLDA	Initial VSL	Final VSL
44.03	33.42	19.57	13.21

Table 5.2: Kullback Leibler Divergence between estimated and groundtruth scene parameters. The least distance measure of the final VSL model implies closer fitting of the model with the groundtruth.

Semantic scene prediction of VSL model: We compare the scene prediction performance of CorrLDA and TSU vis-a-vis VSL model. In CorrLDA, both visual words and textual words in an image depend on the same topic indicator. Hence, for the model to fit, textual labels and visual features must display similar clustering ambiguities. Total Scene Understanding model assumes (a) a single semantic topic for generating textual labels and (b) one-to-one correspondence between textual labels and visual words. We implement and test the supervised versions of both these models vis-a-vis our VSL model. Each model is trained using the ground-truth regions and labels in the SUN09 dataset. Given a new image, we predict the semantic

TSU	CorrLDA	Initial VSL	Final VSL
0.29	0.36	0.63	0.87

Table 5.3: Average accuracy of prediction of most confident label.

context using local bounding boxes as regions. In CorrLDA, the semantic context is equivalent to the posterior estimate of parameter θ, in TSU it is the estimate of c and in the VSL model it is the estimate of θt. The ground-truth semantic context is obtained by fitting the ground-truth labels of regions to the models. We compare the estimated and groundtruth distributions using the symmetric Kullback-Leibler Divergence measure for each of the parameters.

$$D_{KL}(\theta_{est}||\theta_{gt}) = \sum_{K} Pr(\theta_{estk})(log_2 Pr(\theta_{estk}) - log_2 Pr(\theta_{gt_k})) \tag{5.17}$$

$$D_{KLSymm} = \frac{1}{2}(D_{KL}(\theta_{est}||\theta_{gt}) + D_{KL}(\theta_{gt}||\theta_{est})) \tag{5.18}$$

Prediction of most confident objects in an image: The most confident labels predicted in an image can be used as descriptive keywords. Figure 5.13 shows the results of top five label predictions. The accuracy measures the number of times the top N predicted labels actually exist in the image, where N is varied from one to five labels. Overall, our average performance improves over the state-of-the-art hcontext model. Interestingly, the comparative performance gain increases with every new label. We attribute this to (a) our method's ability to detect rare classes which also provide strong cues for detecting general classes. (b) we model visual context of regions that overlap multiple textures as mixtures of multiple labels. Qualitative results are shown in Figure 5.16. We see that both frequent and rare object categories can emerge as top labels if they fit both the contextual models well, e.g., wall (frequent label) in (b) and showcase (rare label) in (g). Table x shows the results of the other baseline generative models in predicting the top label in an image.

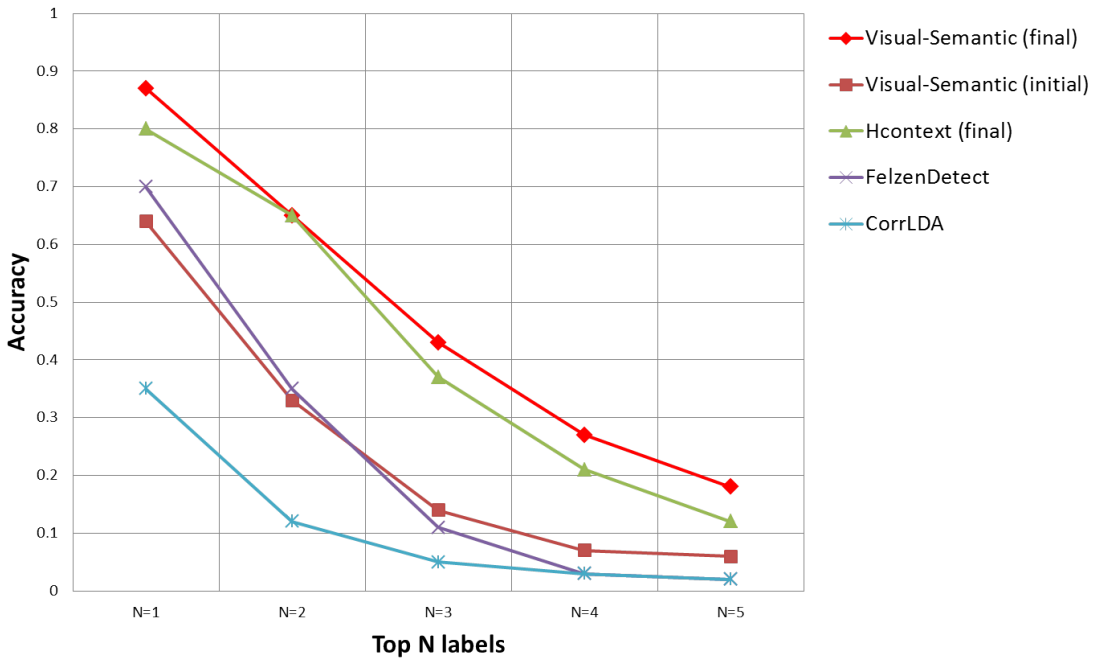


Figure 5.12: Accuracy of prediction of top N labels using VisualSemantic Model on 200 categories versus hcontext on 107 categories.

Object recognition: An efficient object retrieval system should be highly accurate on all types of objects, not just the frequent ones. As shown in Table 5.4, our method successfully recognizes objects from a large variety of categories. We compare our recognition rates to the hcontext [18] based on precision values at 0.25 False Positives Per Image(FPPI). Even though we consider a *larger number of object categories* (200 vs. 107 in hcontext), our recognition rates outperform the hcontext model. To provide more insight into our model's performance, we consider a few specific cases in Table 5.4. In (a), we consider categories with high intraclass variability. In (b), we consider rare categories that gain by sharing appearance information with dominant categories. In (c) we consider categories that occur multiple times within a single image (*self-context*). (d) shows some more examples of rare categories and finally, in (e), we consider categories that do not benefit much from our context model, either because they

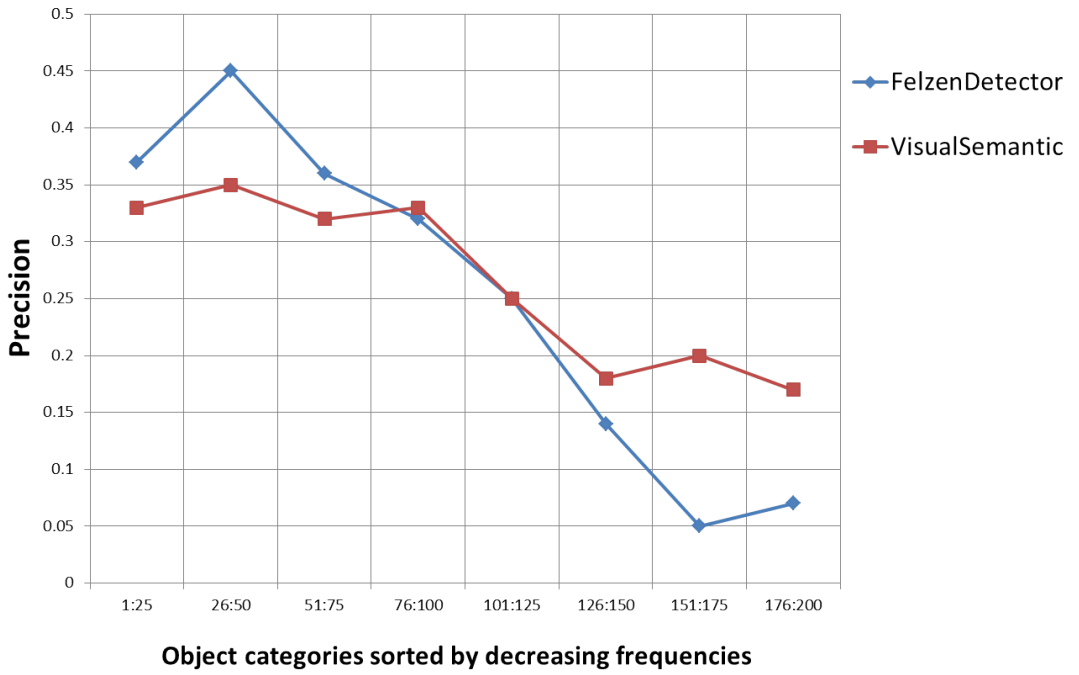
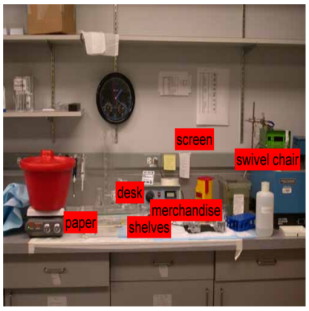


Figure 5.13: Precision of object detections N labels using VisualSemantic Model on 200 categories versus hcontext on 107 categories.

are already well-detected by their appearance or because they remain undetected, even after combining the two contexts.

Qualitative results of object detections based on final label posteriors are shown in Figure 5.16. We see that small objects occupying an insignificant area in the image is detected strongly through by the VSL model e.g., stand in (h) and paper in (a). However, in some cases the iterative method might overfit by linking categories e.g., boat and sea in (d). Another way to interpret this behavior is that the model parallely generates strong responses for multiple scene concepts e.g., warehouse with merchandise as well as a study room in (a).



(a) **cupboard,floor,desk**

table,desk,building

(b) **curtain,wall,seats**

window,building,wall

(c) **car,cars,building**

wall,building,floor

(d) **sky,tree,person**

table,truck,tree

(e) **curtain,shelf,d.lamp**

stool,floor,wall

(f) **c.lamp,chair,floor**

table,chair,floor

(g) **bread,tray,showcs**

sky,bed,countertop

(h) **c.lamp,bkcase,bks**

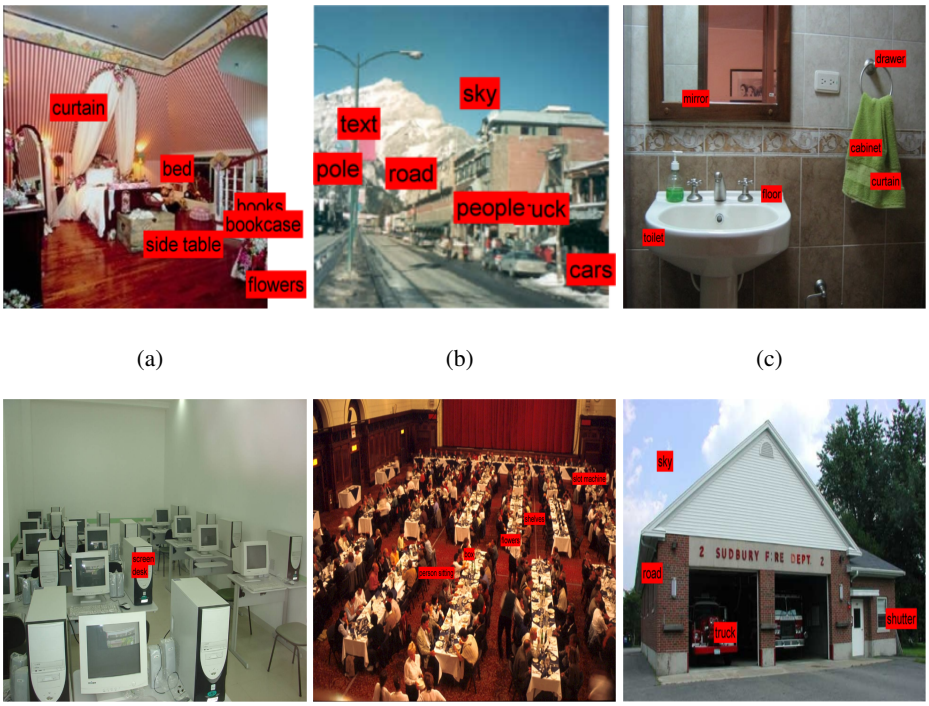
floor,bldg,wall

(i) **column,floor,wall**

floor,wall,road

Figure 5.14: Marked objects (at average window locations) are detections based on thresholded label posteriors. Red captions: Visual Semantic label predictions, blue: top detections by the object detectors used in [18]

.



(a) (b) (c)

(d) (e) (f)

Figure 5.15: Object detections using Composite Scene Detection model (CSD)

(a) **aircraftcarrier, sea** **boat,** **sea** mountain, sea railing, shopwindow

(b) **ceiling, closet** **ceiling,** **c.lamp** ceiling, c.lamp tv,furniture

(c) **blackboard, ceiling** **c.lamp, chair** c.lamp, ceiling bicycle, people

(d) **bed, curtain** **bed, c.lamp** c.lamp, ceiling steps, railing

(e) **book, bookcase** **c.lamp,bookcase** pot, plant bottle, bottles

(f) **building, grass** **building,** **grass** painting, d.lamp tree, pot

Figure 5.16: Black: Groundtruth labels, Red: VSL top 2 labels, Blue: TSU labels, Green: CorrLDA labels.

Precision at .25 FPPI(VSLfinal, VSLInitial, Hcontext [18])	
a)	people $(\mathbf{0.74}, 0.0, -)$, cars $(\mathbf{0.68}, 0.42, -)$ food $(\mathbf{0.63}, 0.0, -)$, picture $(\mathbf{0.71}, 0.25, \mathit{0.76})$
b)	boat $(\mathbf{0.79}, 0.17, -)$, truck $(.\mathbf{82}, 0.3, \mathit{0.85})$ painting $(\mathbf{0.62}, 0.23, -)$, poster $(\mathbf{0.55}, 0.0, \mathit{0.57})$ shop window $(\mathbf{0.73}, 0.11, -)$, balcony $(\mathbf{0.91}, 0.64, \mathit{0.80})$
c)	videos $(\mathbf{0.93}, 0.24, \mathit{1.0})$, bottles $(\mathbf{0.72}, 0.34, \mathit{0.60})$ books $(\mathbf{0.72}, 0.28, \mathit{0.80})$, merchandise $(\mathbf{0.93}, 0.68, -)$
d)	cow $(\mathbf{0.93}, 0.0, -)$, fish $(\mathbf{0.87}, 0.45, -)$ deck chair $(\mathbf{0.67}, 0, -)$, umbrella $(\mathbf{0.59}, 0.07, \mathit{0.57})$
(CSDfinal, gain over CSDInitial)	
e)	flowers $(0.68, +0.08)$, stand $(0.47, -0.03)$ sea $(0.84, +0.07)$, field $(0.66, -0.02)$ microwave $(0.03, +0.03)$, bathtub $(0.04, +0.04)$ clock $(0.06, +0.01)$, outlet $(0.08, +0.07)$

Table 5.4: Object recognition results

5.6 Conclusion

This paper presents an significant advancement over prior work for scene understanding by integrating semantic and visual contexts within a single algorithmic framework. Our Visual Semantic approach detects subscenes within scenes along with both dominant and subordinate objects in images and is a promising approach for other applications. In the future, we hope to extend our method in a semi-supervised setting where new objects and scenes can learnt by transferring knowledge through topic associations in multiple contexts.

Chapter 6

Conclusions

Images and language are inherently ambiguous and contain multiple meanings. There is evidence that humans make sense of the world by combining bottom-up and top-down influences of information flow and connecting together contextually relevant meanings that create a coherent whole. If computers are to match human abilities in this regard, automatic scene understanding methods should not be limited to one-shot, bottom-up analysis. In this dissertation, we have posed scene understanding as an coupled inference problem, in which bottom-up visual context interacts with top-down semantic context to disambiguate identities of image regions.

We presented two Bayesian models that stratify scene generation into a hierarchy of semantic and visual context models for objects in scenes. First, we proposed ViewLDA for object and background separation with minimal supervision and without negative exemplars. It is used to detect objects in images as a two-level hierarchy. In the first level, an object is parameterized as a set of distributions of local appearance features that capture the generic appearance of the object category using and LDA model. This causes each image region to get assigned to a probability of being an object part. At the second level, the object is jointly parameterized with its corresponding background using the object priors from the LDA step and the surrounding regions within the image. This novel ViewLDA model enables a more fine grained, view-specific separation of object and background. We have shown results on a challenging dataset of multi-view car images where our results surpass the state of the art classification when the domain of training and testing images are varied.

Next, we present Visual Semantic Integration Model that predicts complex scene descriptions and infers objects with limited training data. VSIM models scene interpretation as a top-down approach where semantically contextual labels are first created to represent a coherent scene composition. These labels are then re-interpreted with their visually contextual counterparts in the appearance space. Specifically, VSIM is a probabilistic, hierarchical model of latent context and observed features. In the first level, the image is modeled as a distribution over latent semantic contexts which determines the semantic labels that compose the scene. In the next level, each semantic labels visual context determines the appearance features that are finally the observed variables in the model. Inference in VSIM is initiated in a bottom-up manner, where observed image regions are the only cues used to infer the semantic and visual object labels in the image. I.e., the goal of VSIM is to infer the semantic object labels in an image, given its appearance features. We derive an iterative Data Augmentation algorithm that pools the label probabilities and maximizes the joint label posterior of an image. Not only does our method perform favorably as compared to the state-of-the-art methods in several visual recognition tasks, but also surpasses them in hard to tackle objects. Below we outline the main advantages of our framework:

1. **Multiple context approach**: Unlike previous context models that capture either semantic or visual context information, our model captures both modalities within a single framework. Our results show that both object recognition and scene prediction was improved by combining contexts to disambiguate object identity. These findings are also corroborated by experiments on human cognition that show that the information processed by the context network involves both spatial and visual associations [29].

2. **Joint, multiple hypotheses approach**: Previous context models are independent, post-processing step to object recognition. They are mostly applied as filters to remove incompatible object detections in the scene. In contrast, our method is an coupled inference problem, in which bottom-up visual context interacts with top-down semantic context. The two contextual levels activate multiple, complementary hypotheses which are exploited iteratively to improve the quality of the inferred labels.

3. **Handles missing and impoverished data**: Our models are inherently generative due to which they perform reasonably well even when the number of training examples are limited. This advantage on impoverished data is due to a richer set of constraints that prevents overfitting in our model. For example, we show that VSIM better handles the data-imbalance problem frequently seen in many learning problems with natural object categories which follow a power law distribution [18]. VSIM also includes a hybrid learning approach, since a nearest neighbor classifier is used to represent visual context that finds discriminative visual hypotheses. nnLDA exploits the strength of nearest neighbor decisions within a structured generative LDA approach. Mixing unsupervised learning with supervised classification improves generalization and adapts to dataset biases.

4. **Extendable and handle multiple modalities of information**: Our models can handle text and images together via a single generative framework leading to improved inference. One advantage of our iterative learning approach is that we can easily add more knowledge into our framework by augmenting it with new images or labels, without having to relearn any individual component separately.

Our current model is based on prior knowledge about the object categories we want to learn and assume that we have images that contain those categories. One future direction we want to explore is how to learn unlabeled concepts in the wild. With multiple sources of web

information such as Flickr and Wikipedia for images and text available, our method could be a powerful technique for finding meaning visual and semantic associations between data. An important question that we are currently pursuing is whether relations between concepts should be entirely data-driven and probabilistic or if there is a possibility for human guided constrained learning of correlations. Our preliminary studies show that a combination of probabilistic and logical approaches are useful in certain types of environment such as human interactions in constrained situations e.g., hospital rooms and sports events [16, 78]. In future, we want to develop our method to investigate large scale visual analysis using diverse logical, semantic and human guided contexts.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 34(11):2189–2202, November 2012.

[2] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2011.

[3] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. Blei, and M. Jordan. Matching words and pictures. 3, 2003.

[4] K. Barnard and M. Johnson. Word sense disambiguation with pictures. *Artificial Intelligence*, 167(1):13–30, 2005.

[5] R. Bekkerman and J. Jeon. Multi-modal clustering for multimedia collections. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[6] I. Biederman. Recognition by components: A theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.

[7] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.

[8] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Proceedings of Advances in Neural Information Processing (NIPS)*, 2004.

[9] D. Blei and M. Jordan. Modeling annotated data. 2003.

[10] D. Blei and J. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

[11] D. M. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[12] M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, page II: 628, 1998.

[13] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In *Proceedings of the ACM conference on Information and knowledge management*, 2008.

[14] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV04*, pages Vol I: 350–362, 2004.

[15] I. Chakraborty, H. Cheng, and O. Javed. 3d visual proxemics: Recognizing human interactions in 3d from a single image. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[16] I. Chakraborty, A. Elgammal, and R. Burd. Video based activity recognition in trauma resuscitation. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FGR)*, 2013.

[17] Y. Chen, O. Jin, G. Xue, J. Chen, and Q. Yang. Visual contextual advertising: Bringing textual advertisements to images. 2010.

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

[19] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[20] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 1990.

[21] T. Deselaers and V. Ferrari. Visual and semantic similarity in imagenet. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1777–1784, 2011.

[22] J. J. DiCarlo, D. Zoccolan, and N. Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[23] M. Dyer. Connectionist natural language processing: A status report. *Computational Architectures Integrating Neural And Symbolic Processes*, 1994.

[24] S. Eggers and S. Moran. Processing ambiguous words: Multi-sense activation, 2012.

[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[26] A. Farhadi, S. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29, 2010.

[27] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 32(9):1627–1645, September 2010.

[28] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)*, 61(1):55–79, January 2005.

[29] M. Fenske, E. Aminoff, N. Gronau, and M. Bar. Top-down facilitation of visual object recognition: object-based and context-based contributions. *Progress in Brain Research*, 155:3–21, 2006.

[30] M. Fenske, E. Aminoff, N. Gronau, and M. Bar. Top-down facilitation of visual object recognition: object-based and context-based contributions. *Progress in Brain Research*, 155:3–21, 2006.

[31] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *International Journal of Computer Vision (IJCV)*, 71(3):273–303, March 2007.

[32] K. Fragkiadaki and J. Shi. Figure-ground image segmentation helps weakly-supervised learning of objects. In *ECCV*, 2010.

[33] K. Fragkiadaki and J. Shi. Figure-ground image segmentation helps weakly-supervised learning of objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages VI: 561–574, 2010.

[34] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *CVIU*, 114(6):712–722, June 2010.

[35] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *Compter Vision and Image Understanding (CVIU)*, 114(6):712–722, June 2010.

[36] T. Gao and K. D. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *ICCV*, 2011.

[37] A. Geiger, M. Lauer, and R. Urtasun. A generative model for 3d urban scene understanding from movable platforms. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[38] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[39] A. Gupta and L. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, pages I: 16–29, 2008.

[40] A. Gupta, J. Shi, and L. Davis. A 'shape aware' model for semi-supervised learning of objects and its context. In *Proceedings of Advances in Neural Information Processing (NIPS)*, 2008.

[41] X. He, R. Zemel, and M. Carreira Perpinan. Multiscale conditional random fields for image labeling. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages II: 695–702, 2004.

[42] G. Heinrich. Parameter estimation for text analysis. *Web: http://www. arbylon. net/publications/text-est. pdf*, 2005.

[43] G. Heinrich. Parameter estimation for text analysis. 2009. Fraunhofer IGD.

[44] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages I: 30–43, 2008.

[45] T. Hoffman. Probabilistic latent semantic analysis. In *UAI*, 1999.

[46] D. Hoiem, A. Efros, and M. Hebert. Geometric context from a single image. pages I: 654–661, 2005.

[47] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008.

[48] Y. Hu, J. Boyd-Graber, and B. Satinof. Interactive topic modeling. In *Association for Computational Linguistics*, 2011.

[49] S. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. In *CVPR*, 2010.

[50] V. Jain and E. Learned-Miller. Online domain-adaptation of a pre-trained cascade of classifiers. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[51] M. Jamieson, A. Fazly, S. Stevenson, S. Dickinson, and S. Wachsmuth. Using language to learn structured appearance models for image annotation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 32(1):148–164, January 2010.

[52] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[53] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. Berg, and T. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, 2011.

[54] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. pages II: 1284–1291, 2005.

[55] C. Kuo and R. Nevatia. Robust multi-view car detection using unsupervised subcategorization. pages 1–8, 2009.

[56] L. J. Latecki, R. Lakamper, and U. Eckhardt. Shape descriptors for non-rigid shapes with a single closed contour. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000.

[57] B. Leong and R. Mihalcea. Measuring the semantic relatedness between words and images. In *International Conference on Computational Semantics (IWCS)*, 2011.

[58] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, pages II: 524–531, 2005.

[59] L. Li and L. Fei Fei. Optimol: automatic online picture collection via incremental model learning. *International Journal of Computer Vision (IJCV)*, 88(2), June 2010.

[60] L. Li, R. Socher, and L. Fei Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2036–2043, 2009.

[61] L. Li, R. Socher, and L. Fei Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009.

[62] L. Li, C. Wang, Y. Lim, D. Blei, and L. Fei Fei. Building and using a semantivisual image hierarchy. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3336–3343, 2010.

[63] X. Li, X. Zhao, Y. Fu, and Y. Liu. Bimodal gender recognition from face and fingerprint. 2010.

[64] D. Liu and T. Chen. Unsupervised image categorization and object localization using topic models and correspondences between images. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2007.

[65] N. Loeff, C. O. Alm, and D. A. Forsyth. Discriminating image senses by clustering with multimodal features. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 547–554. Association for Computational Linguistics, 2006.

[66] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[67] A. Lucchi and J. Weston. Joint image and word sense discrimination for image retrieval. In *ECCV*, pages I: 130–143, 2012.

[68] S. M. Semi-supervised extraction of entity attributes using topic models. *Master's Thesis*, 2009. Carnegie Mellon University.

[69] J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 918–925. IEEE, 1999.

[70] T. Malisiewicz and A. Efros. Recognition by association via learning per-exemplar distances. pages 1–8, 2008.

[71] D. Marr. Vision: A computational investigation into the human representation and processing of visual information. In *W.H. Freeman*, 1982.

[72] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[73] T. Mensink, J. Verbeek, and G. Csurka. Learning structured prediction models for interactive image labeling. In *CVPR*, pages 833–840, 2011.

[74] D. Mimno, W. Li, and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML*, 2006.

[75] T. Minka. Estimating a dirichlet distribution. *Annals of Physics*, 2003.

[76] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2002.

[77] H. Misra, F. Hopfgartner, A. Goyal, P. Punitha, and J. Jose. Tv news story segmentation based on semantic coherence and content similarity. In *Proceedings of the International Conference on MultiMedia Modeling*, 2010.

[78] V. Morariu and L. Davis. Multi-agent event recognition in structured scenarios. pages 3289–3296, 2011.

[79] G. Mori, S. Belongie, and J. Malik. Efficient shape matching using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 27(11):1832–1837, November 2005.

[80] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: a graphical model relating features, objects and scenes. In *Proceedings of Advances in Neural Information Processing (NIPS)*, 2003.

[81] Y. W. T. D. Newman and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Proceedings of Advances in Neural Information Processing (NIPS)*, 2006.

[82] Z. Niu, G. Hua, X. Gao, and Q. Tian. Spatial-disclda for visual recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[83] J. Philbin, J. Sivic, and A. Zisserman. Geometric lda: A generative model for particular object discovery. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2008.

[84] J. Ponce, T. Berg, M. Everingham, D. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. Russell, A. Torralba, C. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In *Toward Category-Level Object Recognition*, pages 29–48, 2006.

[85] A. Quattoni, M. Collins, and T. Darrell. Learning visual representations using images with captions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007.

[86] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[87] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, pages 1–8, 2007.

[88] X. Ren, C. C. Fowlkes, and J. Malik. *Mid-level cues improve boundary detection*. Citeseer, 2005.

[89] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[90] P. Roth, S. Sternig, H. Grabner, and H. Bischof. Classifier grids for robust adaptive object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2727–2734, 2009.

[91] K. Saenko and T. Darrell. Filtering abstract senses from image search results. In *Advances in Neural Information Processing Systems*, pages 1589–1597, 2009.

[92] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011.

[93] T. Sebastian, P. Klein, and B. Kimia. On aligning curves. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 25(1):116–124, January 2003.

[94] E. Seemann, B. Leibe, and B. Schiele. Multi-aspect detection of articulated objects. In *CVPR*, 2006.

[95] M. Shafiei and E. Milios. Latent dirichlet co-clustering. In *IEEE International Conference on Data Mining (ICDM)*, pages 542–551, 2006.

[96] Y. Shao, Y. Zhou, X. He, D. Cai, and H. Bao. Semi-supervised topic modeling for image annotation. In *Proceedings of the ACM conference on Multimedia(MM)*, 2009.

[97] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision (IJCV)*, 81(1), January 2009.

[98] J. Sivic and A. Zisserman. A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[99] T. M. Strat and M. A. Fischler. Context-based vision: recognizing objects using information from both 2 d and 3 d imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1050–1065, 1991.

[100] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Describing visual scenes using transformed objects and parts. *IJCV*, 77(1-3):291–330, May 2008.

[101] D. Swinney. Lexical access during sentence comprehension: (re) consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 1979.

[102] M. Tanner and W. Wong. The calculation of posterior distributions by data augmentation. In *Journal of the American Statistical Association*, 1987.

[103] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 2006.

[104] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision (IJCV)*, 53(2):169–191, July 2003.

[105] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision (IJCV)*, 53(2):169–191, July 2003.

[106] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(5):854–869, 2007.

[107] J. Varadarajan, R. Emonet, and J. Odobez. A sparsity constraint for topic models - application to temporal activity mining. In *Proceedings of Advances in Neural Information Processing (NIPS)*, 2010.

[108] H. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. In *Proceedings of Advances in Neural Information Processing (NIPS)*, 2009.

[109] C. Wang and D. Blei. A sparsity constraint for topic models - application to temporal activity mining. In *Proceedings of Advances in Neural Information Processing (NIPS)*, 2009.

[110] Y. Wang and G. Mori. Human action recognition by semilatent topic models. *PAMI*, 31(10):1762–1774, October 2009.

[111] Y. Wang and G. Mori. Max-margin latent dirichlet allocation for image classification and annotation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2011.

[112] S. Wermter, E. Riloff, and G. Scheler. *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*. IJCAI Workshop, Lecture Notes in Computer Science, 1995.

[113] M. Wertheimer. Untersuchungen zur lehre der gestalt, ii. *Psychologische Forschung*, 4:301–350, 1923.

[114] S. Yang, J. Bian, and H. Zha. Hybrid generative/discriminative learning for automatic image annotation. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010.

[115] Y. Zheng, M. Zhao, S. Neo, T. Chua, and Q. Tian. Visual synset: Towards a higher-level visual representation. In *CVPR*, pages 1–8, 2008.

[116] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[117] J. Zhu, A. Ahmed, and E. Xing. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[118] J. Zhu, L. Li, L. Fei Fei, and E. Xing. Large margin learning of upstream scene understanding models. In *Proceedings of Advances in Neural Information Processing (NIPS)*, 2010.