Modeling Semantic Relations between Visual Attributes and Object Categories via Dirichlet Forest Prior

Xin Chen[1] Xiaohua Hu[1] Zhongna Zhou[2] Yuan An[1] Tingting He[3] E.K. Park[4]
[1]College of Information Science and Technology, Drexel University, Philadelphia, PA 19104, USA, [2]Dept. of ECE at University of Missouri in Columbia, MO, USA, [3]Dept. of Computer Science at Central China Normal University, Wuhan, China, [4]California State University - Chico, Chico, CA 95929, USA
bruce.chen@drexel.edu thu@ischool.drexel.edu zz3kb@mizzou.edu
yuan.an@ischool.drexel.edu tthe@mail.ccnu.edu.cn ek.ek.park@gmail.com

ABSTRACT
In this paper, we deal with two research issues: the automation of visual attribute identification and semantic relation learning between visual attributes and object categories. The contribution is two-fold, firstly, we provide uniform framework to reliably extract both categorical attributes and depictive attributes. Secondly, we incorporate the obtained semantic associations between visual attributes and object categories into a text-based topic model and extract descriptive latent topics from external textual knowledge sources. Specifically, we show that in mining natural language descriptions from external knowledge sources, the relation between semantic visual attributes and object categories can be encoded as Must-Links and Cannot-Links, which can be represented by Dirichlet-Forest prior. To alleviate the workload of manual supervision and labeling in image categorization process, we introduce a semi-supervised training framework using soft-margin semi-supervised SVM classifier. We also show that the large-scale image categorization results can be significantly improved by combining automatically acquired visual attributes. Experimental results show that the proposed model achieves better ability in describing object-related attributes and makes the inferred latent topics more descriptive.

Categories and Subject Descriptors
I.2.6 [**Artificial Intelligence**]: Learning – *Parameter learning*; H.2.8 [**Database Management**]: Database applications – *Data mining*; *Image databases*; H.1.2 [**Models and Principles**]: User/Machine Systems – *Human factors, Human information processing*

General Terms
Algorithms, Experimentation, Human Factors, Design.

Keywords
Visual attribute identification, topic model, Dirichlet-Forest prior

1. INTRODUCTION
In our daily life, a large amount of our verbal communication describes the scene/environment around us. Also, recent years have seen increasing amount of online visual resources (such as images and videos) with natural language descriptions. Such information may potentially serve as a rich knowledgebase of how people construct natural language to describe visual content. In order that an image annotation system facilitate extracting and understanding the knowledge encoded in the visual content, it is very important to generate descriptive topic models that combines natural languages descriptions with image visual attributes. This work differs from conventional computer vision approaches such as scene recognition and object classification. Instead, it will encode additional semantic information such as the relation between object categories and different visual attributes, which is then linked to natural language descriptions of human knowledge (such as Wikipeida) to generate descriptive topic model regarding object with those visual attributes (Fig. 1).

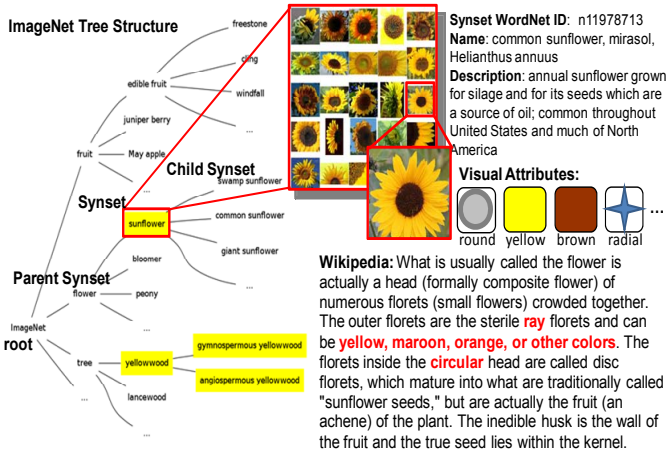


Fig. 1 illustration of lexical concept, narrative natural language description and visual attributes

Image annotation was conventionally solved as nearest-neighbor problem [3, 20]. Similar approaches range from studying the relevance between visual similarity and semantic similarity [15], using language entities to construct visual ontologies [7] or jointly modeling images and tags [11]. However, those approaches are infeasible when labeled reference exemplars are not available. An alternative way is to rely on structured knowledge bases of natural language descriptions (such as Wikipedia). Due to the increasing need of linking visual appearance to structured human knowledge in scalable image categorization/annotation, the extraction of semantic visual attributes has received increasing research focus. By its literal definition, the term "attribute" means "a quality or characteristic inherent in or ascribed to an object". Compared to

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

low-level image features, semantic visual attributes have much stronger relation to both object categories and human knowledge. It should be noted that although various types of attributes can be used to literally describe an object, however, only a small fraction of those attributes may be visible from an object image. Moreover, the usage of textual attributes may differ in different context. For example, in addition to color, texture, shape, body parts, semantic attributes of an animal may also involve its behavior, nutrition, activity, habitat and characters; on the contrary, the attributes about a plant may involve its cultivation and uses, which may be related to botany study. In order that the semantic attributes be useful for image annotation, these attributes should be visible and discriminating among different object categories, also, the union of semantic visual attributes should have sufficient coverage, which means that each object category be covered by at least one attribute.

In our research, we focus on the automation of attribute identification process and semantic relation learning between visual attributes and external textual knowledge sources. The contribution is two-fold, firstly, we provide uniform framework to reliably extract both categorical attributes and depictive attributes. Secondly, we incorporate the obtained semantic associations between visual attributes and object categories into a text-based topic model and extract descriptive latent topics from natural language knowledge base. Specifically, we show that in mining large scale knowledge base of natural language descriptions, the relation between semantic visual attributes and object categories can be encoded as Must-Links and Cannot-Links, which can be represented by Dirichlet-Forest prior. To reduce the amount of manual supervision and labeling in large-scale image categorization, we introduce a semi-supervised training framework using soft-margin semi-supervised SVM classifier (Fig. 2). We also show that the large-scale image categorization results can be significantly improved by combining automatically acquired visual attributes.

The remainder of this paper is organized as follows. In Section 2, we introduce the preliminary task in providing reliable source for attribute learning. In Section 3, we introduce our approach for image attribute classification. In Section 4, we present the framework of associate semantic visual attributes with text-based topic models via Dirichlet Forest prior and provide the Gibbs sampler for model estimation. Section 5 reports the experimental results. We conclude the paper in Section 6.

2. SEMI-SUPERVISED LARGE-SCALE MULTICLASS OBJECT CLASSIFICATION

ImageNet dataset [9] is a recently established large scale image ontology (over 15 million images from more than twenty thousand synsets) built upon the WordNet Structure, covering a subset of the nouns of WordNet. In ImageNet dataset, bounding boxes are available for over 3000 popular synsets. For each synset, there are on average 150 images with bounding boxes. The bounding boxes are manually annotated and verified through Amazon Mechanical Turk (AMT) workers. Comparing to attribute learning from full image (FI), the advantage of attribute learning in bounding boxes is obvious, the concept is much cleaner than the full image, no background clutter and other unrelated objects. Related researches have shown that image visual recognition algorithms significantly benefit from explicitly localizing category instance in the image [15]. Moreover, the association between image categories and visual attributes can also be significantly strengthen when using bounding box annotation. While high-quality manual labeled bounding boxes have led to impressive object recognition results, however, the main drawback of this approach is that it requires labor-intensive manual labeling and is not scalable to new object categories. In our approach, we proposed to robustly classify object categories and learn visible semantic attributes from automatically detected bounding boxes in ImageNet images (Fig. 2).

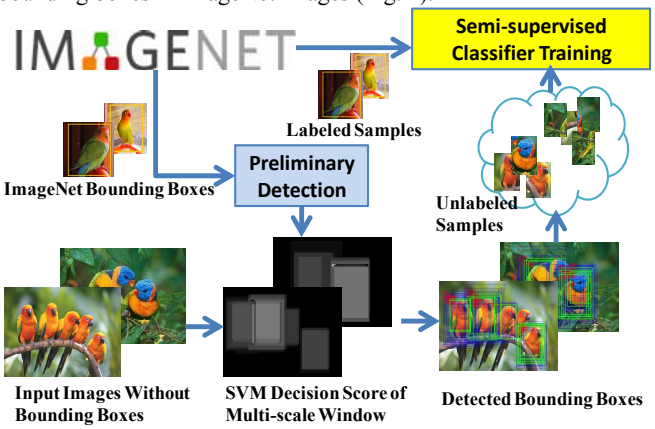


Fig. 2 Bounding boxes as reliable source for attribute learning

2.1 Bounding Box Detection in ImageNet Images

In our approach, we extract the HOG-LBP feature [20] for bounding box detection. We follow the settings in [3] to train the preliminary non-linear SVM classifiers, in which the kernel of categorical classification is the sum of individual χ^2 kernel SVM of each features. We use 80% image data for training and remaining 20% for validation. Achieves average multi-class classification accuracy of 38.5%. Specifically, we densely sample multi-scale detection windows $W_j^T(x, y, s)$ in whole-image range and then perform the 3D mean-shift [17] mode seeking algorithm (in both spatial and scale dimension) on the density map of SVM decision scores across the image to effectively locate the bounding boxes of objects. Given an detected window $\mathbf{x_{j-1}} = (\mathrm{x}, \mathrm{y}, \mathrm{s})$, the 3D mean-shift is calculated as:

$$m_j(x) = \frac{\sum_{i=1}^{M} x_i w(x_i) k\left(\left\|\frac{x_i - \mathbf{x}}{h}\right\|^2\right)}{\sum_{i=1}^{M} w(x_i) k\left(\left\|\frac{x_i - \mathbf{x}}{h}\right\|^2\right)} - \mathbf{x_{j-1}} \qquad (1)$$

In which $\{x_i\}_{i=1}^{M}$ are locations corresponding to sliding windows within the neighborhood of $\mathbf{x_{j-1}}$, $w(x_i)$ is the SVM decision score associated to each location x_i, and $k(x)$ is the profile of kernel K, which satisfies $K(\mathbf{x}) = k(\|\mathbf{x}\|^2)$. We begin with $\mathbf{x_0} = (x_0, y_0, \mathrm{s} = 0)$, iteratively compute j^{th} mean shift vector $m_j(x)$ and move the estimation window by $m_j(x)$ repeat until convergence. We choose a set of kernel scales around the original image scale as $\{\sigma_s = \sigma_0 \times 1.17^s, -n \le s \le n\}$, in which n=2 use the Gaussian kernel $K(\mathbf{x}) = exp\{-\|x_i - x\|^2/(2\sigma_s^2)\}$ for the spatial dimension x,y, use flat kernel for scale dimension s, with its shadow kernel $H(s) = 1 - (s/n)^2$.

2.2 Optimized Kernel Function for Soft-margin Semi-supervised Support Vector Machine

Given the relatively low accuracy (38.5%) in preliminary bounding box detection, directly assigning hard labels to the detected bounding boxes is sub-optimal. Instead, it is reasonable for us to consider the bounding box data as high-quality

'unlabeled' data with balanced positive and negative samples (i.e. accurate bounding boxes and inaccurate bounding boxes, respectively). In order to achieve optimal performance in object categorization, we propose to use soft-margin semi-supervised classifier in training (Fig. 2). However, one of the major challenges is how to appropriately involve unlabeled examples and efficiently update the discriminative model in an online semi-supervised setting. In our approach, we focus on exploring the intrinsic manifold structure of data marginal distribution and studying its role in kernel function optimization.

As shown in [2], general SVM training problem can be extended by considering the ambient space and the marginal distribution of the target function, thus two appropriate penalty terms can be introduced to reflect both the ambient space and the intrinsic structure of the data marginal distribution $\mathcal{P}_x$. Specifically, the target function could be estimated by:

$$f^* = \text{argmin}_{f\in\mathcal{H}_K} \frac{1}{l}\sum_{i=1}^{l} C\left(\mathbf{x_i}, y_i, f(\mathbf{x_i})\right) + \gamma_A\|f\|_{\mathcal{H}_K}^2 + \gamma_I\|f\|_I^2 \quad (2)$$

The $\|f\|_I^2$ can be estimated as the weighted Laplace-Beltrami operator associated with $\mathcal{P}_x$: $\|f\|_I^2 = \int_{\mathbf{x}\in\mathcal{M}} \|\nabla_{\mathcal{M}} f(\mathbf{x})\|^2 \, d\mathcal{P}_x$. It's shown in [12] that given a set of l labeled examples $\{(\mathbf{x_i}, y_i)\}_{i=1}^{l}$ and a set of u unlabeled examples $\{\mathbf{x_j}\}_{j=l+1}^{l+u}$, the Laplace-Beltrami operator on the manifold $\mathcal{M} \subset \mathbb{R}^N$ can be approximated by the graph Laplacian L on the basis of labeled and unlabeled data i.e. $\|\nabla_{\mathcal{M}} f(\mathbf{x})\|^2 := \langle \mathbf{f}, L\mathbf{f} \rangle$, in which $\mathbf{f} = [f(\mathbf{x_1}), \ldots, f(\mathbf{x_{l+u}})]^{\mathbf{T}}$. L is the graph Laplacian given by $L = D - W$, in which W_{ij} is similarity between $\mathbf{x_i}$ and $\mathbf{x_j}$ calculated by kernel function k, $D = diag(D_{1,1}, \ldots, D_{l+u,l+u})$ is a diagonal matrix with the entry $D_{i,i} = \sum_{j=1}^{l+u} W_{ij}$. The optimization problem (2) becomes:

$$f^* = \text{argmin}_{f\in\mathcal{H}_K} \frac{1}{l}\sum_{i=1}^{l} C\left(\mathbf{x_i}, y_i, f(\mathbf{x_i})\right) + \gamma_A\|f\|_{\mathcal{H}_K}^2 + \frac{\gamma_I}{(u+l)^2}\mathbf{f}^{\mathbf{T}}L\mathbf{f} \quad (3)$$

By Representer theorem, the solution is an expansion of kernel functions over both labeled and unlabeled data:

$$f^*(x) = \sum_{i=1}^{l+u} \alpha_i^* \, k(\mathbf{x_i}, \mathbf{x}) \quad (4)$$

According to Riesz Representation theorem, define the Gram kernel matrix K with its entries $K_{i,j} = k(\mathbf{x_i}, \mathbf{x_j})$, we have: $\|f^*\|_{\mathcal{H}_K}^2 = \langle f^*, f^* \rangle_{\mathcal{H}_K} = \sum_{i=1}^{l+u}\sum_{j=1}^{l+u} \alpha_i^* \alpha_j^* k(\mathbf{x_i}, \mathbf{x_j}) = \boldsymbol{\alpha}^{\mathbf{T}}\mathrm{K}\boldsymbol{\alpha}$. Similarly, $\mathbf{f}^{*\mathbf{T}}L\mathbf{f}^* = \langle \mathbf{f}^*, L\mathbf{f}^* \rangle = \boldsymbol{\alpha}^{\mathbf{T}}\mathrm{KLK}\boldsymbol{\alpha}$. By substituting into (3) the hinge loss function $\left(1 - y_i f(\mathbf{x_i})\right)_+ = max\left(0, 1 - y_i f(\mathbf{x_i})\right)$, the optimization problem can be re-written as: $f^* = \text{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^{l+u}, \xi_i\in\mathbb{R}} \frac{1}{l}\sum_{i=1}^{l}\xi_i + \gamma_A\boldsymbol{\alpha}^{\mathbf{T}}\mathrm{K}\boldsymbol{\alpha} + \frac{\gamma_I}{(u+l)^2}\boldsymbol{\alpha}^{\mathbf{T}}\mathrm{KLK}\boldsymbol{\alpha}$ (5)

subject to the relaxed separation constraint:

$$y_i\left(\sum_{j=1}^{l+u} \alpha_j^* \, k\left(\mathbf{x_i}, \mathbf{x_j}\right) + b\right) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, l.$$

The above constraint optimization problem (5) can be solved by introducing the Lagrangian in which two Lagrange multipliers $\beta_i, \zeta_i \geq 0$ are defined for either constraint:

$$L(\boldsymbol{\alpha}, \boldsymbol{\xi}, b, \boldsymbol{\beta}, \boldsymbol{\zeta}) = \frac{1}{l}\sum_{i=1}^{l}\xi_i + \boldsymbol{\alpha}^{\mathbf{T}}\left(\gamma_A\mathrm{K} + \frac{\gamma_I}{(u+l)^2}\mathrm{KLK}\right)\boldsymbol{\alpha} - \sum_{i=1}^{l}\beta_i\left(y_i\left(\sum_{j=1}^{l+u}\alpha_j \, k\left(\mathbf{x_i}, \mathbf{x_j}\right) + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{l}\zeta_i\xi_i \quad (6)$$

Vanishing the derivative of L with respect to b and ξ_i leads to: $\sum_{i=1}^{l}\beta_i y_i = 0$, $\frac{1}{l} - \beta_i - \zeta_i = 0$, $0 \leq \beta_i \leq \frac{1}{l}$. Substituting them into (6) with b and ξ_i removed, it gives:

$$L^R(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}^{\mathbf{T}}\left(\gamma_A\mathrm{K} + \frac{\gamma_I}{(u+l)^2}\mathrm{KLK}\right)\boldsymbol{\alpha} - \boldsymbol{\alpha}^{\mathbf{T}}\mathrm{KJ}^{\mathrm{T}}\mathrm{Y}\boldsymbol{\beta} + \sum_{i=1}^{l}\beta_i \quad (7)$$

In which $\mathrm{J} = [\mathrm{I}\ 0]_{\mathrm{l}\times(\mathrm{l+u})}$, $\mathrm{Y} = diag(y_1, \ldots, y_l)$.

Taking derivative of (8) with respect to $\boldsymbol{\alpha}$ leads to: $\frac{\partial L^R}{\partial \boldsymbol{\alpha}} = \left(\gamma_A\mathrm{K} + \frac{\gamma_I}{(u+l)^2}\mathrm{KLK}\right)\boldsymbol{\alpha} - \boldsymbol{\alpha}^{\mathbf{T}}\mathrm{KJ}^{\mathrm{T}}\mathrm{Y}\boldsymbol{\beta} = 0$, which implies that the $l+u$ expansion coefficients $\alpha_1, \ldots, \alpha_{\mathrm{l+u}}$ can be obtained by solving the following quadratic dual program:

$$\begin{cases} \boldsymbol{\alpha}^* = \left(\gamma_A\mathrm{I} + \frac{\gamma_I}{(u+l)^2}\mathrm{LK}\right)^{-1}\mathrm{J}^{\mathrm{T}}\mathrm{Y}\boldsymbol{\beta}^* \\ \boldsymbol{\beta}^* = \text{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^l}\sum_{i=1}^{l}\beta_i - \boldsymbol{\beta}^{\mathbf{T}}\mathrm{Q}\boldsymbol{\beta} \end{cases} \quad (8)$$

subject to $\sum_{i=1}^{l}\beta_i y_i = 0$, $0 \leq \beta_i \leq \frac{1}{l}, i = 1, \ldots, l$, in which: $\mathrm{Q} = YJK\left(\gamma_A\mathrm{I} + \frac{\gamma_I}{(u+l)^2}\mathrm{LK}\right)^{-1}\mathrm{J}^{\mathrm{T}}\mathrm{Y}$.

(8) is a standard restricted quadratic program which can be solved via conjugate gradient descent in Ch. 6 of [2]. During training, the labeled data $\{(\mathbf{x_i}, y_i)\}_{i=1}^{l}$ and unlabeled data $\{\mathbf{x_j}\}_{j=l+1}^{l+u}$ are used for solving $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ by conjugate gradient descent, where $y_i \in \{-1, +1\}$. By substituting the solution $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ of quadratic program (7) to (4), we obtain the expansion of kernel function over both labeled data $\{(\mathbf{x_i}, y_i)\}_{i=1}^{l}$ and unlabeled data $\{\mathbf{x_j}\}_{j=l+1}^{l+u}$. At the stage of detection, the decision function classified new samples into class +1 or -1 by $y(x) = sign(f^*(x))$..

3. DESCRIPTIVE VISUAL ATTRIBUTE EXTRACTION AND RELATION LINKS

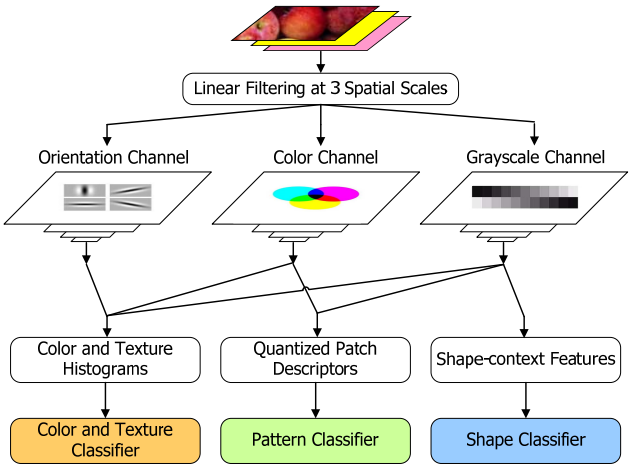


Fig. 3 Illustration of semantic attribute classifier

Previous studies on descriptive visual attributes have shown beneficial to improve the performance of object ategorization and text description generation [14]. The depictive visual attributes involves color attributes, texture attributes (such as furry, wooden, rough), pattern attributes (spotted, striped) and shape attributes (long, round, rectangle). The attributes may also be associated to the visual similarity with known object classes (for example, giant panda and polar bear both have bear-like attribute). Ideally, these attributes should be able to discriminate between object classes (being associated to some but not all of them), provide sufficient coverage (all classes have at least a single attribute association), and be correlated to visual object class properties that can be

observed in images. In our approach, three types (Fig. 3) of features are used for attribute extraction, i.e. GIST feature [4], densely sampled SIFT [6] and HOG-LBP feature [20]. Each of the three feature types is normalized independently to unit length, then a histogram intersection kernel SVMs [14] is performed to train the attribute classifier. For each classifier, we fit a sigmoid function [10] to the SVM decision score and convert the output to a probability. The probability $p(Attribute|Category)$ can be aggregated across the whole dataset and eventually build the semantic relations (such as Must-link and Cannot-link) between visual attributes and object categories.

3.1 Attributes as Unique Signatures of Object Categories

Given images of an object category, some visual attributes may be presented while some may not, which results in unique attribute signatures associated with each category. Let $a^y = \left(a_1^y, \cdots, a_M^y\right)$ be a vector of binary associations $a_m^y \in \{0,1\}$between attributes $a_m$and trained object category y. An aggregation of all results (in which $a_m^y = 1$ for positive samples and 0 for negative samples) from binary classifier of attribute a_m can provide an estimation of the conditional probability $p(a_m|y)$of that attribute being present in category y. By assuming mutual independence among attributes, we have $(a|y) = \prod_{m=1}^{M} p(a_m|y)$.

Following the idea of Direct Attribute Prediction model (DAP) in [3], For an image x with $M = K$ attributes $a_1, \dots, a_K$ where each attribute corresponds to exactly one conditional probability $p(a_m|y)$, the posterior probability of image x belong to object category y is given as

$$p(y|x) \propto \prod_{k=1}^{K} \left(\frac{p(a_k|y)}{p(a_k)}\right) \quad (9)$$

Our experiment results show the performance of object categorization can be significantly improved when the categorization results are smoothed by (9), with $K = 10$ most relevant attributes used (Fig. 7).

By aggregating the results of binary attribute classifiers for all the object categories, we obtain an aggregated attribute-category concurrence map. From which we threshold the aggregation score to produce both Must-Link and Cannot-Link relations for each attribute-category pair (a, y).

4. DESCRIPTIVE TOPIC MODELING VIA DIRICHLET FOREST PRIOR

In this section, we introduce a novel topic model to infer depictive latent topics from both text corpora and attribute-category relations (Fig. 4). Recent studies of large-scale visual classification in ImageNet [8, 15] suggest that visual classification across semantically-defined class boundaries is feasible. In [13], the author proposed to infer object class-attribute association by text-based semantic relatedness on WordNet and Wikipedia. The WordNet is a large scale lexical database of English Language, in which English words are organized into concepts (synonym sets or synsets) according to synonymy and various lexical and semantic relations between lexicalized concepts.

Wikipedia is one of the most comprehensive and well-formed electronic knowledge repositories on the web with millions of articles contributed collaboratively by volunteers. Because of its reliability, accuracy and neutral point of view. Wikipedia has been exploited as external knowledge source in various data mining applications [13, 19]. Although Wikipedia is different from standard WordNet ontology, which is backed up by structured thesaurus, however, each article in Wikipedia only describes one single concept under a hierarchical categorization system. We have found a large amount of Wikipedia articles share the same lexicalized entry as ImageNet synsets, which makes mapping between ImageNet synset to a Wikipedia articles possible. (In our study, about 75% of the ImageNet synsets have corresponding Wikipedia articles). In our approach, the semantic relations between attribute-category pairs (i.e. Must-Links and Cannot-Links) are encoded as Dirichlet Forest prior in the proposed topic model. In order to effectively encode the semantic relations, we explore the WordNet synonym set and extend Must-Links and Cannot-Links between attribute-category terms (i.e. the lexical word of both attribute and object) to their synonyms.

4.1 Preliminary of Dirichlet Tree Distribution and the Modeling of Must-Links

The Dirichlet Tree Distribution[17] is a generalization of Dirichlet distribution that allow to break the mutual independence in words generation process, makes the generative process controlled by word-link such as Must-Link (u, v). The Dirichlet-tree distribution is a tree with the words as leaf nodes; let $r^{(k)}$be the Dirichlet tree edge leading into node k, let $c(k)$be the immediate children of node $k$in the tree, L the leaves of the tree, I the internal nodes, and $L(k)$ the leaves in the subtree under k, to generate a sample $\phi \sim Dirichlet\ Tree(r)$, one first draws a multinomial at each internal node $s \epsilon I$ from $Dirichlet(r^{c(s)})$, i.e. using the weights from s to its children as the Dirichlet parameters.

The probabilityϕ_k of a word $k \in L$ is then simply the product of the multinomial parameters on the edges from $k$to the root.

It can be shown that, the above procedure gives

$$Dirichlettree(r) \equiv p(\phi|r) = \left(\prod_{k \in L} {\phi_k}^{r^{(k)}-1}\right)\left[\prod_{s \in I} \frac{\Gamma\left(\sum_{k \in c(s)} r^{(k)}\right)}{\prod_{k \in c(s)} \Gamma(r^{(k)})} \left(\sum_{k \in L(s)} \phi_k\right)^{\Delta(s)}\right]$$

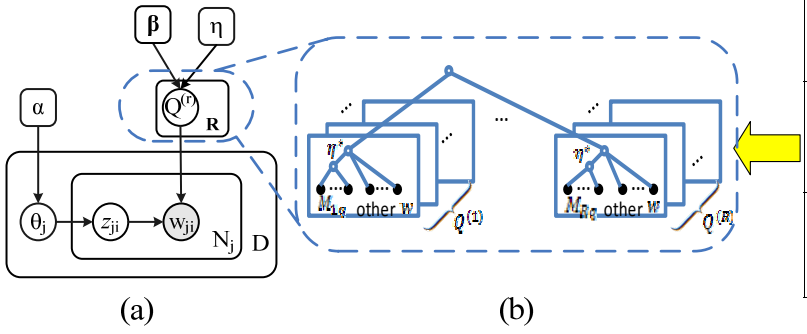
In which $\Gamma(\cdot)$ is the gamma function, and the notation Π_k^L means $\Pi_{k \in L}$; the function $\Delta(s) \equiv r^{(s)} - \sum_{k \in c(s)} r^{(k)}$ is the difference between the in-degree and out-degree at internal nodes$s$. (When the difference $\Delta(s) = 0$, for all internal node s, the Dirichlet tree reduces to a Dirichlet distribution).

Like the Dirichlet distribution, the Dirichlet tree distribution is conjugate to the multinomial. It's possible to integrate out ϕ to get a distribution over word counts directly, similar to the multivariate Polya distribution (a.k.a. Dirichlet-Multinomial) in [16]:

$$p(\mathbf{w}|r) = \prod_{s \in I}\left(\frac{\Gamma(\sum_k^{c(s)} r^{(k)})}{\Gamma(\sum_k^{c(s)}(r^{(k)} + n^{(k)}))} \cdot \prod_{k \in c(s)} \frac{\Gamma(r^{(k)} + n^{(k)})}{\Gamma(r^{(k)})}\right)$$

in which $n^{(k)}$ is the number of word tokens in $\mathbf{w}$ that appear in $L(k)$, $L(k)$ is the leaves in the subtree under k. $c(s)$ is the immediate children of node s.

The definition of Must-Link is transitive. Must-Link (u, v) and Must-Link (u, w) define a transitive closure of and Must-Link (u, v, w). In Dirichlet Tree for Must-Links, each transitive closure is subtree, in which words are leaves nodes with symmetric uniform base measure $\eta\beta$ from one internal node s and each of theinternal node s is connected to the root node with weight $|L(s)|\beta$, in which $|L(s)|$is the size of leavens in sub-tree under s.



Category	Must-Link Attributes	Cannot-Link Attributes
 Polar Bear	white, furry, tough skin, arctic, coastal, ocean, water	black, blue, stripes, hairless, desert, bush, plains, forest, fields, jungle, mountains, tree ...
 Humpback Whale	black, gray, hairless, tough skin, swims, ocean, water	white, brown, orange, red, yellow, spots, stripes, furry, desert, bush, plains, forest, fields, jungle, mountains, ground, tree ...
 Zebra	black, white, stripes, furry, tough skin, long leg, long neck	blue, brown, gray, orange, red, yellow, patches, spots, hairless, fish, arctic, coastal, desert, forest, jungle, mountains, ocean, water...

(c)

(a) Text topic model (b) Dirichlet Forest prior encoding the visual constraint (c)Must-Link and Cannot-Link Attributes
Fig. 4 Graphical representation of the proposed method

If $\eta = 1$, then in-degree equals out-degree for any internal nodes(both are $|L(s)|\beta$), and the tree reduces to a Dirichlet distribution with symmetric prior β. When we take $\eta = |L(s)|$, it will re-distribute the probability mass at node s. Which results in increased concentration, and re-distribute the mass evenly in the transitive closure s. The independence (which is enforced in Dirichlet distribution) among Must-Link words is thus eliminated and allows for similar but not identical probabilities for the Must-Link words.

4.2 Dirichlet Forest Prior and Cliques of Cannot-Links

From the aggregated concurrence map of attribute-category relations, we are able to assign both Must-Links and Cannot-Links to an object category. It should be noted that, given the presence of an object category in an image, the Must-Links and Cannot-Links corresponding to that category should be simultaneity observed, therefore, such Must-Links and Cannot-Links should be encoded in the same latent topic. With this consideration, we propose a clique-based topic sampling process as follows .

In our approach, each 'clique' is associated with one single object category, it is composed of two parts: the first part is a Dirichlet sub-tree corresponding to Must-Links of that object category, the second parts is all other words (other than words in Must-Links) that are allowed to simultaneously have large probability without violating the Cannot-Links of that object category (Fig. 4b). Each clique is also a Dirichlet tree.

For each object category r, we generate a total of $Q^{(r)} = Q$ cliques $q = 1, \dots, Q^{(r)}$, in this way, we create a mixture model of $Q^{(r)}$ Dirichlet subtrees, one for each of the $Q^{(r)}$ cliques. In generating the latent topics, the cliques are sampled according to their probability $p(q), q = 1, \dots, Q^{(r)}$.

The clique's root node connects to an internal node s(root node of Must-Link sub-tree) with weight $\eta \cdot |L(s)| \cdot \beta$, the node s then connects to words in Must-Links with weight β. The clique's root also directly connects to words that is not in Must-Links (but not violating the Cannot-Links of that object category) with weight β. This structure will send majority probability mass down to s and then re-distribute it among words in Must-Links. Which results in strong association among Must-Link words

Let R be the number of object categories. Our Dirichlet Forest prior (β, η) will consist of $\prod_{r=1}^{R} Q^{(r)}$ possible Dirichlet trees (cliques), each Dirichlet tree has R branches under the root, one for each connected component, for the r-th branch, there are $Q^{(r)}$ possible Dirichlet subtrees corresponding to $Q^{(r)}$ cliques, which leads to $\prod_{r=1}^{R} Q^{(r)}$ different Dirichlet trees. Therefore, a Dirichlet tree in the forest is uniquely identified by an index vector $q = (q^{(1)}, \dots, q^{(R)})$, where $q^{(r)} \in \{1, \dots, Q^{(r)}\}$.

In generating a Dirichlet Forest model (Fig. 4a), let $n_j^{(d)}$ be the number of word tokens in document d assign to topic j, integrating out $\boldsymbol{\theta}$, $\mathbf{z}$ can be generated as:

$$p(\mathbf{z}|\alpha) = \left(\frac{\Gamma(\mathrm{T}\alpha)}{\Gamma(\alpha)^{\mathrm{T}}}\right)^{D} \prod_{d=1}^{D} \frac{\prod_{j=1}^{T} \Gamma(\mathrm{n}_{\mathrm{j}}^{(\mathrm{d})} + \alpha)}{\Gamma(n^{(d)} + \mathrm{T}\alpha)}$$

For each topic $j = 1, \dots, \mathrm{T}$, we sampled a Dirichlet tree $q_j = (q_j{}^{(1)}, \dots, q_j{}^{(R)})$ from the Dirichlet Forest prior (β, η)

$$p(q_j) = \prod_{r=1}^{R} p(q_j{}^{(r)})$$

In which each $q_j{}^{(r)}$, $r = 1, \dots R$ is sampled by: $(q_j{}^{(r)}) \propto \left|clique_{q_j{}^{(r)}}\right|$ $(r = 1, \dots R)$.

Finally, the model can be generated by:

$$p(w, z, q_{1:\mathrm{T}}|\alpha, \beta, \eta) = p(w|q_{1:\mathrm{T}}, z, \beta, \eta) \cdot p(z|\alpha) \cdot \prod_{j=1}^{\mathrm{T}} p(q_j)$$

4.3 Gibbs Sampling For Model Estimation

In this section, we introduce the Markov Chain Monte Carlo and Gibbs sampling process of the proposed topic model.

Let $n_{-i,j}^{(d)}$ be the number of word tokens in document d assigned to topic j, excluding the word w_i. Let $n_{-i,j}^{(k)}$ denotes the number of word tokens in the corpus that are under node k in topic j's Dirichlet tree q_j, excluding the word at position i. For candidate topic labels $t = 1, \dots, \mathrm{T}$, we have:

$$p(z_i = t|z_{-i}, q_{1:T}, w) \propto \left(n_{-i,t}^{(d)} + \alpha\right) \prod_{s}^{Parent_i} \frac{r_t^{(c_t(s))} + n_{-i,t}^{(c_t(s))}}{\sum_k^{c_t(s)} \left(r_t^{(k)} + n_{-i,t}^{(k)}\right)}$$

In which $Parent_{\mathrm{i}}$ denotes the subset of internal nodes in topic vDirichlet tree that are ancestors of leaf w_i and $c_t(s)$ is the unique

node that is s's immediate child and is also an ancestor of w_i(including w_i itself).

Since the R branches (each corresponding to an object category, featured by both Must-Links and Cannot-Links) are independent, sampling the Dirichlet tree $q_j$is factorized to sampling the cliques for each ${q_j}^{(r)}$. For candidate cliques of connected component r: $q' = 1, \ldots, Q(r)$ we have:

$$p\left({q_j}^{(r)} = q' | z, q_{-j}, {q_j}^{(-r)}, w\right) \propto \left(\sum_{k}^{\left|clique_{q', {q_j}^{(r)}}\right|} \beta_k \right) \prod_{s}^{\mathrm{I}_{j,r=q'}} \left(\frac{\Gamma\left(\sum_{k}^{c_j(s)} {r_j}^{(k)}\right)}{\Gamma\left(\sum_{k}^{c_j(s)} \left({r_j}^{(k)} + {n_j}^{(k)}\right)\right)} \cdot \prod_{k}^{c_j(s)} \frac{\Gamma\left({r_j}^{(k)} + {n_j}^{(k)}\right)}{\Gamma\left({r_j}^{(k)}\right)} \right)$$

In which $\mathrm{I}_{j,r=q'}$ denotes the internal nodes below the r^{th} branch of tree q_j when clique ${q_j}^{(r)}$ is selected.

5. EXPERIMENTS AND RESULTS

In this section, we evaluate the performance of the proposed methods, including automatic attribute identification, object categorization, and modeling the semantic relations between visual attributes and object categories.

5.1 Datasets and Experimental Setup

In learning the visual attributes of object categories, we use the Animals with Attributes (AwA) dataset introduced by [3] (which is a fraction of ImageNet database). The AwA dataset consists of 50 mammal object categories with a human provided attribute inventory and corresponding object class-attribute associations. In our experiment, we split the dataset into 80% training and 20% testing (i.e. 24,295 training images and 6,180 test images) for learning the attribute classifiers. We map all the 50 AwA categories to the corresponding synsets (identified with WordNet ID) under the ImageNet hierarchical taxonomy, from which we are able to calculate the semantic metric among different categories. Also, with the help of word entities from the corresponding WordNet ID (wnid), we download 75 Wikipedia articles that share the same lexicalized entry as WordNet entities (considering synonyms). The Wikipedia articles are then considered as the knowledge base of natural language description for corresponding ImageNet synsets or AwA categories.

5.2 . Object Categorization and Attribute Identification

The first part of our experiments is semi-supervised learning for object categorization. As mentioned in Section 2, bounding box detection is performed to ensure that we identify clean attributes from each object category. We are the learning process from 50 labeled images bounding box per class (i.e. 2500 image bounding boxes in total) for training, The semi-supervised SVM use 50 labeled bounding boxes and an addition of 50 unlabeled bounding box samples (from preliminary detection in Section 2.1). We use another 100 images from each class (other than the 5000 training image) for testing. For each test image i, if it is correctly classified then the flat error $e_i = 0$, if it is not correctly classified then $e_i = 1$. Fig. 5 shows the Receiver Operating Characteristic (ROC) curves of the object categorization results. Each curve is the result of the one-vs-all semi-supervised SVM categorical classifier on HOG-LBP feature. The Area Under Curves (AUC) are also provided. Higher AUC indicate better classify performance. For example, the AUC scores of giant panda (AUC=0.912) and zebra (AUC=0.975) are among the highest, indicating that these object categories are well represented by HOG-LBP feature and are well separated in the feature space. The categorization results of raccoon and lion is fair, which are possibly caused by the high diversity of object appearance among training samples.

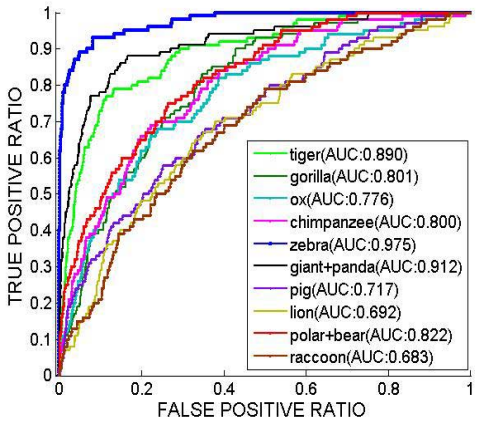
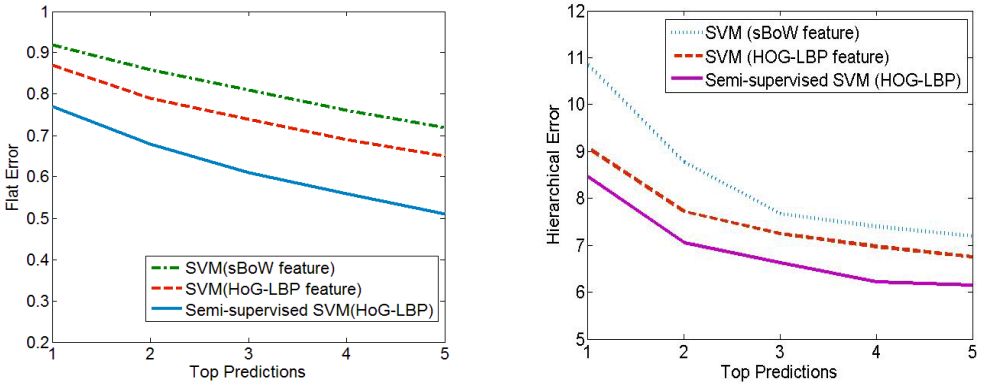


Fig. 5 Part of the object classification results plotted in receiver operating characteristic (ROC) curves. Each curve is the result of the one-vs-all semi-supervised SVM categorical classifier on HOG-LBP feature.



(a) Average Flat Error (AFE) **(b) Average Hierarchical Error (AHE)**

Fig. 6 Performance comparison of proposed method (semi-supervised SVM + HOG-LBP features) with state-of-the-art approaches

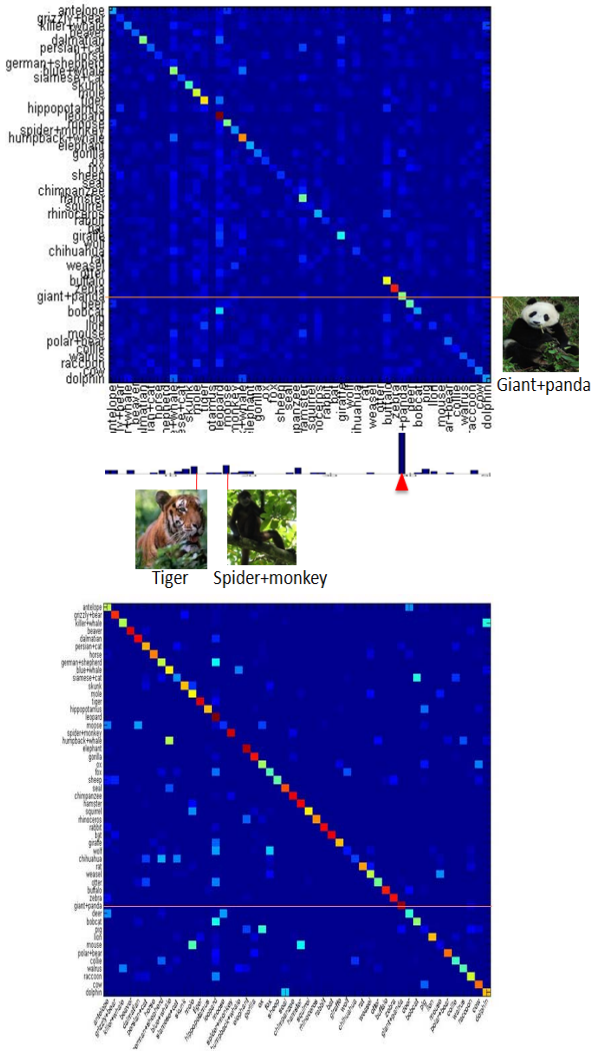


Fig. 7 Confusion matrix of classifying 50 AWA animal classes. Upper: confusion matrix by semi-supervised SVM classifier. Lower: confusion matrix after object-attribute signatures being introduced in, categorization results are smoothed by (eq. 9), with $K = 10$ most relevant visual attributes used.

In Fig. 6, we compare our categorization method (semi-supervised SVM with HOG-LBP features) to two state-of-the-art approaches, i.e. SVM classifier using spatial bag of word (sBoW) features and SVM using HOG-LBP features. Specifically, Fig. 6a shows the Average Flat Error (AFE) with respect test images with top predictive scores. The AFE score is defined as: $e = \frac{1}{N}\sum_{i=1}^{N} e_i$, e $\in [0,1], N = 5000$, the lower AFE score indicates better classification performance. Fig. 6b represents the Average Hierarchical Error (AHE) of different approaches. Supposing that the image from class i is mis-classified as class j, and $\pi(i,j)$is the lowest common ancestor between class i and j in the hierarchy of ImageNet taxonomy. The height $h(\pi(i,j))$ of node $\pi(i,j)$ on the hierarchy is then defined as the length of the longest path to one of its leaf node. Leaf nodes have height 0. The AHE is the average of $h(\pi(i,j))$ for all the testing images, $h = \frac{1}{N}\sum_{i=1}^{N} h(\pi(i,j)), N = 5000$, the lower AHE score indicate higher semantic accuracy in object categorization. As shown in Fig. 6, the proposed object categorization method consistently outperforms the state-of-the-art approaches under both AFE and AHE comparisons.

Fig. 7 shows the confusion matrix of object categorization in all the 50 AWA categories. By comparing the confusion matrix of semi-supervised SVM classification and the confusion matrix categorization results are smoothed by the signature of the most relevant attributes, we can see that, the performance of object categorization can be significantly improved when attribute signatures are introduced. For example, in the original semi-supervised SVM classification results (upper part of Fig.7), the *giant panda* category is to some extends confused with the *tiger* category and the *spider monkey* category. However, after introducing in the object-attribute signatures and smoothing the categorization results with posterior object-attribute prediction model (eq. 9), the categorization ambiguity is mostly eliminated (lower part of Fig. 7). The significantly reduced categorization ambiguity across the 50 AWA animal classes (Fig. 7) evidences the effectiveness of identified attribute-object relations.

5.3 Model Estimation and Illustration

In this section, we perform both quantitative and qualitative evaluation on the performance of proposed topic model. The quantitative evaluation includes comparing both log-likelihood and perplexity, while qualitative evaluation is achieved by visualizing the inferred latent topics and evaluate its relevance to the object-attribute relations.

5.3.1 Log-likelihood and Perplexity Comparison

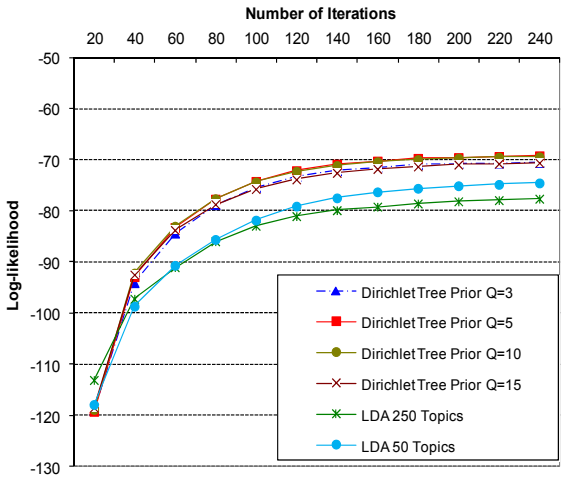
Log-likelihood is one of the standard criteria in generative model evaluation. It provides a quantitative measurement of how well a topic model fits the training data. The score of log-likelihood (which is a negative real number) is the higher the better. In practice, the log-likelihood of words given latent topics can be calculated by integrating out all the latent variables:

$$p(\mathbf{w}\mid\mathbf{z}) = \prod_{t=1}^{T}\left[\int_{\varphi_{z_t}} p(\mathbf{w}\mid z_t,\varphi_{z_t})p(\varphi_{z_t}\mid z_t)d\varphi_{z_t}\right] \tag{10}$$
$$= \left[\frac{\Gamma(W\beta)}{\Gamma(\beta)^{W}}\right]^{T}\cdot\prod_{t=1}^{T}\frac{\prod_{w_i}\Gamma(n_t^{(w_i)}+\beta)}{\Gamma(n_t^{(\cdot)}+W\beta)}\cdot\frac{\Gamma(W\eta)}{\Gamma(\eta)^{W}}\cdot\frac{\prod_{w_i}\Gamma(n_0^{(w_i)}+\eta)}{\Gamma(n_0^{(\cdot)}+W\eta)}$$

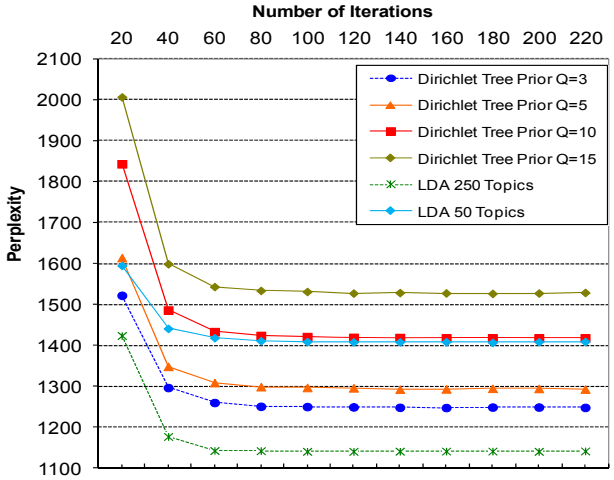
The perplexity is another standard criterion for generative probabilistic models that evaluates how well the model predicts the testing data. The perplexity of a testing dataset D_{test} is:

$$perplexity(D_{test}) = \exp\left[\frac{-\sum_{j=1}^{D_{test}}\log(p(\mathbf{t}_j))}{\sum_{j=1}^{D_{test}}N_j^t}\right] \tag{11}$$

The perplexity score for a model is the lower the better. Fig. 8a represents the log-likelihood comparison results between our proposed model and the LDA model over the iterations. As we can see from Fig. 8a, our proposed topic model has consistently higher log-likelihood than standard LDA model, which can be explained by the introduced Dirichlet Forest priors, which make our model fit better to training data than the LDA model. Fig. 8b shows the comparison of perplexity between our model and the LDA model over the iterations. Our model achieves best perplexity scores when Q=3, while the LDA model achieves best perplexity scores when topic number is 250. Although LDA model has relative lower perplexity score compared to our model, however, as we can see in the next section, the LDA model may not be able to accurately link object category to its attributes.



a) Log-likelihood comparison (using T=50)



b) Perplexity comparison (using T=50)

Fig. 8 Log-likelihood and perplexity comparison between proposed model and LDA model over the iterations

5.3.2 Topic Visualization and Relevance Analysis

On the convergence of the Markov Chain Monte Carlo and Gibbs sampling process, the conditional probability of each word/entity given each inferred latent topic can be obtained.

In Fig. 9 - Fig. 11, we illustrate the qualitative evaluations of 3 ImageNet object categories (i.e. *n02391049: zebra*, *n02129165: lion* and *n02581957: dolphin*), including the category names, the identified visual attributes, the Must-Links and Cannot-Link from aggregated attribute-object concurrence map. We also visualize the most relevant inferred latent topics with respect to each object category name entity. The relevance between object category name entities and the inferred latent topics can be obtained by calculating the Mutual Information (MI) score.

The calculation of MI between a specific word entity and a latent topic is shown as eq. 12, in which R_g and Z_t are binary indicator variables corresponding to the word and the latent topic, respectively. The variable pair (R_g,Z_t) indicates the cases that latent topic Z_t being assigned to word entity R_g.

$$MI(R_g,Z_t)=p(R_g,Z_t)\log\frac{p(R_g,Z_t)}{p(R_g)p(Z_t)} \qquad (12)$$

Given the training data, both the joint probability $p(R_g,Z_t)$ and the marginal probabilities $p(R_g)$ and $p(Z_t)$ can be empirically estimated by counting the number of evidences over the training dataset.

As we can see from Fig. 9 - Fig. 11, for the LDA model, the inferred latent topic that is most relevant to the object category doesn't contain much visual attribute names associated with that object category. On the contrary, the latent topics inferred from our model have a lot of important visual attributes among the top-ranked words of the most relevant latent topic. Specifically, for the object category '*dolphin*' in Fig. 11, the most relevant latent topic inferred from our model involve visual attributes that are highly consistent with the identified Must-Link relations associated with the *dolphin* category such as 'blue', 'water', 'white', 'fish', etc. while the most relevant latent topic inferred by LDA model doesn't involve any visual attributes associated with *dolphin* in its top-ranked words. Similarly, for the object category '*zebra*' in Fig. 9, the most relevant latent topic inferred from our model involve most visual attributes associated with the *dolphin* category such as 'stripes', 'black', 'white', 'large', 'group', etc. , suggesting that the Must-Link relationships between object categories and visual attributes are well preserved by the Dirichlet-tree distribution in our proposed model (Section 4.1).

It's also worth mentioning that, in Fig. 10, one of the attributes (i.e. 'spotted') that cannot be linked to *lion* category is among the top-ranked words of the most relevant latent topic inferred by the conventional LDA model. As a comparison, none of the latent topics inferred by our proposed model violate Cannot-Link relations. The experiment results indicate that the Cannot-Link relations between object categories and visual attributes can be effectively encoded by the Dirhchlet Forest prior introduced in Section 4.2, which enables the topic model to purify the inferred latent topics, filter out the 'noisy' and 'self-contradictory' information from the textual descriptions and produce consistently well topical abstraction of object categories and associated visual attributes.

n02391049: zebra
Must-link Attributes: black, white, stripes, furry, toughskin, large, lean, longleg, longneck, tail, bush, plains, fields, ground, group
Cannot-link Attributes: blue , brown , gray , orange , red , yellow , patches , spots , hairless,…

Most relevant latent topic:

Top words	Probability
zebra	0.19874
stripes	**0.04919**
Mountain	0.04099
plains	**0.03075**
social	0.01436
Stallion	0.01231
predator	0.01026
hybrids	0.00617
bachelor	0.00617
striping	0.00617

(a) LDA Model

Top words	Probability
zebra	0.18024
stripes	**0.02847**
white	**0.02278**
Grevy	0.01899
wild	**0.01441**
large	**0.01330**
black	**0.01108**
plains	**0.00950**
extinct	0.00950
group	**0.00761**

(b) Our Proposed Model

Fig. 9 the most relevant latent topic of object category 'zebra'

N02129165 : lion
Must-link Attributes: brown, yellow, furry, big, lean, pads, paws, tail, claws, walks, muscular, quadrapedal, desert, bush, ground
Cannot-link Attributes : black , white , blue , gray , orange , red , patches , spots , stripes , hairless , toughskin, fly, swim,…

Most relevant latent topic:

Top words	Probability
lion	0.18134
cubs	0.03486
Asiatic	0.01932
subspecies	0.01649
hunting	0.01272
spotted	0.01272
incidents	0.00942
American	0.00926
Barbary	0.00848
ligers	0.00613

(a) LDA model

Top words	Probability
lion	0.17802
wild	**0.01949**
kills	0.01860
cats	0.01861
desert	**0.01684**
Asiatic	0.00975
forest	**0.00798**
food	**0.00798**
big	**0.00709**
away	0.00709

(b) Our Proposed Model

Fig. 10 the most relevant latent topics of object category 'lion

N02581957: dolphin
Must-link Attributes: white, blue, gray, hairless, toughskin, big, lean, flippers, tail, swims, fish, coastal, ocean, water,
Cannot-link Attributes: black , brown , orange , red , yellow , patches , spots , stripes , furry, desert , bush , plains , forest , mountain,…

Most relevant latent topic:

Top words	Probability
dolphins	0.14474
River	0.03321
whale	0.01860
species	0.01329
attacks	0.01296
meat	0.01111
tuna	0.01062
sounds	0.00931
mammals	0.00798
Delphinidae	0.00798

(a) LDA Model

Top words	Probability
blue	**0.15112**
dolphin	0.10289
tons	0.04181
water	**0.03645**
white	**0.02144**
large	**0.01930**
metric	0.01930
brevicauda	0.01501
fish	**0.01287**
harpoon	0.00858

(b) Our Proposed Model

Fig. 11 comparison the most relevant latent topics of object category 'dolphin'

6. CONCLUSIONS

In this paper, we deal with two research issues, i.e. the automation of visual attribute identification and semantic relation learning between visual attributes and object categories. The contribution is two-fold, firstly, we provide uniform framework to reliably extract both categorical attributes and depictive attributes. Secondly, we incorporate the obtained semantic associations between visual attributes and object categories into a text-based topic model and extract descriptive latent topics from natural language knowledge base. Specifically, we show that in mining large scale text corpora of natural language descriptions, the relation between semantic visual attributes and object categories can be encoded as Must-Links and Cannot-Links, which can be represented by Dirichlet-Forest prior. To reduce the amount of manual supervision and labeling in large-scale image categorization, a semi-supervised training framework using soft-margin semi-supervised SVM classifier is introduced. Last but not least, automatically extracted visual attributes are used in a posterior object-attribute prediction model to further improve the performance of object categorization. Experimental results show that the proposed model achieves better ability in describing object-related attributes and makes the inferred latent topics more descriptive.

7. ACKNOWLEDGMENTS

This work was supported in part by NSF IIP 1160960NSF CCF 1049864, NSF CCF 0905291, Project in the National Science & Technology Pillar Program in the Twelfth Five-year Plan Period (No. 2012BAK24B01),and NSFC 90920005.

8. REFERENCES

[1] Andrzejewski D, Zhu X, Craven M. (2009) Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In Proceedings of the 26th Annual international Conference on Machine Learning (Montreal, Quebec, Canada, June 14 - 18, 2009). ICML '09, vol. 382. ACM, New York, NY, 25-32.

[2] B. Schoelkopf, A. Smola, Learning with Kernels, 644, MIT Press, Cambridge, MA (2002).

[3] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In CVPR, 2009.

[4] C. Siagian and L. Itti, Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention, IEEE TPAMI, pp. 300-312, 2007.

[5] C. Wang, L. Zhang, and H. Zhang. Learning to reduce the semantic gap in web image retrieval and annotation. In SIGIR, 2008.

[6] D. G. David G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 2004.

[7] H. Wang, X. Jiang, L.-T. Chia, and A.-H. Tan. Ontology enhanced web image retrieval: aided by wikipedia & spreading activation theory. In MIR, 2008. 2

[8] J. Deng, A. Berg, K. Li and L. Fei-Fei, What does classifying more than 10,000 image categories tell us? Proceedings of the 12th European Conference of Computer Vision (ECCV). 2010.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

[10] J. Platt. Probabilistic outputs for support vector machines and comparison to regularize likelihood methods. In Advances in Large Margin Classifiers, pages 61–74, 2000.

[11] L. Li, C. Wang, Y. Lim, D. Blei and L. Fei-Fei. Building and Using a Semantivisual Image Hierarchy. IEEE Computer Vision and Pattern Recognition (CVPR). 2010.

[12] M. Belkin, P. Niyogi & V. Sindhwani (2004) Manifold Regularization: A Geometric Framework for Learning for Examples. Technical Report TR-2004- 06, Dept. of Computer Science, Univ. of Chicago.

[13] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, What helps where - and why? Semantic relatedness for knowledge transfer. In Proceedings of CVPR. 2010, 910-917.

[14] O. Russakovsky and L. Fei-Fei, Attribute Learning in Large-scale Datasets. Proceedings of the 12th European Conference of Computer Vision (ECCV), 1st International Workshop on Parts and Attributes. 2010.

[15] T. Deselaers and V. Ferrari. Visual and Semantic Similarity in ImageNet. IEEE Computer Vision and Pattern Recognition (CVPR 2011), Colorado Springs, CO, USA, pages 1777-1784 June 2011.

[16] T. P. Minka. Estimating a dirichlet distribution. http://research. Microsoft.com/en-us/um/people/minka/papers/dirichlet, 2009.

[17] T. P. Minka, "The dirichlet-tree distribution," in http://research.microsoft.com/~minka/papers/dirichlet/minkadirtree.pdf, 1999.

[18] X. Chen, X. Hu, Z. Zhou, C. Lu, G. Rosen, T. He, E.K. Park, A Probabilistic Topic-Connection Model for Automatic Image Annotation, the 19th ACM Conference on Information and Knowledge Management (ACM CIKM 2010), Oct 26-29, 2010, Toronto, Canada. pp. 899-908

[19] X. Hu, X. Zhang, C. Lu, E.K. Park, and X. Zhou, Exploiting Wikipedia as external knowledge for document clustering. In Proceedings of KDD. 2009, 389-396

[20] X. Wang, T. X. Han and S. Yan, “An HOG-LBP Human Detector with Partial Occlusion Handling,” IEEE International Conference on Computer Vision (ICCV 2009), Kyoto, 2009.

[21] Y. Cheng: Mean Shift, Mode Seeking, and Clustering. IEEE Transaction on Pattern Analysis and Machine Intelligence 17(8) (1995) 790-799