Computer Sciences Department

The Multimodal Focused Attribute Model: A Nonparametric Bayesian Approach to Simultaneous Object Classification and Attribute Discovery

Jake Rosin
Charles R. Dyer
Xiaojin Zhu

Technical Report #1697

January 2012

The Multimodal Focused Attribute Model: A Nonparametric Bayesian Approach to Simultaneous Object Classification and Attribute Discovery

Jake Rosin, **Charles R. Dyer** and **Xiaojin Zhu**
Department of Computer Sciences
University of Wisconsin–Madison
Madison, WI 53706 , USA
{rosin,dyer,jerryzhu}@cs.wisc.edu

Abstract

A nonparametric Bayesian model for attribute-based object recognition and image-based class attribute inference is presented. This model draws on existing work in Bayesian nonparametrics such as the focused topic model *[25]. The model allows either the classification of objects or the inference of attributes over known classes (or both simultaneously). Attributes inferred from image datasets allow an improvement in classification accuracy when combined with attributes from other sources, including "layperson" knowledge and existing inference methods.*

1. Introduction

Recent object recognition techniques provide effective classification by using machine learning techniques (e.g., support vector machines), which operate over low-level image features. These low-level features can be efficiently extracted from both labeled training examples and arbitrary test images. Unfortunately, training models to recognize a specific object class requires a large number of labeled images, and the work of gathering these examples makes extending recognition systems to cover new object classes a difficult and time-consuming process.

Attribute-based recognition methods resolve this problem by placing an abstraction layer between low-level features and high-level object recognition systems. Object classes are assumed to possess "attributes" (fixed for each class), with each attribute being recognizable through its low-level features. The learned visual models for attributes of the training classes can then be applied to recognize *new* classes, replacing the need to provide explicit class examples with the simpler task of describing new classes in terms of a set of known attributes. Although this approach eases the burden on users interested in detecting novel object classes, existing attribute systems impose restrictions on those new classes - they must be describable in terms of only previously-trained attributes, and their attribute associations must be known in advance of recognition.

We have designed, implemented and tested a probabilistic model for nonparametric Bayesian inference in the setting of multimodal attribute-based object recognition. Our model enables zero-shot recognition through cross-class attribute transfer concurrent with attribute discovery. The nonparametric design allows entirely new visual attributes and object classes to be inferred from image data alone, or with the aid of textual class examples. Our approach combines the learning of both *class-attribute associations* and *attribute appearance models* into a single probabilistic procedure. Among other advantages, this greatly reduces the human effort needed to introduce new object classes to a recognition system by allowing any combination of labeled examples and partial sets of known attribute associations (presence or absence) to be provided as input.

To evaluate this model, we created eight datasets for attribute-based object recognition using images from ImageNet [4] synsets, including known attributes for each category. Preliminary tests show promise in two domains: object recognition, including recognition in the zero-shot learning setting, and attribute inference, wherein visually significant class attributes are inferred from image data. Attributes inferred by our method produced classification results comparable to attributes chosen "by hand," and provide new information not present in either hand-selected attributes or attributes inferred using other existing methods.

2. Related Work

2.1. Attribute Methods

Attribute-based methods are a recent approach to vision tasks such as object recognition and face verification [11, 10, 5]. These methods represent an object class as a unique collection of attributes, possibly with associ-

ated (nonnormalized) weights. The *Direct Attribute Projection* (DAP) method of Lampert et al. [12] provides good results for zero-shot learning by using per-class binary attribute vectors in lieu of labeled images for text classes. Attribute methods allow zero-shot recognition through the use of attribute detectors, which may be trained using examples from other object classes than those to be recognized. For example, a "stripe detector" trained on zebra images could be used to detect a tiger's stripes as well. DAP uses independently trained SVMs as attribute detectors, combining their outputs using a simple probabilistic formula.

Applying these detectors to a new object class requires not only the prior construction of all necessary detectors, but also requires knowledge of which attributes are associated with the new class (and for some methods, the degree of association). Human supervision shifts from the task of finding and labeling class examples to providing expert knowledge of class/attribute associations. Automatic inference of these attribute associations has been studied as a separate task from classification, but not integrated with the recognition step [2, 19, 18]. A major contribution of this paper is to provide a unified model which probabilistically provides both the attributes belonging to a class and the classification of example images (or documents).

Most attribute-based work has treated attributes as binary values, either present or absent for a given class or in a given image. Another approach treats attributes as present "by degrees," either by their prominence and size (for example, "nose" as an attribute of faces) or, more recently, by the uncertainty of their presence when compared against other examples. In particular, recent work by D. Parikh and K. Grauman [17] using relative attributes (e.g., "bears are furrier than giraffes") for classification shows superior accuracy to existing methods using binary attributes. Our work retains the binary assumption of earlier methods; the implications of relative attributes are discussed in Section 6.

2.2. Nonparametric Bayesian Methods

Bayesian inference requires a well-defined probabilistic model for the observable data and any unknown parameters. Unknown variables of interest—latent clusters, class labels, etc.—are examined through their posterior probabilities, conditioned on the observed data. Generative models provide insight into the underlying structure of the data, but this insight comes at the price of necessarily biasing inference towards certain interpretations of that data [15]; for example, by presupposing that the observations fall into a certain number of clusters.

Nonparametric (NP) Bayesian methods gain flexibility and avoid dependence on parametric assumptions by using Bayesian priors over not only parameter values, but over *the number and meaning of the parameters themselves*. These methods provide the same insight into the data's underlying structure, but do so by inferring that structure from the data.

NP Bayesian inference has been applied to many diverse problems, from modeling language acquisition in children [8] to database duplicate detection [14]. A hidden Markov model using Bayesian nonparametrics for speaker diarization gives state-of-the-art results on the benchmark NIST rich transcriptions database [6]. Recent work has begun applying NP Bayesian methods to problems in computer vision, with promising results [16, 26, 22].

3. Generative Model

3.1. Model Overview

Nonparametric Bayesian inference provides a means to automatically infer visually relevant class attributes concurrently with zero-shot classification. We model the visual (and textual) features and underlying classification of (annotated resp.) images, and the attributes associated with each object class, as the output of a unified generative process. In nonparametric fashion, the number of topics and categories is not specified in advance, emerging instead from the data being modeled. We base our *multimodal focused attribute model* (MFAM) on the *IBP compound Dirichlet process* (ICD) of Williamson et al. [25]; the complete generative story and technical details are given in Sections 3.3 and 3.4 respectively.

In short, the model defines a prior over distributions of an infinite number of text- and image-generating *attributes*, and distributions over an infinite number of object *classes*, with each class representing an independent sample of an almost surely finite number of attributes. Classes produce "*documents*," which are represented as an unordered collection of visual and textual "words" (e.g., word tokens, or clustered visual descriptors). Documents may be individual images with or without text annotation, text documents, a web page with both text and images, etc., while words are drawn from one or more fixed vocabularies. Each word in a document derives from one of the class's attributes, with the total number of those words being generated according to the summed "contribution strength" of the included topics. For instance, the attribute "stripes" for the class "tiger" could generate visual words in an image, as well as text (e.g. "camouflage") accompanying the image.

MFAM is applied by taking visual words extracted from images, any accompanying text (and data from any additional modalities), and any known classes or attributes as observed data. Following typical procedure, Gibbs sampling [24] is used to approximate the posterior distribution for unobserved variables: as in other NP Bayesian methods, inference using priors over infinite latent parameters is tractable because only finitely many parameters are responsible for any finite observation. MLE classification uses the approximate per-document class distribution resulting from

sampling. In practice, unobserved class attributes tend to stabilize very quickly, making a single-iteration snapshot sufficient for attribute inference.

3.2. Nonparametric Background

Our generative model uses ICD [25] as its basis for inference. ICD combines elements of a hierarchical Dirichlet process (HDP) and the Indian buffet process (IBP) [9] into an infinite latent topic model that decouples the between-document frequency of a topic from its within-document contribution.

The *Dirichlet process* (DP) [7] determines a distribution over distributions: $G_0 \sim \text{DP}(\zeta_0, H)$, with H the base probability measure and ζ_0 a concentration parameter. Draws from the DP sample infinite mixtures of component distributions, each drawn from H, and mixed according to ζ_0. In a hierarchical Dirichlet process [21], the per-data distribution G_m is sampled from $DP(\zeta_1, G_0)$. In typical modeling scenarios this distribution is over latent topics, which themselves are each associated with a distribution over words (textual, visual, or other). One result of this hierarchical construction is the lack of distinction between a component's weight between data points and its weight within them: a component with low weight in G_0 is unlikely to contribute to many data points, and its contribution to those where it appears will be small. As Williamson et al. argued [25], this correlation is not always desirable: infrequent topics may dominate those few documents where they appear. By integrating an *Indian buffet process* (IBP) [9] into the model this correlation is removed.

The Indian buffet process defines a distribution over binary matrices with infinite columns but almost surely finite non-zero columns. For most IBP methods, including ICD, the rows of these matrices represent documents, and the columns represent latent topics—a 1 indicates that a document contains a topic. In ICD, a Dirichlet process determines the contribution strength ϕ_k of each component k when it is present, with IBP's binary matrix B determining the presence of that component in each document.

One derivation of IBP, with columns in order of strictly decreasing expected sum, is obtained via a "stick-breaking" construction [20]. For binary matrix B having M rows (i.e., documents), and column (i.e., topic) $k = 1, 2, \ldots$:

$$\mu_k \sim Beta(\alpha, 1) \tag{1}$$

$$\pi_k = \prod_{i=1}^{k} \mu_i \tag{2}$$

$$b_{mk} \sim \text{ Bernoulli } (\pi_k) \text{ for } m = 1, 2, \ldots, M \tag{3}$$

Combining these, ICD samples topic distributions θ_i from Dirichlet$(\phi \circ b_i)$, with b_i a row in B and $\circ$ the Hadamard product. As IBP selects finite components for each row, θ_i may be equivalently sampled as the normalization of draws from independent Gamma distributions $\Gamma(\phi_k \cdot b_i k, 1)$ [25]. Williamson et al. used ICD as the basis of their *focused topic model* (FTM); the differences between it and MFAM are the following:

1. MFAM accounts for the presence of data in multiple modalities (for example, text and images) without unwanted statistical effects for documents where some kinds of data are absent.

2. FTM treats each document as an independent combination of latent topics; i.e., each row of B corresponds to a single document. MFAM generalizes this by allowing documents to share a latent class—i.e., a row in B—and thus share topics, where appropriate.

These differences provide MFAM with versatility FTM lacks. Simultaneous operation over documents in different modalities allows information transfer between multiple domains: for example, text descriptions such as encyclopedia entries may provide attribute information applicable to the classification of even unannotated images. By assigning a latent class to each document, MFAM allows similar documents to influence each other's attribute assignments, a feature FTM lacks; additionally, MFAM is directly applicable to image classification problems.

3.3. Multimodal Focused Attribute Model

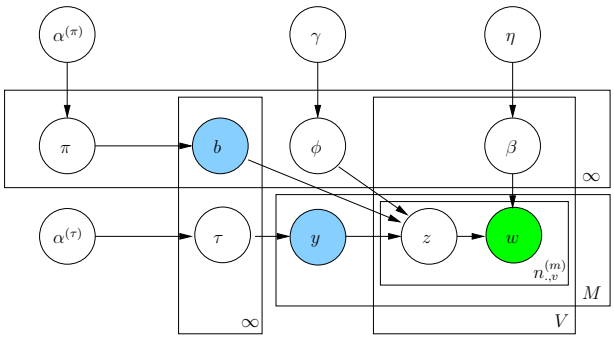

Figure 1. A graphical model of our approach. Observed visual and textual words are shown in green; object labels and class attributes, which are quantities of interest but may also be included as partial observations, are shown in blue.

A graphical model of our approach is given in Figure 1; the complete generative process for M documents is:

1. For latent topics $k = 1, 2, \ldots$,

   (a) Sample topic frequency with stick length π_k using hyperparameter $\alpha^{(\pi)}$

(b) Sample the relative mass $\phi_k \sim \text{Gamma}(\gamma, 1)$

(c) For vocabularies $v = 1, 2, \ldots, V$,

i. Sample the topic word distribution, $\beta_{k,v} \sim \text{Dirichlet}(\eta)$

2. For classes $c = 1, 2, \ldots$,

(a) Sample class probability with stick length τ_c using hyperparameter $\alpha^{(\tau)}$

(b) Sample a binary vector b_c according to $b_{c,k} \sim \text{Bernoulli}(\pi_k)$

3. For posts $m = 1, 2, \ldots, M$,

(a) Draw the class label $y_m \sim \tau$

(b) Sample the topic distribution $\theta_m \sim \text{Dirichlet}(b_c \circ \phi)$

(c) For vocabularies $v = 1, 2, \ldots, V$,

i. Draw $r_{m,v} \sim \text{Bernoulli}(\upsilon_v)$, the presence of v in the post

ii. Draw the number of words, $n_{.,v}^{(m)} \sim \text{NB}(\sum_k b_{y_m k}\phi_k\nu_v r_{m,v}, 1/2)$

iii. For each word $w_{mvi}, i = 1, 2, \ldots n_{.,v}^{(m)}$,

A. Draw the attribute index $z_{mvi} \sim \theta_m$

B. Draw the word $w_{mvi} \sim \beta_{z_{mvi},v}$

Some important features of this generative process are:

- An attribute's prevalence across classes is *not* correlated with its average proportion within documents where it is present. FTM similarly decouples topic prevalence across data and within data, although in that model documents do not have labels.
- The number of word types in a vocabulary is not correlated with the number of word tokens generated in documents representing it.
- Within a document, attributes contribute to all vocabularies in the same underlying proportions.
- Within a class, attributes contribute to documents in proportions that are conditionally independent given $b_c \cdot \phi$.
- Documents in entirely different modalities, for example text and image data, can be generated within the same class and using the same underlying attributes.

We take known class-attributes as partially observed rows in B; inferred attributes will populate columns beyond the observation. When a class including attribute k generates a document, k contributes to that document's words in proportion to $\phi_k \cdot \nu_v$, with v the vocabulary in question. A given post m contains elements in vocabulary v with probability υ_v; i.e., some posts may be text-only, some image-only.

Observed variables are $w_{\cdot}$, the text and visual words in each document; $n_{\cdot}$, the number of those words; and $r_{\cdot}$, representing whether a given document contains an image, text, or both. Partial observations include the class of each labeled document. As with many nonparametric Bayesian models, inference is accomplished using Gibbs sampling over the posterior probabilities of the latent and unobserved variables. Closed-form solutions for the posterior are available for most of the latent variables; for the rest, approximation methods are used. Approximations for otherwise intractable posteriors are used frequently in nonparametric Bayesian inference, with good results [21, 25, 15].

3.4. Sampling Unobserved Variables

To improve mixing time, we integrate out β, θ and τ, and sample only ϕ, π, B, z and y. Given the observed lengths and word tokens of each document, υ and ν are conditionally independent of the unobserved variables; maximum likelihood estimates are taken before sampling begins.

3.4.1 Sampling ϕ and π

ϕ and π are sampled conditioned on z, y and B. Adapting terminology from [25], an attribute is "active" if at least one element of its column in B is 1, and "inactive" otherwise.

The total number of words in the mth document belonging to vocabulary v and assigned to attribute k, assuming that vocabulary v is present in the document (which has independent probability υ_v), is distributed according to $\text{NB}(b_{y_m k}\phi_k\nu_v, 1/2)$. The joint probability of ϕ_k and the total number of words assigned to the kth topic, proportional to the probability ϕ_k, is

$$
\begin{aligned}
&p(\phi_k, n_{k,\cdot}^{(\cdot)} \mid \Psi) \\
&\quad = p(\phi_k \mid \gamma) \prod_{m=1}^{M} \prod_{v=1}^{V} p(n_{k,v}^{(m)} \mid b_{y_m,k}, \phi_k, r_{m,v}) \\
&\quad = \frac{\phi_k^{\gamma-1} e^{-\phi_k}}{\Gamma(\gamma)} \times \\
&\qquad \prod_{m: b_{y_m,k}=1} \prod_{v: r_{m,v}=1} \frac{\Gamma(\phi_k\nu_v + n_{k,v}^{(m)})}{\Gamma(\phi_k\nu_v)\Gamma(n_{k,v}^{(m)} - 1) 2^{\phi_k\nu_v + n_{k,v}^{(m)}}}
\end{aligned}
$$

Monte Carlo methods can be used to sample from the posterior of ϕ_k (and the posterior of γ, if such sampling is desired).

To sample π_k, Williamson et al. used a semi-ordered stick breaking approach similar to that given by Teh et

al. [20]. We modify their approach to conform to our assumption that the rows of B represent object classes, not documents. Active features are distributed according to

$$p(\pi_k \mid B) \sim \text{Beta}\left(\sum_{c=1}^{C} b_{c,k}, 1 + C - \sum_{c=1}^{C} b_{c,k}\right)$$

and inactive features are irrelevant in our sampling procedure.

3.4.2 Sampling y

Our experiments deal primarily with object recognition, so we do not allow our implementation to infer entirely new classes. With this limitation, class labels y_m can be sampled from the multivariate Polya distribution:

$$\begin{aligned} p(y_m = c \mid \Psi) & \\ &\propto \left[\int_\tau p(y_m = c \mid \tau) p(\tau \mid \delta, y_{-m})\right] \\ &\quad \times \left[\int_{\theta_m} p(z_{m..} \mid \theta_m) p(\theta_m \mid b_c, \phi, z_{-m:y=c,..,.})\right] \\ &= MP(y_m = c; \delta, y_{-m}) MP(z_{m..}; b_c \cdot \phi). \end{aligned}$$

However, this formulation gives extremely poor mixing results, especially for large documents—consider for example that the posterior of any class c' that lacks attributes present in $z_{m..}$ will be zero, even if that class could easily explain the observed words $w_{m..}$ using other attributes.

Instead, we integrate over $z_{m..}$ and sample the document class conditioned on the observed words for that document. This gives

$$\begin{aligned} p(y_m =& c \mid \Psi) \\ &\propto \left[\int_\tau p(y_m = c \mid \tau) p(\tau \mid \delta, y_{-m})\right] \\ &\quad \times \left[\int_{\theta_m} p(w_{m,..,.} \Psi, \theta_m) p(\theta_m \mid \phi, b_c)\right] \\ &= MP(y_m = c; \delta, y_{-m}) \\ &\quad \times \int_{\theta_m} p(w_{m..} \Psi, \theta_m) \text{Dirichlet}(\theta_m; \phi \cdot b_c) \end{aligned}$$

where $p(w_{m,..,.}\Psi, \theta_m)$ is the product over vocabularies v of the multinomial p.d.f. with input distribution $\beta^*_{m,v}$,

$$\beta^*_{mvi} = \sum_k p(w_i \mid \eta, z_{-m}) \theta_{m,k}.$$

The integration over θ_m is analogous to one used in Latent Dirichlet Allocation [3], where it is approximated using variational inference. The same technique can be applied here.

3.4.3 Sampling z

The conditional distribution of the topic assignment of word w_{mvi} depends on the attribute proportion for that document θ_m, which we marginalize away:

$$\begin{aligned} p(z_{mvi} = k \mid & z_{\neg(mvi)}, w_{mvi}, w_{\neg(mvi)}, \Psi) \\ &\propto p(w_{mvi} \mid z_{mvi} = k, z_{\neg(mvi)}, w_{\neg(mvi)}) \\ &\quad \times p(z_{mvi} = k \mid z_{\neg(mvi)}, \Psi) \\ &\propto (n^{(w_{mvi})}_{k,\neg(mvi)} + \eta) MP(z_{mvi} = k; b_c \cdot \phi, z_{m.\neg i}) \end{aligned}$$

where $MP(z_{mvi} = k; b_c \cdot \phi, z_{m.\neg i})$ is the probability $z_{mvi} = k$ according to the multivariate Polya distribution with prior $b_c \cdot \phi$ and observations $z_{m.\neg i}$, i.e., the other attribute assignments for document m.

Note that this straightforward sampling procedure differs greatly from the approximation procedure used in FTM; this difference arises from the fact that here we explicitly sample B, whereas in FTM B is integrated out [25].

3.4.4 Sampling B

Active columns of B, i.e., those columns k where ϕ_k and π_k are known, are sampled according to

$$\begin{aligned} p(B_{c,k} = 1 & | \Psi) \\ &\propto \pi_k \prod_{m:y_m=c} \int_{\theta_m} p(w_{m..} \Psi_{\neg c}, \theta_m) \text{Dirichlet}(\theta_m; \phi \cdot b^*_c) \end{aligned}$$

where here b^*_c is b_c with $b_{c,k}$ set to 1. The integration is otherwise the same as used when sampling y, except $p(w_{m..}\Psi_{\neg c}, \theta_m)$ excludes all documents with label c in calculating β^*_{mvi}.

The prior over the number of inactive columns to make active in row c is Poisson($\alpha/|C|$), from the Indian Buffet Process model [9]. The posterior is thus:

$$\begin{aligned} & p(l \text{ new columns for } c) \\ &\propto \text{Poisson}(l \mid \alpha/|C|) \\ &\quad \times \int_{\phi^l} p(\phi^l \mid \gamma) \int_{\theta_m} p(w_{m..} \Psi_{\neg c}, \theta_m) \\ &\quad \text{Dirichlet}(\theta_m; [\phi \phi^l] \cdot [b_c 1^l]) \end{aligned}$$

with the outer integration being over the ϕ values for the l new attributes.

4. Datasets for Empirical Testing

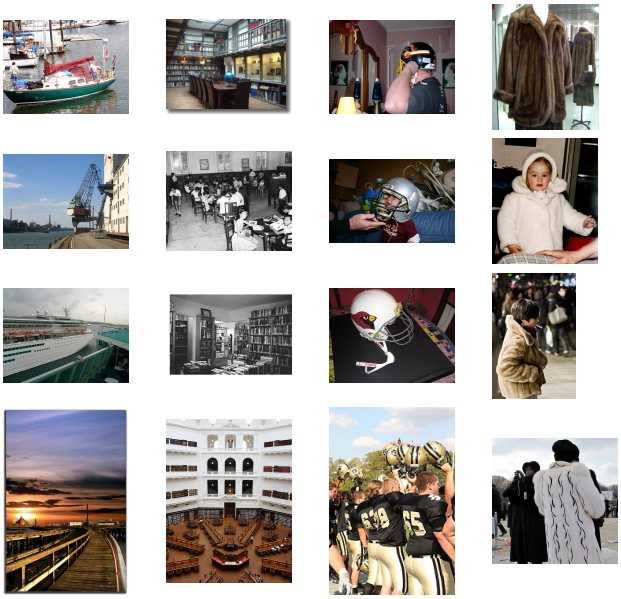

Figure 2. Examples from two of eight datasets constructed from the ILSVRC2010 dataset. Left: images from the "structures" dataset; columns represent the classes "dock" and "library." Right: images from the "clothing" dataset; columns are "football helmet" and "fur coat."

We constructed eight themed image datasets, generally applicable to attribute-based recognition tasks. Each dataset is comprised of approximately 6,000 images taken from 50 object categories in ImageNet [4], along with a set of between 20 and 60 manually-specified, theme-relevant attributes and known class-attribute associations. We refer to individual datasets by their theme, defined as the lowest common ImageNet ancestor of all included categories. Examples are given in Figure 2. Local color histograms [23] were extracted from the images across a 15-pixel grid, and clustered using k-means into a 500 word color vocabulary. SIFT [13] features, extracted from the same images and quantized into 1,000 bins, were provided as part of the ImageNet Large Scale Visual Recognition Challenge 2010 [1] and taken as a second visual vocabulary.

Some examples of theme-relevant attributes include "stripes", "claws" and "aquatic" for the *animals* dataset, and "user-propelled," "enclosed" and "war" for the *vehicles* dataset. Unlike the *Animals with Attributes* (AwA) dataset by Lampert et al. [12], these attributes (and class-attribute associations) represent layperson knowledge of the classes. AwA used attributes and associations drawn from a large existing body of research. Obviously, attributes chosen with care and expertise will be more effective for classification; however, we believe this level of effort would be very rarely justifiable in real-world zero-shot learning applications, as spending the required time collecting labeled examples of the classes of interest would likely produce a better result. Our "layperson knowledge" attributes are thus more representative of real-world zero-shot learning scenarios.

	textbfhand **-selected**	**"Yahoo/Flickr"**	**"parts"**
animal	0.321	0.194	0.145
clothing	0.235	0.156	0.141
device	0.246	0.102	0.129
equipment	0.343	0.360	0.191
implement	0.130	0.135	0.123
produce	0.297	0.175	0.123
structure	0.271	0.264	0.153
vehicle	0.314	0.272	0.211
Mean	0.270	0.207	0.152

Table 1. Multiclass accuracy for Direct Attribute Projection using different attribute sets. Zero-shot classification with 40 training and 10 test classes; chance is 0.1.

We also considered existing methods for automatic attribute discovery as an alternative to our hand-selected attributes. The attribute discovery code provided by Rohrbach et al. [18] was used to generate two attribute sets for each of our datasets. The first, "parts," was comprised of object meronyms discovered within the WordNet hierarchy. These meronyms represent object components which may be represented as distinct image regions (e.g. the meronyms "wing," "tail ," or "beak") or more generally as a non-localized influence on image color and texture (e.g. "thick skin," "hair"). The second attribute set, "Yahoo/Flickr," used the same list of attributes chosen by-hand when constructing the datasets, but with the class-attribute associations inferred using result counts from Flickr and Yahoo image searches. This approach more closely simulates the typical usage case for image classification, where a user may be willing to provide a list of candidate attributes, but not a complete matrix of class-attribute associations. The specific procedure for generating class-attribute associations is described in [18].

We compared DAP classification performance (with the same 40 training, 10 test class split used in all later experiments) over these three attribute sets, for each dataset. Complete results are shown in Table 1; overall our hand-selected attributes gave the best performance by a significant margin, and except where otherwise noted they were used for all further experimental trials.

The images chosen did not have accompanying metadata (e.g., text captions). As described above, our generative model could easily account for such textual data as additional observations.

5. Experimental Results

We implemented MFAM for inference using Gibbs sampling, using several approximations for efficiency and ease of computation. We tested the model in two domains: zero-shot learning via attributes, and automatic attribute inference.

5.1. Approximations Used

In sampling z, the count $n_{k,\neg(mvi)}^{(w_{mvi})}$, the number of times attribute k has produced the word type w_{mvi} across the entire corpus with the exception of token w_{mvi} itself, is updated only once per document. For a large corpus with many documents this should have very little effect on the resulting probabilities, and empirical tests showed this optimization does not effect classification results.

Although variational inference as described in the appendix of [3] can be used to approximate

$$\int_{\theta_m} p(w_{m..}\Psi, \theta_m)\text{Dirichlet}(\theta_m; \phi \cdot b_c)$$

when sampling y and B, the current implementation instead approximated the integration using random sampling of θ_m. Similarly, the outer integration over ϕ^l when sampling new columns was approximated using random sampling.

5.2. Zero-shot Learning

For each dataset described above, we provided category labels for images across 40 categories, and the full set of attribute associations for all 50 categories. The number of additional inferred attributes was limited to at most 10. We ran Gibbs sampling for 400 total iterations; only images from the 40 labeled classes were considered for the first 200, to provide better initialization of latent parameters. All but the last 100 iterations were discarded as burn-in; classification used the mode of the sampled class labels.

The model produced image classification results with an average multiclass accuracy of 0.193 (chance: 0.1). Cross-class zero-shot test accuracy on individual datasets varied from 0.2455 (animals) to 0.104 (articles of clothing). These results, significantly better than chance for most tested datasets, show that our model is capable of zero-shot recognition through attribute transfer. However, applying Direct Attribute Projection [12] to the same data produced better results: an average multiclass accuracy of 0.27, with best performance of 0.34 (equipment). Complete results are shown in Table 2.

One possible advantage of Direct Attribute Projection to account for this is, counter-intuitively, its independent examination of each attribute. DAP determines the likelihood of an attribute's presence through examination of every available image descriptor (visual word), whereas MFAM, by assigning attributes to visual words, makes the implicit assumption that each descriptor is derived from exactly one attribute. In the cases of some attributes, for example "spots" and "fur" for the class "leopard," it is obvious that no image pixel or collection thereof is entirely representative of one attribute and not the other.

	MFAM	**DAP**
animal	0.246	0.321
clothing	0.105	0.235
device	0.147	0.246
equipment	0.233	0.343
implement	0.118	0.130
produce	0.205	0.297
structure	0.239	0.271
vehicle	0.246	0.314
Mean	0.192	0.270

Table 2. Zero-shot multiclass accuracy using 40 training, 10 test classes; chance 0.1. Hand-selected attributes used in both methods.

5.3. Attribute Discovery

We also applied our generative model to the attribute discovery task. We focused on the animal and structure datasets, for which our model gave good classification results. Input data was a 10% subsampling of all 50 image classes, including class labels. No attributes were taken as observations, and the number of inferred attributes was limited to 40. Examination after 10,000 Gibbs sampling iterations revealed that class attributes very quickly stabilized in both cases (with no changes in class-attribute associations over the last 5,000 iterations). These inferred attributes were used to apply DAP to the remaining 90% of images, using the same training / test class division as above.

Zero-shot multiclass accuracy is given in Table 3, with the hand-selected attribute results from Table 2 provided again for comparison. As shown, attributes inferred by MFAM are less effective for classification than those selected by hand, but substantially better than chance, showing that attributes can be effectively inferred from unannotated images.

Although classification using MFAM-inferred attributes did not meet the standard set by hand selected attributes, it is possible that they represent new information not present in the other available attribute sets. With each attribute represented as a binary vector over classes (an attribute is present in a class or it is not), attribute sets can be trivially combined without any information loss, by concatenating attribute matrices and removing redundant columns.

Table 4 shows the result of zero-shot classification using Direct Attribute Projection, with MFAM-inferred attributes combined with each of the hand selected, "parts," and "Ya-

	MFAM inferred	hand-selected
animal	0.292	0.321
clothing	0.243	0.235
device	0.302	0.246
equipment	0.125	0.343
implement	0.105	0.130
produce	0.065	0.297
structure	0.278	0.271
vehicle	0.106	0.314
Mean	0.1895	0.270

Table 3. Zero-shot multiclass accuracy using 40 training, 10 test classes; chance 0.1. DAP used for classification in both cases; MFAM-inferred attributes vs. hand-selected attributes.

	hand-selected	"Yahoo/Flickr"	"parts"
animal	0.362	0.251	0.214
clothing	0.291	0.201	0.214
device	0.295	0.266	0.195
equipment	0.370	0.348	0.236
implement	0.115	0.113	0.115
produce	0.319	0.244	0.065
structure	0.300	0.260	0.241
vehicle	0.301	0.261	0.209
Mean	0.294	0.243	0.207

Table 4. Multiclass accuracy for Direct Attribute Projection using different attribute sets, each combined with the attributes inferred by MFAM. Compare against Table 1.

hoo/Flickr" attribute sets. In each case, including MFAM-inferred attributes improved average classification accuracy over that of the original attribute set. This demonstrates that the attributes inferred using our method represent visual information not present in any of the available attribute sets.

Using the intuition that class attributes will be represented in examples across multiple modalities, previous attribute inference methods have focused on inference from sources such as web search results or WordNet meronyms. While the existence of labeled class images allows object models to be learned directly, these models can be applied only to examples using the same *representation* (for example, the same visual word vocabulary). Class attribute inference allows information transfer between multiple datasets in otherwise incompatible representations. Datasets represented by entirely nonoverlapping "vocabularies" could be combined as input to our generative model, or attributes inferred from one could be applied to others as a separate step. As demonstrated in Table 4, attributes inferred from a visual dataset provide information that may be difficult to acquire in any other way.

6. Future Directions

The MFAM approach presented in this paper for zero-shot object recognition has several advantages over previous methods including (1) it allows probabilistic attribute assignment, (2) attribute inference for new classes is fully integrated, and (3) it allows for the straightforward combination of multimodal documents and expert knowledge in the form of labeled examples and class-attribute associations. Possible applications of this model include its direct use for object classification (including zero-shot classification, as examined here), and the inference of class attributes for use in separate tasks or merely to gain insight into the object classes themselves. Used to infer class attributes for use in DAP, MFAM produces results comparable to manually-defined attributes using layperson knowledge, and is superior to automatic methods that rely on web results for inference. Although DAP using hand-selected attributes produces better classification results for the tested datasets than either MFAM classification or DAP classification using MFAM-inferred attributes, MFAM-inferred attributes provide new class information that can improve classification results using any other tested attribute set.

Further tests are needed to explore other applications. Class discovery from unlabeled image sets is possible, with the Bayesian priors serving to limit the number of new classes. Multimodal datasets should be developed to test attribute transfer across modalities: in addition to information transfer between text and image data, transfer between images in different representations – i.e., different using different feature representations such as SIFT vs. SURF – should be tested.

The nonparametric Bayesian approach could be extended to provide other benefits as well. For example, class discovery from unlabeled image sets is possible, with the Bayesian priors serving to limit the number of new classes. The visual vocabulary itself can be included as a product of the generative process by, e.g., representing visual features as draws from a mixture model, with a number of underlying mixture components determined by the data.

Finally, the work by D. Parikh and K. Grauman [17] demonstrates the utility of examining attributes within an example or class in relative terms, by comparison to other examples and classes (e.g. "bluer," "more furry"), rather than as a simple binary value. Although in this work we examine MFAM using binary attributes (as both input and output), the model itself treats attributes as probabilistic random variables. Parikh and Grauman discuss relative attributes in terms of expert (or lay-person) uncertainty – people may disagree as to whether, for example, a particular person is smiling in a particular picture. This uncertainty is directly represented in MFAM as posterior likelihood, and although this work does not examine that likelihood directly, the results presented in [17] suggest that just as

this work applied MFAM-inferred binary attributes to Direct Attribute Projection, future work should apply MFAM-inferred attribute probabilities as input to relative attribute methods.

7. Acknowledgments

This work was supported in part by the National Science Foundation under Grant No. IIS-1148012. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank Lampert et al. for supplying a Direct Attribute Projection implementation [12] as well as the *Animals with Attributes* dataset, Rohrbach et al. for supplying web-based attribute discovery code [18], and Williamson et al. for supplying the Focused Topic Model [25] implementation upon which our MFAM implementation is based.

References

[1] A. C. Berg, J. Deng, and L. Fei-Fei. ImageNet large scale visual recognition challenge 2010 (ILSVRC2010), 2010.

[2] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *Proc. 11th European Conf. on Computer Vision: Part I*, pages 663–676, 2010.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. Computer Vision and Pattern Recognition Conf.*, pages 248–255, 2009.

[5] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. Computer Vision and Pattern Recognition Conf.*, pages 1778–1785, 2009.

[6] E. Fox, E. Sudderth, M. Jordan, and A. Willsky. A sticky HDP-HMM with application to speaker diarization. *Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

[7] S. Ghoshal. Dirichlet process, related priors and posterior asymptotics. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics (Cambridge Series in Statistical and Probabilistic Mathematics)*, pages 35–79. Cambridge University Press, 2010.

[8] S. J. Goldwater. *Nonparametric Bayesian Models of Lexical Acquisition*. PhD thesis, Brown University, 2007.

[9] T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems 18*, pages 475–482, 2006.

[10] N. Kumar, P. N. Belhumeur, and S. K. Nayar. FaceTracer: A search engine for large collections of images with faces. In *Proc. 10th European Conf. on Computer Vision*, pages 340–353, 2008.

[11] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. 12th Int. Conf. on Computer Vision*, pages 365–372, 2009.

[12] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. Computer Vision and Pattern Recognition Conf.*, pages 951–958, 2009.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Computer Vision*, 60(2):91–110, 2004.

[14] N. E. Matsakis. *Active Duplicate Detection with Bayesian Nonparametric Models*. PhD thesis, MIT, 2010.

[15] P. Muller and F. Quintana. Nonparametric Bayesian data analysis. *Statistical Science*, 19(1):95–110, 2004.

[16] P. Orbanz and J. Buhmann. Nonparametric Bayesian image segmentation. *Int. J. Computer Vision*, 77:25–45, 2008.

[17] D. Parikh and K. Grauman. Relative attributes. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 503–510, 2011.

[18] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where – and why? Semantic relatedness for knowledge transfer. In *Proc. Computer Vision and Pattern Recognition Conf.*, 2010.

[19] O. Russakovsky and L. Fei-Fei. Attribute learning in large-scale datasets. In *Parts and Attributes Workshop at the European Conf. on Computer Vision*, 2010.

[20] Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. *J. Machine Learning Research*, 2:556–563, 2007.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. American Statistical Assoc.*, 101(476):1566–1581, 2006.

[22] M. K. Titsias. The infinite Gamma-Poisson feature model. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, 2008.

[23] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[24] B. Walsh. Markov Chain Monte Carlo and Gibbs Sampling, 2004. University of Arizona, Tucson.

[25] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In J. Fürnkranz and T. Joachims, editors, *Proc. 27th Int. Conf. on Machine Learning*, pages 1151–1158, 2010.

[26] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin. Non-parametric Bayesian dictionary learning for sparse image representations. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2295–2303. 2009.