

Weakly Supervised Learning of Objects, Attributes and their Associations

Zhiyuan Shi, Yongxin Yang, Timothy M. Hospedales, Tao Xiang

Queen Mary, University of London, London E1 4NS, UK
{z.shi,yongxin.yang,t.hospedales,t.xiang}@qmul.ac.uk

Abstract. When humans describe images they tend to use combinations of nouns and adjectives, corresponding to objects and their associated attributes respectively. To generate such a description automatically, one needs to model objects, attributes and their associations. Conventional methods require strong annotation of object and attribute locations, making them less scalable. In this paper, we model object-attribute associations from weakly labelled images, such as those widely available on media sharing sites (e.g. Flickr), where only image-level labels (either object or attributes) are given, without their locations and associations. This is achieved by introducing a novel weakly supervised non-parametric Bayesian model. Once learned, given a new image, our model can describe the image, including objects, attributes and their associations, as well as their locations and segmentation. Extensive experiments on benchmark datasets demonstrate that our weakly supervised model performs at par with strongly supervised models on tasks such as image description and retrieval based on object-attribute associations.

Keywords: Weakly supervised learning, object attribute associations

1 Introduction

Vision research is moving beyond simple classification, annotation and detection to encompass generating more structured and semantic descriptions of images. When humans describe images they use combinations of nouns and adjectives, corresponding to objects and their associated attributes respectively. For example, an image can be described as containing “a person in red clothes and a shiny car”. In order to imitate this ability, a computer vision system needs to learn models about objects, attributes, and their associations. Object-attribute associations is important for avoiding the mistakes such as “a shiny person and a red car”. Learning object-attribute association also provides new query capabilities, e.g., “find images with a furry brown horse and a red shiny car”.

There has been extensive work on detecting and segmenting objects [33, 6, 31] and describing specified objects and images in terms of semantic attributes [9, 36, 3, 32]. However, these tasks have previously been treated separately; jointly learning about and inferring object-attribute association in images with potentially multiple objects is much less studied. The few existing studies on modelling

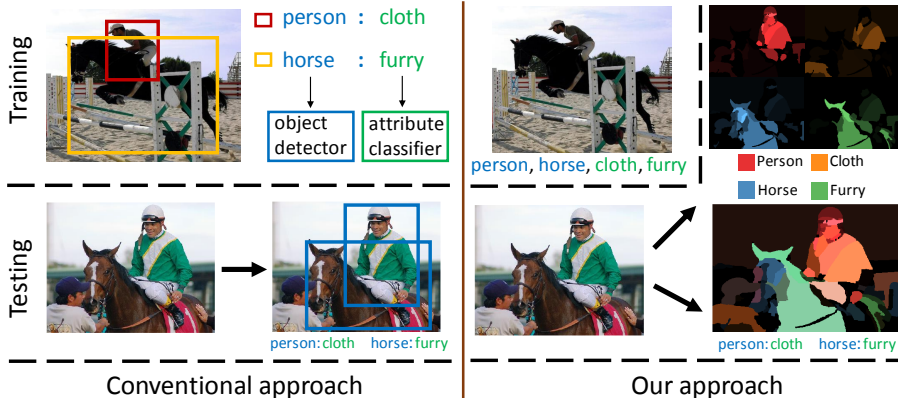


Fig. 1: Comparing our weakly supervised approach to object-attribute association learning to the conventional strongly supervised approach.

object-attribute association [17, 38, 20, 37, 36, 35] use fully annotated data [38, 20, 37] and/or are separately trained [17, 35]. In the conventional pipeline (Fig. 1) images are strongly labelled with object bounding boxes and associated attributes, from which object detectors and attribute classifiers are trained. Given a new image, the learned object detectors are first applied to find object locations, where the attribute classifiers are then applied to produce the object descriptions. However, there is a critical limitation of the existing approach: it requires strongly labelled objects and attributes. Considering there are over 30,000 object classes distinguishable to humans [18], even more attributes to describe them, and an infinite number of combinations, it is not scalable.

In this paper we propose to learn objects, attributes, and their associations from weakly labelled data. That is, images with object and attribute labels but not their associations nor their locations (see Fig. 1). Such weakly labelled images are abundant on media sharing websites such as Flickr. Therefore lack of training data would never be a problem. However, learning strong semantics, i.e. explicit object-attribute association from weakly labelled images is extremely challenging due to the label ambiguity: a real-world image with the tags “dog, white, coat, furry” could contain a furry dog and a white coat or a furry coat and a white dog. Furthermore, the tags/labels typically only describe the foreground/objects. There could be a white building in the background which is ignored by the annotator, and a computer vision model must infer that this is not what the tag ‘white’ refers to. Conventional methods cannot be applied without object locations and explicit object-attribute association being labelled.

To address the challenges of learning strong semantics from weak annotation, we develop a unified probabilistic generative model capable of jointly learning objects, attributes and their associations, as well as their location and segmentation. Our model is also able to learn from realistic images where there are multiple objects of variable sizes per image such as PASCAL VOC. More specif-

ically, our model generalises the non-parametric Indian Buffet Process (IBP) [13]. The IBP is chosen because it is designed for explaining multiple factors that simultaneously co-exist to account for the appearance of a particular image or patch, e.g., such factors can be an object and its particular texture and colour attributes. However, the conventional IBP is limited in that it is unsupervised and, as a flat model, applies to either patches or images, not both; it thus cannot be directly applied to our problem. To overcome these limitations, a novel model termed Weakly Supervised Stacked Indian Buffet Process (WS-SIBP) is formulated in this work. By introducing hierarchy into IBP, WS-SIBP is able to group data, thus allowing it to explain images as groups of patches, each of which has an inferred multi-label description vector corresponding to an object and its associated attributes. We also introduce weak image-level supervision, which is disambiguated into multi-label patch explanations by our WS-SIBP.

Modelling weakly labelled images using our framework provides a number of benefits: (i) By jointly learning multiple objects, attributes and background clutter in a single framework, ambiguity in each is explained away by knowledge of the other. (ii) The infinite number of factors provided by the non-parametric Bayesian framework allows structured background clutter of unbounded complexity to be explained away. (iii) A sparse binary latent representation of each patch allows an unlimited number of attributes to co-exist on one object. The aims and capabilities of our approach are illustrated schematically in Fig. 1, where weak annotation in the form of a mixture of objects and attributes is transformed into object and attribute associations with locations.

2 Related work

Learning objects and attributes A central task in computer vision is understanding image content. Such an understanding has been shown in the form of an image description in terms of nouns (object detection or region segmentation), and more recently adjectives (visual attributes) [9, 27]. Attributes have been used to describe objects [9, 34], people [3], clothing [4], scenes [36], faces [32], and video events [12]. However, most previous studies have learned and inferred object and attribute models separately, e.g., by independently training binary classifiers, and require strong annotations/labels indicating object/attribute locations and/or associations if the image is not dominated by a single object.

Learning object-attribute associations A few recent studies have learned object-attribute association explicitly [35, 16, 17, 36–38, 20]. Different from our approach, [35, 37, 38, 20] only trains and tests on unambiguous data, i.e. images containing a single dominant object, assumes object-attribute association is known at training; and moreover allocates exactly one attribute per object. [17] tests on more challenging PASCAL VOC data with multiple objects and attributes coexisting. However, their model is pre-trained on object and attribute detectors learned on strongly annotated images with object bounding boxes provided. [36] also does object segmentation and object-attribute prediction. But their model is learned from strongly labelled images in that object-attribute

association are given during training; and importantly prediction is restricted to object-attribute pairs seen during training. In summary none of the existing work learns object-attribute association from weakly labelled data as we do here.

Multi-attribute query Some existing work aims to perform attribute-based query [26, 15, 30, 32]. In particular, Recent studies have considered how to calibrate [30] and fuse [26] multiple attribute scores in a single query. We go beyond these studies in supporting conjunction of object+multi-attribute query. Moreover, existing methods either require bounding boxes or assume simple data with single dominant objects, and do not reason jointly about multiple attribute-object association. This means they would be intrinsically challenged in reasoning about (multi)-attribute-object queries on challenging data with multiple objects and multiple attributes in each image (e.g., querying furry brown horse, in a dataset with black horses and furry dogs in the same image). In other words, they cannot be directly extended to solve query by object-attribute association.

Probabilistic models for image understanding Discriminative kernel methods underpin many high performance recognition and annotation studies [9, 27, 3, 36, 28, 17, 5, 21, 6]. However the flexibility of generative probabilistic models has seen them successfully applied to a variety of tasks, especially learning structured scene representations, and weakly-supervised learning [12, 33, 31, 19]. These studies often generalise probabilistic topic models (PTM) [2]. However PTMs are limited for explaining objects and attributes in that latent topics are competitive - the fundamental assumption is that an object is a horse *or* brown *or* furry. They intrinsically do not account for the reality that it is *all* at once.

We therefore generalise instead the Indian Buffet Process (IBP) [7, 13]. The IBP is a latent feature model that can independently activate each latent factor, explaining imagery as a weighted sum of active factor appearances. However, conventional IBP is (i) fully unsupervised, and (ii) only handles flat data. Thus, it could explain patches or images, but not images composed of patches, thereby limiting usefulness for multiple object-attribute association within images. We therefore formulate a novel Weakly Supervised Stacked Indian Buffet Process (WS-SIBP) to model grouped data (images composed of patches), such that each patch has an infinite latent feature vector. This allows us to exploit image-level weak supervision, but disambiguate it to determine the best explanation in terms of which patches correspond to un-annotated background; which patch corresponds to which annotated object; and which objects have which attributes.

Weakly supervised learning Weakly supervised learning (WSL) has attracted increasing attention as the volume of data which we are interested in learning from grows much faster than available annotation. Existing studies have generally focused on WSL of objects alone [31, 19, 6], with limited work on WSL of attributes [36, 12]. Some studies have treated this as a discriminative multi-instance learning (MIL) problem [22, 6], while others leveraged PTMs [31, 19]. Weakly supervised localisation is a particularly challenging variant where images are annotated with objects, but absent bounding boxes means their location is unknown. This has been solved by sampling bounding boxes for MIL treatment

[6], or more ‘softly’ by PTMs [31]. In this paper we uniquely consider WSL of both objects, attributes, their associations and their locations simultaneously.

Our contributions In this paper we make three key contributions: (i) We for the first time jointly learn all object, attribute and background appearances, object-attribute association, and their locations from realistic weakly labelled images; (ii) We formulate a novel weakly supervised non-parametric Bayesian model by generalising the Indian Buffet Process; (iii) From this weakly labelled data, we demonstrate various image description and query tasks, including challenging tasks relying on predicting strong object-attribute association. Extensive experiments on benchmark datasets demonstrate that in each case our model is comparable to the strongly supervised alternatives and significantly outperforms a number of weakly supervised baselines.

3 Weakly Supervised Stacked Indian Buffet Process

We propose a non-parametric Bayesian model that learns to describe images composed of super-pixels/patches from weak object and attribute annotation. Each patch is associated with an infinite latent factor vector indicating if it corresponds to (an unlimited variety of) unannotated background clutter, or an object of interest, and what set of attributes are possessed by the object. Given a set of images with weak labels and segmented into super-pixels/patches, we need to learn: (i) which are the unique patches shared by all images with a particular label, (ii) which patches correspond to unannotated background, and (iii) what is the appearance of each object, attribute and background type. Moreover, since multiple labels (attribute and object) can apply to a single patch, we need to disambiguate which aspects of the appearance of the patch are due to each of the (unknown) associated object and attribute labels. To address all these learning tasks we build on the IBP [7] and introduce a weakly-supervised stacked Indian Buffet process (WS-SIBP) to model data represented as bags (images) of instances (patches) with bag-level labels (image annotations). This is analogous to the notion of documents in topic models [2].

3.1 Model formulation

First, we associate each object category and each attribute to a latent factor. If there are K_o object categories and K_a attributes, then the first $K_{oa} = K_o + K_a$ latent factors correspond to these. An unbounded number of further factors are available to explain away background clutter in the data. At training time, we assume a binary label vector $L^{(i)}$ for objects and attributes is provided for each image i . So $L_k^{(i)} = 1$ if attribute/object k is present, and zero otherwise. Also $L_k^{(i)} = 1$ for all $k > K_{oa}$. That is, without any labels, we assume all background types can be present. With these assumptions, the generative process (illustrated in Fig. 2) for image i represented as bags of patches $\mathbf{X}^{(i)} = \{\mathbf{X}_j^{(i)}\}$ is as follows:

For each latent factor $k \in 1 \dots \infty$:

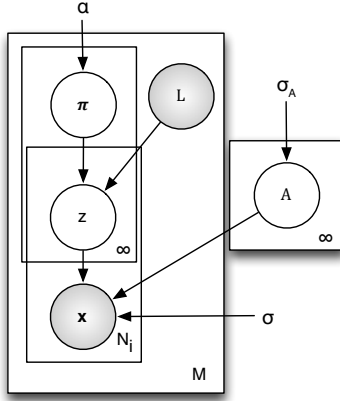


Fig. 2: The graphical model for WS-SIBP. Shaded nodes are observed.

1. Draw an appearance distribution mean $\mathbf{A}_k. \sim \mathcal{N}(0, \sigma_A^2 \mathbf{I})$.

For each image $i \in 1 \dots M$:

1. Draw a sequence of i.i.d. random variables $v_1^{(i)}, v_2^{(i)} \dots \sim \text{Beta}(\alpha, 1)$,
2. Construct an image prior $\pi_k^{(i)} = \prod_{t=1}^k v_t^{(i)}$,
3. Input weak annotation $L_k^{(i)} \in \{0, 1\}$,
4. For each super-pixel patch $j \in 1 \dots N_i$:
 - (a) Sample state of each latent factor $k: z_{jk}^{(i)} \sim \text{Bern}(\pi_k^{(i)} L_k^{(i)})$,
 - (b) Sample patch appearance: $\mathbf{X}_{j.}^{(i)} \sim \mathcal{N}(\mathbf{Z}_{j.}^{(i)} \mathbf{A}, \sigma^2 \mathbf{I})$.

where \mathcal{N} , Bern and Beta respectively correspond to Normal, Bernoulli and Beta distributions with the specified parameters; and the notation $\mathbf{X}_{j.}$ means the vector of row j in matrix \mathbf{X} . The Beta-Bernoulli and Normal-Normal conjugacy are chosen because they allow more efficient inference. α is the prior expected sparsity of annotations and σ^2 is the prior variance in appearance for each factor.

Denote hidden variables by $\mathbf{H} = \{\boldsymbol{\pi}^{(1)}, \dots, \boldsymbol{\pi}^{(M)}, \mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(M)}, \mathbf{A}\}$, images by $\mathbf{X} = \{\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(M)}\}$, and parameters by $\boldsymbol{\Theta} = \{\alpha, \sigma_A, \sigma, \mathbf{L}\}$. Then the joint probability of the variables and data given the parameters is:

$$\begin{aligned}
 p(\mathbf{H}, \mathbf{X} | \boldsymbol{\Theta}) &= \prod_{i=1}^M \left(\prod_{k=1}^{\infty} \left(p(\pi_k^{(i)} | \alpha) \prod_{j=1}^{N_i} p(z_{jk}^{(i)} | \pi_k^{(i)}, L_k^{(i)}) \right) \right. \\
 &\quad \left. \cdot \prod_{j=1}^{N_i} p(\mathbf{X}_{j.}^{(i)} | \mathbf{Z}_{j.}^{(i)}, \mathbf{A}, \sigma) \right) \prod_{k=1}^{\infty} p(\mathbf{A}_k | \sigma_A^2). \quad (1)
 \end{aligned}$$

Learning in our model aims to compute the posterior $p(\mathbf{H} | \mathbf{X}, \boldsymbol{\Theta})$ for: disambiguating and localising all the annotated ($L^{(i)}$) objects and attributes among

the patches (inferring $\mathbf{Z}_j^{(i)}$), inferring the attribute and background prior for each image (inferring $\boldsymbol{\pi}^{(i)}$), and learning the appearance of each factor (inferring \mathbf{A}_k).

3.2 Model learning

Exact inference for $p(\mathbf{H}|\mathbf{X}, \boldsymbol{\Theta})$ in our stacked IBP is intractable, so an approximate inference algorithm in the spirit of [7] is developed. The mean field variational approximation to the desired posterior $p(\mathbf{H}|\mathbf{X}, \boldsymbol{\Theta})$ is:

$$q(\mathbf{H}) = \prod_{i=1}^M (q_{\boldsymbol{\tau}}(\mathbf{v}^{(i)})q_{\boldsymbol{\nu}}(\mathbf{Z}^{(i)})q_{\boldsymbol{\phi}}(\mathbf{A})) \quad (2)$$

where $q_{\boldsymbol{\tau}}(v_k^{(i)}) = \text{Beta}(v_k^{(i)}; \tau_{k1}^{(i)} \tau_{k2}^{(i)})$, $q_{\boldsymbol{\nu}}(z_{jk}^{(i)}) = \text{Bernoulli}(z_{jk}^{(i)}; \nu_{jk}^{(i)})$, $q_{\boldsymbol{\phi}}(\mathbf{A}_k) = \mathcal{N}(\mathbf{A}_k; \boldsymbol{\phi}_k, \boldsymbol{\Phi}_k)$ and the infinite stick-breaking process for latent factors is truncated at K_{max} , so $\pi_k = 0$ for $k > K_{max}$. A variational message passing (VMP) strategy [7] can be used to minimise the KL divergence of Eq. (2) to the true posterior. Updates are obtained by deriving integrals of the form $\ln q(\mathbf{h}) = E_{\mathbf{H} \setminus \mathbf{h}}[\ln p(\mathbf{H}, \mathbf{X})] + C$ for each group of hidden variables \mathbf{h} . These result in the series of iterative updates given in Algorithm 1, where $\varphi(\cdot)$ is the digamma function; and $q_{ms}^{(i)}$ and $E_{\nu}[\log(1 - \prod_{t=1}^k v_t^{(i)})]$ are given in [7]. In practice, the truncation approximation means that our WS-SIBP runs with a finite number of factors K_{max} where truncation factor K_{max} can be freely set so long as it is bigger than the number of factors needed by both annotations and background clutter (K_{bg}), i.e., $K_{max} \gg K_o + K_a + K_{bg}$. Despite the combinatorial nature of the object-attribute association and localisation problem, our model is of complexity $\mathcal{O}(MNDK_{max})$ for M images with N patches, D feature dimension and K_{max} truncation factor.

3.3 Inference for test data

At testing time, the appearance of each factor k , now modelled by sufficient statistics $\mathcal{N}(\mathbf{A}_k; \boldsymbol{\phi}_k, \boldsymbol{\Phi}_k)$, is assumed to be known (learned from the training data), while annotations for each test image $L_k^{(i)}$ will need to be inferred. Thus Algorithm 1 still applies, but without the appearance update terms and with $L_k^{(i)} = 1 \forall k$, to reflect the fact that all the learned object, attribute, and background types could be present without any prior knowledge.

3.4 Applications of the model

Given the learned model applied to testing data, we can perform the following tasks: **Free Annotation:** This is to describe an image using a list of nouns and adjectives corresponding to objects and their associated attributes, as well as locating them. To infer what objects are present in image i , the first K_o latent

Algorithm 1: Variational Inference for WS-SIBP

```

while not converge do
  for k = 1 to Kmax do
     $\phi_k = (\frac{1}{\sigma^2} \sum_{i=1}^M \sum_{j=1}^{N_i} \nu_{jk}^{(i)} (\mathbf{X}_j^{(i)} - \sum_{l:l \neq k} \nu_{jl}^{(i)} \phi_l)) (\frac{1}{\sigma_A^2} + \frac{1}{\sigma^2} \sum_{i=1}^M \sum_{j=1}^{N_i} \nu_{jk}^{(i)})^{-1}$ 
     $\Phi_k = (\frac{1}{\sigma_A^2} + \frac{1}{\sigma^2} \sum_{i=1}^M \sum_{j=1}^{N_i} \nu_{jk}^{(i)})^{-1} \mathbf{I}$ 
  end
  for i = 1 to M do
    for k = 1 to Kmax do
       $\tau_{k1}^{(i)} = \alpha + \sum_{m=k}^{K_{max}} \sum_{j=1}^{N_i} \nu_{jm}^{(i)} + \sum_{m=k+1}^{K_{max}} (N_i - \sum_{j=1}^{N_i} \nu_{jm}^{(i)}) (\sum_{s=k+1}^m q_{ms}^{(i)})$ 
       $\tau_{k2}^{(i)} = 1 + \sum_{m=k}^{K_{max}} (N_i - \sum_{j=1}^{N_i} \nu_{jm}^{(i)}) q_{mk}^{(i)}$ 
      for j = 1 to Ni do
         $\eta = \sum_{t=1}^k (\varphi(\tau_{t1}^{(i)}) - \varphi(\tau_{t2}^{(i)})) - \mathbb{E}_v[\log(1 - \prod_{t=1}^k v_t^{(i)})]$ 
         $-\frac{1}{2\sigma^2} (\text{tr}(\Phi_k) + \phi_k \phi_k^T - 2\phi_k (\mathbf{X}_j^{(i)} - \sum_{l:l \neq k} \nu_{jl}^{(i)} \phi_l)^T)$ 
         $\nu_{jk}^{(i)} = \frac{L_k^{(i)}}{1+e^{-\eta}}$ 
      end
    end
  end
end
end

```

factors of the inferred $\pi^{(i)}$ are thresholded or ranked to obtain a list of objects. This is followed by locating them via searching for the patches j^* maximising $\mathbf{Z}_{jk}^{(i)}$, then thresholding or ranking the K_a attribute latent factors in $\mathbf{Z}_{j^*k}^{(i)}$ to describe them.

Annotation given object names: This is a more constrained variant of the free annotation task above. Given a named (but not located) object k , its associated attributes can be estimated by first finding the location as $j^* = \arg \max_j \mathbf{Z}_{jk}^{(i)}$, then the associated attributes by $\mathbf{Z}_{j^*k}^{(i)}$ for $K_o < k \leq K_o + K_a$.

Object+Attribute Query: Images can be queried for a specified object-attribute conjunction $\langle k_o, k_a \rangle$ by searching for $i^*, j^* = \arg \max_j \mathbf{Z}_{jk_o}^{(i)} \cdot \mathbf{Z}_{j^*k_a}^{(i)}$.

4 Experiments

Datasets:

Various object and attribute datasets are available such as aPascal, ImageNet, SUN [24] and Awa [18]. We use aPascal because it has multiple objects per image; and ImageNet due to sharing attributes widely across categories.

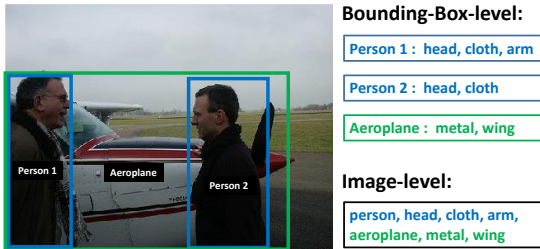


Fig. 3: Strong bounding-box-level annotation and weak image-level annotations for aPascal are used for learning strongly supervised models and weakly supervised models respectively.



Fig. 4: 43 subordinate classes of dog are converted into a single entry-level class ‘dog’.

aPascal: This dataset [9] is an attribute labelled version of PASCAL VOC 2008. There are 4340 images of 20 object categories. Each object is annotated with a list of 64 attributes that describe them by shape (e.g., isBoxy), parts (e.g., hasHead) and material (e.g., isFurry). In the original aPascal, attributes are strongly labelled for 12695 object bounding boxes, i.e. the object-attribute association are given. To test our weakly supervised approach, we merge the object-level category annotations and attribute annotations into a single annotation vector of length 84 for the entire image. This image-level annotation is much weaker than the original bounding-box-level annotation, as shown in Fig. 3. In all experiments, we use the same train/test splits provided by [9].

ImageNet Attribute: This dataset [27] contains 9600 images from 384 ImageNet synsets/categories. We ignore the provided bounding box annotation. Attributes for each bounding box are labelled as 1 (presence), -1 (absence) or 0 (ambiguous). We use the same 20 of 25 attributes as [27] and consider 1 and 0 as positive examples. Many of the 384 categories are subordinate categories, e.g. dog breeds. However, distinguishing fine-grained subordinate categories is beyond the scope of this study. We are interested in finding a ‘black-dog’ or ‘white-car’, rather than ‘black-mutt’ or ‘white-ford-focus’. We thus convert the 384 ImageNet categories to 172 entry-level categories using [23] (see Fig. 4). We evenly split each class to create the training and testing sets.

Features:

We first convert each image i to N_i super-pixels/patches by a recent segmentation algorithm [1]. We set the segmentation threshold to 0.1 to obtain a single over-segmentation from the hierarchical segmentation for each image. Each segmented patch is represented using two types of normalised histogram features: SIFT and Color. (1) SIFT: we extract regular grid (every 5 pixels) colorSIFT [29] at four scales. A 256 component GMM model is constructed on the collection of ColourSIFTs from all images. We compute Fisher Vector + PCA for all regular points in each patch following [14]. The resulting reduced descriptor is 512-D for every segmented region. (2) Colour: We convert the image to quantised LAB

space $8 \times 8 \times 8$. A 512-D color histogram is then computed for each patch. The final 1024-D feature vector concatenates SIFT and Colour features together.

Compared Methods:

We compare our WS-IBP to one strongly supervised model and three weakly supervised alternatives:

Strongly supervised model: A strongly supervised model uses bounding-box-level annotation. Two variants are considered for the two datasets respectively.

DPM+s-SVM: for aPascal, both object detector and attribute classifier are trained from fully supervised data (i.e. Bounding-Box-level annotation in Fig. 3). Specifically, we use the 20 pre-trained DPM detectors from [10] and 64 attribute classifiers from [9]. **GT+s-SVM:** for ImageNet attributes, there is not enough data to learn 172 strong DPM detectors as in aPascal. So we use the ground truth bounding box instead assuming we have perfect object detectors, giving a significant advantage to this strongly supervised model. We train attribute classifiers using our feature and liblinear SVM [8]. These strongly supervised models are similar in spirit to the models used in [17, 36, 35] and can provide a performance upper bound for the weakly supervised models compared.

w-SVM [9, 27]: In this weakly-supervised baseline, both object detectors and attribute classifiers are trained on the weak image-level labels as for our model (see Fig. 3). For aPascal, we train object and attribute classifiers using the feature extraction and model training codes (which is also based on [8]) provided by the authors of [9]. For ImageNet, our features are used, without segmentation.

MIML [41]: This is the multi-instance multi-label (MIML) learning method in [41]. In a way, our model can also be considered as a MIML method with each image a bag and each patch an instance. The MIML model provides a mechanism to use the same super-pixel/patch based representation for images as our model, thus providing the object/attribute localisation capability as our model does.

w-LDA: Weakly-supervised Latent Dirichlet Allocation (LDA) approaches [25, 31] have been used for object localisation. We implement a generalisation of LDA [2, 31] that accepts continuous feature vectors (instead of bag-of-words). Like MIML this method can also accept patch based representation, but w-LDA is more related to our WS-SIBP than MIML since it is also a generative model.

4.1 Image annotation with object-attribute association

An image description can be automatically generated by predicting objects and their associated attributes. Evaluating the performance of a multi-faceted framework covering annotation, association and localisation is non-trivial. To comprehensively cover all aspects of performance of our method and competitors, we perform three annotation tasks with different amount of constraints on test images: (1) *free annotation*, where no constraint is given to a test image, (2) *annotation given object names*, where named but not located objects are known for each test image, and (3) *annotation given locations*, where objects locations are given in the form of bounding boxes, where the attributes can be predicted.

Free annotation: For WS-SIBP, w-LDA and MIML the procedure in Sec. 3.4 is used to detect objects and then describe them using the top t attributes. For

aPascal	w-SVM	MIML	w-LDA	WS-SIBP	DPM+s-SVM
AP@2	24.8	28.7	30.7	38.6	40.6
AP@5	21.2	22.4	24.0	28.9	30.3
AP@8	20.3	21.0	21.5	24.1	23.8
ImageNet	w-SVM	MIML	w-LDA	WS-SIBP	GT+s-SVM
AP@2	46.3	46.6	48.4	58.5	65.9
AP@3	41.1	43.2	43.1	51.8	60.7
AP@4	37.5	38.3	38.4	47.4	53.2

Table 1: Free annotation performance evaluated on t attributes per object.



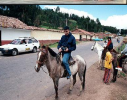
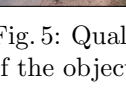
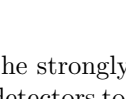
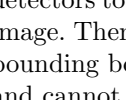
	w-SVM		MIML		w-LDA		WS-SIBP		DPM+s-SVM	
	bicycle	motorbike	person	bicycle	motorbike	bicycle	person	motorbike	person	motorbike
	metal row wind shiny text wool	metal row wind shiny text wool	skin cloth shiny leather foot/shoe	headlight window wheel arm screen	taillight label engine shiny glass	wheel rein beak metal sail	cloth clear arm skin head	wheel shiny pedal exhaust round	cloth skin hair head leg	engine metal label shiny taillight
	cat	sheep	cat	dog	person	cat	person	cat	person	dog
	furry cloth snout leg head	furry cloth snout leg head	furry horn ear occluded leg	mouth furry torso taillight shiny	cloth wood mast torso arm	nose furry arm jet engine foot	skin cloth vegetation torso arm	furry ear beak leg leather	cloth torso skin clear head	beak hand arm furry skin
	car	person	person	train	car	cow	horse	car	car	horse
	window glass pedal 3d boxy metal	window glass pedal 3d boxy metal	hand screen cloth arm flower	metal vert cyl wing door leg	wheel metal propeller label door	furn.leg head furry cloth wool	furry leg head torso rein	3d boxy window wheel handlebars row wind	metal 3d boxy plastic engine side mirror	furry torso leaf saddle feather

Fig. 5: Qualitative results on free annotation. False positives are shown in red. If the object prediction is wrong, the corresponding attribute box is shaded.

the strongly supervised model on aPascal (DPM+s-SVM), we use DPM object detectors to find the most confident objects and their bounding boxes in each test image. Then we use the 64 attribute classifiers to predict top t attributes in each bounding box. In contrast, w-SVM trains attributes and objects independently, and cannot associate objects and attributes. We thus use it to predict only one attribute vector per image regardless of which object label it predicts.

Since there are variable number of objects per image in aPascal, quantitatively evaluating free annotation is not straightforward. Therefore, we evaluate only the most confident object and its associated top t attributes in each image, although more could be described. For ImageNet, there is only one object per image. We follow [11, 39] in evaluating annotation accuracy by average precision (AP), given varying numbers (t) of predicted attributes per object. Note that if the predicted object is wrong, all associated attributes are considered wrong.

Table 1 compares the free annotation performance of the five models. We have the following observations: (1) Our WS-SIBP, despite learned with the weak image-level annotation, yields comparable performance to the strongly supervised model. The gap is particularly small for the more challenging aPascal dataset, whilst for ImageNet, the gap is bigger as the strongly supervised GT+s-SVM has an unfair advantage by using the ground truth bounding boxes during testing. (2) WS-SIBP consistently outperforms the three weakly supervised alternatives. The margin is particularly large for $t = 2$ attributes per object, which is

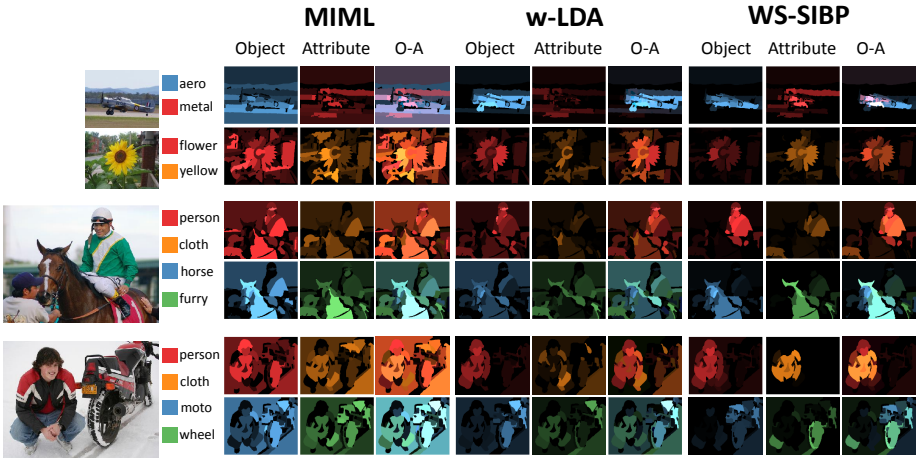


Fig.6: Illustrating the inferred patch-annotation. Object and attributes are coloured, and multi-label annotation blends colours. The bottom two groups each have two rows corresponding to the two most confident objects detected.

closest to the true number of attributes per object. For bigger t , all models must generate some irrelevant attributes thus narrowing the gaps. (3) As expected, the w-SVM model obtains the weakest results, suggesting that the ability to locate objects is important for modelling object-attribute association. (4) Compared to the two generative models, MIML has worse performance because a generative model is more capable of utilising weak labels [31]. (5) Between the two generative models, the advantage of our WS-SIBP over w-LDA is clear; due to the ability of IBP to explain each patch with multiple non-competing factors. (Training two independent w-LDA models for objects and attributes respectively is not a solution: the problem would re-occur for multiple competing attributes.)

Fig. 5 shows qualitative results on aPascal via the two most confident objects and their associated attributes. This is challenging data – even the strongly supervised DPM+s-SVM makes mistakes for both attribute and object prediction. Compared to the weakly supervised models, WS-SIBP has more accurate prediction – it jointly and non-competitively models objects and their attributes so object detection benefits from attribute detection and vice versa. Other weakly supervised models are also more likely to mismatch attributes with objects, e.g. MIML detects a shiny person rather than the correct shiny motorbike.

To gain some insight into what has been learned by our model and why it is better than the weakly supervised alternatives, Fig. 6 visualises the attribute and object factors learned by WS-SIBP model and by the two baselines that also use patches as input. It is evident that without explicit background modelling, MIML suffers greatly by trying to explain the background patches using the weak labels. In contrast, both w-LDA and WS-SIBP have good segmentation of foreground objects, showing that both the learned foreground and background topics are

		w-SVM	MIML	w-LDA	WS-SIBP	strongly supervised
GN	aPascal	–	32.1	35.5	38.9	41.8
	ImageNet	32.4	33.5	39.6	51.5	56.8
GL	aPascal	33.2	35.1	35.8	43.8	42.1
	ImageNet	37.7	39.1	46.8	53.7	56.8

Table 2: Results on annotation given object names (GN) or locations (GL).

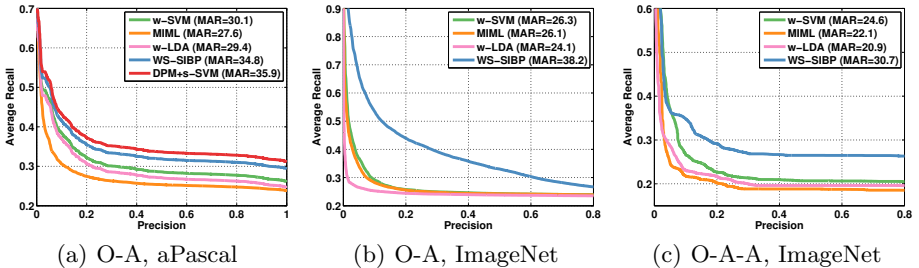


Fig. 7: Object-attribute query results as precision-average recall curve.

meaningful. However, for w-LDA, since object and attributes topics compete for the same patch, each patch is dominated by either an object or attribute topic. In contrast, the object factors and attribute factors co-exist happily in WS-SIBP as they should do, e.g. most person patches have the clothing attribute as well.

Annotation given object names (GN): In this experiment, we assume that object labels are given and we aim to describe each object by attributes, corresponding to tasks such as: “Describe the car in this image”. For the strongly supervised model on aPascal, we use the object’s DPM detector to find the most confident bounding box. Then we predict attributes for that box. Here, annotation accuracy is the same as attribute accuracy, so the performance of different models is evaluated following [40] by mean average precision (mAP) under the precision-recall curve. Note that for aPascal, w-SVM reports the same list of attributes for all co-existing objects, without being able to localise and distinguish them. Its result is thus not meaningful and is excluded. The same set of conclusions can be drawn from Table 2 as in the free annotation task: our WS-SIBP at par with the supervised models and outperforming the weakly supervised ones.

Given object location (GL): If we further know the bounding box of an object in a test image, we can simply predict attributes inside each bounding box. This becomes the conventional attribute prediction task [9, 27] for describing an object. Table 2 shows the results, where similar observations can be made as in the other two tasks above. Note that in this case the strongly supervised model is the method used in [9]. The mAP obtained using our weakly supervised model is even higher than the strongly supervised model (though our area-under-ROC-curve value of 81.5 is slightly lower than the 83.4 figure reported in [9]).

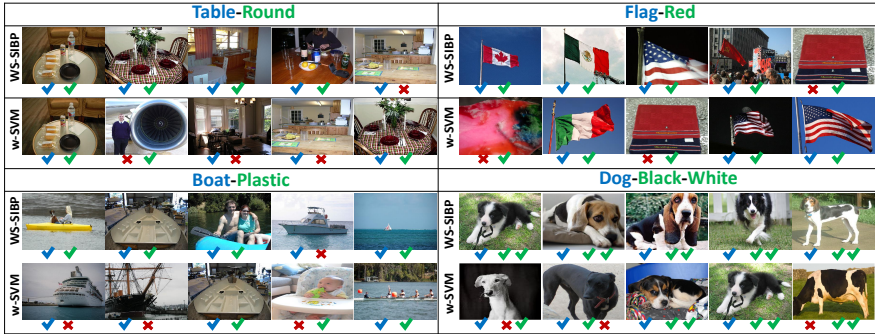


Fig. 8: Object-attribute query: qualitative comparison

4.2 Object-attribute query

In this task object-attribute association is used for image retrieval. Following work on multi-attribute queries [26], we use mean average recall over all precisions (MAR) as the evaluation metric. Note that unlike [26] which requires each queried *combination* to have enough (100) training examples to train conjunction classifiers, our method can query novel never-previously-seen combinations. Three experiments are conducted. We generate 300 random object-attribute combinations for aPascal and ImageNet respectively and 300 object-attribute-attribute queries for ImageNet. For the strongly supervised model, we normalise and multiply object detector with attribute classifier scores. No object detector is trained for ImageNet so no result is reported there. For w-SVM, we use [30] to calibrate the SVM scores for objects and attributes as in [26]. For the three WS models, the procedure in Sec. 3.4 is used to compute the retrieval ranking.

Quantitative results are shown in Fig. 7 and some qualitative examples in Fig. 8. Our WS-SIBP has a very similar MAR values to the strongly supervised DPM+s-SVM, while outperforming all the other models. w-SVM calibration [30] helps it outperform MIML and w-LDA. However, the lack of object-attribute association and background modelling still causes problems for w-SVM. This is illustrated in the ‘dog-black-white’ example shown in Fig. 8 where a white background caused an image with a black dog retrieved at rank 2 by w-SVM.

5 Conclusion

We have presented an effective model for weakly-supervised learning of objects, attributes, their location and associations. Learning object-attribute association from weak supervision is non-trivial but critical for learning from ‘natural’ data, and scaling to many classes and attributes. We achieve this for the first time through a novel weakly-supervised stacked IBP model that simultaneously disambiguates patch-annotation correspondence, as well as learning the appearance of each annotation. Our results show that our model performs comparably with a strongly supervised alternative that is significantly more costly to supervise.

References

1. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. TPAMI (2011)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR (2003)
3. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: ICCV (2011)
4. Chen, H., Gallagher, A., Girod, B.: Describing clothing by semantic attributes. In: ECCV (2012)
5. Chen, X., Shrivastava, A., Gupta, A.: Neil: Extracting visual knowledge from web data. In: ICCV (2013)
6. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. IJCV 100 (2012)
7. Doshi-Velez, F., Miller, K.T., Gael, J.V., Teh, Y.W.: Variational inference for the indian buffet process. Tech. rep., University of Cambridge (2009)
8. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. JMLR (2008)
9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009)
10. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. TPAMI (2010)
11. Feng, Z., Jin, R., Jain, A.: Large-scale image annotation by efficient and robust kernel metric learning. In: ICCV (2013)
12. Fu, Y., Hospedales, T., Xiang, T., Gong, S.: Attribute learning for understanding unstructured social activity. In: ECCV (2012)
13. Griffiths, T.L., Ghahramani, Z.: The indian buffet process: An introduction and review. JMLR (2011)
14. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local image descriptors into compact codes. TPAMI (2011)
15. Kovashka, A., Grauman, K.: Attribute adaptation for personalized image search. In: ICCV (2013)
16. Kovashka, A., Vijayanarasimhan, S., Grauman, K.: Actively selecting annotations among objects and attributes. In: ICCV (2011)
17. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A., Berg, T.: Baby talk: Understanding and generating simple image descriptions. In: CVPR (2011)
18. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE TPAMI (2013)
19. Li, L.J., Socher, R., Fei-Fei, L.: Towards total scene understanding: classification, annotation and segmentation in an automatic framework. In: CVPR (2009)
20. Mahajan, D., Sellamanickam, S., Nair, V.: A joint learning framework for attribute models and object descriptions. In: ICCV (2011)
21. Marchesotti, L., Perronnin, F.: Learning beautiful (and ugly) attributes. In: BMVC (2013)
22. Nguyen, N.: A new svm approach to multi-instance multi-label learning. In: ICDM (2010)
23. Ordonez, V., Deng, J., Choi, Y., Berg, A.C., Berg, T.L.: From large scale image categorization to entry-level categories. In: ICCV (2013)
24. Patterson, G., Hays, J.: Sun attribute database: Discovering, annotating, and recognizing scene attributes. In: CVPR (2012)

25. Rasiwasia, N., Vasconcelos, N.: Latent dirichlet allocation models for image classification. TPAMI (2013)
26. Rastegari, M., Diba, A., Parikh, D., Farhadi, A.: Multi-attribute queries: To merge or not to merge? In: CVPR (2013)
27. Russakovsky, O., Fei-Fei, L.: Attribute learning in large-scale datasets. In: ECCV (2010)
28. Sadeghi, M., Farhadi, A.: Recognition using visual phrases. In: CVPR (2011)
29. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. TPAMI (2010)
30. Scheirer, W., Kumar, N., Belhumeur, P.N., Boult, T.E.: Multi-attribute spaces: Calibration for attribute fusion and similarity search. In: CVPR (2012)
31. Shi, Z., Hospedales, T.M., Xiang, T.: Bayesian joint topic modelling for weakly supervised object localisation. In: ICCV (2013)
32. Siddiquie, B., Feris, R., Davis, L.: Image ranking and retrieval based on multi-attribute queries. In: CVPR (2011)
33. Socher, R., Fei-Fei, L.: Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In: CVPR (2010)
34. Turakhia, N., Parikh, D.: Attribute dominance: What pops out? In: ICCV (2013)
35. Wang, G., Forsyth, D.: Joint learning of visual attributes, object classes and visual saliency. In: ICCV (2009)
36. Wang, S., Joo, J., Wang, Y., Zhu, S.C.: Weakly supervised learning for attribute localization in outdoor scenes. In: CVPR (2013)
37. Wang, X., Ji, Q.: A unified probabilistic approach modeling relationships between attributes and objects. In: ICCV (2013)
38. Wang, Y., Mori, G.: A discriminative latent model of object classes and attributes. In: ECCV (2010)
39. Wu, L., Jin, R., Jain, A.K.: Tag completion for image retrieval. TPAMI (2013)
40. Zhang, N., Farrell, R., Iandola, F., Darrell, T.: Deformable part descriptors for fine-grained recognition and attribute prediction. In: ICCV (2013)
41. Zhou, Z.H., Zhang, M.L., Huang, S.J., Li, Y.F.: Multi-instance multi-label learning. Artificial Intelligence (2012)