Latent Topic Random Fields: Learning using a taxonomy of labels

Xuming He
University of Toronto
hexm@cs.toronto.edu

Richard S. Zemel
University of Toronto
zemel@cs.toronto.edu

Abstract

An important problem in image labeling concerns learning with images labeled at varying levels of specificity. We propose an approach that can incorporate images with labels drawn from a semantic hierarchy, and can also readily cope with missing labels, and roughly-specified object boundaries. We introduce a new form of latent topic model, learning a novel context representation in the joint label-and-image space by capturing co-occurring patterns within and between image features and object labels. Given a topic, the model generates the input data, as well as a topic-dependent probabilistic classifier to predict labels for image regions. We present results on two real-world datasets, demonstrating significant improvements gained by including the coarsely labeled images.

1. Introduction

Many pattern analysis tasks involve labeling high-dimensional structural inputs, such as part-of-speech tagging in text analysis, webpage classification in information retrieval, and detailed object labeling in vision. In particular, great strides have been made in the area of image labeling recently; discriminative learning approaches, such as Conditional Random Fields (CRFs), have been successfully applied and extended to provide impressive performance on this difficult task [7, 17, 21].

However, these learning methods require a training dataset in which many images are labeled at the pixel level. Labeling every pixel of an image using a vocabulary of many object classes is very tedious, and not practical for real-world data sets that include large numbers of images. Few such datasets are currently available, and those that do exist vary considerably in terms of complexity, completeness of labeling, and the object classes.

On the other hand, it is easier to obtain weakly-labeled image data. While "weakly labeled" can have a variety of meanings, including weak or non-existent information for object position within an image (as in captioned or keyworded images), we focus here on image data with multiple

Figure 1. Example image with two levels of labeling. Left: Original image. Middle: Detailed labeling. Right: Coarse labeling. Key: red='animate object', gray='inanimate object', dark='void'; see Fig. 8 for full color key. [All figures best viewed in color.]

levels of labels, including regions with no labels at all. In particular, when labels correspond to object classes, the labels may take on different levels of granularity, and can be grouped into a hierarchy based on their semantics. For example, a region with the label 'rhino' can also be labeled as 'animal' or 'animate object'. Using such more abstract or coarse labels requires less effort in collecting labeled image data, because the coarser label vocabulary is smaller, and also often has a simpler structure in the image. Figure 1 shows a typical example where the coarse label configuration has simpler boundaries. We also explore simplifying the labeling task by allowing the label boundaries to be roughly specified in an image. We thus aim to leverage the process of building label prediction models by incorporating images with different label granularities and boundary specificities into the dataset.

Another key aim of label prediction models is to incorporate *contextual* information, as local image features often cannot provide enough evidence to resolve the true labels. Some methods model context at the label level, based on co-occurrences of objects (e.g., [3, 4]), while others model image-based context, by finding object/scene specific patterns in image feature space (e.g., [15, 12]). However, a general context model would ideally incorporate information within and between the image and object spaces.

In this paper, we develop a novel approach to image labeling in order to address those two aims. We first introduce a context representation in the joint label-and-image space by extending latent topic models (e.g., Latent Dirichlet Allocation (LDA) [1]). In the standard topic model, topics capture co-occurring words (or image features). Our model learns topics not only of input features but also including label information, capturing co-occurrences within and be-

tween image feature patterns and object classes in the data set. Given a topic, our context representation consists of two components: one generates the input data with the feature patterns, and the other is a topic-dependent probabilistic mapping from input data to the output labels. This latter component allows the topic model to predict labels for novel images. Unlike traditional Markov models, our approach does not pre-specify the context's scope (e.g., a 3x3 neighborhood), but instead can flexibly model higher-order joint context. We refer to this extended topic model as a *latent topic random field* (LTRF).

To construct the model, we utilize both a small set of image data with detailed labels and a larger set of images with coarse labels. Both image sets can have different levels of boundary specificity. The detailed-labeled data provide precise information for learning the joint patterns in the input and the detailed label space, whereas the coarsely-labeled data help the system to build a better context model by regularizing those patterns with coarse level contexts.

This paper is organized as follows. The next section discusses related work. In Section 3, we describe the label hierarchy and the architecture of our latent topic random field model. Section 4 presents the inference methods used to label a new image, and for learning the model parameters. The learning procedure is detailed in Section 5. We compare our model with other approaches based on two image datasets in Section 6. Section 7 summarizes the paper and discusses some further issues.

2. Related Work

Many approaches have been proposed for image labeling with object-level information, in which different types of contexts are incorporated. Some capture local context (e.g., [7]), whereas others focus on longer range interactions, such as hierarchical structures [3], extended neighborhoods [17], and geometrical or layout information [5, 21]. While these methods mostly capture interactions between objects or their parts, several approaches exist for modeling context using a combination of object and image information. Murphy et al. [10] use image gists to represent the specific environments that images are taken from, and combine this with a generative model of co-occurring object classes. In [8], multiclass object detection is combined with image segmentation to form a joint labeling problem, in which multiple interactions between regions and objects can be exploited for consistent labeling. However, in order to learn such representations, these methods require labeled images with detailed object class information for training, which may limit their ability to scale up to more object classes.

Our work also relates to a variety of methods proposed for combining a generative model of the input data with a discriminative model for image labeling. For example, Kelm et al. combine generative and discriminative methods for pixel classification, using a modified likelihood for learning the generative model [6]. Lasserre et al. suggest a principled way to integrate generative and discriminative models by constructing separate but complementary models [9]. The generative models in these methods are fairly simple, which restricts the range of contextual information that can be represented. Also, to our knowledge, few methods have examined a variety of labels, ranging from coarse to fine. One example in this space is [16], which detects objects in tiny images, using WordNet to model the set of object labels. Note that this method focuses on object recognition in tiny images rather than labeling at the level of image pixels.

Finally, our work relates to a rapidly growing literature on topic models. These models were first used for document modeling, and recently extended to image modeling [14, 13]. Most topic models are purely generative for images. For example, Sudderth et al. [14] extended the topic model by including a locality constraint, which is based on geometric information about the relative position between object parts. Spatial structure is also incorporated in SLDA [20]. In [18], aspect models are combined with a random field. However, their topics only model image feature patterns and are category-specific, instead of being shared by different categories. Fei-Fei and colleagues cope with unknown number of topics based on hierarchical Dirichlet processes for object recognition [19], and apply topic modeling to video sequence for activity categorization [11]. Topic models are also used to jointly model image features and captions, so that associations between image segments and caption can be learned [2].

3. Model Architecture

3.1. Label Hierarchy

We consider a situation in which we have a set of label values that are not exclusive and can form a hierarchy according to their semantics. The label hierarchy could be derived from some taxonomy of concepts, as in WordNet,or from a hierarchical clustering process. For example, Figure 2 shows a hierarchy of objects in the Microsoft Research Cambridge Image Database [12]. In this tree structure, a parent label value includes its children as special cases. The label values at the leaf nodes correspond to detailed labeling of images, whereas the internal nodes are used in coarse labeling. In this paper, we assume that the hierarchy is given, and each training image is labeled at some level: either using detailed labels at the leaf level, or coarse labels from a single internal level. The construction of this type of label hierarchy is an interesting problem, but is beyond the scope of this paper. Note that the label hierarchy gives two different senses of coarseness: first, the internal label values are coarse in terms of their semantics; second, the spatial con-

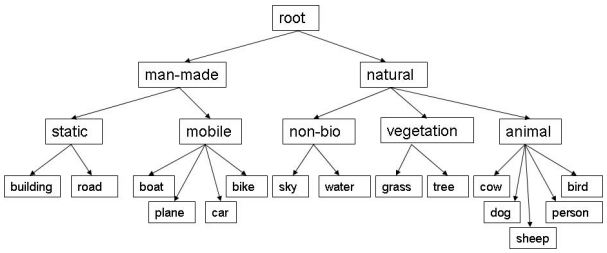

Figure 2. A label hierarchy of objects used in the Microsoft Research Cambridge Image Database. We construct the hierarchy based on the semantics of labels.

figuration of labeling can be coarser due to the merging of subclasses (Figure 1). Finally, to extend the model applicability to include regions of unknown labels, all levels of the hierarchy may include a "catch-all" label, denoted 'void'.

3.2. Topic Model with Labels

We start by building a generative topic model for images based on the LDA model [1]. Suppose each image is represented by a set of image features, and the i^{th} feature has its appearance descriptor a_i in image location x_i. The appearance variable a_i takes values from a vocabulary of visual words. We assume that the appearance of image feature a_i is generated from a finite set of K hidden topics. Each image I is described by a multinomial distribution θ over the hidden topics. To generate a new appearance a_i in an image, we start by first sampling a hidden topic z_i from the θ corresponding to the image. Given the topic z_i, the appearance a_i is sampled from its topic conditional distribution. As in the LDA model, the parameters between different images are tied by drawing θ of all images from a common Dirichlet prior parameterized by α.

We then incorporate the label variables into the latent topic model of images. Two types of labels are associated with each image feature: coarse label c_i and detailed label d_i. The detailed label variable d_i takes values from $\{1, ..., L_d\}$, and the coarse label c_i from $\{1, ..., L_c\}$. Given image feature a_i, its location x_i, and the corresponding topic z_i, the label pair $\{c_i, d_i\}$ is predicted from a conditional multinomial distribution. Viewing each topic as a context, we have a context dependent appearance model $P(a_i|z_i)$, and a context dependent label predictor $P(d_i, c_i|a_i, x_i, z_i)$. Thus, the joint distribution of the model can be written as

$$P(\mathbf{a}, \mathbf{d}, \mathbf{c}, \mathbf{z}, \theta|\alpha, \mathbf{x}) = P(\theta|\alpha) \prod_{i=1}^{N_d} [P(d_i, c_i|a_i, x_i, z_i) \times P(a_i|z_i)P(z_i|\theta)] \quad (1)$$

where N_d is the number of image features in image I_d. Note that the appearance model is position invariant, whereas the label predictor uses the position information. The graphical representation of the model is shown in Figure 3, and each component of the joint is formulated as follows.

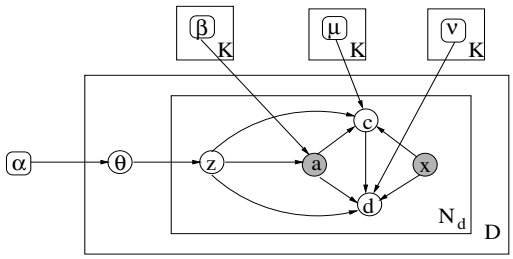

Figure 3. A graphical representation of the extended topic model for image features and their labels. Circular nodes are random variables, rectangular nodes are parameters, and shaded nodes are observed. N_d is the number of image features in each image, and D denotes all the training data.

(a) Label prediction model $P(d_i, c_i|a_i, x_i, z_i)$. The conditional distribution of the pair $\{c_i, d_i\}$ is parameterized as follows: the coarse label c_i is first generated from a multinomial $P(c_i|a_i, x_i, z_i)$, given a_i, x_i and z_i. Conditioned on the coarse label, the detailed label d_i is then generated from $P(d_i|c_i, a_i, x_i, z_i)$ (see Figure 3). The coarse label predictors $P(c_i|a_i, x_i, z_i)$ are modeled by a set of topic-dependent probabilistic classifiers: for each topic k, we have a classifier $P_k^c(f|a, x)$:

$$P(c_i|a_i, x_i, z_i) = P_{z_i}^c(c_i|a_i, x_i; \mu_{z_i}), \quad (2)$$

where $\{\mu_k\}_{k=1}^K$ are the parameters of the coarse-classifiers, and we assume that each classifier produces a properly normalized distribution. To build $P(d_i|c_i, a_i, x_i, z_i)$, we introduce another set of classifiers, in which $P_k^d(d|a, x; \nu_k)$ predicts the detailed label given the topic k and input (a, x). We denote the parameters of the kth detailed-classifier as ν_k, and assume their outputs form normalized distributions. The conditional distribution of the detailed label d_i is written as $P(d_i|c_i, a_i, x_i, z_i) \propto P_{z_i}^d(d_i|a_i, x_i)[c_i = f(d_i)]$, where $[c_i = f(d_i)]$ is 1 if c_i is the parent of d_i (denoted $f(d_i)$) in the label hierarchy and 0 otherwise. Notice that by summing out the coarse label variables, the conditional distribution of the detailed label given input and topic can be written as

$$P(d_i|a_i, x_i, z_i) = \frac{P_{z_i}^d(d_i|a_i, x_i)}{\sum_{d \in s(d_i)} P_{z_i}^c(d|a_i, x_i)} P_{z_i}^c(f(d_i)|a_i, x_i) \quad (3)$$

where $s(d_i)$ includes d_i and its siblings in the hierarchy.

(b) Image appearance model $P(a_i|z_i)$. The topic conditional distributions of the image appearance is multinomial with parameters β_{z_i}, as our image features come from a set of visual words: $P(a_i = v|z_i = k) = \beta_{k,v}$.

(c) Topic prior model $P(z_i|\theta)P(\theta|\alpha)$. The topic distribution θ has a Dirichlet distribution with a symmetric parameter α. Given θ, the topic distribution is multinomial: $P(z_i = k|\theta) = \theta_k$. Note that the topic prior induces a weak correlation between the hidden topic variables, which can be seen

by integrating out θ: $P(\mathbf{z}|\alpha) \propto \prod_k \Gamma(\alpha_k + \sum_i \delta(z_i, k))$, where $\Gamma(\cdot)$ is the Gamma function.

Our model can be viewed as an integrated structure of a generative topic model and a set of discriminative classifiers. Though we introduce label variables into the topic model, the basic assumption of the original LDA model, viewing each image as a bag of features, remains the same. A consequence of this weak assumption is that the topics may not be easy to interpret. However, the advantage is that a topic could correspond to any commonly co-occurring feature/label pattern in the images.

4. Inference and Label Prediction

Given a new image $I = \{\mathbf{a}, \mathbf{x}\}$ and our topic model, we predict its labeling based on the Maximum Posterior Marginals (MPM) criterion:

$$(d_i^*, c_i^*) = \arg\max_{d_i, c_i} P(d_i, c_i|\mathbf{a}, \mathbf{x}), \tag{4}$$

where the marginal distribution can be computed as: $P(d_i, c_i|\mathbf{a}, \mathbf{x}) = \sum_{z_i} P(d_i, c_i|z_i, a_i, x_i)P(z_i|\mathbf{a}, \mathbf{x})$.

The key step in inference is to obtain the conditional distribution of the hidden topic variables $\mathbf{z}$ given observed data components. We integrate out the Dirichlet variable θ, and take a Gibbs sampling approach to estimate that distribution. From Equation 1, we can derive the posterior of each topic variable z_i given other variables:

$$P(z_i = k|\mathbf{z}_{-i}, \mathbf{a}, \mathbf{x}) \propto P(a_i|z_i)(\alpha_k + \sum_{m \in \mathcal{S} \setminus i} \delta(z_m, k))$$

where $\mathbf{z}_{-i}$ denotes all the topic variables in $\mathbf{z}$ except z_i, and $\mathcal{S}$ is the set of all sites. Given the samples of the topic variables, we estimate their posterior marginal distribution $P(z_i|\mathbf{a}, \mathbf{x})$ by simply computing their normalized histograms. To be specific, given a set of J samples $\{\mathbf{z}^{n,j}\}_{j=1}^J$ for image I^n, we can estimate the posterior distribution $P(z_i^n = k|\mathbf{a}^n, \mathbf{x}^n) \propto \sum_j \delta(z_i^{n,j}, k)$.

As we will see in the following section, the posterior of the hidden topic variables $\mathbf{z}$ is also required during the learning procedure. In training, we observe not only the image features, but also their labels at either the coarse or detailed level. Therefore, we want to compute the posterior distribution $P(\mathbf{z}|\mathbf{a}, \mathbf{x}, \mathbf{d})$ or $P(\mathbf{z}|\mathbf{a}, \mathbf{x}, \mathbf{c})$, depending on which type of labeling is observed. We use the same Gibbs sampling approach to estimate these distributions. For instance, the conditional distribution used by the first Gibbs sampler is

$$P(z_i = k|\mathbf{z}_{-i}, \mathbf{a}, \mathbf{x}, \mathbf{d}) \propto P(d_i|a_i, x_i, z_i)P(a_i|z_i)(\alpha_k + \sum_{m \in \mathcal{S} \setminus i} \delta(z_m, k)). \tag{5}$$

5. Parameter Estimation

To build a LTRF model, we assume the following learning scenario: the training data include a small set of image data with detailed labeling, and a large set of image data with only coarse labeling. Let the training dataset have two subsets, $\mathcal{D} = \{\mathcal{D}^d, \mathcal{D}^c\}$: the $\mathcal{D}^d$ denotes the image set with detailed labeling, whereas $\mathcal{D}^c$ is the image set with coarse labeling only. We learn the model by maximizing weighted log data likelihood:

$$\sum_{n \in \mathcal{D}^d} \log P(\mathbf{a}^n, \mathbf{d}^n|\mathbf{x}^n) + \gamma \sum_{t \in \mathcal{D}^c} \log P(\mathbf{a}^t, \mathbf{c}^t|\mathbf{x}^n) \tag{6}$$

where γ controls the relative influence of the coarsely-labeled data. Note that the second term in the objective can be computed straightforwardly by marginalizing out the detailed label variables from the model. (In the detailed-labeled data, the coarse labels $\mathbf{c}^n$ can be derived from the detailed labels $\mathbf{d}^n$ according to the label hierarchy if they are not given.)

We maximize the log likelihood by Monte Carlo EM. In the E step, the posterior distributions of the topic variables are estimated by the Gibbs sampling procedure in Section 4 (e.g., Equation 5). We denote $P(z_i^n|\mathbf{a}^n, \mathbf{x}^n, \mathbf{d}^n)$ as $q^d(z_i^n)$, and $P(z_i^t|\mathbf{a}^t, \mathbf{x}^t, \mathbf{c}^t)$ as $q^c(z_i^t)$ in the following. In the M step, we update the model parameters by maximizing the expected joint likelihood:

$$\mathcal{L} = \sum_{n,i} \langle \log P(a_i^n|z_i^n) + \log P(d_i^n|a_i^n, x_i^n, z_i^n) \rangle_{q^d(z_i^n)} + \gamma \sum_{t,i} \langle \log P(a_i^t|z_i^t) + \log P(c_i^t|a_i^t, x_i^t, z_i^t) \rangle_{q^c(z_i^t)} \tag{7}$$

where $P(d_i|a_i, x_i, z_i)$ and $P(c_i|a_i, x_i, z_i)$ are specified by Equation 3 and 2, respectively.

(1) Learning appearance model

The parameters $\{\beta_k\}_{k=1}^K$ of the multinomial distribution in the appearance model $P(a|z)$ is updated by maximizing the objective in Equation 7. We take the derivative of $\mathcal{L}$, and derive the updating equation from its stationary point:

$$\beta_{k,v}^* \propto \sum_{n,i} q^d(z_i^n = k)\delta(a_i^n, v) + \gamma \sum_{t,i} q^c(z_i^t = k)\delta(a_i^t, v)$$

(2) Learning classifiers for detailed labeling

While directly optimizing $\mathcal{L}$ w.r.t. the classifier parameters is feasible, the required normalization in the distribution $P(d_i|a_i, x_i, z_i)$ (see Equation 3) complicates learning of the detailed and coarse label classifier parameters. However, we notice that the output of a coarse label classifier can be approximated by the detailed label classifier if they are consistent during training. That is,

$$P_{z_i}^c(f(d_i)|a_i, x_i) \approx \sum_{d \in s(d_i)} P_{z_i}^d(d|a_i, x_i). \tag{8}$$

Using this simplification, we update the parameters in the detailed label classifiers by maximizing the following

weighted log likelihood:

$$\nu_k^* = \max_{\nu_k} \sum_{n,i} q^d(z_i^n = k) \log P_k^d(d_i^n | a_i^n, x_i^n; \nu_k). \quad (9)$$

We implement this sub-learning problem by a gradient-based algorithm, in which each example is weighted by the posterior distribution $q^d(z_i^n = k)$. Notice that we need to run the gradient ascent for only a few steps at each iteration, which reduces training time.

(3) Learning classifiers for coarse labeling

Based on the approximation in Equation 8, updating the parameters in the classifiers for coarse labeling is simplified to maximizing a weighted log likelihood:

$$\mu_k^* = \max_{\mu_k} \gamma \sum_{t,i} q^c(z_i^t = k) \log P_k^c(c_i^t | a_i^t, x_i^t; \mu_k). \quad (10)$$

The learning procedure is implemented by the same modified gradient-based algorithm used in the detailed labeling case.

6. Experimental Evaluation

6.1. Data sets

Our first experiments use the Microsoft Research Cambridge (MSRC) Image Database [12]. We select a subset of the database, focusing on outdoor classes, yielding 415 detailed-labeled images and 16 different label classes. We randomly split the dataset into four subsets: 10% is used as the training dataset with detailed labels, 20% is used for validation, and another 20% is used as the test dataset. The remaining 50% is used as a training dataset with coarse labels. The original dataset only has the detailed labeling. To obtain the coarse labeling, we use the label hierarchy in Figure 2. We choose the second level such that the coarse labels have three different values: 'void', 'man-made' and 'natural'.

The second experiments use a labeled subset of the Corel database as in [4]. It includes 305 manually labeled images with 11 classes: 'rhino/hippo', 'tiger', 'horse','polar bear', 'wolf/leopard', 'vegetation', 'sky', 'water', 'snow', 'ground' and 'fence'. We also randomly split the dataset into four subsets with the same proportion as the MSRC case. To obtain the coarse level labels, we group the classes into 'animal' and 'background'.

6.2. Baseline methods

We compare our approach with two baseline systems: a super-pixel-wise classifier and a basic CRF model. The super-pixel-wise classifier is an MLP with one hidden layer, which predicts labels for each super-pixel independently. Based on validation performance, the MLP has 30 hidden units. In the basic CRF, the conditional distribution of the labels of an image is defined as:

$$P(\mathbf{d}|\mathbf{a}, \mathbf{x}) \propto \exp\{\sum_{i,j} g(d_i, d_j) + \alpha \sum_i h(d_i | a_i, x_i)\} \quad (11)$$

where $h(\cdot)$ is the log output from the super-pixel classifier and $g(\cdot)$ is the compatibility function. We train the CRF model using the pseudo-likelihood algorithm, and label the image based on the marginal distribution of each label variable, computed by the loopy belief propagation algorithm.

To utilize coarsely-labeled data in the baseline systems, we modify these models as follows. First, we keep the baseline systems trained using the detailed-labeled data and denote them as P_0. Then we train a separate set of baseline systems based on the coarsely-labeled data. The coarse models are denoted as P_1. The final labeling of a super-pixel i by the baseline systems is given by combining the two sets of models:

$$(d_i^*, c_i^*) = \arg\max_{d_i, c_i} P_0(d_i | \mathbf{a}, \mathbf{x}) P_1(c_i | \mathbf{a}, \mathbf{x}) [c_i = f(d_i)]$$

We evaluate the final performance by two different measures: class accuracy, and F1 score (i.e., the harmonic mean of the precision and the recall). We only consider the prediction performance on the detailed labeling, and the class 'void' is not included in the metrics.

6.3. Feature representation

We use the normalized cut segmentation algorithm to build the super-pixel representation of the images, in which the segmentation algorithm is tuned to generate approximately 1000 segments for each image on average. We extract a set of basic image features, including color, edge and texture information, from each pixel site. For the color information, we transform the RGB values into CIE Lab* color space, which is perceptually uniform. The edge and texture are extracted by a set of filter-banks including a difference-of-Gaussian filter at 3 different scales, and quadrature pairs of oriented even- and odd-symmetric filters at 4 orientations $(0; \pi/4; \pi/2; 3\pi/4)$ and 3 scales. The color descriptor of a super-pixel is the average color over the pixels in that super-pixel. For edge and texture descriptors, we first discretize the edge/texture feature space by K-means, and use each cluster as a bin. Then we compute the normalized histograms of the texture features within a super-pixel as the edge/texture descriptor. In the experiments reported here, we used 10 bins for edge information and 50 bins for texture information. In total, the image descriptor of a super-pixel has 63 dimensions. The image position of a super-pixel is the average position of its pixels.

To compute the vocabulary of visual words in the topic model, we apply K-means to group the super-pixel descriptors into clusters after the descriptors are centered and normalized. The cluster centers are used as visual words and each descriptor is encoded by its word index.

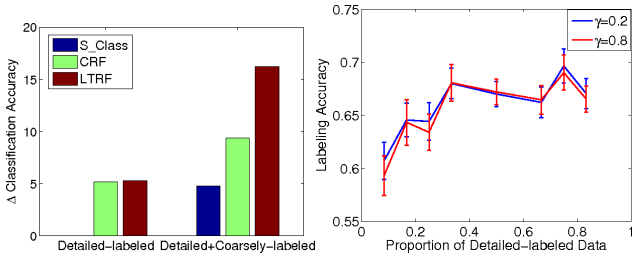

Figure 4. Left: Improvement of classification accuracy over the baseline method, the super-pixel classifier (S_Class) trained only on the detailed-labeled data (52.4% accuracy, MSRC dataset). Right: Classification accuracy of LTRF models with different proportions of detailed- vs. coarsely-labeled images.

Note that the word indices are only used in the appearance model; the topic-dependent classifiers take the original descriptors as input. The size of vocabulary is chosen from $100, 200, 500, 800, 1000$ based on the model performance on the validation set.

6.4. Experimental Results I - MSRC

In this experiment, we set the size of vocabulary to 500, the number of hidden topics to 20, and each symmetric Dirichlet parameter $\alpha_k = 0.3$, based on validation performance. For the topic-dependent classifiers, we use Multilayer Perceptons (MLP) with one hidden layer. The classifiers for detailed labeling have 5 hidden units and the classifiers for coarse labeling are linear logistic regressors. Those classifiers are initialized by training them on the corresponding labeled dataset. The appearance model for topics is initialized randomly. In the learning procedure, the E step uses 1000 samples to estimate the posterior distribution of topics. In the M step, we take a single step in gradient ascent learning of the classifiers per iteration.

The performance of LTRF is first evaluated based on learning from detailed-labeled data only. We compare the performance of LTRF to the super-pixel classifier (S_Class), and the CRF model. The left half of Figure 4 (Left) shows the average classification accuracy rates of our model and the baselines, trained with only the detailed-labeled data, all relative to the super-pixel classifier. The LTRF model achieves almost the same performance as the CRF model in this condition, and has higher average accuracy than the simple classifier.

We then compare the performance of LTRF to the baseline systems when they are learned from both detailed-labeled and coarsely-labeled data. The right half of Figure 4 (Left) shows the average pixel-level classification accuracies of the three models, relative to the super-pixel classifier trained with only the detailed-labeled data. We also report the pixel-level classification accuracies and F1 measures for each class, using LTRF and baseline models selected via validation (see Table 1). These results show that the

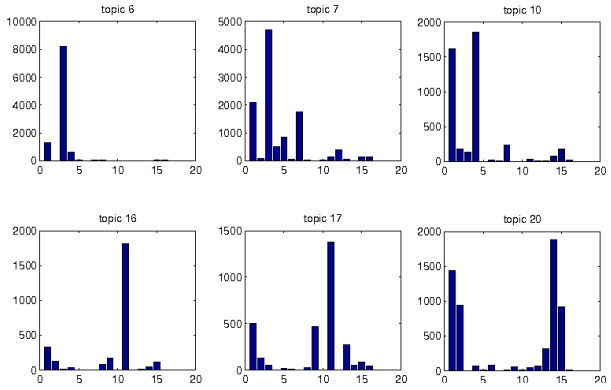

Figure 5. Counts of ground truth labels for 6 randomly-selected topics. Each x-axis tick corresponds to one of the 16 label classes; y values are counts of super-pixels in the MSRC test set.

LTRF model with additional coarsely-labeled data provides a significant improvement over the baseline methods trained on the same data, and over a LTRF trained with detailed-labeled data only. Even for individual classes, our model is always better than or comparable to other approaches, except on a few small classes, typically those with limited examples. We also evaluate the performance of our model with varying proportions of detailed- to coarsely-labeled data. Figure 4 (Right) shows the means and standard deviations of accuracies at eight different proportions (15 runs each), and for two different γ values. These results show that the model is fairly insensitive to γ, and also to the proportion if it exceeds 30%.

Furthermore, we test the robustness of the LTRF model by training it with blurry, or rough labels at object boundaries. To obtain blurred labels, we apply a dilation operator to the ground-truth labeling in the training data (see Figure 7 (Right)). Given such weakly-labeled data, our model achieves an average classification rate of 64%, which is just slightly worse than using the original labeling.

Figure 5 shows the counts of ground truth labels for a randomly-selected subset of 6 topics. Most topics specialize in a subset of labels, while some topics, such as topic 6 and 16, are almost completely focused on one object class. Figure 6 displays some example topics in the test images (note that we show the MAP estimate of the topic variables). While the topic instantiations are slightly noisy, and weakly local, they capture some co-occurring patterns, such as "sky-cow-grass", "tree-grass" and "building-sky". Figure 8 displays outputs of the methods on some test images. These results show that classes with small regions usually have poorer accuracy than others. Overall, LTRF works better than other approaches in terms of accuracy, while the CRF method gets smoother labelings.

6.5. Experimental Results II - Corel

We set the size of vocabulary to 300 and the number of hidden topics to 30 based on validation performance; the

Table 1. A comparison of classification accuracy and F1 measure (in parenthesis) of the LTRF model with a super-pixel classifier and CRF model. The average classification accuracy and F1 measure are at the pixel- and class-level, respectively. The winners for each class, based on both measures, are shown in boldface.

Methods	Overall
S_class	57.2(33.1)
CRF	61.8(39.6)
LTRF	**68.6**(**44.0**)

Label	building	grass	tree	cow	person	sheep	sky	boat
S_class	38.7(33.1)	87.2(84.3)	58.2(56.5)	19.9(23.5)	4.5(4.7)	19.7(23.5)	89.4(77.9)	0.3(0.2)
CRF	**46.8**(45.5)	86.2(**85.0**)	60.2(62.3)	**28.8**(**32.2**)	13.5(**11.8**)	34.7(38.1)	87.6(80.2)	0.3(0.3)
LTRF	46.3(**50.8**)	**92.9**(84.2)	**74.7**(**67.8**)	27.2(27.1)	**14.0**(10.2)	**50.6**(**44.4**)	**95.2**(**80.7**)	**6.4**(**3.9**)

Label	plane	water	dog	car	bike	road	bird	-
S_class	20.0(22.1)	31.6(40.1)	**10.6**(**10.9**)	26.4(26.2)	46.2(36.3)	51.7(51.5)	4.7(5.4)	-
CRF	26.0(30.1)	37.2(49.0)	4.0(4.3)	43.3(39.0)	71.6(49.0)	**62.9**(**59.6**)	**7.6**(**8.0**)	-
LTRF	**74.5**(**48.8**)	**56.6**(**63.2**)	5.4(5.4)	**64.6**(**65.2**)	**81.0**(**55.5**)	43.2(49.2)	4.8(4.2)	-

Figure 6. Three examples of learned topics. The lower image displays the super-pixels in a test image accounted for by the particular topic.

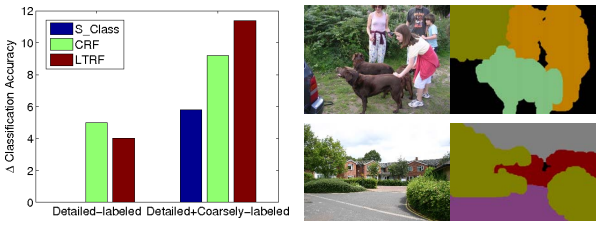

Figure 7. Left: Improvement of classification accuracy over the super-pixel classifier (S_Class) trained on detailed-labeled data only (63.0%, Corel); Right: Examples of blurred labelings.

classifiers for detailed labeling have 2 hidden units, and other settings are the same as the previous experiment. The performance of LTRF learned from both detailed-labeled data and coarsely-labeled data are compared to the baseline systems in Figure 7 (Left).

Note that the dataset sizes in our experiments are still relatively small, so the accuracy is limited compared to other methods, which utilize many training examples. However, our focus here is on the performance gained by augmenting detailed-labeled images with coarsely-labeled data.

7. Conclusion and Discussion

In this paper, we have presented a novel approach that relaxes a limitation of discriminative labeling models due to their reliance on detailed training data. Our method integrates a generative topic model with discriminative label classifiers for image labeling. One main contribution of our approach is that the extended topic model, LTRF, is capable of utilizing both detailed and more coarsely labeled images. This is a step towards extending image labeling to learn from a larger database of images, and potentially combining databases with differing label sets.

The proposed framework is able to capture high-order image contexts, in that the topics model co-occurring configurations of image features in the entire image, and in conjunction with labels. Our learning method uses the coarsely-labeled data to regularize the topic model, which would otherwise easily overfit the small detailed-labeled set. The results of applying our method to a real-world image dataset suggest that this integrated approach may extend to a variety of image types and databases. The labeling system consistently out-performs alternative discriminative approaches, such as a standard classifier and a standard CRF.

An important limitation of our model concerns the bag-of-features assumption. While the topics can potentially detect high-order configurations of features, the model is unable to learn and utilize spatial relations between parts of an object in the labeling procedure. In addition, the proposed model is the first step towards using weakly labeled data for learning labeling models. We are exploring an extension of the model that can handle other types of weakly-labeled images that are easier to obtain, including sparsely labeled and captioned images.

References

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

[2] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *ECCV*, 2004.

[3] X. Feng, C. K. Williams, and S. Felderhof. Combining belief networks and neural networks for scene segmentation. *IEEE Trans. PAMI*, 24(4):467–483, 2002.

[4] X. He, R. S. Zemel, and D. Ray. Learning and incorporating top-down cues in image segmentation. In *ECCV*, 2006.

[5] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

[6] M. Kelm, C. Pal, and A. McCallum. Combining generative and discriminative methods for pixel classification with multi-conditional learning. In *ICPR*, 2006.

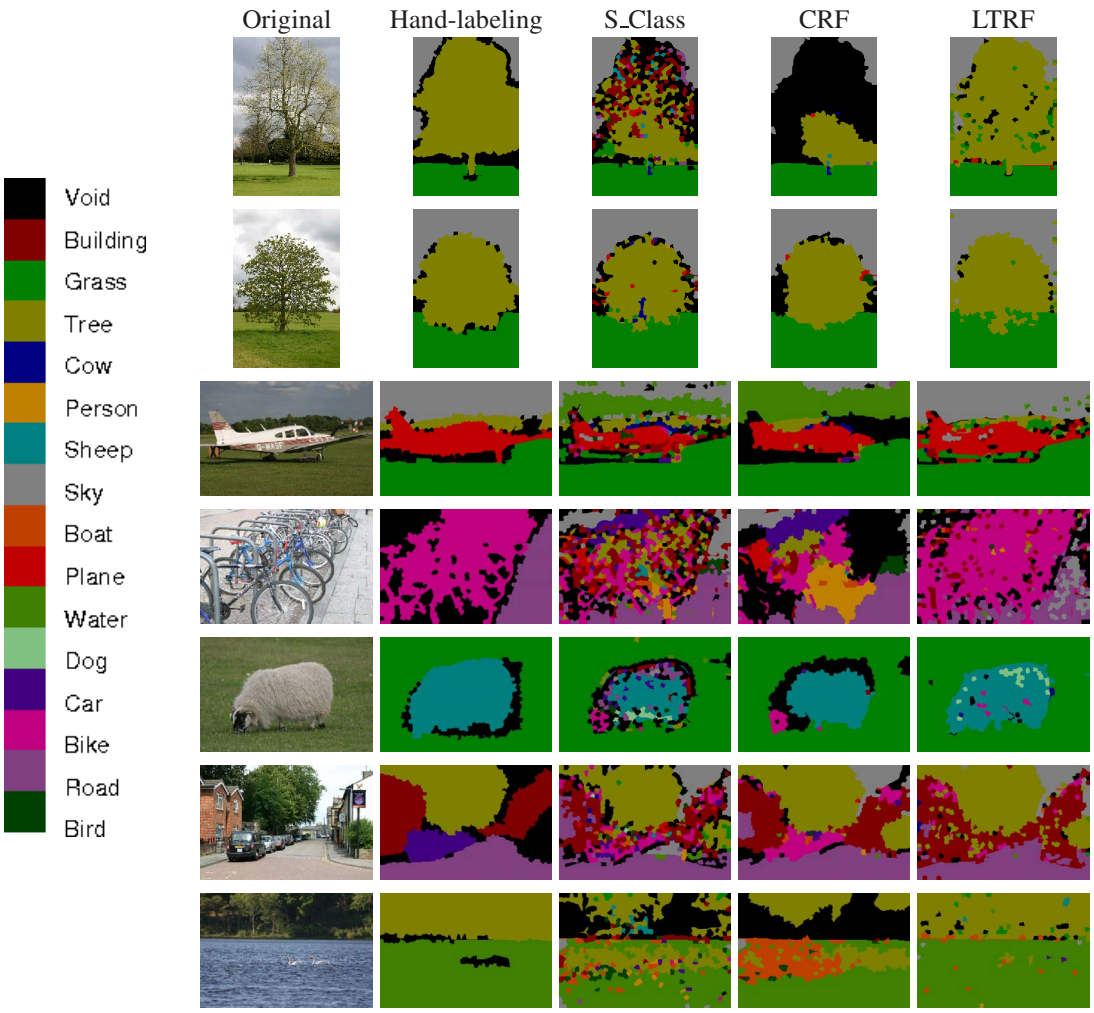

Figure 8. Example labeling results on the MSRC test set, using the super-pixel-wise classifier, CRF, and LTRF.

[7] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *ICCV*, 2003.

[8] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, 2005.

[9] J. A. Lasserre, C. M. Bishop, and T. P. Minka. Principled hybrids of generative and discriminative models. In *CVPR*, 2006.

[10] K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: A graphical model relating features, objects and scenes. In *NIPS*, 2003.

[11] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[12] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006.

[13] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *ICCV*, 2005.

[14] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005.

[15] A. Torralba. Contextual priming for object detection. *IJCV*, 53:169–191, 2003.

[16] A. Torralba, R. Fergus, and W. T. Freeman. Object and scene recognition in tiny images. *J. Vis.*, 7:193–193, 2007.

[17] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *NIPS*, 2005.

[18] J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *CVPR*, 2007.

[19] G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In *CVPR*, 2006.

[20] X. Wang and E. Grimson. Spatial latent Dirichlet allocation. In *NIPS*, 2007.

[21] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *CVPR*, 2006.

Appendix

Label prediction model (Section 3.2.a, derivation of Eqn. 3)

Label prediction model $P(d_i, c_i|a_i, x_i, z_i)$ is defined as follows:

$$P(d_i, c_i|a_i, x_i, z_i) = P(d_i|c_i, a_i, x_i, z_i)P(c_i|a_i, x_i, z_i)$$

1. Coarse-level label distribution

   $$P(c_i|a_i, x_i, z_i) = P^c_{z_i}(c_i|a_i, x_i)$$

   where $P^c_{z_i}(c_i|a_i, x_i)$ is a classifier with outputs normalized to 1, i.e., $\sum_{c_i} P^c_{z_i}(c_i|a_i, x_i) = 1$.

2. Detailed-level conditional label distribution

   $$P(d_i|c_i, a_i, x_i, z_i) \propto P^d_{z_i}(d_i|a_i, x_i)[c_i = f(d_i)]$$

   where $P^d_z(d|a, x)$ is a classifier with output normalized to 1, i.e., $\sum_d P^d_z(d|a, x) = 1$ (sum over all the detailed label values), and $[c = f(d)] = 1$ if c is the parent of d in the label hierarchy, and 0 otherwise. In doing so, we share a single detailed classifier across different coarse label classes. The normalizing constant of $P(d_i|c_i, a_i, x_i, z_i)$ can be written as

   $$\sum_d P^d_{z_i}(d|a_i, x_i)[c_i = f(d)] = \sum_{d \in s(d_i)} P^d_{z_i}(d|a_i, x_i).$$

   where the first sum is over all the detailed label values, and the second sum is over all the siblings of d_i, which all share the same parent, c_i.

3. Detailed-level label distribution can be derived by summing out the coarse-level label variable:

   $$P(d_i|a_i, x_i, z_i) = \frac{P^d_{z_i}(d_i|a_i, x_i)}{\sum_{d \in s(d_i)} P^d_{z_i}(d|a_i, x_i)} P^c_{z_i}(c[d_i]|a_i, x_i)$$

Inference algorithm (Section 4, derivation of Eqn. 5)

The joint distribution of the model can be written as

$$P(\mathbf{a}, \mathbf{d}, \mathbf{c}, \mathbf{z}, \theta|\alpha, \mathbf{x}) = P(\theta|\alpha) \prod_i P(d_i, c_i|a_i, x_i, z_i)P(a_i|z_i)P(z_i|\theta).$$

- We can integrate out θ due to the conjugacy property

$$\begin{aligned} P(\mathbf{z}, \mathbf{a}, \mathbf{d}, \mathbf{c}|\alpha, \mathbf{x}) &= \int_\theta P(\mathbf{a}, \mathbf{d}, \mathbf{c}, \mathbf{z}, \theta) d\theta \\ &= \prod_i P(d_i, c_i|x_i, a_i, z_i)P(a_i|z_i) \times \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(\alpha_k + \sum_i \delta(z_i, k))}{\Gamma(\sum_k \alpha_k + N)}. \end{aligned}$$

- To use Gibbs sampling, we want to derive $P(z_i|\mathbf{z}_{-i}, \mathbf{a}, \mathbf{x})$, $P(z_i|\mathbf{z}_{-i}, \mathbf{a}, \mathbf{d}, \mathbf{x})$, and $P(z_i|\mathbf{z}_{-i}, \mathbf{a}, \mathbf{c}, \mathbf{x})$. In the following, we use $P(z_i|\mathbf{z}_{-i}, \mathbf{a}, \mathbf{d}, \mathbf{x})$ as an example (the other two are similar). Note that

$$\begin{aligned} &P(z_i = k|\mathbf{z}_{-i}, \mathbf{a}, \mathbf{d}, \mathbf{x}) \propto P(z_i = k, \mathbf{z}_{-i}, \mathbf{a}, \mathbf{d}, \mathbf{x}) \\ &= C \cdot P(d_i|a_i, x_i, z_i = k)P(a_i|z_i = k)\Gamma(\alpha_k + \sum_{j \in S \setminus i} \delta(z_j, k) + 1) \prod_{l \neq k} \Gamma(\alpha_l + \sum_{j \in S \setminus i} \delta(z_j, l)) \\ &= C \cdot P(d_i|a_i, x_i, z_i = k)P(a_i|z_i = k)(\alpha_k + \sum_{j \in S \setminus i} \delta(z_j, k)) \prod_l \Gamma(\alpha_l + \sum_{j \in S \setminus i} \delta(z_j, l)) \\ &\propto P(d_i|a_i, x_i, z_i = k)P(a_i|z_i = k)(\alpha_k + \sum_{j \in S \setminus i} \delta(z_j, k)) \end{aligned}$$

  in which we make use of the property of the Gamma function: $\Gamma(x + 1) = x\Gamma(x)$.

Learning label prediction models (Section 5, derivation of Eqns. 7, 9, 10)

In M-step, we have the following likelihood function (ignoring other irrelevant terms):

$$\begin{aligned}\mathcal{L} &= \sum_{n,i} \langle \log P(d_i^n | a_i^n, x_i^n, z_i^n) \rangle_{q^d(z_i^n)} + \gamma \sum_{t,i} \langle \log P(c_i^t | a_i^t, x_i^t, z_i^t) \rangle_{q^c(z_i^t)} \\ &= \sum_{n,i} \{ \langle \log P^d_{z_i^n}(d_i^n | a_i^n, x_i^n) \rangle_{q^d(z_i^n)} + \langle \log P^c_{z_i^n}(c[d_i^n] | a_i^n, x_i^n) \rangle_{q^d(z_i^n)} \\ &- \langle \log \sum_{d' \in c[d_i^n]} P^d_{z_i^n}(d' | a_i^n, x_i^n) \rangle_{q^d(z_i^n)} \} + \gamma \sum_{t,i} \langle \log P^c_{z_i^t}(c_i^t | a_i^t, x_i^t) \rangle_{q^c(z_i^t)}\end{aligned}$$

- Assume that the output of the coarse label classifier can be approximated by the detailed label classifier (i.e., they are consistent during training), we have

$$P^c_{z_i}(f(d_i) | a_i, x_i) \approx \sum_{d \in s(d_i)} P^d_{z_i}(d | a_i, x_i).$$

  Then the log likelihood function can be simplified as

$$\mathcal{L} \approx \sum_{n,i} \langle \log P^d_{z_i^n}(d_i^n | a_i^n, x_i^n) \rangle_{q^d(z_i^n)} + \gamma \sum_{t,i} \langle \log P^c_{z_i^t}(c_i^t | a_i^t, x_i^t) \rangle_{q^c(z_i^t)}$$

- Denote the parameters in the kth detailed-label classifier as ν_k, and consider the gradient of $\mathcal{L}$ w.r.t. ν_k:

$$\frac{\partial \mathcal{L}}{\partial \nu_k} = \sum_{n,i} q^d(z_i^n = k) \frac{\partial}{\partial \nu_k} \log P^d_k(d_i^n | a_i^n, x_i^n).$$

- Denote the parameters in the kth coarse-label classifier as μ_k, then the gradient of $\mathcal{L}$ w.r.t. μ_k can be written as

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = \sum_{t,i} q^c(z_i^t = k) \frac{\partial}{\partial \mu_k} \log P^c_{z_i^t}(c_i^t | a_i^t, x_i^t).$$