Incorporating Semantic Constraints into a Discriminative Categorization and Labelling Model.

Ariadna Quattoni Michael Collins Trevor Darrell
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139
{ariadna, mcollins, trevor}@csail.mit.edu

Abstract

This paper describes an approach to incorporate semantic knowledge sources within a discriminative learning framework. We consider a joint scene categorization and region labelling task and assume that some semantic knowledge is available. For example we might know what objects are allowed to appear in a given scene. Our goal is to use this knowledge to minimize the number of fully labelled examples (i.e. data for which each region in the image is labelled) required for learning. For each scene category the probability of a given labelling of image regions is modelled by a Conditional Random Field (CRF). Our model extends the CRF framework by incorporating hidden variables and combining class conditional CRFs into a joint framework for scene categorization and region labelling.

We integrate semantic knowledge into the model by constraining the configurations that the latent region label variable can take, i.e. by constraining the possible region labelling for a given scene category. In a series of synthetic experiments, designed to illustrate the feasibility of the approach, adding semantic constraints about object entailment increased the region labelling accuracy given a fixed amount of fully labelled data.

1. Introduction

In this paper we consider the problem of joint scene categorization and region labelling, and investigate how semantic knowledge can be used to reduce the number of fully labelled examples needed for training. That is, we wish to make use of additional semantic knowledge about the structure of the domain. Different from generative approaches where the semantics of the model can be easily understood in terms of an underlaying process, discriminative models don't offer such an interpretation making the incorporation of certain types of semantic knowledge more challenging.

We define the problem as follows, given an image we wish to predict a label for the scene (e.g. forest, street) as well as labels for each image region (e.g. tree, sun, car). We assume that we are given a set of partially labelled examples that consist of pairs of images (images are represented as sets of image regions) and corresponding scene labels. We are also given a small set of fully labelled examples, i.e triples where in addition to the scene labels we are given a label for some or all image regions. In contrast to conventional semi-supervised learning approaches, we have some semantic knowledge available that specifies the region labels allowed for each scene category. For example, our semantic database could tell us that cars appear on streets and highways but not in offices. Our goal is to incorporate this knowledge into our model for joint scene categorization and region labelling so that we can minimize the number of fully labelled examples required for training.

The motivation for this approach is that while labelling images is a costly task semantic knowledge bases are becoming increasingly available, examples of such databases are WordNet, ConceptNet and Cyc (1,2,5). While these databases provide a variety of lexical relationships, we decided to focus on object entailment relationships of the type "object A appears in scene Y" because they seem to be the most useful for the scene recognition task. Notice that potentially such a database could be built automatically from a set of captioned images.

To solve the task of joint scene categorization and region labelling we decided to extend the model developed in [8] because it allows us to combine region labelling and scene classification in a single discriminative framework. Under this framework images are represented as sets of local features and we model the conditional distribution $p(scene|image)$ directly. For each scene category the probability of a given image region labelling is modelled by a class Conditional Random Field.

A key difference of our approach from previous work on CRFs is that we make use of a hidden variable that intuitively models missing image region labels. More specifically, our model defines conditional probabilities $P(y, \mathbf{r} \mid \mathbf{x})$, and hence indirectly $P(y \mid \mathbf{x}) = \sum_{\mathbf{r}} P(y, \mathbf{r} \mid \mathbf{x})$, using a CRF [8] where y is an image label variable and $\mathbf{r}$ is a hidden variable.

In this paper we address the question of how to incorpo-

Figure 1: A high level view of our learning paradigm

rate semantic knowledge into our learning framework. We follow a simple approach in which we constrain the hidden region labelling variables to be consistent with the object entailment information provided by the semantic knowledge source. In other words, we incorporate the semantic information as a hard prior on the possible configurations of the hidden region labelling variable.

In a series of synthetic experiments, we illustrate with a concrete example how incorporating these constraints can increase the region labelling accuracy given a fixed amount of fully labelled data.

2. Background: joint scene categorization and region labelling

We define the joint scene and region labelling task following the part-based object recognition method in [8]. We are given a training set of n partially labelled pairs. Each such example is of the form $(\mathbf{x}, y)$, where $y \in Y$ is a scene category , and $\mathbf{x} = [x_1..x_m]$ is an image where x_j is the j-th image region. For example, the image regions can be obtained with a bottom up segmentation and represented with a gradient based feature descriptor such as SIFT.

We also assume we are given l fully labelled triples of the form $(\mathbf{x}, y, \mathbf{r})$ where $\mathbf{r} = \{r_1, \ldots, r_m\}$ and $\mathbf{r} \in R$ is a hidden region labelling variable. Intuitively, this variable assigns a region label to each image region in $\mathbf{x}$.

From this training set we would like to learn models that map images $\mathbf{x}$ to scene categories y in the scene categorization task, and that map images $\mathbf{x}$ to region labels $\mathbf{r}$ in the region labelling task.

Our approach combines scene categorization with region labelling. To label an image region we first decide the scene category of the image (modelled by the scene variable y) by summing over all possible region labellings. Then we label each image region with the best region labelling given the known scene (modelled by the region labelling variable $\mathbf{r}$).

Given the above definitions, we define a conditional model:

$$P(y, \mathbf{r}|\ \mathbf{x}, \theta) = \frac{e^{\Psi(y,\mathbf{r},\mathbf{x};\theta)}}{\sum_{y',\mathbf{r}} e^{\Psi(y',\mathbf{r},\mathbf{x};\theta)}}. \quad (1)$$

where θ are the parameters of the model and $e^{\Psi(y',\mathbf{r},\mathbf{x};\theta)}$ is a potential function which can be thought as an energy function that measures the compatibility between a scene category y, a region labelling $\mathbf{r}$, and an image $\mathbf{x}$.

Given a new test image $\mathbf{x}$, and parameter values θ^* induced from a training set, we will perform scene categorization, by taking the scene label for the image to be $y^* = argmax_y P(y|\ \mathbf{x}, \theta)$ where $P(y|\ \mathbf{x}, \theta) = \sum_{\mathbf{r}} P(y, \mathbf{r}|\ \mathbf{x}, \theta)$. We then label each image region by finding $\mathbf{r}^* = argmax_{\mathbf{r}} P(\mathbf{r}|y^*, \mathbf{x}; \theta)$.

We use the following objective function in training the parameters θ of the model:

$$L(\theta) = \sum_{t \in\ TrainingSet} \log P(t|\mathbf{x}, \theta) - \frac{1}{2\sigma^2} \parallel \theta \parallel^2 \quad (2)$$

where $P(t|\mathbf{x}, \theta) = P(y|\mathbf{x}, \theta)$ if t is a partially-labelled example and $P(t|\mathbf{x}, \theta) = P(y, \mathbf{r}|\mathbf{x}, \theta)$ if t is a fully-labelled example. The first term in equation (2) is the log-likelihood of the data and the second is a regularization term, we will use gradient ascent to search for the optimal parameters values $\theta^* = argmax_\theta L(\theta)$ under this criterion.

We encode spatial constraints between region labels with an undirected graph structure, where the hidden variables $\{r_1, \ldots, r_m\}$ correspond to vertices in the graph.E denotes the set of edges in the graph and $(j, k) \in E$ denotes that there is an edge in the graph between variables r_j and r_k. E can be an arbitrary graph; intuitively it should capture any domain specific knowledge that we have about the structure of $\mathbf{r}$. For example in our case it could encode spatial consistency between region labels. For this paper the tree E

is formed by running a minimum spanning tree algorithm where the cost of an edge in the graph between r_j and r_k is taken to be the distance between x_j and x_i in the image.

Following [8] we define Ψ to take the following form:

$$\begin{aligned}\Psi(y, \mathbf{r}, \mathbf{x}; \theta) = &\sum_{j=1}^{m} \sum_{l} f_l^1(j, y, r_j, \mathbf{x}) \theta_l^1 \\ &+ \sum_{(j,k) \in E} \sum_{l} f_l^2(j, k, y, r_j, r_k, \mathbf{x}) \theta_l^2\end{aligned} \tag{3}$$

where f_l^1, f_l^2 are functions defining the features in the model, and θ_l^1, θ_l^2 are the components of θ. The f^1 features depend on single hidden variable values in the model, the f^2 features can depend on pairs of values.

More specifically the compatibility between a scene category y, an image $\mathbf{x}$ and a region labelling $\mathbf{r}$ is defined by:

$$\begin{aligned}\Psi(y, \mathbf{r}, \mathbf{x}; \theta) = &\sum_{j} \phi(x_j) \cdot \theta(r_j) + \sum_{j} \theta(y, r_j) \\ &+ \sum_{(j,k) \in E} \theta(y, r_j, r_k)\end{aligned} \tag{4}$$

In this definition, $\phi(x_j) \in \Re^d$ is a feature–vector representing the image region x_j; $\theta(r_j) \in \Re^d$ is a parameter vector; $\theta(y, r_j)$ is a value in the reals modelling the compatibility between a scene category y and a region label r_j; $\theta(y, r_j, r_k)$ is a real value modelling the compatibility between scene category y and pairs of region labels.

The gradient of the objective function with respect to the parameters θ_l^1 corresponding to features $f_l^1(j, y, h_j, \mathbf{x})$ that depend on single hidden variables for a single partially labelled example in the training set can be written as:

$$\begin{aligned}\frac{\partial L(\theta)}{\partial \theta_l^1} = &\sum_{j,a} P(r_j = a \mid y, \mathbf{x}, \theta) f_l^1(j, y, a, \mathbf{x}) \\ - &\sum_{y', j, a} P(r_j = a, y' \mid \mathbf{x}_i, \theta) f_l^1(j, y', a, \mathbf{x}_i)\end{aligned} \tag{5}$$

Since $P(r_j = a \mid \mathbf{x}, \theta)$ and $P(y \mid \mathbf{x}, \theta)$, can be calculated using belief propagation, provided that the graph E forms a tree structure, we can do efficient inference and parameter estimation in the model. A similar calculation shows that $\partial L(\theta)/\partial \theta_l^2$ can also be expressed in terms of expressions that can be calculated using belief propagation.

Similarly the gradient of the objective function with respect to the parameters θ_l^1 for a single fully labelled example in the training set can be written as:

$$\begin{aligned}\frac{\partial L(\theta)}{\partial \theta_l^1} = &\sum_{j,a} P(r_j^u = a \mid y, \mathbf{r}^o \mathbf{x}, \theta) f_l^1(j, y, a, \mathbf{x}) \\ - &\sum_{y', j, a} P(r_j = a, y' \mid \mathbf{x}_i, \theta) f_l^1(j, y', a, \mathbf{x}_i)\end{aligned} \tag{6}$$

where r^u is an unobserved region label and $\mathbf{r^o}$ are the observed region labels. It follows that these can also be calculated using belief propagation and the same holds for $\partial L_i(\theta)/\partial \theta_l^2$.

3 Incorporating Semantic Knowledge

In order to minimize the number of fully labelled examples required for learning we would like to incorporate semantic knowledge. We follow a simple approach where we constrain the hidden region labels to be consistent with the information provided by the knowledge source. That is, we incorporate the semantic knowledge as a hard prior on the configurations of the hidden region label variables.

More formally, let's assume that for every scene category $y \in Y$ we are given information of the form : $S(y) = \{r_1..r_m\}$, that is we are told the allowed region labels for every scene category. For example, if $Y = \{forest, street\}$ and $R = \{tree, sun, water, car, buildings\}$ our semantic knowledge could be of the form $S_{forest} = \{tree, sun, water\}$ and $S_{street} = \{tree, sun, sky, car, building\}$.

Notice that our approach assumes that the valid region labels are known a-priori since the goal of our paper is to show how such a knowledge can be incorporated in a discriminative latent variable model as a hard prior on the configurations of the latent variables. The question of wether semantic knowledge of the type that we assume in this work can be automatically learned from data is another interesting question that we do not intend to address in this paper.

Given the semantic knowledge we define the set S_y to be the set of all region labellings that are consistent with $S(y)$. For example, let's assume that S_{forest} is as defined above and that we wish to label and image of a forest scene that contains four image regions. Then given our definition of consistency $\mathbf{r} = [tree, tree, tree, tree]$ and $\mathbf{r} = [tree, water, water, sun]$ are consistent with $S(forest)$ and belong to S_{forest}, but $\mathbf{r} = [tree, computer, water, sun]$ is not consistent with $S(forest)$ and therefore is not in S_{forest}.

Given the above definitions we incorporate the semantic knowledge by defining the following conditional model:

$$P(y, s_y | \, \mathbf{x}, \theta) = \frac{\sum_{\mathbf{r} \in S_y} e^{\Psi(y, \mathbf{r}, \mathbf{x}; \theta)}}{\sum_{y', \mathbf{r}} e^{\Psi(y', \mathbf{r}, \mathbf{x}; \theta)}}. \tag{7}$$

where s_y represents our semantic knowledge about scene category y. We include s_y in equation 5 to make explicit that we are jointly optimizing for the scene category and its corresponding semantic knowledge. The above formulation effectively constrains the hidden region labellings of

the partially labelled examples for scene y to those that are consistent with $S(y)$.

The gradient for the constraint model for a single partially labelled example in the training set can be written as:

$$\frac{\partial L(\theta)}{\partial \theta_l^1} = \sum_{j,a} P(r_j = a \mid y, s_y, \mathbf{r}, \mathbf{x}, \theta) f_l^1(j, y, a, \mathbf{x}_i) - \sum_{y',j,a} P(r_j = a, y' \mid \mathbf{x}, \theta) f_l^1(j, y', a, \mathbf{x}) \quad (8)$$

where $P(r_j = a \mid y, s_y, \mathbf{x}, \theta) = \frac{\sum_{r_j = a \wedge \mathbf{r} \in s_y} P(\mathbf{r} \mid y, \mathbf{x}, \theta)}{\sum_{\mathbf{r} \in S_y} P(\mathbf{r} \mid y, \mathbf{x}, \theta)}$

which if E is a tree can be calculated using belief propagation, a similar calculation shows that the same is true for $P(r_j = a, r_k = b \mid y, c_y, \mathbf{x}, \theta)$ thus like in the unconstrained model we can do efficient inference and training.

4 Experiments

To show the feasibility of our approach we conducted a set of experiments with a synthetic dataset. The goal of the experiments is to illustrate with a concrete example how our idea can be applied to incorporate partially labelled data and semantic constraints into a joint object recognition and scene categorization model in a principled manner. These experiments are not to be taken as a demonstration of our approach. We are aware that such a demonstration could only be achieved by running experiments on natural images, since that is the only way of testing wether our assumptions about natural image generation are correct. We are currently working on such experiments, as this is a work in progress.

Each example for the different scene categories was generated assuming the semantic constraints were satisfied. Notice that we never created actual imagery, the dataset was simply derived from our assumptions. That is, we use the term image region to denote a set of appearance and location features that we have generated following our assumptions. For experiments with natural images these regions could be obtained by running a bottom-up segmentation over the image. Each region could then be represented with some appearance descriptor such as SIFT.

More specifically, for each region label i in R we assume there is a multivariate Gaussian distribution A_i over the appearance features, and we take each component of this distribution to be independent. In addition we assume that for each region label there is a corresponding L_i multivariate Gaussian distribution over the x and y coordinates of the location of the image region.

To generate an example from a given scene, that is to generate a set of image regions x_k we sample from the appearance and location distributions corresponding to region labels that are allowed according to the semantic knowledge for that scene.

Model	ML	MUL	MULSC
Scene-Categorization Error Rate	25 %	3 %	4 %

Figure 2: Comparative scene categorization error rates

For example if as before $S_{forest} = \{tree, sun, water\}$ we will generate a forest scene by creating a set of image regions that are each sampled from one of the following distributions: $Distribution_{trees} = \{A_{tree}, L_{tree}\}$, $Distribution_{sun} = \{A_{sun}, L_{sun}\}$ and $Distribution_{water} = \{A_{water}, L_{water}\}$.

For these experiments we generated a dataset of 4 scene categories and 13 region labels, we set the semantic knowledge so that there is a significant amount of shared region labels across different scenes. More specifically, scenes 1 and 2 shared 5 region labels and scenes 3 and 4 shared 4 region labels, also one region label was shared across all scenes.

We divided the data into training and testing set containing 20 and 80 examples of each scene category respectively.

We conducted 3 sets of experiments to evaluate the usefulness of incorporating constraints. For all the 3 experiments we used a single fully labelled example for each scene category.

We obtain the labels for the fully labelled examples in the following manner: If an image region was generated by sampling from distributions A_i and L_i we set its label to be r_i. For the partially labelled examples we obtain scene labels in the following manner: If an image (i.e. a set of image regions) was generated by sampling from the image region distributions $Distribution_i = \{A_i, L_i\}$ corresponding to scene y, we set the scene label for the image to be y.

For the first experiment we trained a model using the fully labelled data only (model ML), for the second experiment we trained a model with the partially labelled and fully labelled data (model MUL) and finally for the third experiment we trained using fully and partially labelled data as well as the semantic constraints (model $MULSC$).

Figure 2 shows scene categorization error rates (percentage of misclassified scenes) for each of the models and Figure 3 shows region labelling error rates (percentage of misclassified image regions). To make a fair comparison of the region labelling performance of each model for the results in Figure 3 we assume that we know the scene category y. We do this so that we can factor out any region labelling performance improvement due to improved scene categorization performance.

From Figure 2 we see that as we would expect the model that was trained with a single training example per class (ML) does significantly worst in terms of scene categoriza-

Model	ML	MUL	MULSC
Region Labelling Error Rate	17 %	15 %	9 %

Figure 3: Comparative region labelling error rate

tion than the other two models. We also see that including semantic constraints doesn't seem to affect scene categorization performance significantly.

From Figure 3 we observe that incorporating semantic knowledge seems to improve region labelling performance significantly. While by just adding unlabelled examples (MLU) the error rate is reduced from 17 % to 15 %, by adding semantic constraints the error rate is further reduced to 9 %.

5 Conclusions and Further Work

In this paper we have addressed the question of how to use a semantic knowledge source to minimize the amount of fully labelled data required for learning a joint scene categorization and region labelling task.

More specifically, we have shown how such knowledge can be incorporated in a discriminative latent variable model as constraints on the allowed region labelling configurations for each scene category. The potential of our approach was illustrated in a set of synthetic experiments that show that incorporating semantic constraints can increase region labelling accuracy given a fixed amount of fully labelled data.

This work is a preliminary study to investigate these ideas, and the synthetic experiments showed the feasibility of the approach. These experiments involved making some intuitive assumptions about the way in which image data is generated. Currently we are starting to run experiments with natural images to test the extent to which such assumptions hold in practice.

References

[1] Christiane Fellbaum. *WordNet , and electoronic lexical database. MIT Press, Cambridge Massachussets*, 1998.

[2] Liu, H. Singh, P. ConceptNet: A Practical Commonsense Reasoning Toolkit. In BT Technology Journal, To Appear. Volume 22,2004.

[3] S. Kumar and M. Hebert. Discriminative random fields: A framework for contextual interaction in classification. In *IEEE Int Conference on Computer Vision*,volume 2, pages 1150-1157, June 2003.

[4] J. Lafferty,A. McCallum and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. Int Conf. on Machine Learning*, 2001.

[5] Matuszek, Cynthia, M. Witbrock, R. Kahlert, J. Cabral, D. Schneider, P. Shah and D. Lenat. Searching for Common Sense: Populating Cyc from the Web. In *Proceedings of the Twentieth National Conference on Artificial Intelligence, Pittsburgh, Pennsylvania* July 2005.

[6] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML-2000*, 2000.

[7] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann,1988.

[8] A. Quattoni, M. Collins and T Darrell. Conditional Random Fields for Object Recognition. In *Neural Information Processing Systems*, 2004.