Tell Me What You See and I will Show You Where It Is

Jia Xu[1] Alexander G. Schwing[2] Raquel Urtasun[2,3]
[1]University of Wisconsin-Madison [2]University of Toronto [3]TTI Chicago
jiaxu@cs.wisc.edu {aschwing, urtasun}@cs.toronto.edu

1. Introduction

Traditional approaches to semantic segmentation require a large collection of training images labeled at the pixel level. Despite the existence of crowd-sourcing systems such as Amazon Mechanical Turk (MTurk), densely labeling images is still a very expensive process, particularly since multiple annotators are typically employed to label each image. Furthermore, a quality control process is frequently required in order to sanitize the annotations.

Here, we are interested in leveraging weak annotations in order to reduce the labeling cost. In particular, we exploit image tags capturing which classes are present in the scene as our sole source of annotation (see Fig. 1 for an illustration). This is an interesting setting as tags are either readily available within most online photo collections or they can be easily obtained at a lesser cost than annotating semantic segmentation. This task is, however, very challenging, as an appearance model cannot be trained due to the fact that the assignment of superpixels to semantic labels is unknown, even at training time.

Several approaches have investigated this setting. In early work, Verbeek and Triggs [9] proposed the latent aspect model, which employs probabilistic latent semantic analysis (PLSA) to model each image as a finite mixture of latent classes also referred to as aspects. The authors extended this approach to capture spatial relationships via a Markov random field (MRF). This model was further extended in a series of papers by Vezhnevets *et al*. [10, 11], for example, to leverage information between multiple images. However, the resulting optimization problem is very complex and non-smooth, making learning a very difficult task. As a consequence, several heuristics were employed to make the problem computationally tractable.

In our recent work [12], we show that this problem can be formalized as the one of learning in a latent structured prediction framework, where the graphical model encodes the presence/absence of a class as well as the assignments of semantic labels to superpixels. As a consequence, we are able to leverage algorithms with good theoretical properties which have been developed for this more general setting. Under our model, different levels of supervision can be simply expressed by specifying which variables are latent and which are observed, without changing the learning and inference algorithms. We demonstrate the effectiveness of our approach using the challenging SIFT-flow dataset [2], showing improvements of 7% in terms of mean class accuracy over the state-of-the-art [12].

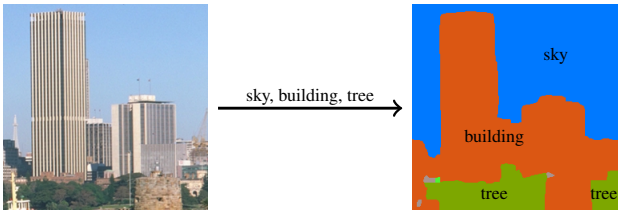

Figure 1. Our approach takes labels in the form of which classes are present in the scene during training, and learns a segmentation model, even though no annotations at the pixel-wise are available.

2. Weakly Labeled Semantic Segmentation

In this paper we investigate how weak supervision can be used in order to perform semantic segmentation. In particular, we focus on the case where the supervision is given by means of a set of tags, describing which classes are present in the image. Towards this goal, we frame the problem as the one of learning in a graphical model encoding the presence and absence of each class $y_i \in \{0, 1\}$ as well as the semantic class h_i of each superpixel i. We define the potentials to be the sum of unary terms encoding the likelihood of the tags, unary potentials encoding the appearance model for segmentation and pairwise potentials ensuring compatibility between both types of variables. Fig. 2 shows the graphical model encoding the dependencies introduced by this probabilistic model, with gray-colored nodes depicting observed variables.

During learning, we are interested in estimating a linear combination of features such that the distribution is able to discriminate between 'good' and 'bad' assignments for variables $\mathbf{y} = (y_1, \dots, y_C)$ and $\mathbf{h} = (h_1, \dots, h_N)$. This problem is particularly challenging due to the non-convexity of the objective function and summations over exponentially sized sets $\mathbf{h}$ and $\mathbf{y}$. However, we note the cost function of our program is a difference of terms, each being

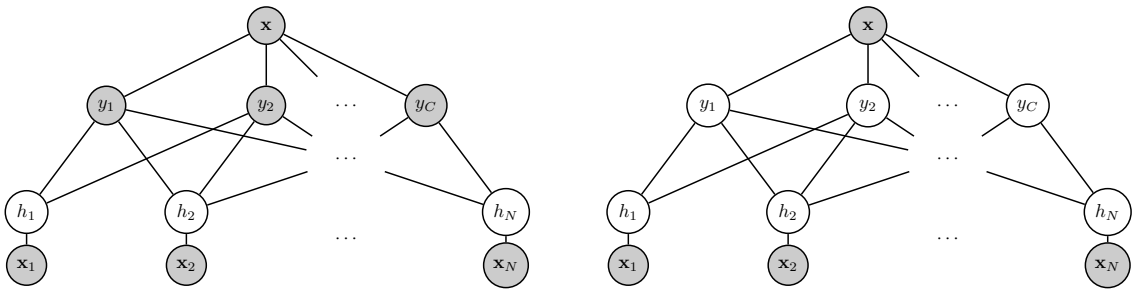

Figure 2. **Graphical Model:** (Left) Graphical model for learning as well as inference when the tags are provided. (Right) Graphical model for inference when the tags are not provided at test time. $y_i \in \{0, 1\}$ describes whether the i-th class is present in the image. $h_j \in \{1, \ldots, C\}$ denotes the semantic label associated with the j-th superpixel. x is the image evidence.

Method	Supervision	Per-Class (%)
Tighe et al. [7]	full	39.2
Tighe et al. [8]	full	30.1
Liu et al. [2]	full	24
Vezhnevets et al. [10]	weak	14
Vezhnevets et al. [11]	weak	21
Ours (CNN-Tag)	weak	**27.9**
Ours (Truth-Tag)	weak	**44.7**

Table 1. Comparison to state-of-the-art on the SIFT-flow dataset. We outperformed the state-of-the-art in the weakly supervised setting by 7%.

convex in the parameters. We exploit this fact and design an iterative optimization scheme via the concave-convex procedure (CCCP) [13]. Due to the bi-partite graphical model structure, estimating the unobserved $\mathbf{h}$ is efficiently possible in each learning step [5] and is guaranteed to converge to a stationary point [6]. For inference, we use distributed convex belief propagation (dcBP) [4], which also has convergence guarantees.

In our experiments, we exploit two settings. In the first case, we follow the standard weakly labeled setting, in which only image level tags are given for training and no annotations are given at the pixel-level. During testing, no source of annotation is provided. Learning in this setting corresponds to the graphical model in Fig. 2 (left), while inference is shown on Fig. 2 (right). We build an image-tag classifier which leverages deep convolutional neural network (CNN) features [1], and a linear SVM per class to form the final potential. We refer to this setting as "Ours (CNN-Tag)."

In the second setting we assume that tags are given both at training and test time, and thus the graphical model in Fig. 2 (left) depicts both learning and inference. This is a natural setting when employing image collections where tags are readily available. We denote this setting as "Ours (Truth-Tag)."

Comparison to the state-of-the-art: Tab. 1 compares our approach to state-of-the-art weakly labeled approaches. For reference, we also include the state-of-the-art when pixel-wise labels are available at training (fully labeled setting). We would like to emphasize that our approach outperforms significantly (7% higher) all weakly labeled approaches. Furthermore, we even outperform the fully supervised method developed by Liu *et al*. [2]. Our Truth-Tag setting almost doubles the per-class accuracy of the previous setting. Surprisingly, we outperformed all fully labeled approaches while not requiring any example to be labeled at the pixel-level.

Our novel view of the problem can be used to incorporate other types of supervision without changing the learning or inference algorithms. In the future we plan to exploit other annotations such as the type of scene or bounding boxes as well as other forms of learning such as active learning [3] to further reduce the need of supervision.

References

[1] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013. 2

[2] C. Liu, J. Yuen, and A. Torralba. Nonparametric Scene Parsing via Label Transfer. *PAMI*, 2011. 1, 2

[3] W. Luo, A. Schwing, and R. Urtasun. Latent structured active learning. In *NIPS*, 2013. 2

[4] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed Message Passing for Large Scale Graphical Models. In *Proc. CVPR*, 2011. 2

[5] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient Structured Prediction with Latent Variables for General Graphical Models. In *Proc. ICML*, 2012. 2

[6] B. Sriperumbudur and G. Lanckriet. On the convergence of the concave-convex procedure. In *Proc. NIPS*, 2009. 2

[7] J. Tighe and S. Lazebnik. Finding Things: Image Parsing with Regions and Per-Exemplar Detectors. In *Proc. CVPR*, 2013. 2

[8] J. Tighe and S. Lazebnik. Superparsing - Scalable Nonparametric Image Parsing with Superpixels. *IJCV*, 2013. 2

[9] J. Verbeek and B. Triggs. Region classification with Markov field aspect models. In *Proc. CVPR*, 2007. 1

[10] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised semantic segmentation with a multi image model. In *Proc. ICCV*, 2011. 1, 2

[11] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly Supervised Structured Output Learning for Semantic Segmentation. In *Proc. CVPR*, 2012. 1, 2

[12] J. Xu, A. G. Schwing, and R. Urtasun. Tell Me What You See and I will Show You Where It Is. In *Proc. CVPR*, 2014. 1

[13] A. L. Yuille and A. Rangarajan. The Concave-Convex Procedure (CCCP). *Neural Computation*, 2003. 2