Learning to Segment Under Various Forms of Weak Supervision

Jia Xu[1], Alexander G. Schwing[2], Raquel Urtasun[2]
[1]University of Wisconsin-Madison [2]University of Toronto

Despite the promising performance of conventional fully supervised algorithms, semantic segmentation has remained an important, yet challenging task. Due to the limited availability of full annotations, it is of great interest to design solutions for semantic segmentation that take into account weakly labeled data, which is readily available at a much larger scale.

Supervision in the form of partial labels has been effectively utilized in interactive object segmentation (e.g. graph-cuts [1]). Recursively propagation of segmentations from labeled masks to unlabeled images has also been employed [2]. An alternative weak supervision are bounding boxes. Grabcut has been a great success for binary object segmentation when only a bounding box is provided around the foreground object [3]. Recent research has extended this idea to semantic segmentation by building object detectors from bounding boxes [7]. A more challenging setting is to infer pixel-wise labeling when only image level tags are given. Researchers have shown encouraging results for this weakly supervised semantic segmentation problem by connecting super-pixels across images, and jointly inferring pixel labels for all images [6], or learning a Markov random field with latent variables representing the super-pixels and observed variables representing tags [8].

In this work, we tackle the problem of semantic segmentation under various modes of weak supervision in the form of image level tags, strokes (*i.e.*, partial labels) as well as bounding boxes. We refer the reader to Fig. 1 for an illustration of the problem. We are interested in inferring pixel-level semantic labels for all the images, as well as learning an appearance model for each semantic class that will allow us to make predictions in new images (test examples). Note that this is extremely difficult as in most of our settings we never observe a single pixel labeled. We formulate the problem as a max margin clustering, where supervision comes as additional constraints in the assignments of pixels to class labels. This allows us to have a unified formulation that can exploit arbitrary combinations of different types of supervision.

Following recent research [6, 8], we first over-segment all images into a total of n super-pixels. For each super-pixel $p \in \{1,\dots,n\}$, we then extract a d dimensional feature vector $\mathbf{x}_p \in \mathbb{R}^d$. Let the matrix $H = [\mathbf{h}_1,\dots,\mathbf{h}_n]^T \in \{0,1\}^{n\times C}$ contain the hidden semantic labels for all super-pixels. We use a 1-of-C encoding, and thus a C-dimensional column vector $\mathbf{h}_p \in \{0,1\}^C$, with C denoting the number of semantic classes.

Our objective is motivated by the fully supervised setting and the success of max-margin classifiers. As the assignments of super-pixels to semantic labels is not known, not even for the training set, supervised learning is not possible. Instead, we take advantage of max-margin clustering (MMC) [9] which searches for those assignments that maximize the margin. We therefore aim at minimizing the regularized margin violation

$$\begin{aligned}\min_{W,H} \quad & \frac{1}{2}\mathrm{tr}(W^TW)+\lambda\sum_{p=1}^{n}\xi(W;\mathbf{x}_p,\mathbf{h}_p) \\ \text{s.t.} \quad & H\mathbf{1}_C=\mathbf{1}_n, \quad H\in\{0,1\}^{n\times C}, \quad H\in\mathcal{S},\end{aligned} \qquad (1)$$

where $W=[\mathbf{w}_1,\dots,\mathbf{w}_C]\in\mathbb{R}^{d\times C}$ is a weight matrix encoding the learned appearance model, $\mathbf{1}_C$ is an all ones vector of length C, and $\mathbf{1}_n$ is an all ones vector of length n. The parameter λ balances the regularization term $\mathrm{tr}(W^TW)$ and the loss contribution $\xi(W;\mathbf{x}_p,\mathbf{h}_p)$. $\mathcal{S}$ is the constrained space for each form of weak annotation (*i.e.*, tags, partial labels, bounding boxes), and it turns out to be linear.

During learning, we jointly optimize for the feature weight matrix W encoding the appearance model, and the semantic labels H for all n super-pixels. Eq. (1) is a non-convex mixed integer programing problem, which is challenging to directly optimize. Taking a close look at Eq. (1), we observe that our optimization problem is bi-convex, *i.e.*, it is convex w.r.t. W if H is fixed, and convex w.r.t. H if W is fixed. Further, our constraints only involve H, and they are linear. We thus employ an alternating procedure to do learning where we alternate between optimizing H and W for fixed values of W and H respectively. It is easy to see that for fixed H, the resulting problem is equivalent to the fully supervised setting, where the labels come from the current estimate of H. In our formulation this decomposes into C different 1-vs-all SVMs which can be trained in parallel.

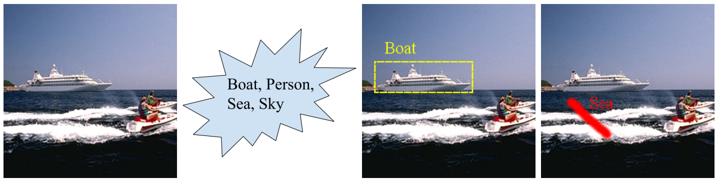

Figure 1: Our semantic segmentation algorithm learns from various forms of weak supervision (image level tags, bounding boxes, partial labels), and produces pixel-wise labels.

When optimizing w.r.t. the assignment matrix H for a fixed appearance model W, we need to solve a constrained optimization problem where both the objective and the constraints are linear. In addition, H is required to be binary, resulting in an integer linear program (ILP). Such optimization problems are generally NP-hard. However, we show that in our case we can decompose the problem into smaller tasks that can be optimally solved in parallel via an LP relaxation. This LP relaxation is guaranteed to retrieve an integer solution, and thus an optimal integral point.

Method	Supervision	Per-class	Per-pixel
Vezhnevets et al. [5]	weak (tags)	14	N/A
Vezhnevets et al. [6]	weak (tags)	22	51
Rubinstein et al. [4]	weak (tags)	29.5	**63.3**
Xu et al. [8]	weak (tags)	27.9	N/A
Ours	weak (tags)	**41.4**	62.7

Table 1: Comparison to state-of-the-art on the SIFT-flow dataset.

To sum up, our approach when compared to existing weakly labeled methods [4, 6, 8] is very efficient, taking only 20 minutes for learning and a fraction of a second for inference. We conduct a rigorous evaluation on the challenging Siftflow dataset for various weakly labeled settings, and demonstrate that our approach outperforms the state-of-the-art by 12% in per-class accuracy (as shown in Tab. 1), while maintaining comparable per-pixel accuracy.

[1] Y. Boykov and M.-P. Jolly. Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images. In *Proc. ICCV*, 2001.

[2] Matthieu Guillaumin, Daniel KÃijttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *IJCV*, 110(3):328–348, 2014. ISSN 0920-5691.

[3] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": interactive foreground extraction using iterated graph cuts. *Siggraph*, 2004.

[4] Michael Rubinstein, Ce Liu, and William T. Freeman. Annotation propagation in large image databases via dense image correspondence. In *Proc. ECCV*, 2012.

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

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

[7] Wei Xia, Csaba Domokos, Jian Dong, Loong-Fah Cheong, and Shuicheng Yan. Semantic segmentation without annotating segments. In *Proc. ICCV*, 2013.

[8] Jia Xu, Alexander G. Schwing, and Raquel Urtasun. Tell me what you see and i will show you where it is. In *Proc. CVPR*, 2014.

[9] Bin Zhao, Fei Wang, and Changshui Zhang. Efficient multiclass maximum margin clustering. In *Proc. ICML*, 2008.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.