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

### Motivation

Observations:
- Full annotations are expensive to collect
- Weak labelings are easy to obtain and available at larger scale
- Different algorithms have been developed for different forms of weak supervision

Unified pixel-wise semantic segmentation algorithm to learn from various forms of weak supervision like image level tags, bounding boxes and partial labels

### Problem Formulation

**Task:** Segment \(m\) images into \(n\) super-pixels with \(C\) categories

**Max-Margin Objective:**

\[
\min_{W,H} \frac{1}{2} \text{tr}(W^T W) + \lambda \sum_{c=1}^{C} \sum_{j=1}^{n} \xi(w_j; x_p, h_p^c) \quad \text{s.t.} \quad H 1_c = 1_n, \quad H \in S
\]

- Feature matrix: \(X = [x_1^T, x_2^T, \ldots, x_n^T] \in \mathbb{R}^{n \times d}\)
- Latent assignment matrix: \(H = [h_1^T, h_2^T, \ldots, h_n^T] \in \{0, 1\}^{n \times C}\)
- Appearance model matrix: \(W \in \mathbb{R}^{d \times C}\)
- Surrogate loss:

\[
\xi(w_j; x_p, h_p^c) = \begin{cases} 
\text{max}(0, 1 + (w_j^T x_p) h_p^c) & h_p^c = 0 \\
\text{max}(0, 1 - (w_j^T x_p)) & h_p^c = 1
\end{cases}
\]

Inference:

\[
\hat{H} = \arg \max_{H} W^T H
\]

Asymmetric loss:

\[
\mu_c = \frac{\sum_{i=1}^{n} (h_i^c - 1)}{\sum_{i=1}^{n} (h_i^c = 1)}
\]

Penalizes according to ratio of negative vs. positive examples

**Supervision as Constraints:**
- Unlabeled: \(S = \emptyset\)
- Image level tags: \(S = \{H \leq BZ, B^T H \geq Z\}\)
- Bounding boxes: \(S = \{H \leq BZ, B^T H \geq Z\}\)
- Semi-supervised: \(S = \{H_1 = H_2\}\)

**An Example:** 2 images of 2 and 3 super-pixels, tagged with classes \([1, 2]\) and \([2, 3]\)

\[
B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad BZ = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}
\]

\[
Z = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad B^T H = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}
\]

### Unified Segmentation Algorithm

**Model:**

\[
\min_{W,H} \frac{1}{2} \text{tr}(W^T W) + \lambda \sum_{j=1}^{n} \xi(W; x_p, h_p) \quad \text{s.t.} \quad H 1_c = 1_n, \quad H \in S
\]

**Challenges:** non-convex mixed integer programming

**Observations:**
- 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
- Constraints are linear and they only involve the super-pixel assignment matrix \(H\)

**Learning to segment by alternating optimization:**

1. Fix \(H\) solve for \(W\) independently for all classes (1-vs-all linear SVM)
2. Fix \(W\) infer super-pixel labels \(H\) in parallel for all images (small LP instances, inference task)

**Inference:**

\[
\max_{H} \text{tr}(X W^T) \quad \text{s.t.} \quad H 1_c = 1_n, \quad H \in \{0, 1\}^{n \times C}, \quad H \in S
\]

**Proposition:** The inference task (integer linear program) can be solved to global optimality in polynomial time

**Reason:** Constraint matrix is totally unimodular

**Model Nature:**
- Decomposable and easily parallelizable
- Theoretical guarantee of relaxation quality

**Computation Efficiency:**
- Orders of magnitude faster than the state-of-the-art for training 20 min vs. 24 hours
- 10 ms for inference on one image

### Results on SIFT-FLOW

**Comparison to state-of-the-art on MSRC:**

<table>
<thead>
<tr>
<th>Method</th>
<th>Supervision</th>
<th>Per-class</th>
<th>Per-pixel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu et al. 2011 (PAMI)</td>
<td>full</td>
<td>24</td>
<td>76.7</td>
</tr>
<tr>
<td>Farabet et al. 2012 (ICML)</td>
<td>full</td>
<td>29.5</td>
<td>78.5</td>
</tr>
<tr>
<td>Farabet et al. 2012 (ICML balanced)</td>
<td>full</td>
<td>46.0</td>
<td>74.2</td>
</tr>
<tr>
<td>Tzeng et al. 2012 (CVPR)</td>
<td>full</td>
<td>32.5</td>
<td>77.0</td>
</tr>
<tr>
<td>Liche et al. 2014 (CVPR)</td>
<td>full</td>
<td>38.3</td>
<td>78.6</td>
</tr>
<tr>
<td>Yang et al. 2014 (CVPR)</td>
<td>full</td>
<td>48.7</td>
<td>79.8</td>
</tr>
<tr>
<td>Vezhnevets et al. 2011 (ICCV)</td>
<td>weak (tags)</td>
<td>14</td>
<td>N/A</td>
</tr>
<tr>
<td>Vezhnevets et al. 2012 (CVPR)</td>
<td>weak (tags)</td>
<td>22</td>
<td>51</td>
</tr>
<tr>
<td>Xu et al. 2014 (CVPR)</td>
<td>weak (tags)</td>
<td>27.9</td>
<td>N/A</td>
</tr>
<tr>
<td>Ours (ILT)</td>
<td>weak (tags)</td>
<td>32.0</td>
<td>64.4</td>
</tr>
<tr>
<td>Ours (ILT + transductive)</td>
<td>weak (tags)</td>
<td>40.0</td>
<td>69.0</td>
</tr>
<tr>
<td>Ours (ILT + transductive)</td>
<td>weak (tags)</td>
<td>41.4</td>
<td>62.7</td>
</tr>
</tbody>
</table>

**Results: Connected to state-of-the-art on MSRC:**

**Visual results:**

- Input
- Truth
- Ours

**Other forms of weak supervision:**

- Unlabeled
- Sky
- Mountain
- Road

**Model behavior:**

- A unified model to learn semantic segmentation under various forms of weak supervision
- An efficient algorithm achieving state-of-the-art results

**Summary:**

- A unified model to learn semantic segmentation under various forms of weak supervision
- An efficient algorithm achieving state-of-the-art results