Introduction

- Semantic Segmentation is the task of labelling every pixel with its object category.
- Fully convolutional networks (FCNs) classify pixels independently of each other, and produce noisy predictions which do not respect image edges.
- As a result, Conditional Random Fields (CRF) with pairwise terms \cite{Krahenbuhl11}, encouraging spatial and appearance consistency, are usually used as post-processing.
- We formulate a richer and more expressive CRF model which utilises Higher Order Potentials (potentials defined over cliques of more than two variables).
- We use the differentiable Mean Field inference algorithm to obtain the most probable labelling, and incorporate it as a layer of our neural network.
- This allows us to train our Higher Order CRF end-to-end with an FCN.

Higher Order Potentials

- We introduce two types of higher order potential into our differentiable CRF:
  - Detection potential uses the complementary cues of an object detector to improve segmentations. It helps in cases where initial unaries are poor.
  - Superpixel potential encourages consistency over larger regions, and removes spurious noise from the output.

Formulation

A Conditional Random Field is defined as

\[
\Pr(X = \mathbf{x}|I) = \frac{1}{Z(I)} \exp(-E(x|I)).
\]

In our case, the energy (ignoring conditioning on Image \(I\)) is:

\[
E(x) = \sum_{\alpha} v^u_{\alpha}(x_\alpha) + \sum_{\alpha < \beta} v^p_{\alpha\beta}(x_\alpha, x_\beta) + \sum_{\alpha} v^{\text{Detection}}(x_\alpha) + \sum_{\alpha} v^{\text{Superpixel}}(x_\alpha).
\]

Detection Potential

Our detection uses the output of an object detector as additional cues for segmentation. Intuitively, object detectors can help when our pixelwise predictions are incorrect.

- Assume we have \(D\) object detections for a given image.
- The \(d\)th detection is of the form \((\alpha, s_\alpha, F_\alpha)\).
- \(\alpha \in \mathcal{L}\) is the class label of the \(d\)th detection.
- \(s_\alpha\) is the detection score. \(F_\alpha\) is the set of pixels belonging to the detection foreground, obtained using GrabCut.
- \(N\) is the number of foreground pixels in the \(d\)th detection.
- Introduce binary latent variables, \(Y_1, Y_2, \ldots, Y_N\) — one for each detection
- Models whether detection is accepted or not.
- \(\Pr(Y_d = 1)\) initialised with \(s_\alpha\), the score of the object detector.
- \(w_{\text{Detection}}(\mathbf{x})\) is a learnable weight parameter that is a function of the class label.

This potential encourages consistency between detections, \(Y\), and labelled pixels, \(X\):

\[
\psi^{\text{Detection}}(X_\alpha = x_\alpha, Y_d = y_d) = \begin{cases} 
  w_{\text{Detection}}(\alpha) \prod_{\beta \in F_\alpha} s_\beta^{y_d} & \text{if } y_d = 0 \\
  w_{\text{Detection}}(\alpha) \prod_{\beta \in F_\alpha} s_\beta^{1-y_d} & \text{if } y_d = 1
\end{cases}
\]

Superpixel Potential

Our learnable superpixel potential enforces consistency over regions obtained by superpixels. This is a soft constraint using a Potts type energy \cite{Kokkinos15}. We use superpixels over multiple scales, which do not necessarily have to form a hierarchy.

\[
\psi^{\text{Superpixel}}(X_\alpha = x_\alpha) = \begin{cases} 
  w_{\text{Superpixel}}(\alpha) & \text{if all } x_\beta^{y_\beta} = 1 \\
  w_{\text{Superpixel}} & \text{otherwise.}
\end{cases}
\]

Experimental Results on PASCAL VOC

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean IoU (%)</th>
<th>Method</th>
<th>Mean IoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>77.9</td>
<td>Baseline (Unary + Pairwise)</td>
<td>72.9</td>
</tr>
<tr>
<td>DPN \cite{Liu17}</td>
<td>77.5</td>
<td>Centrale \cite{Krahenbuhl11}</td>
<td>75.7</td>
</tr>
<tr>
<td>Dilated \cite{Yu17}</td>
<td>75.3</td>
<td>BoxSup \cite{Felzenszwalb04}</td>
<td>75.2</td>
</tr>
<tr>
<td>Attention \cite{Liu17}</td>
<td>75.1</td>
<td>CRF-RNN \cite{P3Beyond}</td>
<td>74.7</td>
</tr>
</tbody>
</table>

Extension to Instance Segmentation

We have recently extended our detection potentials for the task of Instance Segmentation \cite{Arnab15}. The detections inform us about possible object instances, and the problem is then to assign each pixel to an instance represented by a detection.

Conclusion

- Introduced two higher order potentials for a CRF which can be integrated into a deep neural network and trained end-to-end.
- Achieved the best performance on the PASCAL VOC dataset.
- In subsequent work \cite{Arnab15}, we have showed how our Detection potential can be used for the task of Instance Segmentation.