

FCN: (Baseline)

\[ \mathcal{L}(\Theta) = \sum p e(X_{\Theta}(p), l(p)) \rightarrow \text{Objective Function} \]

\( p = \text{pixel index}, l(p) = \text{gt label}, X_{\Theta}(p) = \text{net labeling} \)

\[ e(l(p), X_{\Theta}(p)) = \text{per pixel loss} \]

\( \Theta = \text{network parametrization; updated w/ SGD and backprop} \)
Box Sup:  
Overlapping  
Objective Function  
\[ E_0 = \frac{1}{N} \sum_s (1 - \text{IoU}(\mathbf{B}, s)) \delta(\mathbf{B}, s) \]  

\[ S = \text{candidate segment mask} \]  
\[ \mathbf{B} = \text{gt bounding box annotation} \]  
\[ \text{IoU}(\mathbf{B}, s) \in [0, 1] \rightarrow \text{intersection-over-union ratio} \]  
\[ \uparrow \text{IoU} \Rightarrow \uparrow \text{box-candidate mask overlap} \]  

\[ \delta(\mathbf{B}, s) = \begin{cases} 1 & \text{if } \mathbf{B} = s \text{ \hspace{1cm} } L_B = \text{semantic label of bounding box } \mathbf{B} \\ 0 & \text{otherwise} \text{ \hspace{1cm} } L_s = \text{semantic label of candidate segment } s \end{cases} \]  

Minimizing \( E_0 \) implies higher \( \text{IoU}s \) for consistent semantic labels  

\[ N = \# \text{ of candidate segments} \]  

\[ E_r = \sum_p e(X_\theta(p), L_s(p)) \]  

\( L_s(p) = \text{semantic label at pixel } p \text{ used for network training} \)  

Target of regression: estimated candidate segment  

Overarching Objective Function:  
\[ \varepsilon = \min_{\Theta, \mathbf{E}s_3} \sum_i (E_0 + \lambda E_r) \]  

\[ \Sigma = \text{sum over all images} \]  

\[ \lambda = 3 \text{ (fixed weighting parameter)} \]  

Parameters to optimize:  
\[ a) \text{ net parameters } \Theta \]  
\[ b) \text{ labelling of all candidate segments } \mathbf{E}s_3 \]
Full Supervision Loss Function:

Let:
- \( I \) = set of pixels of image \( j \)
- \( N \) = # of pixels
- \( S_{ic} \) = CNN score for pixel \( i \) and class \( c \)
- Softmax probability of \( c \) at \( i \): \( S_{ic} = \frac{e^{S_{ic}}}{\sum_{k=1}^{N} e^{S_{ik}}} \in [0,1] \)
- \( G \) = ground truth map
- \( L \) \( \rightarrow \) pixel \( i \) belongs to class \( G_i \)

Cross-entropy loss

\[
L_{\text{pix}}(S, G) = - \sum_{i \in I} \log(S_{Gi})
\]

(If \( G_i \) undefined, set \( \log(S_{Gi}) = 0 \) for that value of \( i \))

Image-Level Supervision Loss Function:

- \( E_1, \ldots, N_3 \) = set of all classes CNN trained to recognize
- \( L \subseteq E_1, \ldots, N_3 \) = classes present in image
- \( L' \subseteq E_1, \ldots, N_3 \) = classes not present in image

\[
L_{\text{img}}(S, L, L') = - \frac{1}{|L|} \sum_{c \in L} \log(S_{Ec}) - \frac{1}{|L'|} \sum_{c \in L'} \log(1 - S_{Ec})
\]

where \( E_c = \arg \max_{i \in I} S_{ic} \)

Single-image cross-entropy loss

Point-Level Supervision Loss Function:

- \( I_s \) = set of pixels w/ known class; supervised pixels

\[
L_{\text{point}}(S, G, L, L') = L_{\text{img}}(S, L, L') - \sum_{i \in I_s} a_i \log(S_{Gi})
\]

\( a_i \) = relative importance of each supervised pixel

Combines \((1)\) and \((2)\)

\((1)\) only for supervised points
Point-level Supervision w/ Object Prior:

$p_i$ = probability pixel $i$ belongs to an object

$\mathcal{O}$ = set of object classes; $\mathcal{O}'$ = set of background classes

e.g. PASCAL VOC $\Rightarrow$ $\mathcal{O}$ = set of 20 object classes

$\mathcal{O}'$ = generic background class

$$L_{obj}(S, P) = - \frac{1}{|I|} \sum_{i \in I} p_i \log \left( \sum_{c \in \mathcal{O}} s_{ic} \right) + (1 - p_i) \log \left( 1 - \sum_{c \in \mathcal{O}} s_{ic} \right)$$