FCN: (Baseline)

$$\varepsilon(\Theta) = \sum e(X_\Theta(p), l(p))$$ → Objective Function

p = pixel index, $l(p)$ = gt label, $X_\Theta(p)$ = net labelin

$e(l(p), X_\Theta(p))$ = per pixel loss

Θ = network parametrization; updated w/ SGD and backprop

BoxSup:

Overlapping Objective Function $\mathcal{E}_o = \frac{1}{N} \sum_S (1 - IoU(B, S))\, \delta(l_B, l_S)$ (1)

→ Find S w/ greatest overlap w/ B and $l_B = l_S$

S = candidate segment mask

B = gt bounding box annotation

$IoU(B, S) \in [0, 1]$ → intersection-over-union ratio

↑ IoU ⇒ ↑ box-candidate mask overlap

$\delta(l_B, l_S) = \begin{cases} 1 & \text{if } l_B = l_S \\ 0 & \text{otherwise} \end{cases}$

l_B = semantic label of bounding box B

l_S = semantic label of candidate segment S

Minimizing $\mathcal{E}_o$ implies higher IoUs for consistent semantic labels

N = # of candidate segments

$\mathcal{E}_r = \sum_p e(X_\theta(p), l_S(p))$, (2)

$l_S(p)$ = semantic label at pixel p used for network training

Target of regression: estimated candidate segment

Overarching Objective Function: $\mathcal{E} = \min_{\theta, \{l_S\}} \sum_i (\mathcal{E}_o + \lambda \mathcal{E}_r)$ (3)

$\sum_i$ = sum over all images

$\lambda = 3$ (fixed weighting parameter)

Parameters to optimize: a) net parameters θ

b) labelling of all candidate segments $\{l_S\}$

Full Supervision Loss Function:

I = set of pixels of image ; N = # of pixels

s_{ic} = CNN score for pixel i and class c

Softmax probability of c at i: $S_{ic} = \dfrac{e^{s_{ic}}}{\sum_{k=1}^{N} e^{s_{ik}}} \in [0,1]$

G = ground truth map

↳ pixel i belongs to class G_i

Loss on single training image: (Cross-entropy loss) ①

$$L_{pix}(S, G) = -\sum_{i \in I} \log(S_{iG_i})$$

(if G_i undefined, set $\log(S_{iG_i}) = 0$ for that value of i)

Image-Level Supervision Loss Function:

$\{1, \ldots, N\}$ = set of all classes CNN trained to recognize

$L \subseteq \{1, \ldots, N\}$ classes present in image

$L' \subseteq \{1, \ldots, N\}$ classes not present in image

$$L_{img}(S, L, L') = -\frac{1}{|L|}\sum_{c \in L} \log(S_{t_c c}) - \frac{1}{|L'|}\sum_{c \in L'} \log(1 - S_{t_c c}) \quad ②$$

↳ Single-image cross-entropy loss

, where $t_c = \underset{i \in I}{\operatorname{argmax}}\, S_{ic}$

Point-Level Supervision Loss Function:

Combines ① and ② ↑ (① only for supervised points)

I_s = set of pixels w/ known class; supervised pixels

$$L_{point}(S, G, L, L') = L_{img}(S, L, L') - \sum_{i \in I_s} a_i \log(S_{iG_i})$$

a_i = relative importance of each supervised pixel

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 ⇒ $\mathcal{O}$ = set of 20 object classes

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

$$\boxed{L_{obj}(S, P) = -\frac{1}{|I|} \sum_{i \in I} \left[P_i \log\left(\sum_{c \in \mathcal{O}} S_{ic}\right) + (1 - P_i) \cdot \log\left(1 - \sum_{c \in \mathcal{O}} S_{ic}\right) \right]}$$