Supplementary Material: Constrained Convolutional Neural Networks for Weakly Supervised Segmentation

Deepak Pathak Philipp Krähenbühl Trevor Darrell
University of California, Berkeley
{pathak,philkr,trevor}@cs.berkeley.edu

This document provides detailed derivation for the results used in the paper and additional quantitative experiments to validate the robustness of CCNN algorithm.

In the paper, we optimize constraints on CNN output by introducing latent probability distribution $P(X)$. Recall the overall main objective as follows:

$$\begin{aligned} \underset{\theta,P}{\text{minimize}} \quad & -H_P \underbrace{-\mathbb{E}_{X\sim P}\left[\log Q(X|\theta)\right]}_{H_{P|Q}} \\ \text{subject to} \quad & A\vec{P} \geq \vec{b}, \qquad \sum_X P(X) = 1, \end{aligned} \tag{1}$$

where $H_P = -\sum_X P(X)\log P(X)$ is the entropy of latent distribution, $H_{P|Q}$ is the cross entropy and $\vec{P}$ is the vectorized version of $P(X)$. In this supplementary material, we will show how to minimize the objective with respect to P. For the complete block coordinate descent minimization algorithm with respect to both P and θ, see the main paper.

Note that the objective function in (1) is KL Divergence of network output distribution $Q(X|\theta)$ from $P(X)$, which is convex. Equation (1) is convex optimization problem, since all constraints are linear. Furthermore, Slaters condition holds as long as the constraints are satisfiable and hence we have strong duality [1]. First, we will use this strong duality to show that the minimum of (1) is a fully factorized distribution. We then derive the dual function (Equation (4) in the main paper), and finally extend the analysis for the case when objective is relaxed by adding slack variable.

1. Latent Distribution

In this section, we show that the latent label distribution $P(X)$ can be modeled as the product of independent marginals without loss of generality. This is equivalent to showing that the latent distribution that achieves the global optimal value factorizes, while keeping the network parameters θ fixed. First, we simplify the cross entropy term in the objective function in (1) as follows:

$$\begin{aligned} H_{P|Q} &= -\mathbb{E}_{X\sim P}\left[\log Q(X|\theta)\right] \\ &= -\mathbb{E}_{X\sim P}\left[\sum_{i=1}^{n}\log q_i(x_i|\theta)\right] \\ &= -\sum_{i=1}^{n}\mathbb{E}_{X\sim P}\left[\log q_i(x_i|\theta)\right] \\ &= -\sum_{i=1}^{n}\mathbb{E}_{x_i\sim P}\left[\log q_i(x_i|\theta)\right] \\ &= -\sum_{i=1}^{n}\sum_{l\in\mathcal{L}} P(x_i=l)\log q_i(l|\theta) \end{aligned} \tag{2}$$

We used the linearity of expectation and the fact that q_i is independent of any variable X_j for $j \neq i$ to simplify the objective, as shown above. Here, $P(x_i = l) = \sum_{X:x_i=l} P(X)$ is the marginal distribution.

Let's now look at the Lagrangian dual function

$$\begin{aligned} \mathcal{L}(P,\lambda,\nu) =& -H_P + H_{P|Q} \\ &+ \lambda^\top(\vec{b} - A\vec{P}) + \nu\left(\sum_X P(X) - 1\right) \\ =& -H_P + H_{P|Q} - \sum_{i,l}\lambda^\top A_{i;l}\vec{P}(x_i = l) \\ &+ \lambda^\top\vec{b} + \nu\left(\sum_X P(X) - 1\right) \\ =& -H_P \underbrace{-\sum_{i=1}^{n}\sum_{l\in\mathcal{L}} P(x_i=l)\left(\log q_i(l|\theta) + A_{i;l}^T\lambda\right)}_{\tilde{H}_{P|Q}} \\ &+ \lambda^T\vec{b} + \nu\left(\sum_X P(X) - 1\right), \end{aligned} \tag{3}$$

where $\tilde{H}_{P|Q}$ is a biased cross entropy term and $A_{i;l}$ is the column of A corresponding to $p_i(l)$.. Here we use the fact

that the linear constraints are formulated on the marginals to rephrase the dual objective. We will now show this objective can be rephrased as a KL-divergence between a biased distribution $\tilde{Q}$ and P.

The biased cross entropy term can be rephrased as a cross entropy between a distribution $\tilde{Q}(X|\theta,\lambda) = \prod_i \tilde{q}_i(x_i|\theta,\lambda)$, where $\tilde{q}_i(x_i|\theta,\lambda) = \frac{1}{\tilde{Z}_i} q_i(x_i|\theta)\exp(A_{i;x_i}^\top \lambda)$ is the biased CNN distribution and $\tilde{Z}_i$ is a local partition function ensuring $\tilde{q}_i$ sums to 1. This partition function is defined as

$$\tilde{Z}_i = \sum_l \exp\left(\log q_i(l|\theta) + A_{i;l}^\top \lambda\right)$$

The cross entropy between P and $\tilde{Q}$ is then defined as

$$\begin{aligned} H_{P|\tilde{Q}} &= -\sum_X P(X)\log\tilde{Q}(X|\theta,\lambda) \\ &= -\sum_i\sum_l P(x_i{=}l)\log\tilde{q}_i(l|\theta,\lambda) \\ &= -\sum_i\sum_l P(x_i{=}l)(\log q_i(l|\theta,\lambda){+}A_{i;l}^\top\lambda{-}\log\tilde{Z}_i) \\ &= \tilde{H}_{P|Q} + \sum_i \log\tilde{Z}_i \end{aligned} \tag{4}$$

This allows us to rephrase (3) in terms of a KL-divergence between P and $\tilde{Q}$

$$\begin{aligned} &\mathcal{L}(P,\lambda,\nu) \\ &= -H_P + H_{P|\tilde{Q}} - \sum_i \log\tilde{Z}_i + \lambda^T\vec{b} + \nu\left(\sum_X P(X) - 1\right) \\ &= D(P||\tilde{Q}) - C + \lambda^T\vec{b} + \nu\left(\sum_X P(X) - 1\right), \end{aligned} \tag{5}$$

where $C = \sum_i \log\tilde{Z}_i$ is a constant that depends on the local partition functions of $\tilde{Q}$.

The primal objective (1) can be phrased as $\min_P \max_{\lambda\geq 0,\nu} \mathcal{L}(P,\lambda,\nu)$ which is equivalent to the dual objective $\max_{\lambda\geq 0,\nu}\min_P \mathcal{L}(P,\lambda,\nu)$, due to strong duality.

Maximizing the dual objective can be phrased as maximizing a dual function

$$\begin{aligned} \mathcal{L}(\lambda) &= \lambda^\top\vec{b} - C + \max_\nu \min_P D(P\|\tilde{Q}) + \nu\left(\sum_X P(X) - 1\right) \\ &= \lambda^\top\vec{b} - C + \underbrace{\min_{P:\sum_X P(X)=1} D(P\|\tilde{Q})}_{0} \\ &= \lambda^\top\vec{b} - \sum_i \log\sum_l \exp\left(\log q_i(l|\theta) + A_{i;l}^\top\lambda\right), \end{aligned} \tag{6}$$

where the maximization of ν can be rephrased as a constraint on P i.e. $\sum_X P(X) = 1$. Maximizing (6) is equivalent to minimizing the original constraint objective (1).

Factorization The KL-divergence $D(P||\tilde{Q})$ is minimized at $P = \tilde{Q} = \prod_i \tilde{q}_i$, hence the minimal value of P fully factorizes over all variables for any assignment to the dual variables λ.

Dual function Using the definition $q_i(l|\theta) = \frac{1}{Z_i}\exp(f_i(l;\theta))$ we can define the dual function with respect to f_i

$$\mathcal{L}(\lambda) = \lambda^\top\vec{b} - \sum_i \log\sum_l \exp(f_i(l;\theta) + A_{i;l}^\top\lambda) + \underbrace{\sum_i \log Z_i}_{\text{const.}},$$

where the log partition function is constant and falls out in the optimization.

2. Optimizing Constraints with Slack Variable

The slack relaxed loss function is given by

$$\begin{aligned} \underset{\theta,P,\xi}{\text{minimize}} \quad & -H_P \underbrace{-\mathbb{E}_{X\sim P}\left[\log Q(X|\theta)\right]}_{H_{P|Q}} + \beta^\top\xi \\ \text{subject to} \quad & A\vec{P} \geq \vec{b} - \xi, \qquad \xi \geq 0, \qquad \sum_X P(X) = 1 \end{aligned} \tag{7}$$

For any $\beta \leq 0$, a value of $\xi \to \infty$ will minimize the objective and hence invalidate the corresponding constraints. Thus, for the remainder of the section we assume that $\beta > 0$. The Lagrangian dual to this loss is defined as

$$\begin{aligned} \mathcal{L}(P,\lambda,\nu,\gamma) = &-H_P + H_{P|Q} + \beta^\top\xi + \lambda^\top(\vec{b} - A\vec{P} - \xi) \\ &+ \nu\left(\sum_X P(X) - 1\right) - \gamma^\top\xi. \end{aligned} \tag{8}$$

We know that the dual variable $\gamma \geq 0$ is strictly non-negative, as well as

$$\frac{\partial}{\partial\xi}\mathcal{L}(P,\lambda,\nu,\gamma) = \beta - \lambda - \gamma = 0. \tag{9}$$

This leads to the following constraint on λ:

$$0 \leq \gamma = \beta - \lambda.$$

Hence the slack weight forms an upper bound on $\lambda \leq \beta$. Substituting Equation (9) into Equation (8) reduces the dual objective to the non-slack objective in Equation (3), and the rest of the derivation is equivalent.

3. Ablation Study for Parameter Selection

In this section, we present results to analyze the sensitivity of our approach with respect to the constraint parameters i.e. the upper and lower bounds. We performed line search along each of the bounds while keeping others fixed. The method is quite robust with a standard deviation of 0.73% in accuracy, averaged over all parameters, as shown in Table 1. These experiments are performed in the setting where image-level tag and 1-bit size supervision is available during training, as discussed in the paper. We attribute this robustness to the slack variables that are learned per constraint per image.

Fgnd lower a_l	Bgnd lower a_0	Bgnd upper b_0	mIoU w/o CRF
0.1	**0.2**	**0.7**	**40.5**
0.2	0.2	0.7	40.6
0.3	0.2	0.7	40.6
0.4	0.2	0.7	39.6
0.1	0.1	0.7	40.5
0.1	0.3	0.7	40.4
0.1	0.4	0.7	40.5
0.1	0.5	0.7	40.4
0.1	0.2	0.5	36.6
0.1	0.2	0.6	38.9
0.1	0.2	0.8	39.7

Table 1: Ablation study for sensitivity analysis of the CCNN optimization with respect to the chosen parameters. The paramters mentioned here are defined in Equations (8) and (9) in the main paper. Parameter values used in all other experiments are shown in bold.

4. Ablation Study for Semi-Supervised Setting

In this section, we experiment by incorporating fully supervised images in addition to our weak objective. The accuracy curve is depicted in Figure 1.

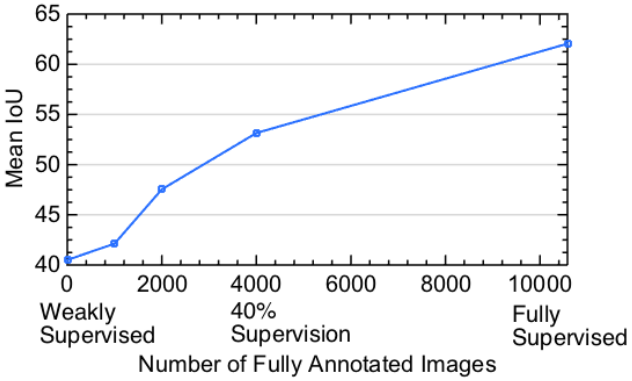

Figure 1: Ablation study with varying amount of fully supervised images. Our model makes good use of the additional supervision.

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 1