Appendix for Holistic Context Models for Visual Recognition

Appendix I
Generalized Expectation maximization (GEM)

The parameters $\Lambda^w = \{\beta_k^w, \boldsymbol{\alpha}_k^w\}$ of the contextual class models of (5) are learned using GEM. This is an extension of the well known EM algorithm, applicable when the M-step of the latter is intractable. It consists of two steps. The E-Step is identical to that of EM, computing the expected values of the component probability mass β_k. The generalized M-step estimates the parameters $\boldsymbol{\alpha}_k$. Rather than solving for the parameters of maximum likelihood, it simply produces an estimate of higher likelihood than that available in the previous iteration. This is known to suffice for convergence of the overall EM procedure [2]. We resort to the Newton-Raphson algorithm to obtain these improved parameter estimates, as suggested in [3] for single component Dirichlet distributions. Omitting the dependence on the concept index w for brevity, the algorithm iterates between two steps,

E-step: compute

$$h_{dk} = \frac{\mathcal{D}ir(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_k)\beta_k}{\sum_l \beta_l \mathcal{D}ir(\boldsymbol{\pi}_d; \boldsymbol{\alpha}_l)} \tag{1}$$

M-step: set

$$(\beta_k)^{new} = \frac{N_k}{N}, \quad \text{where} \quad N = \sum_{dk} h_{dk}, N_k = \sum_d h_{dk} \tag{2}$$

$$(\boldsymbol{\alpha}_k)^{new} = (\boldsymbol{\alpha}_k)^{old} + \mathcal{H}^{k^{-1}}\mathbf{g}^k \tag{3}$$

$$\text{where} \quad \mathbf{g}_i^k = N_k(\Psi(\sum_{p=1}^L \alpha_p) - \Psi(\alpha_i)) + \sum_d h_{dk} \log \pi_{id} \tag{4}$$

$$\text{and} \quad \mathcal{H}_{ii}^k = N_k(\Psi'(\sum_{p=1}^L \alpha_p) - \Psi'(\alpha_i)) \tag{5}$$

$$\mathcal{H}_{ij}^k = N_k(\Psi'(\sum_{p=1}^L \alpha_p)), \tag{6}$$

Ψ and Ψ' are the Digama and Trigamma functions [3].

Appendix II
Computation of Image-SMNs

Given N patch-based SMNs, $\boldsymbol{\pi}^{(n)}$, the Image-SMN $\boldsymbol{\pi}^*$ is

$$\begin{aligned}\boldsymbol{\pi}^* &= \arg\min_{\boldsymbol{\pi}} \frac{1}{N}\sum_{n=1}^N KL(\boldsymbol{\pi}||\boldsymbol{\pi}^{(n)}) \\ &= \arg\min_{\boldsymbol{\pi}} \frac{1}{N}\sum_{n=1}^N\sum_{i=1}^L \pi_i \log\frac{\pi_i}{\pi_i^{(n)}} \\ &= \arg\min_{\boldsymbol{\pi}} \frac{1}{N}\sum_{n=1}^N\sum_{i=1}^L \left[\pi_i\log\pi_i - \pi_i\log\pi_i^{(n)}\right]\end{aligned}$$

subject to $\sum_{i=1}^L \pi_i = 1$. This has Lagrangian

$$\mathcal{L}(\pi,\lambda) = \sum_{i=1}^L \pi_i \log\pi_i - \frac{1}{N}\sum_{i=1}^L \pi_i \sum_{n=1}^N \log\pi_i^{(n)} + \frac{\lambda}{N}(1-\sum_{i=1}^L \pi_i).$$

Setting derivatives with respect to π_i to zero leads to

$$1 + \log\pi_i - \frac{1}{N}\sum_{n=1}^N \log\pi_i^{(n)} - \frac{\lambda}{N} = 0, \tag{7}$$

$$\text{or} \qquad \pi_i = \exp\left(\hat{\lambda} + <\log\pi_i>\right) \tag{8}$$

where $<\log\pi_i> = \frac{1}{N}\sum_{n=1}^N \log\pi_i^{(n)}$ and $\hat{\lambda} = \frac{\lambda}{N} - 1$. Summing over i and using the constraint $\sum_i \pi_i = 1$,

$$1 = \exp(\hat{\lambda})\sum_{i=1}^L \exp <\log\pi_i> \tag{9}$$

$$\exp(\hat{\lambda}) = \frac{1}{\sum_{i=1}^L \exp <\log\pi_i>}. \tag{10}$$

Substituting (10) in (8),

$$\pi_i^* = \frac{\exp <\log\pi_i>}{\sum_{i=1}^L \exp <\log\pi_i>} \tag{11}$$

$$= \frac{\exp\frac{1}{N}\sum_n \log\pi_i^{(n)}}{\sum_i \exp\frac{1}{N}\sum_n \log\pi_i^{(n)}}. \tag{12}$$

Appendix III
Variational Approximation

Variational methods approximate the posterior $P(\boldsymbol{\pi}, w_{1:N}|x_{1:N})$ by a mean-field variational distribution $q(\boldsymbol{\pi}, w_{1:N})$, indexed by free variational parameters, within some class of tractable probability distributions $\mathcal{F}$. These distributions usually assume independent factors,

$$q(\boldsymbol{\pi}, w_{1:N}) = q(\boldsymbol{\pi};\boldsymbol{\gamma})\prod_n q(w_n;\boldsymbol{\phi}_n) \tag{13}$$

where $q(y)$ and $q(z_i)$ are categorical models, and $q(\boldsymbol{\pi})$ a Dirichlet distribution. Given an observation $x_{1:N}$, the optimal variational approximation minimizes the Kullback-Leibler (KL) divergence between the two posteriors

$$q^* = \arg\min_{q\in\mathcal{F}} KL(q(\boldsymbol{\pi}, w_{1:N})||P(\boldsymbol{\pi}, w_{1:N}|x_{1:N})) \tag{14}$$

$$= \mathcal{L}(q(\boldsymbol{\pi}, w_{1:N})) + \log P(x_{1:N}) \tag{15}$$

where,

$$\mathcal{L}(q(\boldsymbol{\pi}, w_{1:N})) = E_q[\log q(\boldsymbol{\pi}, w_{1:N})] - E_q[\log P(\boldsymbol{\pi}, w_{1:N}, x_{1:N})]. \tag{16}$$

Since the data likelihood $P(x_{1:N})$ is constant for an observed image, the optimization problem is identical to

$$q^*(\boldsymbol{\pi}, w_{1:N}) = \arg\min_{q\in\mathcal{F}} \mathcal{L}(q(\boldsymbol{\pi}, w_{1:N})), \tag{17}$$

From Appendix A.3 of [1], the update rule for coordinate descent of the variational parameters is

$$\gamma_i^* = \sum_n \phi_{ni} + \alpha_i \tag{18}$$

$$\phi_{ni}^* \propto P_{X|W}(x_n|w_n = i)\ e^{\psi(\gamma_i)-\psi(\sum_j \gamma_j)} \tag{19}$$

such that $\sum_i \phi_{ni} = 1$ and, where α_i are the parameters of the prior class distribution $P(\boldsymbol{\pi};\boldsymbol{\alpha})$ and ψ is the Digamma function [3]. Once the parameters of the variational distribution are obtained, the SMN for an image can be computed as,

$$\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}} q(\boldsymbol{\pi};\boldsymbol{\gamma}) \tag{20}$$

$$= \arg\max_{\boldsymbol{\pi}} \log q(\boldsymbol{\pi};\boldsymbol{\gamma}) \tag{21}$$

$$= \arg\max_{\boldsymbol{\pi}} \sum_j^L (\gamma_j - 1)\log \pi_j \tag{22}$$

$$\text{such that, } \sum_j \pi_j = 1 \tag{23}$$

Using the Lagrange multiplier, λ, we get

$$J(\boldsymbol{\pi}, \lambda) = \sum_j^L (\gamma_j - 1)\log \pi_j + \lambda(1 - \sum_j^L \pi_j) \tag{24}$$

Taking partial derivatives with respect to, π_j and λ and setting them to zero we get,

$$\frac{\partial J}{\partial \pi_j} = \frac{(\gamma_j - 1)}{\pi_j} - \lambda = 0, \forall j \tag{25}$$

$$\frac{\partial J}{\partial \lambda} = 1 - \sum_j^L \pi_j = 0 \tag{26}$$

From (25) and (26) we get,

$$\pi_j = \frac{\gamma_i - 1}{\sum_j \gamma_j - L} \tag{27}$$

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, B-39, 1977.

[3] T. Minka. Estimating a dirichlet distribution. *http://research.microsoft.com/ minka/papers/dirichlet/*, 1:3, 2000.