Appendix

A.1. Variational Inference

In both the variational prediction rule (5) and the parameter estimation, for each document we need to infer the variational distribution $q_s(\psi, o)$ for each category s. This can be efficiently done by using the mean field method. Specifically, we assume

$$q_s(\psi, o) = q(\psi|\nu_s, \text{diag}(\tau_s^2)) \prod_n q(o_n|\phi_n),$$

where $q(\psi|\nu_s, \text{diag}(\tau_s^2)) = \mathcal{N}(\nu_s, \text{diag}(\tau_s^2))$, and $q(o_n|\phi_{sn})$ is a multinomial distribution with parameter ϕ_{sn}. Then, we have the variational lower bound

$$\mathcal{L}_{-\theta}(q_s, \Theta) = \mathbb{E}[\log p(\psi|\mu_s, \Sigma_s)] + \sum_n \mathbb{E}[\log p(o_n|\psi) + \log p(\mathbf{r}_n|o_n, \beta) + \log p(\mathbf{x}_n|o_n, \eta)] - H(q),$$

where all the terms except the second one can be efficiently computed. Similar as in CTMs [3], we need to further approximate the second term as[5]

$$\mathbb{E}[\log p(o_n|\psi)] \geq \sum_k \phi_{nk}\nu_{sk} - \frac{\sum_k \exp(\nu_{sk} + \frac{1}{2}\tau_{sk}^2)}{\zeta_s} + 1 - \log \zeta_s$$

where ζ_s is an additional variational parameter. Now, posterior inference is to find a q that maximizes the lower bound $\mathcal{L}_{-\theta}$, whose second term is approximated with the above inequality.

The first term of $\mathcal{L}_{-\theta}$ is $\mathbb{E}[\log p(\psi|\mu_s, \Sigma_s)]$, and it can be computed similarly as in the CTM model [3]. The last two terms of $\mathcal{L}_{-\theta}$ are $\mathbb{E}[\log p(\mathbf{r}_n|o_n, \beta)] = \sum_m \sum_k \phi_{snk} \log \beta_{mk\mathbf{r}_{nm}}$ and $\mathbb{E}[\log p(\mathbf{x}_n|o_n, \eta)] = \sum_m \sum_k \phi_{snk} \log \eta_{k\mathbf{x}_{nm}}$.

For each document, optimize $\mathcal{L}_{-\theta}$ over ϕ_{nk} and we can get

$$\phi_{snk} \propto \exp\big(\nu_{sk} + \sum_m \log \beta_{mk\mathbf{r}_{nm}} + \sum_m \log \eta_{k\mathbf{x}_{nm}}\big).$$

For the parameters ν_s and τ_s^2, we do not have a closed form solution. Thus we apply gradient descent methods. Instead of optimizing over the positivity-constrained τ_s^2, we perform the optimization in the log-space. Here, we use the L-BFGS method [15] to optimize over μ and $\log \tau^2$ jointly, which is much faster than the coordinate-wise Newton method [3]. The gradients of ν_s and τ_s^2 are

$$\nabla_{\nu_s} L = -\Sigma_s^{-1}(\nu_s - \mu_s) + \textstyle\sum_n \phi_n - \frac{N}{\zeta_s} \exp(\nu_s + \frac{1}{2}\tau_s^2)$$
$$\nabla_{\tau_{sk}^2} L = \frac{1}{2\tau_{sk}^2} - \frac{\Sigma_{skk}^{-1}}{2} - \frac{N}{2\zeta_s} \exp(\nu_{sk} + \frac{1}{2}\tau_{sk}^2).$$

For the variational parameter ζ, the optimum solution is $\zeta_s = \sum_k \exp(\nu_{sk} + \frac{1}{2}\tau_{sk}^2)$.

A.2. Estimating Gaussian and Multinormial Parameters

For the parameters $(\mu, \Sigma, \beta, \eta)$, the optimal solution is achieved by solving the sub-problem

$$\min_{\mu,\Sigma,\beta,\eta} \sum_d \max_s [-\theta^\top \Delta \mathbf{f}_d(s) + \Delta \ell_d(s) + \mathcal{L}_{-\theta}(q_s^\star)] - (\frac{\lambda}{C} + 1) \sum_d \mathcal{L}_{-\theta}(q_{s_d}^\star).$$

By using the loss-augmented prediction $\hat{s}_d \triangleq \arg\max_s \theta^\top \mathbf{f}(\mathbf{g}_d, s) + \Delta \ell_d(s) + \mathcal{L}_{-\theta}(q_s^\star)$ to replace the maximum operator, this problem has closed-form solutions. For μ and Σ, the solutions are

$$\mu_s = \frac{(1 + \frac{\lambda}{C}) \sum_d \mathbb{I}_{(s_d=s)} \nu_s - \sum_d \mathbb{I}_{(\hat{s}_d=s)} \nu_s}{(1 + \frac{\lambda}{C}) N_s - \hat{N}_s}, \; \Sigma_s = \frac{(1 + \frac{\lambda}{C}) \sum_d \mathbb{I}_{(s_d=s)} (\Lambda + \mathcal{C}_s) - \sum_d \mathbb{I}_{(\hat{s}_d=s)} (\Lambda + \mathcal{C}_s)}{(1 + \frac{\lambda}{C}) N_s - \hat{N}_s},$$

where $\mathbb{I}$ is an indicator function; $N_s = \sum_d \mathbb{I}_{(s_d=s)}$ is the number of training examples that are in category s; $\hat{N}_s = \sum_d \mathbb{I}_{(\hat{s}_d=s)}$ is the number of training data that are predicted to be in category s; $\Lambda = \text{diag}(\tau_s^2)$; and $\mathcal{C}_s = (\nu_s - \mu_s)(\nu_s - \mu_s)^\top$. Although in principle the covariance matrix can be non-positive semidefinite, in practice this can be avoided by choosing a large enough λ, which is much larger than C. For β and η, the optimal solutions are

$$\beta_{mkr} \propto \sum_d \sum_s (\mathbb{I}_{(s=s_d)} + (1 + \frac{\lambda}{C}) \mathbb{I}_{(s=\hat{s})}) \sum_n \phi_{snk}^d \mathbb{I}_{(\mathbf{r}_{dnm}=r)}$$
$$\eta_{kx} \propto \sum_d \sum_s (\mathbb{I}_{(s=s_d)} + (1 + \frac{\lambda}{C}) \mathbb{I}_{(s=\hat{s})}) \sum_{nm} \phi_{snk}^d \mathbb{I}_{(\mathbf{x}_{dnm}=x)}.$$

[5] By the inequality: $\log x \leq a^{-1}x - 1 + \log a, \; \forall a > 0$, where the equality holds when $a = x$.

Table 3: Annotation accuracy of several example objects detected in the sports dataset.

		athlete	grass	tree	horse	water	floor	rock
Object-Level	P	0.369	0.397	0.510	0.394	0.447	0.251	0.266
	R	0.528	0.243	0.512	0.238	0.427	0.370	0.325
	F1	0.435	0.301	0.511	0.297	0.437	0.299	0.292
Pixel-Level	P	0.279	0.577	0.631	0.567	0.633	0.343	0.496
	R	0.613	0.651	0.688	0.511	0.742	0.629	0.648
	F1	0.384	0.612	0.659	0.537	0.683	0.444	0.562

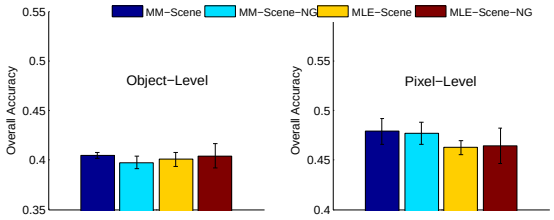

Figure 6: (Left) object-level; and (Right) pixel-level overall accuracy of object annotation.

A.3. Object Annotation on the Sports Dataset

Although our main goal is to improve the imbalanced prediction rule achieved by MLE for scene categorization, the joint scene and object model can also be used to annotate objects when *human annotated* training examples of objects are available, as explained in Sec. 2.2. For the two datasets, only the sports dataset [13] contains manually labeled objects. Here, we report some results of object annotation on the sports dataset by the scene model learned with different methods. For each object class, we randomly sample 40 percent of the regions in testing images as examples with labeled objects and use the k-Nearest Neighbor (kNN) to annotate the rest regions based on the cosine similarity of their latent object representations. We compute two scores: (1) object-level accuracy–the percentage of correctly annotated objects; and (2) pixel-level accuracy–the percentage of the area of correctly annotated regions.

Fig. 6 shows the average overall accuracy of object annotation in both object-level and pixel-level, and Table 3 presents the *P*recision, *R*ecall and *F1* scores for several example objects. Note that the MLE-Scene-NG model is similar to the spatially coherent latent topic model [6]. For the kNN algorithm, the optimum parameter k is chosen for each case. From the results, we can see that the joint scene models learned with partially labeled images can achieve promising results on annotating some objects, and the max-margin method can result in slightly better performance than the MLE-based methods. From these results, we can also identify some challenging problems for future investigation, for example, how to effectively incorporate global features, and how to explore max-margin learning to improve object annotation accuracy.