# Supplementary Material of Evidential Concept Embedding Models: Towards Reliable Concept Explanations for Skin Disease Diagnosis

Yibo Gao, Zheyao Gao, Xin Gao, Yuanye Liu, Bomin Wang, Xiahai Zhuang

Fudan University
https://zmiclab.github.io

In this supplementary material, we provide the detailed derivation of the variational concept loss.

We denote $C_k$ to be the $k$-th concept of the target concepts and $c_k$ to be its label. To derive the variational concept loss $\mathcal{L}_{Beta}$, we assume that the concept label $c_k$ follows *Binomial* distribution $c_k \sim Bin(c_k|p_k)$, where $p_k$ represent the probability supporting concept $C_k$ from the network. $p_k$ follows the *Beta* distribution $p_k \sim \mathcal{B}(\alpha_k, \beta_k)$, which is also the conjugate prior of Binomial distribution. Here, $\alpha_k$ and $\beta_k$ are the evidence generated by the network. Therefore, the marginal log likelihood $p_(c_k|\mathbf{x})$ has an Evidence Lower BOund (ELBO),

$$\begin{aligned}
\log p(c_k|\mathbf{x}) =& \log \int p(c_k, p_k|\mathbf{x}) \mathrm{d}p_k \\
=& \log \int q(p_k|\mathbf{x}) \frac{p(c_k, p_k|\mathbf{x})}{q(p_k|\mathbf{x})} \mathrm{d}p_k \\
=& \log \mathbb{E}_{q(p_k|\mathbf{x})} \left[ \frac{p(c_k, p_k|\mathbf{x})}{q(p_k|\mathbf{x})} \right] \\
\geq& \mathbb{E}_{q(p_k|\mathbf{x})} \left[ \log \frac{p(c_k, p_k|\mathbf{x})}{q(p_k|\mathbf{x})} \right] \\
=& \mathbb{E}_{q(p_k|\mathbf{x})} \left[ \log p(c_k|p_k) \right] - \mathrm{KL}(q(p_k|\mathbf{x}) || p(p_k|\mathbf{x})),
\end{aligned}$$

where the inequality is due to Jensen's inequality and $q(p_k|\mathbf{x})$ is the variational distribution $\mathcal{B}(\alpha_k, \beta_k)$. Minimizing the negative ELBO, we obtain the variational concept loss for the $k$-th concept:

$$\mathcal{L}^k_{Beta} = \mathbb{E}_{q(p_k|\mathbf{x})} \left[ -\log p(c_k|p_k) \right] + \mathrm{KL}(q(p_k|\mathbf{x}) || p(p_k|\mathbf{x}))$$

Xiahai Zhuang is the corresponding author.

The first term of $\mathcal{L}_{Beta}$ can be regarded as the Bayes risk of binary cross-entropy loss with respect to the variational distribution,

$$\begin{aligned}
&\mathbb{E}_{q(p_k|\mathbf{x})}\left[\log p(c_k|p_k)\right] \\
=&\mathbb{E}_{\mathcal{B}(\alpha_k,\beta_k)}[-c_k \log p_k - (1-c_k)\log(1-p_k)] \\
=&-c_k \mathbb{E}_{\mathcal{B}(\alpha_k,\beta_k)}[\log p_k] - (1-c_k)\mathbb{E}_{\mathcal{B}(\alpha_k,\beta_k)}[\log(1-p_k)] \\
=&-c_k[\psi(\alpha_k) - \psi(\alpha_k+\beta_k)] - (1-c_k)[\psi(\beta_k) - \psi(\alpha_k+\beta_k)] \\
=&\psi(\alpha_k+\beta_k) - c_k\psi(\alpha_k) - (1-c_k)\psi(\beta_k).
\end{aligned}$$

The second term can be seen as the prior constraints for evidence. In order to penalizing the evidence of incorrect prediction to 1, we set $\tilde{\alpha}_k = c_k\alpha_k + (1-c_k)$ and $\tilde{\beta}_k = c_k + (1-c_k)\beta_k$, and the second term becomes

$$\begin{aligned}
&\mathrm{KL}(\mathcal{B}(\tilde{\alpha}_k,\tilde{\beta}_k)||\mathcal{B}(1,1)) \qquad (*) \\
=&\mathbb{E}_{\mathcal{B}(\tilde{\alpha}_k,\tilde{\beta}_k)}\left[\log\frac{\Gamma(\tilde{\alpha}_k+\tilde{\beta}_k)}{\Gamma(\tilde{\alpha}_k)\Gamma(\tilde{\beta}_k)} + (\tilde{\alpha}_k-1)p_k + (\tilde{\beta}_k-1)(1-p_k)\right] \\
=&\log\frac{\Gamma(\tilde{\alpha}_k+\tilde{\beta}_k)}{\Gamma(\tilde{\alpha}_k)\Gamma(\tilde{\beta}_k)} + (\tilde{\alpha}_k-1)[\psi(\tilde{\alpha}_k) - \psi(\tilde{\alpha}_k+\beta_k)] \\
&+ (\tilde{\beta}_k-1)[\psi(\tilde{\beta}_k) - \psi(\tilde{\alpha}_k+\tilde{\beta}_k)],
\end{aligned}$$

where $\Gamma(\cdot)$ and $\psi(\cdot)$ denotes *gamma* and *digamma* function respectively. When $c_k = 1$, we have $\tilde{\alpha}_k = \alpha_k$ and $\tilde{\beta}_k = 1$,

$$(*) = \log\frac{\Gamma(\alpha_k+1)}{\Gamma(\alpha_k)} + (\alpha_k-1)[\psi(\alpha_k) - \psi(\alpha_k+1)] = \log\alpha_k + \frac{1-\alpha_k}{\alpha_k}.$$

Similarly, when $c_k = 0$, we have $\tilde{\alpha}_k = 1$ and $\tilde{\beta}_k = \beta_k$,

$$(*) = \log\frac{\Gamma(\beta_k+1)}{\Gamma(\beta_k)} + (\beta_k-1)[\psi(\beta_k) - \psi(\beta_k+1)] = \log\beta_k + \frac{1-\beta_k}{\beta_k}.$$

Adding the Bayes risk term and the KL term together, we obtain

$$\begin{aligned}
\mathcal{L}_{Beta}^{k} =&\psi(\alpha_k+\beta_k) + c_k\left[\log\beta_k + \frac{1-\beta_k}{\beta_k} - \psi(\alpha_k)\right] \\
&+ (1-c_k)\left[\log\alpha_k + \frac{1-\alpha_k}{\alpha_k} - \psi(\beta_k)\right].
\end{aligned}$$