Open-World Semi-Supervised Learning under Compound Distribution Shifts (Supplementary Material)

BMVC 2024 Submission # 762

A Derivation of the Upper Bound of Mutual Information

The interaction information[3] among $\mathbf{x}$, f_{di} and f_{ds} is represented as follows:

$$I(f_{di};f_{ds};\mathbf{x}) = I(f_{di};f_{ds}) - I(f_{di};f_{ds}|\mathbf{x}) = I(f_{di};\mathbf{x}) - I(f_{di};\mathbf{x}|f_{ds}). \tag{1}$$

Therefore, the mutual information between f_{di} and f_{ds} can be expressed as:

$$I(f_{di};f_{ds}) = I(f_{di};\mathbf{x}) - I(f_{di};\mathbf{x}|f_{ds}) + I(f_{di};f_{ds}|\mathbf{x}). \tag{2}$$

We have $I(\mathbf{x};f_{di},f_{ds}) = I(\mathbf{x};f_{ds}) + I(\mathbf{x};f_{di}|f_{ds})$ according to the chain rule of mutual information, and hence the second term of Equation 2 is expressed as:

$$I(\mathbf{x};f_{di}|f_{ds}) = I(\mathbf{x};f_{di},f_{ds}) - I(\mathbf{x};f_{ds}). \tag{3}$$

Since f_{di} is independent of f_{ds}, the posterior distribution satisfies: $q(f_{di}|\mathbf{x}) = q(f_{di}|\mathbf{x},f_{ds})$, and then we have: $H(f_{di}|\mathbf{x}) = H(f_{di}|\mathbf{x},f_{ds})$, where $H(\cdot)$ denotes the information entropy. Therefore, we can write the third term of Equation 2 as:

$$I(f_{di};f_{ds}|\mathbf{x}) = H(f_{di}|\mathbf{x}) - H(f_{di}|\mathbf{x},f_{ds}) = 0. \tag{4}$$

By applying Equation 3 and Equation 4, we can rewrite Equation 2 as:

$$\begin{aligned} I(f_{di};f_{ds}) &= I(\mathbf{x};f_{di}) - I(\mathbf{x};f_{di}|f_{ds}) \\ &= I(\mathbf{x};f_{di}) + I(\mathbf{x};f_{ds}) - I(\mathbf{x};f_{di},f_{ds}). \end{aligned} \tag{5}$$

However, directly minimizing Equation 5 is intractable, so we need to minimize its variational upper bound. Specifically, we need to obtain the variational upper bound of $I(\mathbf{x};f_{di})$ and $I(\mathbf{x};f_{ds})$, as well as the variational lower bound of $I(\mathbf{x};f_{di},f_{ds})$. When $\mathbf{x}$ is an input data, and f_{di} is a feature, similar to VIB[1], we can construct a tractable variational upper bound of $I(\mathbf{x};f_{di})$ by introducing the variational approximation $r(f_{di})$ to approximate the true marginal $p(f_{di})$.

$$\begin{aligned} I(\mathbf{x}; f_{di}) &= E_{p(\mathbf{x}, f_{di})}\left[\log \frac{p(f_{di}|\mathbf{x})}{p(f_{di})}\right] \\ &= E_{p(\mathbf{x}, f_{di})}\left[\log \frac{p(f_{di}|\mathbf{x})r(f_{di})}{r(f_{di})p(f_{di})}\right] \\ &= E_{p(\mathbf{x}, f_{di})}\left[\log \frac{P(f_{di}|\mathbf{x})}{r(f_{di})}\right] - KL(p(f_{di})\|r(f_{di})) \\ &\leq E_{p(\mathbf{x})}\left[KL(p(f_{di}|\mathbf{x})||r(f_{di})\right]. \end{aligned} \tag{6}$$

Likewise, by introducing the variational approximation $r(f_{ds})$, the upper bound of $I(\mathbf{x}; f_{ds})$ can be expressed as:

$$\begin{aligned} I(\mathbf{x}; f_{ds}) &= E_{p(\mathbf{x}, f_{ds})}\left[\log \frac{p(f_{ds}|\mathbf{x})}{p(f_{ds})}\right] \\ &= E_{p(\mathbf{x}, f_{ds})}\left[\log \frac{p(f_{ds}|\mathbf{x})r(f_{ds})}{r(f_{di})p(f_{di})}\right] \\ &= E_{p(\mathbf{x}, f_{ds})}\left[\log \frac{P(f_{ds}|\mathbf{x})}{r(f_{ds})}\right] - KL(p(f_{ds})\|r(f_{ds})) \\ &\leq E_{p(\mathbf{x})}\left[KL(p(f_{ds}|\mathbf{x})||r(f_{ds})\right]. \end{aligned} \tag{7}$$

For the variational lower bound of $I(\mathbf{x}; f_{di}, f_{ds})$, since conditional distributions $p(\mathbf{x}|f_{di}, f_{ds})$ is intractable, we leverage the variational approximation $q(\mathbf{x}|f_{di}, f_{ds})$ to approximate $p(\mathbf{x}|f_{di}, f_{ds})$ similar to IIAE [2], so the lower bound of $I(\mathbf{x}; f_{di}, f_{ds})$ is expressed as:

$$\begin{aligned} I(\mathbf{x}; f_{di}, f_{ds}) &= E_{p(\mathbf{x}, f_{di}, f_{ds})}\left[\log \frac{p(\mathbf{x}|f_{di}, f_{ds})}{p(\mathbf{x})}\right] \\ &= E_{p(\mathbf{x}, f_{di}, f_{ds})}\left[\log q(\mathbf{x}|f_{di}, f_{ds})\right] + E_{p(f_{di}, f_{ds})}\left[KL(p(\mathbf{x}|f_{di}, f_{ds})||q(\mathbf{x}|f_{di}, f_{ds}))\right] + H(X) \\ &\geq E_{p(\mathbf{x}, f_{di}, f_{ds})}\left[\log q(\mathbf{x}|f_{di}, f_{ds})\right] + H(X) \\ &= E_{p(\mathbf{x}, f_{di}, f_{ds})}[\log q(\mathbf{x}|f_{di}, f_{ds})] + E_{\mathbf{x}\sim p(\mathbf{x})} \log p(\mathbf{x}) \\ &= E_{p(\mathbf{x}, f_{di}, f_{ds})}[\log q(\mathbf{x}|f_{di}, f_{ds})] + C, \end{aligned} \tag{8}$$

where C is a constant. And then we use $q(f_{di}|\mathbf{x})$ and $q(f_{ds}|\mathbf{x})$ to be the approximation of $p(f_{di}|\mathbf{x})$ and $p(f_{ds}|\mathbf{x})$, respectively, so we have:

$$\begin{aligned} E_{p(\mathbf{x}, f_{di}, f_{ds})}[\log q(\mathbf{x}|f_{di}, f_{ds})] &= E_{p(\mathbf{x})p(f_{di}|\mathbf{x})p(f_{ds}|\mathbf{x})}[\log q(\mathbf{x}|f_{di}, f_{ds})] \\ &= E_{p(\mathbf{x})}[E_{q(f_{di}|\mathbf{x})q(f_{ds}|\mathbf{x})}[\log q(\mathbf{x}|f_{di}, f_{ds})]]. \end{aligned} \tag{9}$$

In summary, the upper bound of mutual information between f_{di} and f_{ds} can be written as:

$$\begin{aligned} I(f_{di}; f_{ds}) \leq\ & E_{p(\mathbf{x})}[KL(q(f_{di}|\mathbf{x})||r(f_{di})) + KL(q(f_{ds}|\mathbf{x})||r(f_{ds}))] \\ &+ E_{p(\mathbf{x})}[E_{q(f_{di}|\mathbf{x})q(f_{ds}|\mathbf{x})}[\log q(\mathbf{x}|f_{di}, f_{ds})]]. \end{aligned} \tag{10}$$

References

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[2] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Variational interaction information maximization for cross-domain disentanglement. *Advances in Neural Information Processing Systems*, 33:22479–22491, 2020.

[3] W. McGill. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4):93–111, 1954.