Appendix of Knowledge Graph Information Bottleneck for Drug-Drug Interaction Prediction

Shun Liu[1], Gaoqi He[2,*], Kai Zhang[3] , and Honglin Li[4]

[1,2,3]School of Computer Science and Technology, East China Normal University, Shanghai, China, 200062, [3,4]Innovation Center for AI and Drug Discovery, East China Normal University, Shanghai 200062, and [4]Shanghai Key Laboratory of New Drug Design, East China University of Science and Technology, Shanghai, China, 200237

* To whom correspondence should be addressed.

Contents

Proof of Equation (2)

According to the definition of Mutual Information (MI), term $I(Y;G_s,M)$ equals:

$$I(Y;G_s,M) = \mathbf{E}_{G_s,M,Y}\left[\log\frac{\mathbb{P}(Y|G_s,M)}{\mathbb{P}(Y)}\right]$$

The variational approximation of $\mathbb{P}(Y|G_s,M)$ is modelled by learnable parameters θ: $\mathbb{P}_\theta(Y|G_s,M)$. Then, $I(Y;G_s,M)$ can be reformulated as:

$$\begin{aligned}
&I(Y;G_s,M)\\
&= \mathbf{E}_{G_s,M,Y}\left[\log\frac{\mathbb{P}_\theta(Y|G_s,M)}{\mathbb{P}(Y)}\right] + \mathbf{E}_{G_s,M}\left[\mathrm{KL}(\mathbb{P}(Y|G_s,M)|\left|\mathbb{P}_\theta(Y;G_s,M)\right)\right]\\
&\geq \mathbf{E}_{G_s,M,Y}\left[\log\frac{\mathbb{P}_\theta(Y|G_s,M)}{\mathbb{P}(Y)}\right]\\
&= \mathbf{E}_{G_s,M,Y}[\log\mathbb{P}_\theta(Y|G_s,M)] + H(Y)\\
&\geq \mathbf{E}_{G_s,M,Y}[\log\mathbb{P}_\theta(Y|G_s,M)]
\end{aligned}$$

Since $H(Y)$ is a constant, $\mathbf{E}_{G_s,M,Y}[\log\mathbb{P}_\theta(Y|G_s)]$ serve as the lower bound of $I(Y;G_s,M)$.

Proof of Equation (4)

As proof of Equation (2), $I(G_s;M)$ can be reformulated as:

$$I(G_s;M) = \mathbf{E}_{G_s,M}\left[\log\frac{\mathbb{P}(M|G_s)}{\mathbb{P}(M)}\right]$$

Leveraging $\mathbb{P}_\phi(Y|G_s,M)$ as the variational approximation of $\mathbb{P}(M|G_s)$, the following equation holds:

$$\begin{aligned}
&I(G_s;M)\\
&= \mathbf{E}_{G_s,M}\left[\log\frac{\mathbb{P}_\phi(Y|G_s,M)}{\mathbb{P}(Y)}\right] + \mathbf{E}_{G_s,M}\left[\mathrm{KL}(\mathbb{P}(Y|G_s,M)|\left|\mathbb{P}_\phi(Y;G_s,M)\right)\right]\\
&\geq \mathbf{E}_{G_s,M}\left[\log\frac{\mathbb{P}_\phi(M|G_s)}{\mathbb{P}(M)}\right]\\
&= \mathbf{E}_{G_s,M}\left[\log\mathbb{P}_\phi(M|G_s)\right] + H(M)\\
&\geq \mathbf{E}_{G_s,M}\left[\log\mathbb{P}_\phi(M|G_s)\right]
\end{aligned}$$

Therefore, $\mathbf{E}_{G_s,M}\left[-\log\mathbb{P}_\phi(M|G_s)\right]$ forms the upper bound of $-I(G_s;M)$.

Proof of the lower bound for $I(Y;G_s,M)$

As derived in Equation (3), $I(Y;G_s,M)$ is decomposed as:

$$I(Y;G_s,M) = I(M;Y,G_s) - I(G_s;M)$$

Assuming the second term is well optimized, there is an optimal subgraph $G_s{}'$ that determines M. Thus, Y cannot provide more information on M. We have $I(M;Y,G_s{}') = I(M;G_s)$. In this case, term $I(Y;G_s,M)$ reaches 0. Therefore, optimizing $-I(G_s;M)$ can lead $I(Y;G_s,M)$ to its lower bound.

Proof of Equation (5)

With the perfect encoder assumption [1], the representation $\mathbf{g}_s$ of G_s are regarded as a lossless encoding, that is, $I(\mathbf{g}_s;G) \approx I(G_s;G)$. Next, the upper bound of $I(\mathbf{g}_s;G)$ can be derived by introducing the variational approximation $\mathbb{Q}(\mathbf{g}_s)$ of $\mathbb{P}(\mathbf{g}_s)$:

$$I(\mathbf{g}_\mathrm{s}; G) = \mathbf{E}_{G,\mathbf{g}_\mathrm{s}}\left[\log\frac{\mathbb{P}_\phi(\mathbf{g}_\mathrm{s}|G)}{\mathbb{P}(\mathbf{g}_\mathrm{s})}\right] = \mathbf{E}_{G,\mathbf{g}_\mathrm{s}}\left[\log\frac{\mathbb{P}_\phi(\mathbf{g}_\mathrm{s}|G)}{\mathbb{Q}(\mathbf{g}_\mathrm{s})}\right] - \mathbf{E}_{G,\mathbf{g}_\mathrm{s}}\left[\mathrm{KL}\left(\mathbb{P}_\phi(\mathbf{g}_\mathrm{s}|G)||\mathbb{Q}(\mathbf{g}_\mathrm{s})\right)\right]$$

Due to the non-negativity of KL divergence, we can derive that:

$$I(\mathbf{g}_\mathrm{s}; G) \le \mathbf{E}_{G,\mathbf{g}_\mathrm{s}}\left[\log\frac{\mathbb{P}_\phi(\mathbf{g}_\mathrm{s}|G)}{\mathbb{Q}(\mathbf{g}_\mathrm{s})}\right]$$

Following VGIB [2], the noise is sampled from a Gaussian distribution: $N(\mu_\mathbf{H}, \sigma_\mathbf{H}^2)$, where $\mathbf{H}$ is the node embedding matrix of G. Here, sum pooling is used as the readout function. Since the sum of Gaussian distribution is another Gaussian distribution, the following equation is derived:

$$\mathbb{Q}(\mathbf{g}_\mathrm{s}) = N(n\mu_\mathbf{H}, n\sigma_\mathbf{H}^2)$$

where n is the node count of G. Then, we can derive the following equation for $\mathbb{P}_\phi(\mathbf{g}_\mathrm{s}|G)$:

$$\mathbb{P}_\phi(\mathbf{g}_\mathrm{s}|G) = N\left(n\mu_\mathbf{H} + \sum_{v=1}^{n}\lambda_v\mathbf{h}_v - \sum_{v=1}^{n}\lambda_v\mu_\mathbf{H}, \sum_{v=1}^{n}(1-\lambda_v)^2\sigma_\mathbf{H}^2\right)$$

where $\mathbf{h}_v$ is the embedding of node v. Combining the above equations, the following upper bound can be derived for $I(\mathbf{g}_\mathrm{s}; G)$:

$$I(\mathbf{g}_\mathrm{s}; G) \le \mathbf{E}_{G,\mathbf{g}_\mathrm{s}}\left[-\frac{1}{2}\log A + \frac{1}{2n}A + \frac{1}{2n}B^2\right]$$

where $A = \sum_{v=1}^{n}(1-\lambda_v)^2$, and $B = \frac{\sum_{v=1}^{n}\lambda_v(\mathbf{h}_v - \mu_\mathbf{H})}{\sigma_\mathbf{H}}$.

Proof the Statement in IV.F

Using the chain rule of MI, the objective (1) can be reformulated as follows:

$$\begin{aligned}
&-I(Y; G_s, M) + I(Y; M|G_s) + \beta I(G_s; G) \\
&= -I(Y; G_s) + \beta I(G_s; G) \\
&= -I(Y; G, G_s) + I(G; Y|G_s) + \beta I(G_s; G) \\
&= -I(Y; G, G_s) + I(G; Y|G_s) + \beta I(G_s, Y; G) - \beta I(G; Y|G_s) \\
&= -I(Y; G, G_s) + (1-\beta) I(G; Y|G_s) + \beta I(G_s, Y; G)
\end{aligned}$$

Since G_s is the subgraph of G, it cannot provide more information than G. Therefore, it holds $I(Y; G, G_s) = I(Y; G)$. Then, the above equation is derived as:

$$\begin{aligned}
&-I(Y; G_s, M) + I(Y; M|G_s) + \beta I(G_s; G) \\
&= -I(Y; G) + (1-\beta) I(G; Y|G_s) + \beta I(G_s, Y; G) \\
&= -I(Y; G) + (1-\beta) I(G; Y|G_s) + \beta I(Y; G) + \beta I(G; G_s|Y) \\
&= (\beta - 1) I(Y; G) + (1-\beta) I(G; Y|G_s) + \beta I(G; G_s|Y)
\end{aligned}$$

Given that there are optimal subgraph patterns G_s' that determines labels Y and vice versa, it holds $Y = f(G_s{}')$ and $G_s{}' = f^{-1}(Y)$, where f is a mapping function. In this case, $I(G; Y|G_s{}')$ and $I(G; G_s{}'|Y)$ terms can both reach 0, which are the lower bounds of MI, and $I(G; Y)$ is a constant that does not affect the model. Therefore, it proves that if $G_s{}'$ exists, it will be the solution of the above equation.

Reference

[1] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

[2] Junchi Yu, Jie Cao, and Ran He. Improving subgraph recognition with variational graph information bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19396–19405, 2022.