

ADDITIONAL FILE

Proof of Theorem 1 for Enhancing Drug Property Prediction with Dual-Channel Transfer Learning based on Molecular Fragment

Yue Wu¹, Xinran Ni², Zhihao Wang³ and Weike Feng^{1*}

*Correspondence:

fengweike315@163.com

¹College of Traditional Chinese Medicine, Shandong University of Traditional Chinese Medicine, Jinan, CN

Full list of author information is available at the end of the article

Theorem 1: The sum of mutual information can be lower bounded by

$$I(\tilde{\mathbf{h}}_g^i, \mathbf{h}_g^i) + I(\mathbf{x}, \hat{\mathbf{z}}) \geq -\mathcal{L} + \text{const} \quad (1)$$

where $I(\tilde{\mathbf{h}}_g^i, \mathbf{h}_g^i)$ is the mutual information between positive pair in contrastive learning, and $I(\mathbf{x}, \hat{\mathbf{z}})$ is the mutual information between raw input and the encoded representation in generative learning.

$$I(\tilde{\mathbf{h}}_g^i, \mathbf{h}_g^i) + I(\mathbf{x}, \hat{\mathbf{z}}) \geq -\mathcal{L} + \text{const} \quad (2)$$

where $I(\tilde{\mathbf{h}}_g^i, \mathbf{h}_g^i)$ is the mutual information between positive pair in contrastive learning, and $I(\mathbf{x}, \hat{\mathbf{z}})$ is the mutual information between raw input and the encoded representation in generative learning.

Proof For simplicity of notation, we use term $I(\mathbf{x}_1, \mathbf{x}_2)$ to represent $I(\tilde{\mathbf{h}}_g^i, \mathbf{h}_g^i)$ in the following derivation. According to previous research [1], it is proved that expression $\exp(\theta(\mathbf{x}_1, \mathbf{x}_2))$ is proportional to $\frac{p(\mathbf{x}_1|\mathbf{x}_2)}{p(\mathbf{x}_1)}$. Thus,

$$\mathcal{L}_C = -\mathbb{E}_{\mathbf{X}} \log \left[\frac{\frac{p(\mathbf{x}_1|\mathbf{x}_2)}{p(\mathbf{x}_1)}}{\frac{p(\mathbf{x}_1|\mathbf{x}_2)}{p(\mathbf{x}_1)} + \sum_{\mathbf{x}_j \in \mathbf{X}_{neg}} \frac{p(\mathbf{x}_j|\mathbf{x}_2)}{p(\mathbf{x}_j)}} \right], \quad (3)$$

$$= \mathbb{E}_{\mathbf{X}} \log \left[1 + \frac{p(\mathbf{x}_1)}{p(\mathbf{x}_1|\mathbf{x}_2)} \sum_{\mathbf{x}_j \in \mathbf{X}_{neg}} \frac{p(\mathbf{x}_j|\mathbf{x}_2)}{p(\mathbf{x}_j)} \right], \quad (4)$$

$$\approx \mathbb{E}_{\mathbf{X}} \log \left[1 + \frac{p(\mathbf{x}_1)}{p(\mathbf{x}_1|\mathbf{x}_2)} (|\mathcal{B}| - 1) \mathbb{E}_{\mathbf{x}_j} \frac{p(\mathbf{x}_j|\mathbf{x}_2)}{p(\mathbf{x}_j)} \right], \quad (5)$$

$$= \mathbb{E}_{\mathbf{X}} \log \left[1 + \frac{p(\mathbf{x}_1)}{p(\mathbf{x}_1|\mathbf{x}_2)} (|\mathcal{B}| - 1) \right], \quad (6)$$

$$\geq \mathbb{E}_{\mathbf{X}} \log \left[\frac{p(\mathbf{x}_1)}{p(\mathbf{x}_1|\mathbf{x}_2)} |\mathcal{B}| \right], \quad (7)$$

$$= -I(\mathbf{x}_1, \mathbf{x}_2) + \log(|\mathcal{B}|) \quad (8)$$

For generative loss, minimizing the reconstruction loss is proved to be equivalent to maximize a lower bound of the mutual information between the raw input $\mathbf{x}$ and

encoded representation $\hat{\mathbf{z}}$ [2], which can be formalized as:

$$\mathcal{L}_{\mathcal{G}} \propto \mathbb{E}_{p(\mathbf{x}_v, \hat{\mathbf{z}}_v)} \left[\|\mathbf{x}_v - \hat{\mathbf{z}}_v\|_2^2 \right] \quad (9)$$

$$\geq -\mathbb{E}_{p(\mathbf{x}_v, \hat{\mathbf{z}}_v)} \log \frac{p(\mathbf{x}_v | \hat{\mathbf{z}}_v)}{p(\mathbf{x}_v)} + H(\mathbf{x}) \quad (10)$$

$$= -I(\mathbf{x}, \hat{\mathbf{z}}) + H(\mathbf{x}) \quad (11)$$

where $H(\mathbf{x})$ is the constant entropy of $\mathbf{x}$. Combining Equation (8) and Equation (11), we arrive at the main theorem.

$$I(\tilde{\mathbf{h}}_g^i, \mathbf{h}_g^i) + I(\mathbf{x}, \hat{\mathbf{z}}) \geq -(\mathcal{L}_{\mathcal{C}} + \mathcal{L}_{\mathcal{G}}) + \log(|\mathcal{B}|) + H(\mathbf{x}) \quad (12)$$

$$= -\mathcal{L} + \text{const} \quad (13)$$

Author details

¹College of Traditional Chinese Medicine, Shandong University of Traditional Chinese Medicine, Jinan, CN.

²College of Pharmacy, Shandong University of Traditional Chinese Medicine, Jinan, CN. ³College of Intelligence and Information Engineering, Shandong University of Traditional Chinese Medicine, Jinan, CN.

References

1. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
2. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.-A., Bottou, L.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research **11**(12) (2010)