Appendix

1 PROOF OF THEOREM 1

We first recall the loss function of UnG-MoCha:

$$\mathcal{L}_{\Theta}(\mathcal{M}) = \alpha\mathcal{L}(m,\hat{m}) + (1-\alpha)\mathcal{L}(v,\hat{v}) + \gamma\mathcal{L}_{CCA} \tag{1}$$

where $\mathcal{L}(m,\hat{m})$ is:

$$\mathcal{L}(m,\hat{m}) = \|m-\hat{m}\|^2 \tag{2}$$

where $\hat{m} = MLP([\mathbf{h}_{\mathcal{M}}||\mathbf{h}_{\mathcal{G}}])$.

The definition of multi-layer perceptron(MLP) is below:

Definition 1.1 (multi-layer perceptron [2]). *A K-layer multi-layer perceptron $f_{MLP}: \mathbb{R}^d \to \mathbb{R}^n$ is the function*

$$f_{MLP}(x) = T_K \circ \rho_K \circ \cdots \circ \rho_1 \circ T_1(x) \tag{3}$$

where $T_k : x \mapsto W_k x + b_k$ is an affine function and $\rho_k : x \mapsto (g_k(x))$ is the non-linear activation function.

Similar to $\mathcal{L}(m,\hat{m})$, $\mathcal{L}(v,\hat{v})$ is:

$$\mathcal{L}(v,\hat{v}) = \|v-\hat{v}\|^2, \tag{4}$$

and $\hat{v}$ is obtained from the same multi-layer perceptron MLP.

$\mathcal{L}_{CCA}$ is written as:

$$\begin{aligned}&\mathcal{L}_{dist}(MLP(\mathbf{h}_{\mathcal{M}}), MLP(\mathbf{h}_{\mathcal{G}}))\\ &+\lambda(\mathcal{L}_{dl}(MLP(\mathbf{h}_{\mathcal{M}})) + \mathcal{L}_{dl}(MLP(\mathbf{h}_{\mathcal{G}})))\end{aligned} \tag{5}$$

correlation loss $\mathcal{L}_{dist}$ is

$$\mathcal{L}_{dist} = \frac{1}{2}\|MLP(\mathbf{h}_{\mathcal{M}}) - MLP(\mathbf{h}_{\mathcal{G}})\|_F^2, \tag{6}$$

and decorrelation loss $\mathcal{L}_{dl}$ is

$$\mathcal{L}_{dl}(\mathbf{U}) = \|\mathbf{U}^T\mathbf{U} - I\|_F^2 \tag{7}$$

To obtain the upper bound of estimation error, the upper bound of every component of Equation (1) needs to be derived. Firstly, we derive the upper bound of $\mathcal{L}(m,\hat{m})$:

$$\begin{aligned}\mathcal{L}(m,\hat{m}) &= \|m-\hat{m}\|^2\\ &= \|m - MLP([\mathbf{h}_{\mathcal{M}}||\mathbf{h}_{\mathcal{G}}])\|^2\end{aligned} \tag{8}$$

Assume that all the operations in MLP are all locally Lipschitz-continuous, and that their partial derivatives $\partial g_k(x)$ can be computed and efficiently maximized.

Theorem 1.1 (Rademacher Theorem [1]). *If $f : \mathbb{R}^d \to \mathbb{R}^n$ is a locally Lipschitz continuous function, then f is differentiable almost everywhere. Moreover, if f is Lipschitz continuous, then*

$$\mathcal{L}_f \le \sup_{x\in R^d} \|\nabla f(x))\|_2. \tag{9}$$

With Theorem 1.1 and the assumption, the MLP can be considered as locally Lipschitz-continuous and the upper bound of each component in Equation (1) can be derived. With the definition of MLP and the assumption, it can be derived that

$$\begin{aligned}\mathcal{L}(m,\hat{m}) \le \frac{\partial\mathcal{L}(m,\hat{m})}{\partial(\mathbf{h}_{\mathcal{M}}||\mathbf{h}_{\mathcal{G}})} &= \frac{\partial\|m - MLP([\mathbf{h}_{\mathcal{M}}||\mathbf{h}_{\mathcal{G}}])\|^2}{\partial(\mathbf{h}_{\mathcal{M}}||\mathbf{h}_{\mathcal{G}})}\\ &\le 2(m - MLP([\mathbf{h}_{\mathcal{M}}||\mathbf{h}_{\mathcal{G}}]))\prod_{k=1}^{K}\|W_k\|_2\end{aligned} \tag{10}$$

where K is the layer number of multi-layer perceptron.

Substitute Equation (10) into Equation (2), Equation (2) can be rewritten as:

$$\|m-\hat{m}\|^2 \le 2(m - MLP([\mathbf{h}_{\mathcal{M}}||\mathbf{h}_{\mathcal{G}}]))\prod_{k=1}^{K}\|W_k\|_2 \tag{11}$$

Through Equation (11), the bound of $\hat{m}$ can be derived as:

$$m - 2\tau \le \hat{m} \le \tau, \tag{12}$$

where $\tau \in \mathbb{R}$ is a constant and $\tau \le \prod_{k=1}^{K}\|W_k\|_2$.

Similarly, we can get the upper bound of $\mathcal{L}(v,\hat{v})$, which is

$$\mathcal{L}(v,\hat{v}) \le 2(v-\epsilon)\prod_{k=1}^{K}\|W_k\|_2 \tag{13}$$

The bound of $\hat{v}$ is:

$$v - 2\tau \le \hat{v} \le \tau \tag{14}$$

For $\mathcal{L}_{CCA}$, we prove the upper bound from variance-covariance perspective. After transformation by MLP, $\mathbf{h}_{\mathcal{M}}$ and $\mathbf{h}_{\mathcal{G}}$ have the same dimension and can be considered as augmented by $s \sim p_{aug}(x)$. $\widetilde{z}$ denotes the representation of s. Correlation loss $\mathcal{L}_{dist}$ can be written as:

$$\begin{aligned}\mathcal{L}_{dist} &= \frac{1}{2}\|MLP(\mathbf{h}_{\mathcal{M}}) - MLP(\mathbf{h}_{\mathcal{G}})\|_F^2\\ &= \sum_{i=1}^{N}\sum_{j=1}^{D}(\widetilde{z}_{i,j}^{\mathbf{h}_{\mathcal{M}}} - \widetilde{z}_{i,j}^{\mathbf{h}_{\mathcal{G}}})^2\\ &\cong N * \mathbb{E}_{\mathbf{h}_{\mathcal{M}};\mathbf{h}_{\mathcal{G}}}(\sum_{k=1}^{D}\mathbb{V}_s[\widetilde{z}_k])\end{aligned} \tag{15}$$

The decorrelation loss can be transformed to the sum of Pearson correlation coefficient [3]:

$$\begin{aligned}\mathcal{L}_{dl}(U) &= \|U^T U - I\|_F^2\\ &= \|Cov[U] - I\|_F^2\\ &\cong \sum_{i\neq j}\rho_{i,j}^{U}\end{aligned} \tag{16}$$

Therefore, the decorrelation loss $\mathcal{L}_{dl}$ can be considered as a constant $\psi \ge 0$.

Therefore, the upper bound of $\mathcal{L}_{CCA}$ is a constant $C \in \mathbb{R}$, where $C \ge 0$. Combining each component's upper bound, the final upper bound is

$$\mathcal{L}_{\Theta}(\mathcal{M}) \le 2(\alpha m + (1-\alpha)v + \tau)\prod_{k=1}^{K}\|W_k\|_2 + C \tag{17}$$

where $\tau \le \prod_{k=1}^{K}\|W_k\|_2$.

REFERENCES

[1] Herbert Federer. 2014. *Geometric measure theory*. Springer.

[2] Aladin Virmaux and Kevin Scaman. 2018. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems* 31 (2018).

[3] Hengrui Zhang, Qitian Wu, Junchi Yan, David Wipf, and Philip S Yu. 2021. From canonical correlation analysis to self-supervised graph neural networks. *Advances in Neural Information Processing Systems* 34 (2021), 76–89.