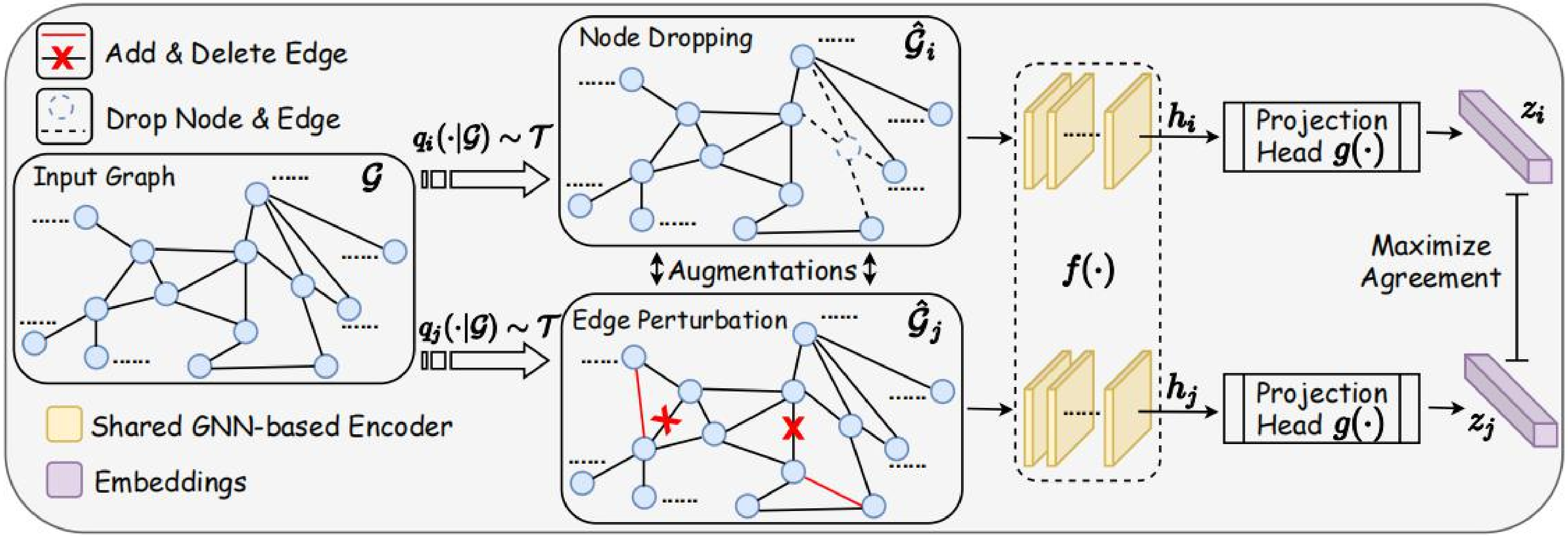
Graph Contrastive Learning



(4) **Contrastive loss function.** A contrastive loss function $\mathcal{L}(\cdot)$ is defined to enforce maximizing the consistency between positive pairs z_i, z_j compared with negative pairs. Here we utilize the normalized temperature-scaled cross entropy loss (NT-Xent) [51, 25, 52].

B.1 Latent Class

GCL Pretext Task GCL is a typical unsupervised learning method which has no information of downstream real classes. However, it can be assumed that there exists a set of latent classes $\mathcal{C}$ whose probabilities are described by distribution ρ. Moreover, a node distribution $\mathcal{D}_c$ is defined over the input graph space $\mathcal{G}$ within each latent class $c \in \mathcal{C}$. GCL samples are generated according to the following scheme,

(1) Sample two latent classes $c^+ c^- \sim \rho^2$.

(2) Sample positive node pair $(v, v^+) \sim (\mathcal{D}_{c^+})^2$.

(3) Sample negative node $v^- \sim \mathcal{D}_{c^-}$.

In most cases of GCL, the positive pair are comprised of two augmentations of the same node v_i, while the negative samples are other nodes $v_{j\ j \neq i}$. We claim that it can be viewed as a set of more fine-grained latent semantic classes.

Downstream Task As for the true label from downstream, it can always be a subset of previous latent classes. We will consider the standard supervised learning task of classifying a data point into one of the classes in $\mathcal{C}$. Formally, downstream samples are generated according to the following scheme,

B.2 Mean Classifier

The Mean Classifier q_ψ^μ is a classifier, whose k^{th} row parameter is $\boldsymbol{\mu}_k \triangleq \underset{v \sim \mathcal{D}_{c_k}}{\mathbb{E}} [f_\theta(\mathcal{G}_v)]$, namely the mean $\boldsymbol{\mu}_{c_k}$ of representations of nodes with label c_k, for GNN encoder f_θ from GCL and downstream classification task $\mathcal{T} = (c_1, \dots, c_k)$, where node v_i belongs to ground truth label y_i.

$$L_{\mathcal{T}}^{\mu}(\mathcal{G}, f_\theta) \triangleq \sum_{i=1}^{N} \ell_{sup}(f_\theta(\mathcal{G}_{v_i}), y_i, q_\psi^\mu) \tag{21}$$

For GCL, the Mean Classifier can be view as a node clustering, where the parameters of downstream new model are not learned freely. Instead, they are fixed to be the mean representations of nodes with same true label from downstream task. With the restrictions on the freedom of downstream new model, the model difficulty can be solved.

Appendix D

$$\operatorname{gen}\left(\mathcal{T}, P_{F|\mathcal{P}}\right) \triangleq \mathbb{E}\left[L_{\mathcal{T}}(\mathcal{G}, F) - s \cdot L_{\mathcal{P}}(\mathcal{G}, F)\right] \tag{7}$$

Definition[eq(7)]:

Inspired by the generalization of supervised learning, the GCL generalization error is the difference between the population risk and the empirical risk over all hypothesis F

s: aimed to adjust the scale

$$\operatorname{gen}\left(\mathcal{T}, P_{F|\mathcal{P}}\right) \triangleq \mathbb{E}\left[L_{\mathcal{T}}(\mathcal{G}, F) - \frac{L_{\mathcal{P}}(\mathcal{G}, F)}{\gamma(1-\tau)}\right] \leq \frac{\sqrt{2\sigma^2 I(\mathcal{G}, F)} - (\beta\tau + \epsilon)}{\gamma(1-\tau)}$$

Another expression of the ineqaulity above:

(A combination of eq(34)(35)(36)(37))

$$\mathrm{E}\left[\gamma(1-\tau)L_{\mathrm{T}}(G, F) - L_P(G, F) + \beta\tau + \varepsilon\right] \leq \sqrt{2\sigma^2 I(G, F)}$$

Actually a extention of Xu and Raginsky [2017]

$$\left|\mathbb{E}[f(X,Y)] - \mathbb{E}[f(\bar{X},\bar{Y})]\right| \leq \sqrt{2\sigma^2 D(P_{X,Y} \| P_X \otimes P_Y)}.$$

Information-theoretic analysis of generalization capability of learning algorithms / v2 2019

Equation 34

Following Russo and Zou [2016] as well as Xu and Raginsky [2017], we start with the Donsker-Varadhan variational representation of the relative entropy Boucheron et al. [2013]. It says, for any two probability measures π, ρ on a common measurable space $(\Omega, \mathcal{K})$, inequality Eq.(32) holds, where $D(\pi \| \rho)$ stands for the KL divergence between two distributions π and ρ. And the supremum is over all measurable functions $K : \Omega \to \mathbb{R}$, such that $e^K \in L^1(\rho)$.

$$D(\pi \| \rho) = \sup_K \left\{ \int_\Omega K \, \mathrm{d}\pi - \log \int_\Omega e^K \, \mathrm{d}\rho \right\} \tag{32}$$

$$\begin{aligned}
&D\left(P_{\mathcal{G},F} \| P_{\mathcal{G}} \otimes P_F\right) \overset{a}{\geq} \mathbb{E}[\lambda L_{\mathcal{P}}(\mathcal{G}, F)] - \log \mathbb{E}\left[e^{\lambda L_{\mathcal{P}}(\bar{\mathcal{G}}, \bar{F})}\right] \\
&\overset{b}{\geq} \lambda(\mathbb{E}[L_{\mathcal{P}}(\mathcal{G}, F)] - \mathbb{E}[L_{\mathcal{P}}(\bar{\mathcal{G}}, \bar{F})]) - \frac{\lambda^2 \sigma^2}{2} \\
&\overset{c}{\geq} \lambda(\mathbb{E}[L_{\mathcal{P}}(\mathcal{G}, F)] - \gamma(1-\tau)\mathbb{E}[L_{\mathcal{T}}^{\mu}(\mathcal{G}, F)] + \beta\tau + \epsilon) - \frac{\lambda^2 \sigma^2}{2}
\end{aligned}$$

Equation 34--Step b

For brevity,we denote $E(Q)$、$E(\overline{Q})$ as $E[L_P(G,F)]$、$E[L_P(\overline{G},\overline{F})]$

All inequality in this page only serve to estimate,not to compare

Eq(34) is equivalent to $D \geq E[\lambda Q] - \log E\left[e^{\lambda \overline{Q}}\right]$

Step1:Taylor expansion of$\left[e^{\lambda \overline{Q}}\right]$about t=0

$$E\left[\lambda \overline{Q}\right] = E\left[1 + \overline{Q} + \frac{\lambda^2 \overline{Q}^2}{2!} + \frac{\lambda^3 \overline{Q}^3}{3!} + \cdots\right]$$

Take first two items for approximation

$$E\left[\lambda \overline{Q}\right] \approx 1 + \lambda E\left(\overline{Q}\right) + \frac{\lambda^2}{2} E\left(\overline{Q}^2\right)$$

$$\log E\left[e^{\lambda \overline{Q}}\right] = \log\left[1 + \lambda E\left(\overline{Q}\right) + \frac{\lambda^2}{2} E\left(\overline{Q}^2\right)\right]$$

Step2:Taylor expansion $ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + o(x^3)$

$$\log\left[1 + \lambda E\left(\overline{Q}\right) + \frac{\lambda^2}{2} E\left(\overline{Q}^2\right)\right] = \lambda E\left(\overline{Q}\right) + \frac{\lambda^2}{2} E\left(\overline{Q}^2\right) - \frac{\left[\ \lambda E\left(\overline{Q}\right) + \frac{\lambda^2}{2} E\left(\overline{Q}^2\right)\ \right]^2}{2} + \cdots$$

$$\log E\left[e^{\lambda \overline{Q}}\right] \approx \lambda E\left(\overline{Q}\right) + \frac{\lambda^2}{2} E\left(\overline{Q}^2\right) - \frac{\left(\lambda E\left[\overline{Q}\right]\right)^2}{2} = \lambda E\left[\overline{Q}\right] + \frac{\lambda^2}{2}\left(\ E\left[\overline{Q}^2\right] - E\left[\overline{Q}\right]^2\ \right)$$

Equation 34

Finally we get that $\log E\left[e^{\lambda\overline{Q}}\right]=\lambda E\left[\overline{Q}\right]+\frac{\lambda^2\sigma^2}{2}$

Then carry out simplification and followed Lemma 1

$$D \geq \lambda\left(E[Q]-E\left[\overline{Q}\right]\right)-\frac{\lambda^2\sigma^2}{2}=\lambda\left(E[L_P(G,F)]-E\left[L_P\left(\overline{G},\overline{F}\right)\right]\right)-\frac{\lambda^2\sigma^2}{2}$$

Lemma 1 $L_{\mathcal{P}}(\mathcal{G}, f_\theta) \leq \gamma(1-\tau)L_{\mathcal{T}}^{\mu}(\mathcal{G}, f_\theta)+\beta\tau+\epsilon$

Simultaneous expectation on both sides

$$D\left(P_{\mathcal{G},F}\|P_{\mathcal{G}}\otimes P_F\right) \overset{a}{\geq} \lambda(\mathbb{E}[L_{\mathcal{P}}(\mathcal{G},F)]-\gamma(1-\tau)\mathbb{E}[L_{\mathcal{T}}^{\mu}(\mathcal{G},F)]+\beta\tau+\epsilon)-\frac{\lambda^2\sigma^2}{2}$$

$$=\mathrm{H}(\lambda)$$

$$D \geq H(\lambda)_{\max} \qquad H(\lambda)=\Delta\cdot\lambda-\frac{\sigma^2}{2}\lambda^2 \qquad H(\lambda)_{\max}=H\left(\frac{\Delta}{\sigma^2}\right)=\frac{\Delta^2}{2\sigma^2}$$

$$D \geq \frac{\Delta^2}{2\sigma^2}\alpha \qquad \cong \qquad 2\sigma^2 D \geq \Delta^2$$

Experiment

$$\mathrm{gen}\left(\mathcal{T}, P_{F|\mathcal{P}}\right) \leq \frac{\sqrt{2\sigma^2 I(\mathcal{G}, F)} - (\beta\tau + \epsilon)}{\gamma(1-\tau)}$$

$$I(G,F) \leq \min\left\{I(G,G_{view}), I(G_{view}, f_\theta)\right\}$$

To decrease both $I(G,G_{view})$ and $I(G_{view}, f_\theta)$

Appendix C

$$L_{\mathcal{P}}(\mathcal{G}, f_\theta) = (1-\tau) L^{\neq}_{\mathcal{P}}(\mathcal{G}, f_\theta) + \tau L^{=}_{\mathcal{P}}(\mathcal{G}, f_\theta) \tag{26}$$

$$L^{\neq}_{\mathcal{P}}(\mathcal{G}, f_\theta) \leq \gamma L^{\mu}_{\gamma,\mathcal{T}}(f_\theta) + \epsilon \tag{27}$$

$$s(f_\theta) \triangleq \mathop{\mathbb{E}}_{c\sim\nu}\left[\sqrt{\|\Sigma(f_\theta, c)\|_2} \mathop{\mathbb{E}}_{v\sim\mathcal{D}_c} \|f_\theta(\mathcal{G}_v)\|\right] \tag{28}$$

$$L^{=}_{\mathcal{P}}(\mathcal{G}, f_\theta) - 1 \leq c' s(f_\theta) \tag{29}$$

$$L^{=}_{\mathcal{P}}(\mathcal{G}, f_\theta) \leq c' \sigma R + 1 \tag{30}$$

Eq(26) → Eq(8) in Saunshi et al. [2019]

Eq(27) → Lemma 5.1 in Saunshi et al. [2019]

Eq(29) → Lemma 4.4 in Saunshi et al. [2019]

Eq(28) → Eq(9) in Saunshi et al. [2019]

Eq(8) = Eq(31) = Lemma 1

$$L_{\mathcal{P}}(\mathcal{G}, f_\theta) \leq \gamma(1-\tau) L^{\mu}_{\mathcal{T}}(\mathcal{G}, f_\theta) + \beta\tau + \epsilon$$

A theoretical analysis of contrastive unsupervised representation learning / ICML 2019

Equation 27

Equation 29