Semi-Implicit Graph Variational Auto-Encoders

Introduction

- SIG-VAE employs a hierarchical variational framework to enable **neighboring node sharing** for better generative modeling of graph dependency structure, together with a Bernoulli-Poisson link decoder.

Problems with VGAE

- The Gaussian assumption restricts its variational
  inference flexibility when the true posterior distribution given a graph clearly violates the Gaussian assumption
- The adopted inner-product decoder restricts its generative model flexibility

Semi Implicit Variational Inference (SIVI)

- Enriches mean-field variational inference with a flexible (implicit) mixing distribution
- No constraint on **explicit** probability density function

Normalizing Flow (NF)

- Transforms a simple Gaussian random variable through a sequence of invertible differentiable functions with tractable Jacobians.
- Restriction -> **explicit** probability density function

Background

Variational graph auto-encoder (VGAE)

Problem 1.
Given adjacency matrix A, M-dimensional node attributes $X \in \mathbb{R}^{N \times M}$, find probability distribution of latent representation $Z \in \mathbb{R}^{N \times L}$, i.e., $p(Z|X, A)$

Finding the true posterior, $p(Z|X, A)$ is difficult, hence, it is approximated by a Gaussian distribution, $q(Z|\psi)$, our goal is to find optimum ψ where $\psi_i = \{\mu_i, diag(\sigma_i^2)\}$ so that $q(Z|\psi)$ is close to the true posterior.

The parameters, $\{\mu_i, diag(\sigma_i^2)\}$, are modeled
and learned using two GCNs. Besides, the parameters are optimized by maximizing the lower bound (ELBO):

$$L = \mathbb{E}_{q(Z|\psi)}[p(A|Z)] - KL[q(Z|\psi)||p(Z)].$$

However, a well-known issue in
variational inference is underestimating the variance of the posterior. Thus, semi-implicit variational inference is introduced.

Semi-implicit variational inference (SIVI)

SIVI assumes that ψ, the parameters of the posterior, are drawn from an implicit distribution rather than being analytic. More specifically, $Z \sim q(Z|\psi)$ and $\psi \sim q_\phi(\psi)$ with ϕ denoting the distribution parameters to be inferred.

Marginalizing ψ out leads to the random variables Z drawn from a distribution family H indexed by variational parameters ϕ, expressed as

$$H = \{h_\phi(Z) : h_\phi(Z) = \int_\psi q(Z|\psi)q_\phi(\psi)d\psi$$

The original posterior $q(Z|\psi)$ is **explicit**, the marginal distribution, $h_\phi(Z)$ is often **implicit**. Semi-implicit does not assumes **indenpent latent dimensions**.

SIVI derives a lower bound for ELBO, as follows, to optimize the variational parameters:

$$\begin{aligned}
\mathcal{L} = \mathbb{E}_{\mathbf{Z} \sim h_{\boldsymbol{\phi}}(\mathbf{Z})}\left[\log \frac{p(\mathbf{Y}, \mathbf{Z})}{h_{\boldsymbol{\phi}}(\mathbf{Z})}\right] &= -\mathbf{KL}(\mathbb{E}_{\psi \sim q_{\boldsymbol{\phi}}(\psi)}[q(\mathbf{Z} \,|\, \psi)] \,||\, p(\mathbf{Z} \,|\, \mathbf{Y})) + \log p(\mathbf{Y}) \\
&\geq -\mathbb{E}_{\psi \sim q_{\boldsymbol{\phi}}(\psi)} \mathbf{KL}(q(\mathbf{Z} \,|\, \psi) \,||\, p(\mathbf{Z} \,|\, \mathbf{Y})) + \log p(\mathbf{Y}) \\
&= \mathbb{E}_{\psi \sim q_{\boldsymbol{\phi}}(\psi)}\left[\mathbb{E}_{\mathbf{Z} \sim q(\mathbf{Z} \,|\, \psi)}\left[\log\left(\frac{p(\mathbf{Y}, \mathbf{Z})}{q(\mathbf{Z} \,|\, \psi)}\right)\right]\right] = \underline{\mathcal{L}}(q(\mathbf{Z} \,|\, \psi), q_{\boldsymbol{\phi}}(\psi)),
\end{aligned} \tag{2}$$

Optimizing this lower bound, however, could drive the mixing distribution $q_\phi(\psi)$ towards a point mass density. To address the degeneracy issue, SIVI adds a nonnegative regularization term, leading to a surrogate ELBO that is asymptotically exact.

Normalizing flow (NF)

NF imposes explicit density functions for the mixing distributions in the hierarchy while SIVI only requires q_ϕ to be reparameterizable

Baseline: Variational Inference with VGAE

SIVI-VGAE

Naive method: apply SIVI in VGAE

$$Z \sim q(Z|\psi), \psi \sim q_\phi(\psi|X, A)$$

Advantages:

1. flexible mixture modeling of the posterior
2. efficient model inference

For example, with ϕ being parameterized by deep neural networks

$$
\begin{aligned}
\mathbf{h}_u &= \mathrm{GNN}_u(\mathbf{A}, \mathrm{CONCAT}(\mathbf{X}, \mathbf{h}_{u-1})), \quad \text{for } u = 1, \dots, L, \ \mathbf{h}_0 = \mathbf{0} \\
\boldsymbol{\ell}_t^{(i)} &= \boldsymbol{T}_t(\boldsymbol{\ell}_{t-1}^{(i)}, \boldsymbol{\epsilon}_t, \mathbf{h}_L^{(i)}), \quad \text{where } \boldsymbol{\epsilon}_t \sim q_t(\boldsymbol{\epsilon}) \text{ for } t = 1, \dots, C, \ \boldsymbol{\ell}_0^{(i)} = \mathbf{0} \\
\boldsymbol{\mu}_i(\mathbf{A}, \mathbf{X}) &= \boldsymbol{g}_\mu(\boldsymbol{\ell}_C^{(i)}, \mathbf{h}_L^{(i)}), \ \boldsymbol{\Sigma}_i(\mathbf{A}, \mathbf{X}) = \boldsymbol{g}_\Sigma(\boldsymbol{\ell}_C^{(i)}, \mathbf{h}_L^{(i)}), \\
q(\mathbf{Z} \mid \mathbf{A}, \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \textstyle\prod_{i=1}^N q(\mathbf{z}_i \mid \mathbf{A}, \mathbf{X}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad q(\mathbf{z}_i \mid \mathbf{A}, \mathbf{X}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \mathcal{N}(\boldsymbol{\mu}_i(\mathbf{A}, \mathbf{X}), \boldsymbol{\Sigma}_i(\mathbf{A}, \mathbf{X})),
\end{aligned}
$$

Given the GNN_L output h_L, $\mu_i(A, X)$, $\sum_i(A, X)$ are now random variables. However, the constructed implicit distributions **may not capture the dependency** between neighboring nodes completely

NF-VGAE
NF requires **deterministic transform functions** whose
Jacobians shall be easy to compute, which limits the flexibility when considering complex dependency structures in graph analytic tasks.

$$
\begin{aligned}
\mathbf{h}_u &= \mathrm{GNN}_u(\mathbf{A}, \mathrm{CONCAT}(\mathbf{X}, \mathbf{h}_{u-1})), \quad \text{for } u = 1, \dots, L, \ \mathbf{h}_0 = \mathbf{0} \qquad (4) \\
\boldsymbol{\mu}(\mathbf{A}, \mathbf{X}) &= \mathrm{GNN}_\mu(\mathbf{A}, \mathrm{CONCAT}(\mathbf{X}, \mathbf{h}_L)), \quad \boldsymbol{\Sigma}(\mathbf{A}, \mathbf{X}) = \mathrm{GNN}_\Sigma(\mathbf{A}, \mathrm{CONCAT}(\mathbf{X}, \mathbf{h}_L)), \\
q_0(\mathbf{Z}^{(0)} \mid \mathbf{A}, \mathbf{X}) &= \textstyle\prod_{i=1}^N q_0(\mathbf{z}_i^{(0)} \mid \mathbf{A}, \mathbf{X}), \quad \text{with} \quad q_0(\mathbf{z}_i^{(0)} \mid \mathbf{A}, \mathbf{X}) = \mathcal{N}(\boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2)), \\
q_K(\mathbf{Z}^{(K)} \mid \mathbf{A}, \mathbf{X}) &= \textstyle\prod_{i=1}^N q_0(\mathbf{z}_i^{(K)} | \mathbf{A}, \mathbf{X}), \quad \ln(q_K(\mathbf{z}_i^{(K)} \mid -)) = \ln(q_0(\mathbf{z}_i^{(0)})) - \displaystyle\sum_k \ln|\det \frac{\partial f_k}{\partial \mathbf{z}_i^{(k)}}|,
\end{aligned}
$$

The GNN output layers are deterministic without neighborhood distribution sharing due to the deterministic nature of the initial density parameters in q_0.

Semi-implicit graph variational auto-encoder (SIG-VAE)

Trivial combinations of the **SIVI-VGAE** and **NF-VGAE** may fail to fully exploit graph dependency structure because they are not capable of propagating uncertainty between neighboring nodes. Therefore, we proposed carefully designed **SIG-VAE**.

Specifically, the first stochastic layer $q(Z|X, A)$ is reparameterizable and has an analytic probability density function. We adopt a hierarchical encoder in SIG-VAE that injects random noise at *L* different stochastic layers:

$$
\begin{aligned}
\mathbf{h}_u &= \mathrm{GNN}_u(\mathbf{A}, \mathrm{CONCAT}(\mathbf{X}, \boldsymbol{\epsilon}_u, \mathbf{h}_{u-1})), \quad \text{where } \boldsymbol{\epsilon}_u \sim q_u(\boldsymbol{\epsilon}) \text{ for } u = 1, \dots, L, \ \mathbf{h}_0 = \mathbf{0} \qquad (5) \\
\boldsymbol{\mu}(\mathbf{A}, \mathbf{X}) &= \mathrm{GNN}_\mu(\mathbf{A}, \mathrm{CONCAT}(\mathbf{X}, \mathbf{h}_L)), \quad \boldsymbol{\Sigma}(\mathbf{A}, \mathbf{X}) = \mathrm{GNN}_\Sigma(\mathbf{A}, \mathrm{CONCAT}(\mathbf{X}, \mathbf{h}_L)), \qquad (6) \\
q(\mathbf{Z} \mid \mathbf{A}, \mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \textstyle\prod_{i=1}^N q(\mathbf{z}_i \mid \mathbf{A}, \mathbf{X}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad q(\mathbf{z}_i \mid \mathbf{A}, \mathbf{X}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \mathcal{N}(\boldsymbol{\mu}_i(\mathbf{A}, \mathbf{X}), \boldsymbol{\Sigma}_i(\mathbf{A}, \mathbf{X})).
\end{aligned}
$$

SIG-VAE incorporates the distributions of
the neighboring nodes to better capture graph dependency.

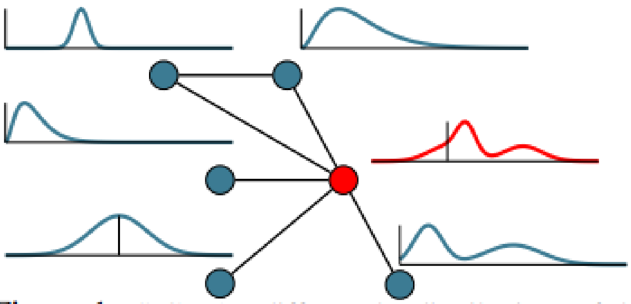

Figure 1: SIG-VAE diffuses the distributions of the neighboring nodes, which is more informative than sharing deterministic features, to infer each node's latent distribution.

The increasing flexibility of variational inference is not enough to better model real-world graph data. Therefore, **Bernoulli-Poisson link** is adopted for the decoder. Hence:
Let $A_{i,j} = \delta(m_{ij} > 0), m_{ij} \sim Poisson(exp(\sum_{k=1}^{l} r_k z_{ik} z_{jk}))$

$$p(\mathbf{A} \mid \mathbf{Z}, \mathbf{R}) = \prod_{i=1}^{N} \prod_{j=1}^{N} p(A_{i,j} \mid \mathbf{z}_i, \mathbf{z}_j, \mathbf{R}), \quad p(A_{i,j} = 1 \mid \mathbf{z}_i, \mathbf{z}_j, \mathbf{R}) = 1 - e^{-\exp(\sum_{k=1}^{\mathscr{L}} r_k z_{ik} \, z_{jk})}, \quad (7)$$

where $\mathbf{R} \in \mathbb{R}_+^{\mathscr{L} \times \mathscr{L}}$ is a diagonal matrix with diagonal elements r_k.

Lower bound ELBO

$$\underline{\mathcal{L}} = -\mathbb{E}_{\psi \sim q_\phi(\psi \mid \mathbf{X}, \mathbf{A})} [\mathbf{KL}(q(\mathbf{Z} \mid \psi) \,||\, p(\mathbf{Z}))] + \mathbb{E}_{\psi \sim q_\phi(\psi \mid \mathbf{X}, \mathbf{A})} \left[\mathbb{E}_{\mathbf{Z} \sim q(\mathbf{Z} \mid \psi)} [\log p(\mathbf{A} \mid \mathbf{Z})]\right] \leq \mathcal{L}.$$

Reference

[Original paper]