[ICML 2020] Contrastive Multi-View Representation Learning on Graphs. [paper] [code]

Node/Graph Tasks: Node classification and graph classification

Training Type: self-supervised pretraining by constrastive learning through maximizing the mutual information between node and graph representations in two graph views.

Pretext task data: graph topology, node features
The pretext task here is to pretrain a GNN-based encoder to extract node or graph representations of high quality by maximizing the mutual information between node representations from one view and the graph representation from another view.

Initial short summary here
This work is to learn node and graph representations by maximizing mutual information between node representations of one view and graph representations of another view. Two different views of the graph is the original graph adjacency matrix and the graph diffusion matrix. Two normal graph diffusion matrix are heat and PPR diffusion matrix, which are:

$$S^{\text{heat}} = \exp(tAD^{-1} - t), \tag{32}$$

$$S^{\text{PPR}} = \alpha(\mathbf{I}_n - (1-\alpha)D^{-1/2}AD^{-1/2})^{-1}, \tag{33}$$

where $alpha$ denotes teleport probability and t is diffusion time, A, D are the adjacency and the diagonal degree matrix. Then we randomly sample nodes and their associative edges from one view and select the exact nodes and edges from the other view.

Node representations $\mathbf{H}^{\alpha}, \mathbf{H}^{\beta} \in \mathbb{R}^{n \times d}$ in each of these two views are learnt by two different graph encoders followed by a shared projection head. The graph representations of these two views are obtained by applying a graph pooling function to the node representations learned by GNNs (before the projection head) and followed by another shared projection head. They also found that the following readout function:

$$\mathbf{h}_g = \sigma(\|_{l=1}^{L}(\sum_{i=1}^{n}\mathbf{h}_i^l)W) \in \mathbb{R}^{h_d}, \tag{34}$$

achieves the best performance.

The mutual information between nodes and graphs is maximized through:

$$\max \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}(\frac{1}{|g|}\sum_{i=1}^{|g|}\text{MI}(\mathbf{h}_i^{\alpha}, \mathbf{h}_g^{\beta}) + \text{MI}(\mathbf{h}_i^{\beta}, \mathbf{h}_g^{\alpha})), \tag{35}$$

where $|\mathcal{G}|$ is the number of graphs in train set or number of sub-sampled graphs, $|g|$ is the number of nodes in graph g, and $\mathbf{h}_i^{\alpha}, \mathbf{h}_g^{\beta}$ are representations of node i and graph g encoded from views α, β. The MI represents the mutual information estimator and four estimators are explored, which are noise-contrastive estimation, Jensen-Shannon estimator, normalized temperature-scaled cross-entropy and Donsker-Varadhan representation of the KL-divergence. Results of both node and graph classification disclose several important insights. Firstly, Jensen-Shannon estimator achieves better results across all graph classification benchmarks, whereas in node classification, noise contrastive estimation achieves better results. Contrasting node and graph encodings consistently perform better across benchmarks. Contrastiving representations from adjacency and PPR views perform better across the benchmarks. Increasing the number of views does not increase the performance on down-stream tasks.

Bibtex:

@inproceedingshassani2020contrastive, title=Contrastive multi-view representation learning on graphs, author=Hassani, Kaveh and Khasahmadi, Amir Hosein, booktitle=International Conference on Machine Learning, pages=4116–4126, year=2020, organization=PMLR