Student ID : 20194293
Name : Go, Kyeong Ryeol

[AI 602] Graph Transformer Networks

1. Paper Summary

Conventionally, Graph neural networks assume the graph structure to be fixed and homogeneous where one type of nodes and edges is allowed. Recent approaches tried to extend the idea into heterogeneous graph as the noisy graph may have unspecified nodes and edges. Meta-paths that are, by definition, the paths connected with heterogeneous edges are exploited to transform the graph into homogeneous graph. However, it was hand-crafted and required two-stage approach which ends up being non-robust. Here, the author proposed the Graph Transformer Network (GTN) to obtain the learnable meta-paths using composite relations on the set of adjacency matrices and the node representations from the graph convolutional network in an end-to-end fashion.

Meta-paths adjacency matrix $\mathrm{A_P}$ is computed by the matrix multiplication between the softly chosen adjacency matrices $Q_l = \sum_{t_l \in T^e} \alpha_{t_l}^{(l)} A_{t_l}$ where A_{t_l} is the adjacency matrix such that $A_{t_l}[i,j]$ is non-zero when t_l-type edge from j to i exists.

$$A_p = Q_1 \cdot Q_2 \cdot \ldots \cdot Q_l$$

To include the original edges, the identity matrix I is considered as an additional element of the pre-defined set of the adjacency matrix $\mathbb{A}$, which enables GTN to learn any length of meta-paths up to $l+1$ when l layers are stacked. Moreover, to simultaneously consider multiple types of meta-paths, the adjacency matrices can be extended to the adjacency tensors where the multiple node representations Z is now concatenated as

$$\mathrm{Z} = \|_{\mathrm{i}=1}^{\mathrm{C}} \sigma(\widetilde{D}_i^{-1} \tilde{A}_i^{(l)} XW)$$

where each component is from the graph convolutional network, C denotes the number of channels, σ is the non-linear activation function, $\widetilde{D}_i = \sum_i \tilde{A}_{ij}$ is the degree matrix of $\widetilde{\mathrm{A}}_{\mathrm{i}}^{(l)} = \mathrm{A} + \mathrm{I}$, X is the feature matrix, and W is the shared weight matrix across channels. For node classification task trained by the cross-entropy loss, two dense layers followed by the soft-max layer are applied to the derived node representations.

2. Discussion

I. What is the role of the degree matrix $\widetilde{\mathrm{D}}_i$ and the identity matrix I when computing $\widetilde{\mathrm{A}}_i^{(l)}$? To my guess, $\widetilde{\mathrm{D}}_i$ may in charge of the normalization to avoid training instability and I is for considering the relationship with itself, but still I don't understand how it affects the overall performance.

II. Would the product of the multiple softly chosen adjacency matrices $\mathrm{Q_l}$ be problematic if the number of edge type is large? How can we resolve that issue? Any better aggregator?