Self-Supervised Graph Transformer on Large-Scale Molecular Data

Zepeng Zhang

This paper [3] introduced a pre-training model for molecular graph called GROVER, which stands for Graph Representation frOm self-superVised mEssage passing tRansformer. The model architecture is presented below.

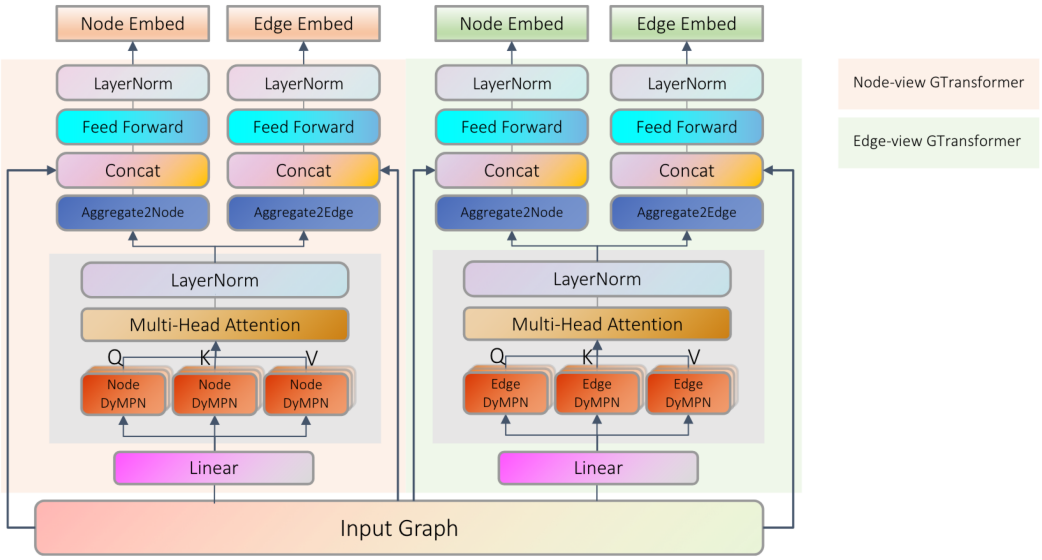

Figure 1: Overview of GROVER architecture

GROVER integrates message passing networks into the transformer-style architecture.

Transformer-style architecture. The architecture has two differences from the original transformer architecture. First, the input node features are passed through a dynamic message passing network (DyMPN) to generate vectors as queries, keys, and values. Second, instead of using short-range residual connections, GROVER only applies one single long-range residual connection, which can alleviate over-smoothing. Considering that the transformer encoder can be viewed as a variant of GAT on a fully connected graph, GROVER actually performs a bi-level information extraction, that is, the DyMPN extracts the local subgraph structure information and the GAT (or transformer architecture) extracts global relations between nodes.

Dynamic message passing network. This paper said that the general message passing process has two hyper-parameters: number of layers and number of hops. In each layer, the same set of parameters are applied for a number of hops. This paper suggests using a randomized strategy for choosing the number of message passing hops, which gives better generalization performance as the experimental results indicate.

Besides the new model architecture, another contribution of the paper is to investigate new self-supervised tasks. For node/edge-level tasks, instead of predicting the node/edge type alone, GROVER predicts the contextual property of nodes. For the graph-level tasks, by incorporating the domain

Preprint. Under review.

knowledge, GROVER extracts the semantic motifs existing in molecular graphs and predicts the occurrence of these motifs for a molecule from graph embeddings.

After pre-training, the GROVER models (with additional parts, e.g., readout function, MLPs) can be used for many downstream with a fine-tuning step.

Something interesting paper worth reading in the future:

- Transformer can be seen as GAT on a fully connected graph [2];
- Message passing over edges [1].

References

[1] Zhengdao Chen, Lisha Li, and Joan Bruna. Supervised community detection with line graph neural networks. In *International Conference on Learning Representations*, 2019. (document)

[2] Chaitanya Joshi. Transformers are graph neural networks. *The Gradient*, 2020. (document)

[3] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020. (document)