Self-Supervised Graph Transformer on Large-Scale Molecular Data

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, Junzhou Huang

Presented by
Monjura Afrin Rumi
Motivation

• Use of deep learning in drug discovery, molecule property prediction
• Issues in using deep learning
  • insufficient labeled data for molecular tasks
    • Time-consuming and resource-costly
  • poor generalization capability of models in the enormous chemical space
• Pre-train using unlabeled data in self-supervised manner
Motivation

• Representation of molecules
  • SMILES – not topology aware
    • BERT
    • N-gram approach
  • Graph – preserves rich structural information
    • Context prediction
    • node-level self-supervised learning
    • graph property prediction for graph-level pre-training
Proposed Method

• **GROVER**: Graph Representation from self-supervised Message passing Transformer

• node/edge-level tasks
  • masks a local subgraph of the target node/edge
  • predicts this contextual property from node embeddings

• graph-level tasks
  • Extracts the semantic motifs existing in molecular graphs
  • predicts the occurrence of these motifs for a molecule from graph embeddings.
GROVER Architecture: GTransformer

• Attention mechanism

$$
\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d})V
$$

• Multi-head attention

$$
\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_k)W_Q^O,
\text{head}_i = \text{Attention}(QW^O_i, KW^K_i, VW^V_i)
$$
GROVER Architecture: dyMPN

• Graph Neural Network
  • Message passing/ neighborhood aggregation
  • Update hidden state

\[
\begin{align*}
    m_v^{(l,k)} &= \text{AGGREGATE}^{(l)}(\{(h_v^{(l,k-1)}), h_u^{(l,k-1)}, e_{uv} \mid u \in \mathcal{N}_v\}), \\
    h_v^{(l,k)} &= \sigma(W^{(l)} m_v^{(l,k)} + b^{(l)}),
\end{align*}
\]

• Randomized strategy for choosing KL
  I. Uniform distribution \( K_t \sim U(a, b) \)
  II. Truncated normal distribution \( \phi(\mu, \sigma, a, b) \)

Figure 1: Overview of GTransformer.
GROVER Architecture - continued

• Aggregate hidden states of nodes
  \[ m_{\text{node-embedding-from-node-states}} = \sum_{u \in N_v} \bar{h}_u \]
  \[ m_{\text{edge-embedding-from-node-states}} = \sum_{u \in N_v \setminus w} \bar{h}_u. \]

• Aggregate hidden states of edges
  \[ m_{\text{node-embedding-from-edge-states}} = \sum_{u \in N_v} \bar{h}_{uv}, \]
  \[ m_{\text{edge-embedding-from-edge-states}} = \sum_{u \in N_v \setminus w} \bar{h}_{uv}. \]

• long-range residual connection
  • Vanishing gradient
  • Over-smoothing
Self-supervised Task Construction

- Contextual Property Prediction
Self-supervised Task Construction

- **Graph-level Motif Prediction**
  - Motifs - recurrent sub-graphs among the input graph data
  - functional groups encodes the rich domain knowledge of molecules
  - Motif extraction tool: RDKit
  - Multi-label classification
Related Work

• **Molecular Representation Learning**
  • chemical fingerprint: represent molecules in the vector space
    • encode the neighbors of atoms in the molecule into a fix-length vector – ECFP
    • Neural fingerprints using convolutional layer - TF_Roubust
  • SMILES
    • RNN-based models to produce molecular representations
  • Graph representation
    • Graph Convolutional Network – GraphConv, Weave, SchNet
    • Graph Attention Network - AttentiveFP
    • GNN – MPNN, DMPNN
    • Hierarchical GNN - MGCN
Related Work

• **Self-supervised Learning on Graphs**
  • Learning objective – vertex proximity relationship
  • Vertex embedding - N-gram model
  • node/edge type prediction - Hu et.al.
Experiments

• Pre-training
  • 11 million (M) unlabelled molecules 10% for validation
  • Context radius k =1, node = 2518, edge = 2686
  • Randomly mask 15% of node and edge labels for prediction
  • RDKit - extract 85 functional groups as the motifs of molecules
  • GROVER_{base} - ~48M parameters
  • GROVER_{large} - ~100M parameters

• Fine-tuning tasks
  • train/validation/test - 8:1:1
  • scaffold splitting
Results

- 6.1% relative improvement
- 2.2% classification
- 10.8% regression
Ablation study

- Pre-training
  - an average AUC increase of 3.8%

- GTransformer Backbone
  - GIN and MPNN
  - toy data set with 600K molecules

- Effect of dyMPN and GTransformer
  - GROVER w/o dyMPN
  - GROVER w/o GTrans
Thank You