Contrastive Pre-Training of GNNs on Heterogeneous Graphs

Xunqiang Jiang¹, Yuanfu Lu², Yuan Fang³, Chuan Shi¹∗

¹Beijing University of Posts and Telecommunications
²WeChat Search Application Department
³Singapore Management University
1 Introduction
2 CPT-HG
3 Experiments
4 Conclusions
CONTENTS

1 Introduction
2 CPT-HG
3 Experiments
4 Conclusions
Introduction

Graph Neural Networks

Graph Neural Network (GNN)

- GCN
- GraphSAGE
- GAT

Various Applications of GNN

- Recommendation System
- Polypharmacy
GNNs need abundant task-specific labeled → Better results

- However, labeled data is usually *expensive* or *infeasible* to obtain

Learning from Unlabeled Data → Pre-training

- *Unlabeled* data (i.e., the whole graph) is abundant
- Recent progresses of pre-training in CV and NLP relieve the *reliance* on labeled data, and some recent works propose to pre-train GNNs in a *self-supervised* manner
Introduction

Heterogeneous Graphs

Existing pre-training methods for GNNs

- They are mainly designed for **homogeneous** graphs

Heterogeneous Graphs and meta graph

Large heterogeneous graphs with **different relations** and **rich semantics**
Two fundamental problems

1. How to distinguish and tailor to different types of nodes and edges during pre-training
   - Defining characteristics of a heterogeneous graph …… → different node and edge types

2. How to further preserve high-order semantic contexts during pre-training
   - High-order semantics → advanced structures for pre-training (i.e., meta-graph)
1 Introduction
2 CPT-HG
3 Experiments
4 Conclusions
Leverage contrastive pre-training tasks on relation- and subgraph-level, which captures both the semantic and structural properties for pre-training GNNs on heterogeneous graph.

- **Relation-level**: differentiate the relation type between two nodes
- **Subgraph-level**: captures high-order structure and embodies complex multiparty relationships
Relation-level Pre-training Task

Encode the relational semantics which constitute the basis of heterogeneity

1. Select negative samples from the unrelated node

   ![Diagram showing the selection of negative samples from the unrelated node]

   - <author1, “writes”, paper1>
   - <author1, “writes”, paper2>
   - <author1, “writes”, institute1>

2. Select negative samples from the inconsistent relation

   ![Diagram showing the selection of negative samples from the inconsistent relation]

   - <author1, “writes”, paper1>
   - <author1, “writes”, paper2>
   - <author1, “joins”, conference>
Relation-level Pre-training Task

Persevere the high-order semantic contexts on a heterogeneous graph.

- Structural Positive Samples capture local connectivity and semantic contexts.
- Queued Negative Samples: negative samples from the latest positive sample, which leads the contrastive procedure more efficient.
PT-HGNN

Schema-level pre-training task

Subgraph-level pre-training task

sampling according to meta-graph

center node with context nodes

Encoder

Pooling

Node embedding

Context embedding

e_{v_i} = \text{Enc}_{\phi}(v_i)(h_{v_i})

\{ Same meta-graph instance, Positive, Different meta-graph instance, Negative \}
### Baselines

- GAE
- DGI
- No-Pretrain
- GPT-GNN
- EdgePred

### Datasets and statistics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Node type</th>
<th>#Nodes</th>
<th>#Edge type</th>
<th>#Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLP</td>
<td>Author (A)</td>
<td>4,057</td>
<td>P-A</td>
<td>19,645</td>
</tr>
<tr>
<td></td>
<td>Paper (P)</td>
<td>16,670</td>
<td>P-V</td>
<td>16,670</td>
</tr>
<tr>
<td></td>
<td>Term (T)</td>
<td>13,420</td>
<td>P-T</td>
<td>133,039</td>
</tr>
<tr>
<td></td>
<td>Venue (V)</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yelp</td>
<td>Business (B)</td>
<td>7,474</td>
<td>B-T</td>
<td>132,928</td>
</tr>
<tr>
<td></td>
<td>Location (L)</td>
<td>65</td>
<td>B-L</td>
<td>7,474</td>
</tr>
<tr>
<td></td>
<td>Star (S)</td>
<td>9</td>
<td>B-S</td>
<td>7,474</td>
</tr>
<tr>
<td></td>
<td>Term (T)</td>
<td>36,412</td>
<td>T-T</td>
<td>360,676</td>
</tr>
<tr>
<td>Aminer</td>
<td>Paper (P)</td>
<td>614,209</td>
<td>P-A</td>
<td>2,311,822</td>
</tr>
<tr>
<td></td>
<td>Author (A)</td>
<td>737,621</td>
<td>P-C</td>
<td>764,246</td>
</tr>
<tr>
<td></td>
<td>Conference (C)</td>
<td>842</td>
<td>P-P</td>
<td>4,665,400</td>
</tr>
<tr>
<td></td>
<td>Terms (T)</td>
<td>80,589</td>
<td>P-T</td>
<td>7,722,124</td>
</tr>
</tbody>
</table>
Experiments

Experiment results on Node classification and Link Prediction

- The improvements suggest that the contrastive pre-training on heterogeneous graphs is capable of learning transferable knowledge.
- Both self-supervised tasks in GPT-GNN help the pre-training framework:
  - Relation-level contrastive task
  - Subgraph-level contrastive task

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Link Type</th>
<th>No pre-train</th>
<th>GAE</th>
<th>EgePred</th>
<th>DGI</th>
<th>GPT-GNN</th>
<th>CPT-HG</th>
<th>Improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBLP</td>
<td>Paper-Term</td>
<td>12.34 ± 1.43</td>
<td>12.51 ± 0.71</td>
<td>12.61 ± 0.44</td>
<td>12.47 ± 0.68</td>
<td>12.71 ± 0.35</td>
<td>13.06 ± 0.42</td>
<td>2.75%</td>
</tr>
<tr>
<td>Yelp</td>
<td>Business-Location</td>
<td>45.83 ± 0.42</td>
<td>45.92 ± 0.52</td>
<td>46.10 ± 0.31</td>
<td>45.57 ± 0.64</td>
<td>46.04 ± 0.75</td>
<td>47.04 ± 0.71</td>
<td>2.03%</td>
</tr>
<tr>
<td>Aminer</td>
<td>Paper-Conference</td>
<td>39.23 ± 1.75</td>
<td>40.31 ± 0.78</td>
<td>39.86 ± 1.17</td>
<td>40.74 ± 1.35</td>
<td>41.37 ± 0.76</td>
<td>42.17 ± 1.23</td>
<td>1.93%</td>
</tr>
<tr>
<td></td>
<td>Paper-Author</td>
<td>5.63 ± 0.73</td>
<td>5.71 ± 0.41</td>
<td>5.62 ± 0.87</td>
<td>5.84 ± 0.52</td>
<td>6.02 ± 0.45</td>
<td>6.43 ± 0.54</td>
<td>6.81%</td>
</tr>
</tbody>
</table>

Table 2: Experiment results (MRR ± std) in link prediction task on the three datasets. The best method is bolded, and the second best is underlined.

<table>
<thead>
<tr>
<th>Downstream Task</th>
<th>Dataset</th>
<th>Link Type</th>
<th>Link Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Paper-Term</td>
<td>Paper-Conference</td>
</tr>
<tr>
<td>No pre-train</td>
<td>12.34 ± 1.43</td>
<td>45.83 ± 0.42</td>
<td>39.23 ± 1.75</td>
</tr>
<tr>
<td>CPT-HG_{sub}</td>
<td>12.65 ± 0.42</td>
<td>47.15 ± 0.44</td>
<td>41.54 ± 0.33</td>
</tr>
<tr>
<td>CPT-HG_{rel}</td>
<td>12.79 ± 0.56</td>
<td>46.74 ± 0.65</td>
<td>41.75 ± 0.65</td>
</tr>
<tr>
<td>CPT-HG</td>
<td><strong>13.06 ± 0.42</strong></td>
<td>47.04 ± 0.71</td>
<td><strong>42.17 ± 1.23</strong></td>
</tr>
</tbody>
</table>

Table 4: Analysis of different ablated models in various downstream tasks.
Ablation study

Experiments

The semantic relations encoded in CPT-HG$_{rel}$ seem more important for node representations in heterogeneous graph.

- CPT-HG$_{sub}$ significantly improves the link prediction performance by focusing on modeling subgraph structures in a graph.

<table>
<thead>
<tr>
<th>Ablation Model</th>
<th>DBLP Paper-Term</th>
<th>Yelp Business-Location</th>
<th>Aminer Paper-Conference</th>
<th>Aminer Paper-Author</th>
<th>DBLP Paper</th>
<th>Aminer Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>No pre-train</td>
<td>12.34 ± 1.43</td>
<td>45.83 ± 0.42</td>
<td>39.23 ± 1.75</td>
<td>5.63 ± 0.73</td>
<td>87.45 ± 0.43</td>
<td>92.17 ± 0.56</td>
</tr>
<tr>
<td>CPT-HG$_{sub}$</td>
<td>12.65 ± 0.42</td>
<td><strong>47.15 ± 0.44</strong></td>
<td>41.54 ± 0.33</td>
<td>6.04 ± 0.51</td>
<td>89.57 ± 0.61</td>
<td>94.14 ± 0.54</td>
</tr>
<tr>
<td>CPT-HG$_{rel}$</td>
<td>12.79 ± 0.56</td>
<td>46.74 ± 0.65</td>
<td>41.75 ± 0.65</td>
<td>6.24 ± 0.15</td>
<td><strong>92.45 ± 0.54</strong></td>
<td>95.16 ± 0.32</td>
</tr>
<tr>
<td>CPT-HG</td>
<td><strong>13.06 ± 0.42</strong></td>
<td>47.04 ± 0.71</td>
<td><strong>42.17 ± 1.23</strong></td>
<td><strong>6.43 ± 0.54</strong></td>
<td>91.45 ± 0.54</td>
<td><strong>96.32 ± 0.43</strong></td>
</tr>
</tbody>
</table>

Table 4: Analysis of different ablated models in various downstream tasks.

Figure 3: Improvement of different ablated models over no pre-train baseline.
CONTENTS

1 Introduction
2 CPT-HG
3 Experiments
4 Conclusions
CPT-HG, which is a pre-training framework, enables the GNN to capture heterogeneous semantics and structural properties.

Extensive experiments three real-world heterogeneous graphs.