Representing Long-Range Context for Graph Neural Networks with Global Attention

Zhanghao Wu*, Paras Jain*, Matthew A. Wright, Azalia Mirhoseini, Joseph E. Gonzalez, Ion Stoica

UC Berkeley RISELab, Google Brain

github.com/ucbrise/graphtrans

NeurIPS 2021
Graphs are an important representation of natural structures

Social networks  Molecules  Meshes
Graph neural networks aggregate local neighborhood structure

By iteratively pooling immediate neighborhood, GNNs slowly learn to represent local structure
Challenge: long-range dependencies are not represented in GNNs

Signal decreases exponentially with GCN depth!
Long-range dependencies important to graph classification

Graph classification considers **pooling embeddings** into single prediction vector.

Small interactions in molecules may result in large changes in function!

(S)-thalidomide = **deadly**

(R)-thalidomide = **safe**
Learning global interactions with **GraphTrans**

Can we use global attention w/ **Transformers** to identify long-range dependencies?

---

Representing Long-Range Context for Graph Neural Networks with Global Attention

[github.com/ucbrise/graphtrans]
Transformer alone cannot model graph structure

Transformer results in -14% test acc. drop!
Learning global interactions with **GraphTrans**

We leverage a SOTA GNN backbone to learn local structures.
Learning global interactions with GraphTrans

We leverage a SOTA GNN backbone to learn local structures

We add a modified Transformer to pool local embeddings to extract global structures

Representing Long-Range Context for Graph Neural Networks with Global Attention

github.com/ucbrise/graphtrans
**GraphTrans** recovers accuracy w/ local + global structure

Local (GNN) + Global (Transformer) results in +15% test accuracy over global only, +1.12% over SOTA
Understanding GraphTrans by visualizing attention

Graph from OGB-code2 dataset
Understanding GraphTrans by visualizing attention

Nodes 8 and 17 are related, yet far in graph

Graph from OGB-code2 dataset
Understanding GraphTrans by visualizing attention

Graph from OGB-code2 dataset
Understanding GraphTrans by visualizing attention

Transformer attention map models distant interaction between nodes 8 and 17
Learning global information with `<CLS>` token
Learning global information with `<CLS>` token

Addition of a single global CLS token aggregates global information into a single vector.
Learning global information with <CLS> token

<CLS> token learns how to aggregate global information together
### Evaluation: Biological Benchmark

<table>
<thead>
<tr>
<th></th>
<th>GNN Type</th>
<th>GNN Layers</th>
<th>NCI1</th>
<th>NCI109</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAGPool$_g$</td>
<td>GCN</td>
<td>3</td>
<td>74.2</td>
<td>74.1</td>
</tr>
<tr>
<td>Strong Baseline</td>
<td>GCN</td>
<td>8</td>
<td>81.5</td>
<td>-</td>
</tr>
<tr>
<td>GraphTrans (small)</td>
<td>GCN</td>
<td>3</td>
<td>80.2</td>
<td>79.0</td>
</tr>
<tr>
<td>GraphTrans (large)</td>
<td>GIN</td>
<td>4</td>
<td><strong>83.0</strong> (+8.8)</td>
<td><strong>82.5</strong> (+8.4)</td>
</tr>
</tbody>
</table>
**Evaluation**: Chemical Benchmark

<table>
<thead>
<tr>
<th>Method</th>
<th>Valid</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCN-Virtual</td>
<td>0.250</td>
<td>0.242</td>
</tr>
<tr>
<td>GIN-Virtual</td>
<td>0.280</td>
<td>0.270</td>
</tr>
<tr>
<td>GraphTrans (GIN)</td>
<td>0.288</td>
<td>0.272</td>
</tr>
<tr>
<td>GraphTrans (GIN-Virtual)</td>
<td>0.286</td>
<td>0.282</td>
</tr>
</tbody>
</table>

GraphTrans (GIN-Virtual) improves over GIN by 0.006 (Valid) and 0.012 (Test).
**Evaluation: Computer Programming Benchmark**

Test Score on OGBG-Code2 Dataset

- GIN: 0.150
- GCN: 0.151
- PNA: 0.159
- DAGNN (SOTA): 0.175
- GraphTrans (GCN): 0.175
- GraphTrans (PNA): 0.182
## Scalability: Train on Large Graphs

<table>
<thead>
<tr>
<th>#Nodes</th>
<th>Model</th>
<th>Edge Density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>20%</td>
</tr>
<tr>
<td>500</td>
<td>GCN-Virtual</td>
<td>44.3</td>
</tr>
<tr>
<td></td>
<td>GraphTrans</td>
<td>48.4</td>
</tr>
<tr>
<td>1000</td>
<td>GCN-Virtual</td>
<td>90.1</td>
</tr>
<tr>
<td></td>
<td>GraphTrans</td>
<td>96.9</td>
</tr>
<tr>
<td>2000</td>
<td>GCN-Virtual</td>
<td>131.8</td>
</tr>
<tr>
<td></td>
<td>GraphTrans</td>
<td>127.9</td>
</tr>
</tbody>
</table>
Representing Long-Range Context for Graph Neural Networks with Global Attention

Zhanghao Wu*, Paras Jain*, Matthew A. Wright, Azalia Mirhoseini, Joseph E. Gonzalez, Ion Stoica

- Long-range dependencies crucial to model for graph classification.
- GNNs struggle to learn long-range dependencies.
- GraphTrans combines local representations (from GNN) with global (from Transformer) to learn long-range dependencies.
- We achieve SOTA results on molecular, biological and code prediction datasets.