Data Augmentation for Graph Neural Networks

Tong Zhao†, Yozen Liu‡, Leonardo Neves‡, Oliver Woodford‡, Meng Jiang†, and Neil Shah‡

† Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
‡ Snap Research, Snap Inc., Santa Monica, CA 90405, USA
Backgrounds on Data Augmentation

• Why data augmentation is important for machine learning?
  • Provides more training data.
  • Reduce overfitting.
  • Improves generalization.
Backgrounds on Data Augmentation

• Why data augmentation is important for machine learning?
  • Provides more training data.
  • Reduce overfitting.
  • Improves generalization.

A boy is holding a bat. Translation: A boy is holding a backpack.

Background on Graph Neural Networks

• Given: graph $G = (V, E)$, node features $\mathbf{x}_v \in \mathbb{R}^m$, $\forall v \in V$.
• Learn: low dimensional node representations $\mathbf{z}_v \in \mathbb{R}^d$, $\forall v \in V$.
• Neighborhood aggregation: generate node representations based on local neighborhoods.
Data Augmentation for GNNs

- **Goal:** use graph data augmentation to improve the performance of GNNs on the task of node classification.

- **Challenges:**
  - There’s no direct analogs of traditional data augmentation operations (flipping, rotating, blurring, etc.) on graphs.
  - Very limited operations exist for perturbing graphs.
    - Any manipulation would affect the whole graph (dataset).
    - Adding/removing edges are the best strategy available.
    - But *which* edges to add/remove?
Manipulating Edges to Augment Graph Data

• For node classification task,
  • There could be noise edges generated by spammers, anomalies, adversarial attacks, etc.

• Adding(removing) intra(inter)-class edges improves node classification performance.
Zackary's Karate Club (ZKC) Graph

Original graph. GCN Performance: 92.4

Omniscient modified graph. GCN Performance: 98.6
Random Initialized Features of ZKC Graph

Random initialized features
Embeddings of ZKC Graph after GCN Layer

Random initialized features

Original Graph
Embeddings of ZKC Graph after GCN Layer

Random initialized features

Original Graph

Modified Graph
Embeddings of ZKC Graph after GCN Layer

Random initialized features

Original Graph

Modified Graph

Ideal Graph
Evaluating on Modified or Original Graph

- How traditional data augmentation methods in CV works:
  - Generate augmented data variants for each data object.
  - Train model with both original and augmented data.

- On graphs: augmentation results with a new graph.
  - Training & inference on different graphs train-test gap.
  - For real-life social networks that are consistently growing/changing, ability of inferencing with original graph is preferred.
GAug-M

1. Use an edge predictor to predict edge probabilities for all node pairs.

2. Based on the edge probabilities, deterministically add (remove) new (existing) edges to create a modified graph.
GAug-M

1. Use an edge predictor to predict edge probabilities for all node pairs.

2. Based on the edge probabilities, deterministically add (remove) new (existing) edges to create a modified graph.
GAug-O for Evaluating on Original Graph

$$A'_{ij} = \left[ \frac{1}{1 + e^{-(\log P_{ij} + G) / \tau}} + \frac{1}{2} \right], \quad \text{where} \quad P_{ij} = \alpha M_{ij} + (1 - \alpha) A_{ij}$$

$$\mathcal{L} = \mathcal{L}_{nc} + \beta \mathcal{L}_{ep}, \quad \text{where} \quad \mathcal{L}_{nc} = CE(\hat{y}, y) \quad \text{and} \quad \mathcal{L}_{ep} = BCE(\sigma(f_{ep}(A, X)), A)$$
<table>
<thead>
<tr>
<th>GNN Arch.</th>
<th>Method</th>
<th>CORA</th>
<th>CITESEER</th>
<th>PPI</th>
<th>BLOGC</th>
<th>FLICKR</th>
<th>AIR-USA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>81.6±0.7</td>
<td>71.6±0.4</td>
<td>43.4±0.2</td>
<td>75.0±0.4</td>
<td>61.2±0.4</td>
<td>56.0±0.8</td>
</tr>
<tr>
<td>GCN</td>
<td>+ADAEDGE</td>
<td>81.9±0.7</td>
<td>72.8±0.7</td>
<td>43.6±0.2</td>
<td>75.3±0.3</td>
<td>61.2±0.5</td>
<td>57.2±0.8</td>
</tr>
<tr>
<td></td>
<td>+GAUG-M</td>
<td>83.5±0.4</td>
<td>72.3±0.4</td>
<td>43.5±0.2</td>
<td><strong>77.6±0.4</strong></td>
<td><strong>68.2±0.7</strong></td>
<td>61.2±0.5</td>
</tr>
<tr>
<td></td>
<td>+DROPEDGE</td>
<td>82.0±0.8</td>
<td>71.8±0.2</td>
<td>43.5±0.2</td>
<td>75.4±0.3</td>
<td>61.4±0.7</td>
<td>56.9±0.6</td>
</tr>
<tr>
<td></td>
<td>+GAUG-O</td>
<td><strong>83.6±0.5</strong></td>
<td><strong>73.3±1.1</strong></td>
<td><strong>46.6±0.3</strong></td>
<td>75.9±0.2</td>
<td>62.2±0.3</td>
<td><strong>61.4±0.9</strong></td>
</tr>
<tr>
<td></td>
<td>Original</td>
<td>81.3±0.5</td>
<td>70.6±0.5</td>
<td>40.4±0.9</td>
<td>73.4±0.4</td>
<td>57.4±0.5</td>
<td>57.0±0.7</td>
</tr>
<tr>
<td>GSAGE</td>
<td>+ADAEDGE</td>
<td>81.5±0.6</td>
<td>71.3±0.8</td>
<td>41.6±0.8</td>
<td>73.6±0.4</td>
<td>57.7±0.7</td>
<td>57.1±0.5</td>
</tr>
<tr>
<td></td>
<td>+GAUG-M</td>
<td><strong>83.2±0.4</strong></td>
<td>71.2±0.4</td>
<td>41.1±1.0</td>
<td><strong>77.0±0.4</strong></td>
<td><strong>65.2±0.4</strong></td>
<td><strong>60.1±0.5</strong></td>
</tr>
<tr>
<td></td>
<td>+DROPEDGE</td>
<td>81.6±0.5</td>
<td>70.8±0.5</td>
<td>41.1±1.0</td>
<td>73.8±0.4</td>
<td>58.4±0.7</td>
<td>57.1±0.5</td>
</tr>
<tr>
<td></td>
<td>+GAUG-O</td>
<td>82.0±0.5</td>
<td><strong>72.7±0.7</strong></td>
<td><strong>44.4±0.5</strong></td>
<td>73.9±0.4</td>
<td>56.3±0.6</td>
<td>57.1±0.7</td>
</tr>
<tr>
<td></td>
<td>Original</td>
<td>81.3±1.1</td>
<td>70.5±0.7</td>
<td>41.5±0.7</td>
<td>63.8±5.2</td>
<td>46.9±1.6</td>
<td>52.0±1.3</td>
</tr>
<tr>
<td>GAT</td>
<td>+ADAEDGE</td>
<td>82.0±0.6</td>
<td>71.1±0.8</td>
<td>42.6±0.9</td>
<td>68.2±2.4</td>
<td>48.2±1.0</td>
<td>54.5±1.9</td>
</tr>
<tr>
<td></td>
<td>+GAUG-M</td>
<td>82.1±1.0</td>
<td>71.5±0.5</td>
<td>42.8±0.9</td>
<td>70.8±1.0</td>
<td><strong>63.7±0.9</strong></td>
<td><strong>59.0±0.6</strong></td>
</tr>
<tr>
<td></td>
<td>+DROPEDGE</td>
<td>81.9±0.6</td>
<td>71.0±0.5</td>
<td><strong>45.9±0.3</strong></td>
<td>70.4±2.4</td>
<td>50.0±1.6</td>
<td>52.8±1.7</td>
</tr>
<tr>
<td></td>
<td>+GAUG-O</td>
<td><strong>82.2±0.8</strong></td>
<td><strong>71.6±1.1</strong></td>
<td><strong>44.9±0.9</strong></td>
<td><strong>71.0±1.1</strong></td>
<td>51.9±0.5</td>
<td>54.6±1.1</td>
</tr>
<tr>
<td></td>
<td>Original</td>
<td>78.8±1.5</td>
<td>67.6±1.8</td>
<td>44.1±0.7</td>
<td>70.0±0.4</td>
<td>56.7±0.4</td>
<td>58.2±1.5</td>
</tr>
<tr>
<td>JK-NET</td>
<td>+ADAEDGE</td>
<td>80.4±1.4</td>
<td>68.9±1.2</td>
<td>44.8±0.9</td>
<td>70.7±0.4</td>
<td>57.0±0.3</td>
<td>59.4±1.0</td>
</tr>
<tr>
<td></td>
<td>+GAUG-M</td>
<td><strong>81.8±0.9</strong></td>
<td>68.2±1.4</td>
<td>47.4±0.6</td>
<td><strong>71.9±0.5</strong></td>
<td><strong>65.7±0.8</strong></td>
<td>60.2±0.6</td>
</tr>
<tr>
<td></td>
<td>+DROPEDGE</td>
<td>80.4±0.7</td>
<td>69.4±1.1</td>
<td>46.3±0.2</td>
<td>70.9±0.4</td>
<td>58.5±0.7</td>
<td>59.1±1.1</td>
</tr>
<tr>
<td></td>
<td>+GAUG-O</td>
<td>80.5±0.9</td>
<td><strong>69.7±1.4</strong></td>
<td><strong>53.1±0.3</strong></td>
<td>71.0±0.6</td>
<td>55.7±0.5</td>
<td><strong>60.4±1.0</strong></td>
</tr>
</tbody>
</table>
Results
Thank you!

Any Questions?