Representational Power of Graph Neural Networks

Stefanie Jegelka
MIT

joint work with
Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe,
Ken-ichi Kawarabayashi, Weihua Hu, Jure Leskovec, Jingling Li,
Mozhi Zhang, Simon S. Du
Learning with Graphs

- Graph Molecule
- Property: Solubility, Toxicity, Drug efficacy
- Node Item → Node Item
- Pair of Nodes: Drugs
- Edge: Interaction

(Duvenaud et al, 2015, …)
(Ying et al, 2018)
(Battaglia et al. 2016)
(Allamanis et al. 2018)
Outline

- Discriminative Power of Graph Neural Networks
  K. Xu, W. Hu, J. Leskovec, S. Jegelka, ICLR 2019

- Network Depth and Graph Structure
  K. Xu, C. Li, Y. Tian, T. Sonobe, K. Kawarabayashi, S. Jegelka, ICML 2018

- Neural Networks for Reasoning
Graph Neural Networks

In each round:

**Aggregate** over neighbors

\[ a_v^{(k)} = \text{AGGREGATE}^{(k)} \left( \{ h_u^{(k-1)} : u \in \mathcal{N}(v) \} \right) \]

**Combine** with current node

\[ h_v^{(k)} = \text{COMBINE}^{(k)} \left( h_v^{(k-1)}, a_v^{(k)} \right) \]

Graph-level **readout**

\[ h_G = \text{READOUT} \left( \{ h_v^{(K)} \mid v \in G \} \right) \]

(Battaglia et al., 2016; Defferrard et al., 2016; Duvenaud et al., 2015; Hamilton et al., 2017a; Kearnes et al., 2016; Kipf & Welling, 2017; Li et al., 2016; Velickovic et al., 2018; Verma & Zhang, 2018; Ying et al., 2018b; Zhang et al., 2018, …)
Graph Neural Networks

Intuition: Nodes aggregate information from their neighbors using neural networks.

INPUT GRAPH

TARGET NODE

layer 0 (input)

shared weights

layer 1

layer 2

(illustrations: J. Leskovec)
Graph Neural Networks

Which graphs can GNNs distinguish? What does this depend on?
**Lemma (XHLJ19)**

Aggregation-based GNNs are at most as powerful as the 1-dim. Weisfeiler-Lehman graph isomorphism test*. 

Reach maximum power?

Aggregate from neighbors

\[ g(X) = \phi \left( \text{MEAN}\{ f(x) : x \in X \} \right) \]  
(Kipf et al. 2017)

\[ g(X) = \phi \left( \text{MAX}\{ f(x) : x \in X \} \right) \]  
(Hamilton et al. 2017)

Failure: same node representations

max / mean pooling fail

max pooling fails
Discriminative Schemes

Theorem \((XHLJ19)\)
If Aggregate, Combine and Readout are \textit{injective}, then the network is as discriminative as the 1-dim. WL test.

- max / mean fail
- max fails
A powerful GNN (GIN)

**Lemma**

Any multi-set function $g$ can be decomposed as

$$g(X) = \phi \left( \sum_{x \in X} f(x) \right)$$

- Aggregation: sum & appropriate nonlinearity:

$$h_{v}^{(k)} = \text{MLP}^{(k)} \left( (1 + \epsilon^{(k)}) \cdot h_{v}^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_{u}^{(k-1)} \right)$$
A powerful GNN (GIN)

**Lemma**

Any multi-set function \( g \) can be decomposed as

\[
g(X) = \phi \left( \sum_{x \in X} f(x) \right)
\]

- Aggregation: sum & appropriate nonlinearity:

\[
h_{v}^{(k)} = \text{MLP}^{(k)} \left( (1 + \epsilon^{(k)}) \cdot h_{v}^{(k-1)} + \sum_{u \in \mathcal{N}(v)} h_{u}^{(k-1)} \right)
\]

- Readout: concatenation or sum

\[
h_{G} = \text{CONCAT} \left( \text{READOUT} \left( \left\{ h_{v}^{(k)} | v \in G \right\} \right) \mid k = 0, 1, \ldots, K \right)
\]
Example Empirical Results

\[ g(X) = \phi \left( \sum_{x \in X} f(x) \right) \]

PROTEINS

Training accuracy

Epoch

Sum — MLP (injective)

Sum — linear+ReLU

Mean/Max — MLP/linear+ReLU
Towards Understanding Graph Neural Networks

- **Discriminative Power of Graph Neural Networks**
  
  K. Xu, W. Hu, J. Leskovec, S. Jegelka, ICLR 2019

- **Network Depth and Graph Structure**
  
  K. Xu, C. Li, Y. Tian, T. Sonobe, K. Kawarabayashi, S. Jegelka, ICML 2018

- **Neural Networks for Reasoning**
Why is deeper not always better here?

Intuition: Nodes aggregate information from their neighbors using neural networks.
Neighborhood Aggregation

**Influence distribution**

\[ I_x(y) = e^T \left[ \frac{\partial h_x^{(k)}}{\partial h_y^{(0)}} \right] e / \left( \sum_{z \in V} e^T \left[ \frac{\partial h_x^{(k)}}{\partial h_z^{(0)}} \right] e \right) \]

*how much is node x's feature influenced by node y*

**Theorem (XLTSKJ18)** Under simplifying assumptions*:


* as in (Choromanska et al 2015, Kawaguchi 2016)
Implications

Random Walk convergence / size of feature neighborhood depends on **expansion properties of the graph**

Conjecture:
Different depths work for different subgraph structures
Examples

3 and 4-layer models make incorrect prediction

(d) 2-layer

(e) 3-layer

(f) 4-layer

(j) 2-layer

(k) 3-layer

(l) 4-layer
Examples

2-layer models make incorrect prediction
Adaptive depth

Adapting to subgraph structures empirically improves performance by 2-30%

JK-Concat adapts once for entire dataset/graph:
*best for more regular graphs (images: DenseNet)*

JK-LSTM adapts for each node/subgraph:
*best for large graphs with diverse structures*
Towards Understanding Graph Neural Networks

» Discriminative Power of Graph Neural Networks
  K. Xu, W. Hu, J. Leskovec, S. Jegelka, ICLR 2019

» Network Depth and Graph Structure
  K. Xu, C. Li, Y. Tian, T. Sonobe, K. Kawarabayashi, S. Jegelka, ICML 2018

» Neural Networks for Reasoning
Reasoning Tasks

**summary statistics**

What is the maximum value difference among treasures?

**“relational argmax”**

What are the colors of the farthest pair of objects?

**dynamic programming**

What is the cost to defeat monster X by following the optimal path?

---

(Johnson et al., 2017a; Weston et al., 2015; Hu et al., 2017; Fleuret et al., 2011; Antol et al., 2015; Battaglia et al., 2016; Watters et al., 2017; Fragkiadaki et al., 2016; Chang et al., 2017; Saxton et al., 2019; Chang et al., 2019; Santoro et al., 2018; Zhang et al., 2019, …)
Algorithmic Alignment: Network can mimic algorithm via few, easy-to-learn modules

Bellman-Ford

\[
\text{for } k = 1 \ldots |S| - 1:\n\]

\[
\text{for } u \text{ in } S:\n\]

\[
d[k][u] = \min_v \ d[k-1][v] + \text{cost}(v, u)
\]

GNN

\[
\text{for } k = 1 \ldots \text{GNN iter}:\n\]

\[
\text{for } u \text{ in } S:\n\]

\[
h_u^{(k)} = \sum_v \text{MLP}(h_v^{(k-1)}, h_u^{(k-1)})
\]
Alignment and Learning

More generally: GNNs align with a class of DP

\[ \text{Answer}[k][i] = \text{DP-Update}(\{ \text{Answer}[k-1][j], j = 1 \ldots n \}) \]

Includes many reasoning tasks: visual question answering, physical reasoning,…
Alignment & Implications

A neural network \((M, \epsilon, \delta)\)-aligns with an algorithm if it can mimic the algorithm via \(n\) different (shared) network modules, each of which can be learned with at most \(M/n\) samples.

**Theorem**

If a network and task algorithm \((M, \epsilon, \delta)\)-align, then, under assumptions, the task is \((M, O(\epsilon), O(\delta))\)-learnable by the network.
Representational Power of GNNs

• Discriminative power of graph neural networks
  ‣ Aggregation schemes important!

• Local expansion and network depth
  ‣ Adaptive depths for different graph structures

• Alignment for reasoning tasks
