Journal Club: Graph Attention Networks

Jitian Zhao

University of Wisconsin Madison

February 21, 2020

Outline

- Existed methods
- Graph attentional layer
- Comparisons
- Results

Existed Literatures for Node classification

- **Recursive methods**:
  - Recursive neural network(DAG)
  - Graph neural network(general graphs including cyclic/directed/undirected)
  - improve the propagation by using GRU
- **Convolution methods**:
  - Spectral: eigendecomposition of graph Laplacian to find filter
  - Non-spectral: define convolution on graph using spacially close neighbors
- **Graph Attention Network**: compute hidden representation of each node by considering its neighbors(self-attention)

Graph attentional layer

- input: $h = \{h_1, ..., h_N\}, h_i \in \mathbb{R}^F$
- output: $h^{'} = \{h_1^{'}, ..., h_N^{'}\}, h_i^{'} \in \mathbb{R}^{F^{'}}$
- attention mechanism:

$$\alpha_{ij} = \frac{\exp\left(\text{ LeakyReLU } \left(\overrightarrow{\mathbf{a}}^T \left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j\right]\right)\right)}{\sum_{k\in\mathcal{N}_i} \exp\left(\text{ LeakyReLU } \left(\overrightarrow{\mathbf{a}}^T \left[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k\right]\right)\right)}$$

attention mechanism

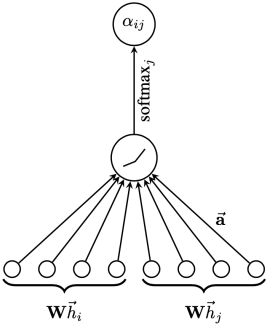

Figure: attention mechanism

- $e_{ij} = a\left(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j\right)$
- $\alpha_{ij} = \text{softmax}_j\left(e_{ij}\right) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}$

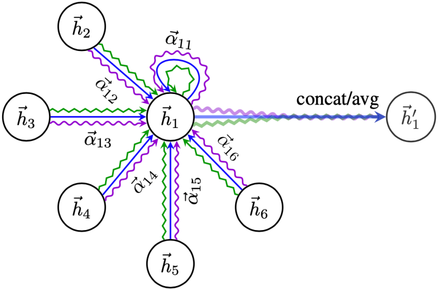

Figure: multi-head attention

- stabilize the learning process
- **concatenate**: $\vec{h}'_i = \|_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \vec{h}_j\right)$
- **average**: $\vec{h}'_i = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \vec{h}_j\right)$

Comparisons

- computationally efficient(paralleled computation through edges but might cause redundant computation because of overlapping)
- assign different importances to nodes in same neighborhood
- dose not depend on global graph structure
- flexible neighborhood size

Results

- **transductive**: unsupervised tasks: label propagation, semi-supervised embedding, manifold regularization, skip-gram based graph embeddings, iterative classification algorithm,
- **inductive**: supervised

Transductive

Method	Cora	Citeseer	Pubmed
MLP	55.1%	46.5%	71.4%
ManiReg (Belkin et al., 2006)	59.5%	60.1%	70.7%
SemiEmb (Weston et al., 2012)	59.0%	59.6%	71.7%
LP (Zhu et al., 2003)	68.0%	45.3%	63.0%
DeepWalk (Perozzi et al., 2014)	67.2%	43.2%	65.3%
ICA (Lu & Getoor, 2003)	75.1%	69.1%	73.9%
Planetoid (Yang et al., 2016)	75.7%	64.7%	77.2%
Chebyshev (Defferrard et al., 2016)	81.2%	69.8%	74.4%
GCN (Kipf & Welling, 2017)	81.5%	70.3%	**79.0%**
MoNet (Monti et al., 2016)	$81.7 \pm 0.5\%$	—	$78.8 \pm 0.3\%$
GCN-64*	$81.4 \pm 0.5\%$	$70.9 \pm 0.5\%$	$\mathbf{79.0} \pm 0.3\%$
GAT (ours)	$\mathbf{83.0} \pm 0.7\%$	$\mathbf{72.5} \pm 0.7\%$	$\mathbf{79.0} \pm 0.3\%$

Figure: transductive