2022 IEEE International Conference on Acoustics, Speech, and Signal Processing
icassp 2022
human-centric signal processing
May 22-27, 2022, Sands Expo & Convention Centre
IEEE

Exploring Deeper Graph Convolutions For Semi-Supervised Node Classification

Ashish Tiwari, Richeek Das, and Shanmuganathan Raman
Computer Vision, Imaging, and Graphics (CVIG) Lab, Indian Institute of Technology Gandhinagar, India
E-mail: {ashish.tiwari, shanmuga} @ iitgn.ac.in, das.richeek01@iitb.ac.in

Paper No. 3673

Introduction

- Graph convolutional networks (GCNs) continue to suffer from *oversmoothing* - performance reduction with an increasing number of layers.
- We introduce a simple yet effective idea of **feature gating** over graph convolution layers to facilitate deeper graph neural networks and address oversmoothing.
- The proposed feature gating is easy to implement without changing the underlying network architecture and is broadly applicable to GCN and almost any of its variants.
- We demonstrate the use of feature gating in assigning importance to node features and the nodes for the node classification task.

Background

- The graph convolution layer is defined as:

$$\mathbf{H}^{(l+1)} = \sigma\left(\widetilde{\mathbf{P}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right), \text{ where } \widetilde{\mathbf{P}} = \widetilde{\mathbf{D}}^{-1/2}\widetilde{\mathbf{A}}\widetilde{\mathbf{D}}^{-1/2}$$

- Such a fixed polynomial filter $\widetilde{\mathbf{P}}^{K}\boldsymbol{x}$ converges to a distribution that is distant from the input feature $\mathbf{x}$ and hence, incurs vanishing gradients.
- Let $d_{\mathcal{M}}(\mathbf{H}) := \inf\{\|\mathbf{H} - \mathbf{Y}\|_F \mid \mathbf{Y} \in \mathcal{M}\}$ denote the distance between $\mathbf{H}$ and $\mathcal{M}$.
- For any initial value $\mathbf{H}^{(0)}$, the output of l^{th} layer $\mathbf{H}^{(l)}$ satisfies

$$d_{\mathcal{M}}\left(\mathbf{H}^{(l)}\right) \leq (s\lambda)^{l} d_{\mathcal{M}}\left(\mathbf{H}^{(0)}\right)$$

- Here, s and λ are the maximum singular and eigen values of the weight matrix $\mathbf{W}$ and is the normalised laplacian matrix $\widetilde{\mathbf{P}}$, respectively.
- For, $\lambda = 1$, we can have $s \geq 1$ and for $\lambda \neq 1$, the sweet spot falls in the range $1 \leq s \leq \lambda^{-1}$.

Proposed Formulation

- Combining initial residual and identity mapping:

$$\mathbf{H}^{(l+1)} = \sigma\left(\left(\alpha_l \widetilde{\mathbf{P}}\mathbf{H}^{(l)} + (1-\alpha_l)\mathbf{H}^{(0)}\right)\left((1-\beta_l)\mathbf{I} + \beta_l \mathbf{W}^{(l)}\right)\right)$$

- Feature gating in GCN:

$$\mathbf{H}^{(l+1)} = \widetilde{\mathbf{P}}\left(\phi\left(\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right) \odot \sigma\left(\mathbf{H}^{(l)}\boldsymbol{G}^{(l)}\right)\right)$$

- GCN-IR-FG:

$$\mathbf{H}^{(l+1)} = \beta_l\left(\widetilde{\mathbf{P}}\left(\phi\left(\mathbf{Z}^{(l)}\mathbf{W}^{(l)}\right) \odot \sigma\left(\mathbf{Z}^{(l)}\boldsymbol{G}^{(l)}\right)\right)\right) + (1-\beta_l)\left(\widetilde{\mathbf{P}}\mathbf{Z}^{(l)}\mathbf{I}\right)$$

such that, $\mathbf{Z}^{(l)} = (1-\alpha_l)\mathbf{H}^{(l)} + \alpha\mathbf{H}^{(0)}$

- For w_{ij} in $\mathbf{W}^{(\mathbf{l})}$ and s_{ij} in $\mathbf{S}^{(\mathbf{l})} = \sigma(\mathbf{H}^{(\mathbf{l})}\mathbf{G}^{(\mathbf{l})})$, we have.

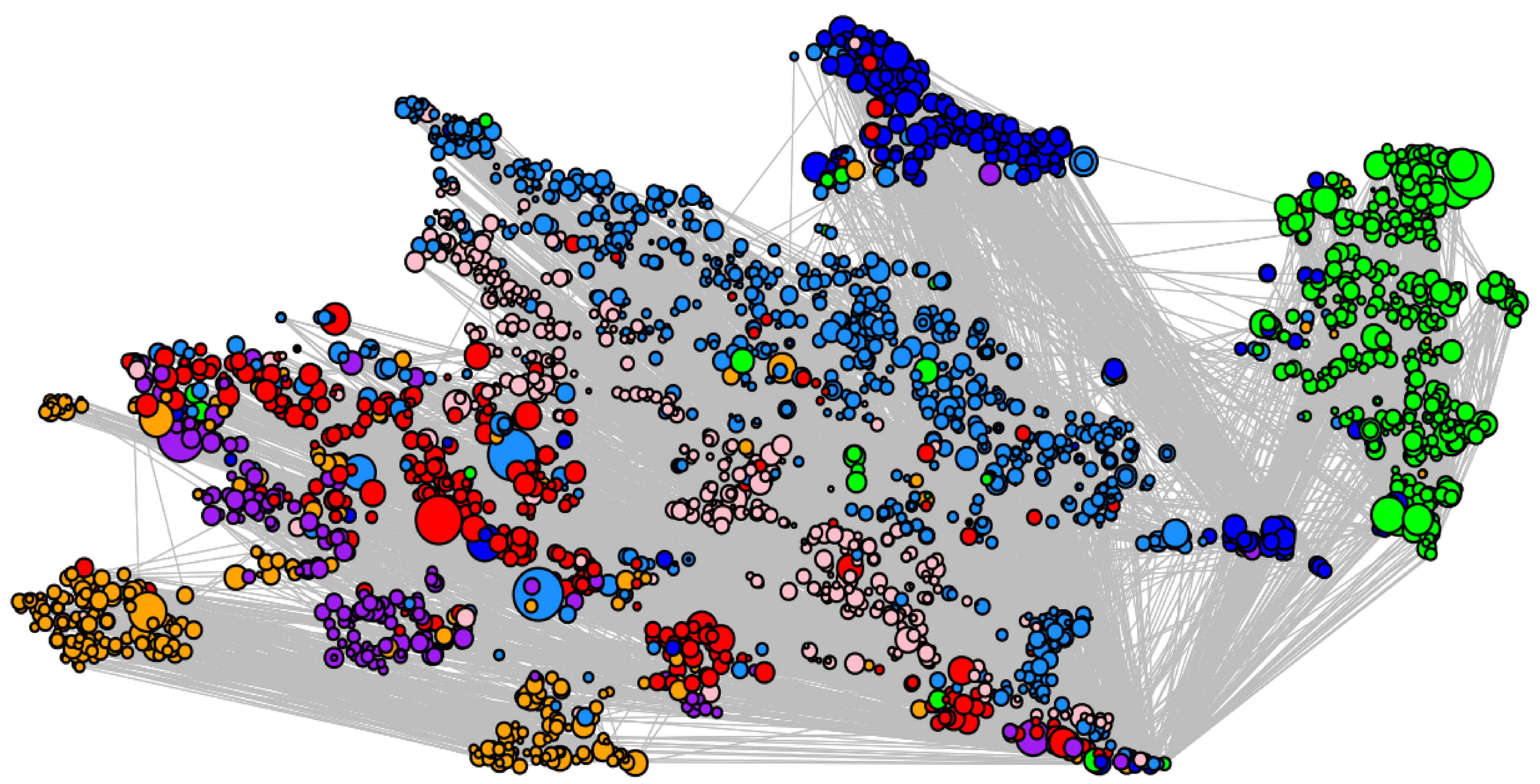
1. Node feature importance $(\mathcal{I}_{n,m})$ **:** The importance of the m^{th} feature of the n^{th} node can be obtained as follows

$$\mathcal{I}_{n,m} = \sum_{j=1}^{F} w_{mj} s_{nj}$$

2. Node importance $(\mathcal{I}_n)$ **:** The importance of an n^{th} node can be obtained as follows

$$\mathcal{I}_n = \sum_{j=1}^{F} \left\|\boldsymbol{w}_j\right\|_2 s_{nj}$$

Here, $\boldsymbol{w}_j = [w_{1j}, w_{2j}, \ldots, w_{dj}]$



Results

Test accuracy and standard deviation over 50 random initializations.

Method	Cora	Citeseer	PubMed	Coauthor CS	Coauthor Physics	Amazon Photos	Amazon Computers
SGC	80.7 ± 0.8	72.8 ± 0.6	77.3 ± 1.1	78.72 ± 0.7	76.28 ± 1.2	75.14 ± 0.5	79.85 ± 0.7
GCN	81.1 ± 0.7 (2)	70.9 ± 0.9 (2)	78.1 ± 0.5 (2)	81.87 ± 0.8 (2)	83.99 ± 0.7 (2)	81.63 ± 0.7 (2)	84.67 ± 0.8 (2)
GCN-DE	82.3 ± 0.6 (2)	71.1 ± 0.5 (2)	78.4 ± 0.3 (2)	82.28 ± 0.6 (2)	85.08 ± 0.6 (2)	82.07 ± 0.8 (2)	84.91 ± 0.6 (2)
GCN-FG	82.8 ± 0.6 (2)	73.1 ± 0.6 (2)	79.1 ± 0.5 (2)	83.05 ± 0.6 (2)	85.56 ± 0.5 (2)	83.86 ± 0.5 (2)	85.29 ± 0.5 (2)
GAT	81.9 ± 0.5 (2)	70.6 ± 0.6 (2)	77.4 ± 0.9 (2)	82.01 ± 0.5 (2)	84.57 ± 1.1 (2)	82.53 ± 0.9 (2)	85.09 ± 0.7 (2)
GAT-DE	82.1 ± 0.6 (2)	70.8 ± 0.4 (2)	77.5 ± 0.8 (2)	82.84 ± 0.7 (2)	85.22 ± 0.6 (2)	83.09 ± 0.7 (2)	85.57 ± 0.5 (2)
GAT-FG	82.0 ± 0.5 (2)	70.9 ± 0.7 (2)	77.8 ± 0.9 (2)	82.56 ± 0.4 (2)	84.98 ± 0.5 (2)	82.84 ± 0.6 (2)	84.11 ± 0.6 (2)
JKNet	81.1 ± 0.8 (4)	69.8 ± 0.6 (16)	78.1 ± 0.5 (16)	84.54 ± 0.7 (8)	85.52 ± 0.5 (16)	81.67 ± 0.6 (16)	82.14 ± 0.6 (16)
APPNP	83.3 ± 0.7	71.8 ± 0.8	79.8 ± 0.7	85.37 ± 0.5	85.98 ± 0.6	82.62 ± 0.4	83.81 ± 0.8
GRAND*	84.3 ± 0.6 (8)	72.8 ± 0.5 (8)	78.5 ± 0.6 (4)	88.31 ± 0.5 (8)	87.89 ± 0.6 (8)	85.31 ± 0.5 (8)	86.66 ± 0.7 (4)
GCNII	84.8 ± 0.5 (32)	72.9 ± 0.3 (64)	79.8 ± 0.3 (16)	88.83 ± 0.8 (16)	88.51 ± 0.7 (32)	88.27 ± 0.6 (32)	87.25 ± 0.6 (32)
GFGN*	84.9 ± 0.6 (4)	73.4 ± 0.4 (4)	80.4 ± 0.4 (8)	89.03 ± 0.6 (4)	89.45 ± 0.3 (4)	89.13 ± 0.8 (8)	87.92 ± 0.5 (4)
GCN-IR	85.3 ± 0.7 (32)	73.2 ± 0.6 (32)	$\mathbf{80.1} \pm \mathbf{0.5}$ (32)	89.98 ± 0.7 (32)	89.75 ± 0.6 (32)	89.16 ± 0.7 (64)	$\mathbf{88.91} \pm \mathbf{0.6}$ (32)
GCN-IR-FG	$\mathbf{85.7} \pm \mathbf{0.5}$ (32)	$\mathbf{73.6} \pm \mathbf{0.4}$ (32)	80.0 ± 0.6 (64)	$\mathbf{90.79} \pm \mathbf{0.6}$ (64)	$\mathbf{90.21} \pm \mathbf{0.4}$ (32)	$\mathbf{89.94} \pm \mathbf{0.4}$ (64)	88.49 ± 0.5 (32)

Performance analysis with increasing number of layers

Dataset	Method	Layers 2	4	8	16	32	64
Cora	GCN	**81.1**	80.4	69.5	64.9	60.3	28.7
	GCN-DE	**82.3**	82.0	75.8	75.7	62.5	49.5
	GCN-FG	**82.8**	78.6	69.9	72.7	59.3	47.4
	GAT	**81.9**	79.8	69.5	47.8	45.1	25.3
	GAT-DE	**82.1**	80.8	72.9	66.3	51.3	43.4
	GAT-FG	**82.0**	68.4	51.3	59.7	45.1	29.2
	GRAND	72.8	79.6	**84.3**	80.1	77.2	73.4
	GCNII	82.1	82.4	83.6	83.9	**84.8**	84.6
	GFGN	83.7	**84.9**	82.3	79.6	65.2	61.9
	GCN-IR	82.4	83.9	84.6	84.8	**85.3**	85.2
	GCN-IR-FG	83.2	82.9	85.1	85.4	**85.7**	85.5
PubMed	GCN	**78.1**	75.5	62.2	41.7	21.4	34.2
	GCN-DE	**78.4**	77.9	78.1	78.2	77.0	61.5
	GCN-FG	**79.1**	74.7	65.1	54.5	41.7	35.1
	GAT	**77.4**	69.4	48.3	39.5	41.3	40.7
	GAT-DE	**77.5**	76.8	76.4	67.1	54.7	68.2
	GAT-FG	**77.8**	68.7	51.8	48.2	49.1	41.6
	GRAND	72.4	74.7	**78.5**	76.2	71.9	68.6
	GCNII	74.9	76.2	77.5	**79.8**	78.1	74.2
	GFGN	**80.4**	79.7	77.6	72.4	67.6	62.8
	GCN-IR	78.1	78.3	79.1	79.6	**80.1**	79.5
	GCN-IR-FG	77.8	78.1	79.7	79.9	**80.0**	**80.0**

Key References

- T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," International Conference on Learning Representations (ICLR), 2017.
- M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li, "Simple and deep graph convolutional networks," in International Conference on Machine Learning. PMLR, 2020, pp. 1725–1735.
- K. Oono and T. Suzuki, "Graph neural networks exponentially lose expressive power for node classification," International Conference on Learning Representations (ICLR), 2020..
- J. Klicpera, S. Weißenberger, and S. Gunnemann, "Diffusion improves graph learning," Advances in Neural Information Processing Systems (NeurIPS), 2019.
- M. Hardt and T. Ma, "Identity matters in deep learning", International Conference on Learning Representations (ICLR), 2017.