ICML Workshop on Graph Representation Learning and Beyond

Deep Graph Contrastive Representation Learning

Presented by Yanqiao ZHU

yanqiao.zhu@cripac.ia.ac.cn

Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences

Joint work with Yichen XU, Feng YU, Qiang LIU, Shu WU, and Liang WANG

Outline

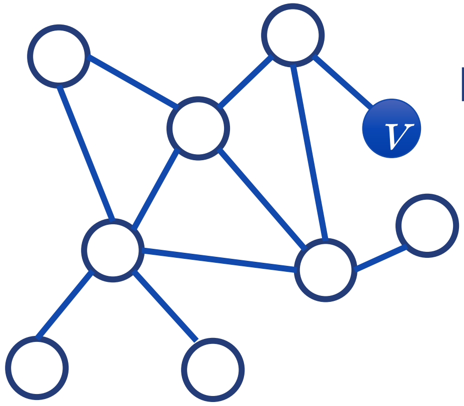
1. Preamble
2. The Proposed Method
3. Experiments
4. Concluding Remarks

Outline

1. **Preamble**
2. The Proposed Method
3. Experiments
4. Concluding Remarks

Representation Learning on Graphs

- Goal: efficient feature learning for machine learning on graphs



node 2 vec

$$f : \mathbb{R}^F \times \{0,1\}^N \to \mathbb{R}^{F'}$$

$\mathbb{R}^{F'}$ $\quad F' \ll F$

Node embeddings

- In reality, labels are not always available to models, which calls for training GNN in a self-supervised manner.

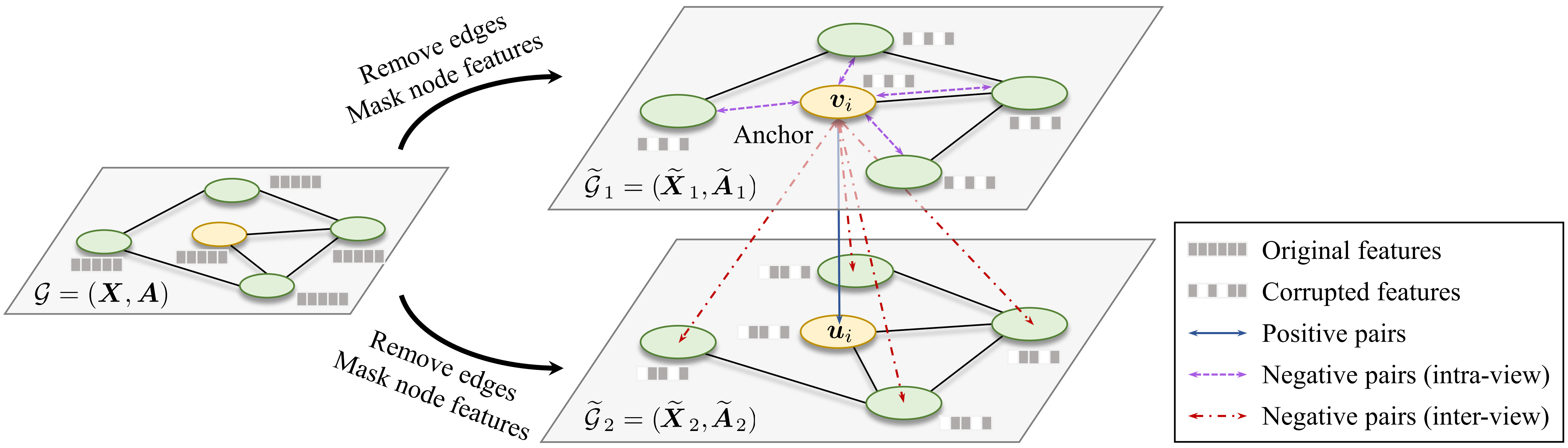
Contrastive Learning for GRL

- Node embedding approaches
  - Pioneering work of node embedding follows a contrastive framework originated in the skip-gram model.
  - For example, node2vec first samples short random walks and then enforces neighboring nodes on the same walk to share similar embeddings by contrasting them with other nodes.
- GNN-based approaches
  - GraphSAGE connects reconstruction objectives to GNN models, which excessively relies on the preset graph proximity matrix.
  - DGI firstly revitalizes InfoMax principle in the graph domain, which maximizes mutual information between node representations and global summary vectors.

Outline

1. Preamble
2. **The Proposed Method**
3. Experiments
4. Concluding Remarks

Model Illustration



Contrastive Learning Across Views

- We first generate two correlated graph views by randomly performing corruption.
- Then, we train the model using a contrastive loss to maximize the agreement between node embeddings in these two views.
  - Rather than contrasting node-level embeddings to global ones, we primarily focus on contrasting embeddings at the node level.

$$\ell(\boldsymbol{u}_i, \boldsymbol{v}_i) = \log \frac{e^{\theta(\boldsymbol{u}_i, \boldsymbol{v}_i)/\tau}}{\underbrace{e^{\theta(\boldsymbol{u}_i, \boldsymbol{v}_i)/\tau}}_{\text{the positive pair}} + \underbrace{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} e^{\theta(\boldsymbol{u}_i, \boldsymbol{v}_k)/\tau}}_{\text{inter-view negative pairs}} + \underbrace{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} e^{\theta(\boldsymbol{u}_i, \boldsymbol{u}_k)/\tau}}_{\text{intra-view negative pairs}}}$$

Hybrid Graph View Generation

- Appropriately choosing negative samples is important for InfoMax-based methods.
- We corrupt the original graph at both structure and attribute levels to construct diverse node contexts.
- **Removing edges (RE)**: randomly remove a portion of edges in the original graph.

$$\widetilde{\boldsymbol{A}} = \boldsymbol{A} \circ \widetilde{\boldsymbol{R}}$$

- **Masking node features (MF)**: randomly mask a fraction of dimensions with zeros in node features.

$$\widetilde{\boldsymbol{X}} = [\boldsymbol{x}_1 \circ \widetilde{\boldsymbol{m}}; \boldsymbol{x}_2 \circ \widetilde{\boldsymbol{m}}; \cdots; \boldsymbol{x}_N \circ \widetilde{\boldsymbol{m}}]^\top$$

Outline

1. Preamble
2. The Proposed Method
3. **Experiments**
4. Concluding Remarks

Experiment Setup

- Datasets

Dataset	Type	#Nodes	#Edges	#Features	#Classes
Cora	Transductive	2,708	5,429	1,433	7
Citeseer	Transductive	3,327	4,732	3,703	6
Pubmed	Transductive	19,717	44,338	500	3
DBLP	Transductive	17,716	105,734	1,639	4
Reddit	Inductive	231,443	11,606,919	602	41
PPI	Inductive	56,944 (24 graphs)	818,716	50	121 (multilabel)

Experiment Setup (cont.)

- Baselines:
  - Traditional methods DeepWalk and node2vec
  - GNN-based methods GAE, VGAE, GraphSAGE, and DGI
  - Representative semi-supervised methods
    - Transductive: GCN and SGC
    - Inductive: FastGCN and GaAN-mean

Transductive Node Classification

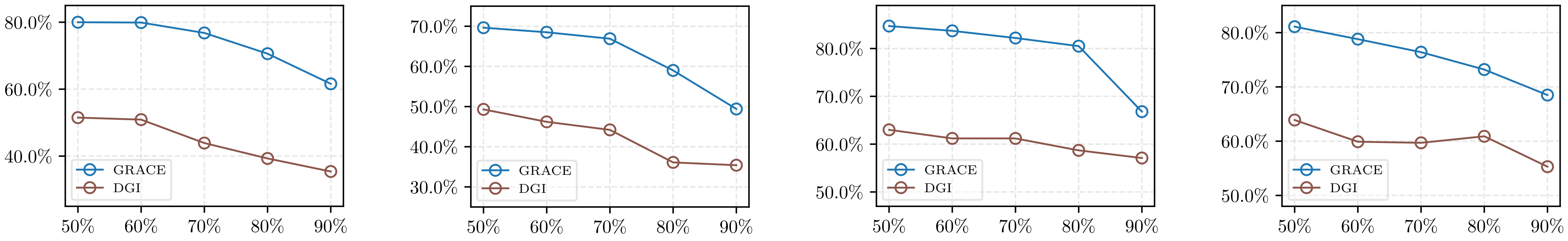
Method	Training Data	Cora	Citeseer	Pubmed	DBLP
Raw features	$\boldsymbol{X}$	64.8	64.6	84.8	71.6
node2vec	$\boldsymbol{A}$	74.8	52.3	80.3	78.8
DeepWalk	$\boldsymbol{A}$	75.7	50.5	80.5	75.9
DeepWalk + features	$\boldsymbol{X}, \boldsymbol{A}$	73.1	47.6	83.7	78.1
GAE	$\boldsymbol{X}, \boldsymbol{A}$	76.9	60.6	82.9	81.2
VGAE	$\boldsymbol{X}, \boldsymbol{A}$	78.9	61.2	83.0	81.7
DGI	$\boldsymbol{X}, \boldsymbol{A}$	82.6±0.4	68.8±0.7	86.0±0.1	83.2±0.1
GRACE	$\boldsymbol{X}, \boldsymbol{A}$	**83.3±0.4**	**72.1±0.5**	**86.7±0.1**	**84.2±0.1**
SGC	$\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}$	80.6	69.1	84.8	81.7
GCN	$\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}$	82.8	72.0	84.9	82.7

Inductive Node Classification

Method	Training Data	Reddit	PPI
Raw features	$\boldsymbol{X}$	58.5	42.2
DeepWalk	$\boldsymbol{A}$	32.4	—
DeepWalk + features	$\boldsymbol{X}, \boldsymbol{A}$	69.1	—
GraphSAGE-GCN	$\boldsymbol{X}, \boldsymbol{A}$	90.8	46.5
GraphSAGE-mean	$\boldsymbol{X}, \boldsymbol{A}$	89.7	48.6
GraphSAGE-LSTM	$\boldsymbol{X}, \boldsymbol{A}$	90.7	48.2
GraphSAGE-pool	$\boldsymbol{X}, \boldsymbol{A}$	89.2	50.2
DGI	$\boldsymbol{X}, \boldsymbol{A}$	94.0±0.1	63.8±0.2
GRACE	$\boldsymbol{X}, \boldsymbol{A}$	**94.2±0.0**	**66.1±0.1**
FastGCN	$\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}$	93.7	—
GaAN-mean	$\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{Y}$	95.8±0.1	96.9±0.2

Robustness to Sparse Features

- Experiments with randomly contaminating the training data by masking a certain portion of the node features to zeros.
  - We vary the contamination rate of node features from 0.5 to 0.9 on four citation networks.



(a) Cora (b) Citeseer (c) Pubmed (d) DBLP

Outline

1. Preamble
2. The Proposed Method
3. Experiments
4. **Concluding Remarks**

Wrapping Up

1. We have developed a novel graph contrastive representation learning framework based on maximizing the agreement at the node level.
2. GRACE learns representations by first generating graph views using a hybrid scheme, removing edges and masking node features, and then applying a contrastive loss to maximize the agreement of node embeddings in these two views.
3. Experimental results demonstrate that GRACE can outperform existing state-of-the-art methods by large margins and even surpass supervised counterparts on transductive tasks.

Acknowledgements

- We are grateful to Mr. Tao Sun and Mr. Sirui Lu for their insightful comments.
- This work is jointly supported by
  - National Key Research and Development Program under grants No. 2018YFB1402600 and No. 2016YFB1001000
  - National Natural Science Foundation of China under grants No. U19B2038 and No. 61772528


THANKS