HC-GAE: The Hierarchical Cluster-based Graph Auto-Encoder for Graph Representation Learning

Lu Bai, Zhuo Xu, Lixin Cui*, Ming Li, Yue Wang, Edwin R. Hancock

School of Artificial Intelligence, Beijing Normal University, Beijing, China

School of Information, Central University of Finance and Economics, Beijing, China

Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, China

Department of Computer Science, University of York, York, UK

Background

- In recent years, graph structure data has been widely used for characterizing pairwise relationships among the components of complex systems. GNNs and their related approaches (e.g. GAEs) have been proposed for graph-based tasks.

Challenges

- **The limitation for multiple downstream tasks.**

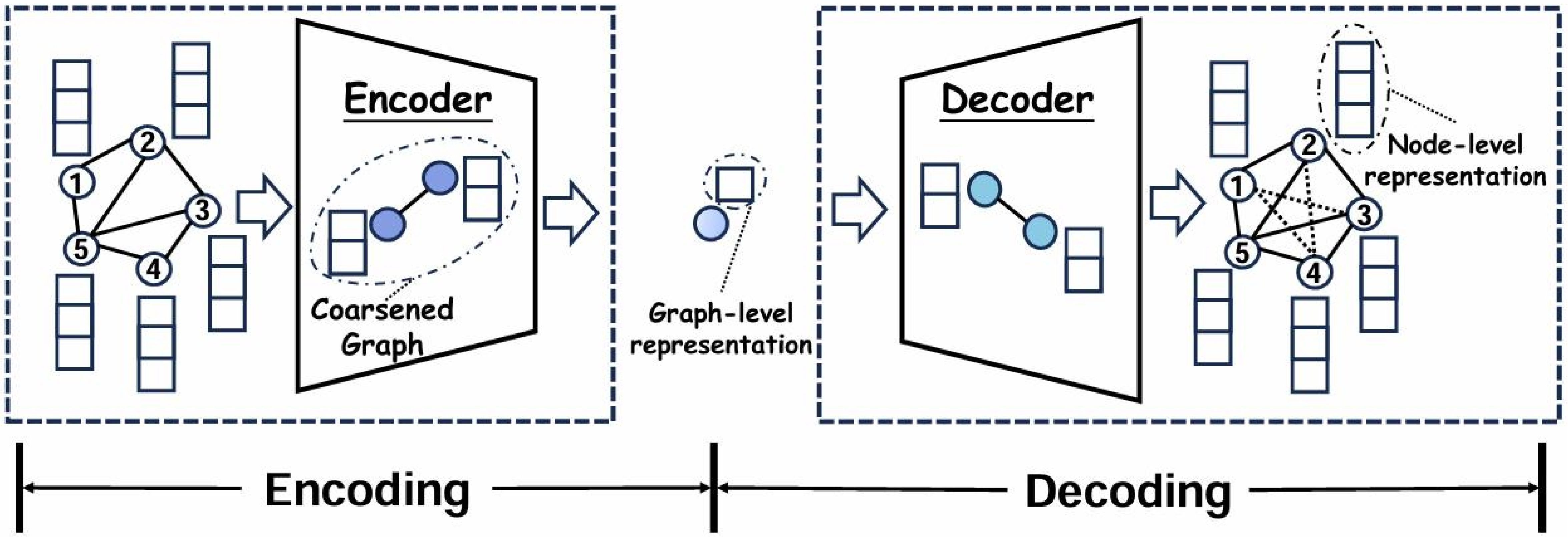
- **The over-smoothing problem.**

NEURAL INFORMATION PROCESSING SYSTEMS

Contributions

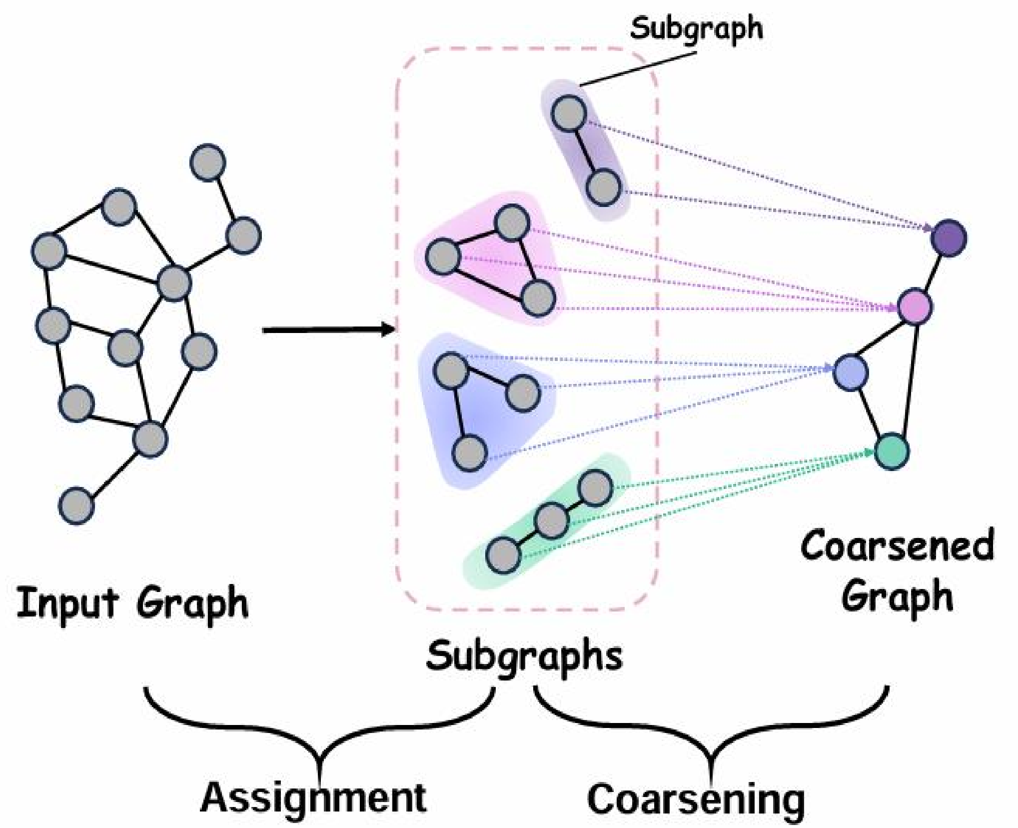
- We propose a novel **Hierarchical Cluster-based GAE (HC-GAE)** for **graph representation learning**.
- We propose a new **loss function** for training the proposed HC-GAE model.
- We evaluate the performance of the proposed HC-GAE model on both node and graph classification tasks, demonstrating the effectiveness of the proposed model.

Framework

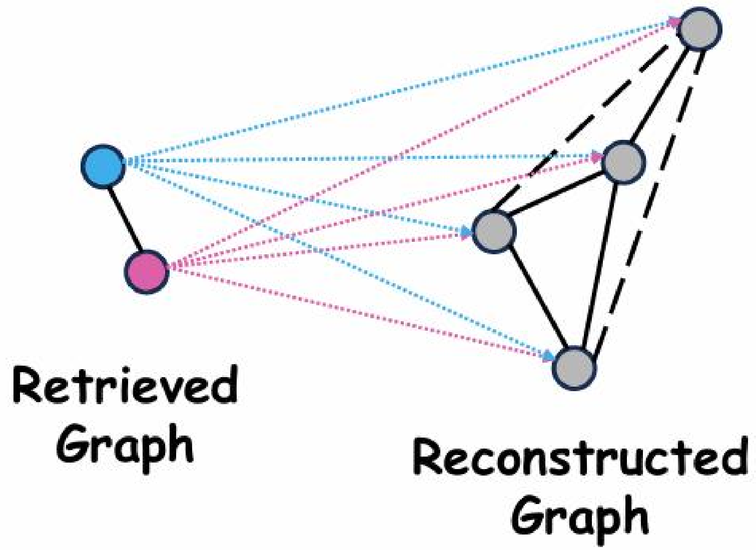


The architecture of our proposed HC-GAE model.

Encoder & Decoder



our proposed layer in the GNN encoder.



our proposed layer in the GNN decoder.

NEURAL INFORMATION PROCESSING SYSTEMS

Loss Function

$$\mathcal{L} = \mathrm{KL}[q(Z \mid V, E) \| p(Z)] - \mathbb{E}_{q(Z|V,E)}[\log p(E' \mid Z)],$$

Update

$$\mathcal{L}_{\text{local}} = \sum_{l=1}^{L} \sum_{j=1}^{n_{(l+1)}} \mathrm{KL}[q(Z_j^{(l)} \mid X_j^{(l)}, A_j^{(l)}) \| p(Z^{(l)})],$$

$$\mathcal{L}_{\text{global}} = -\sum_{l=1}^{L} \mathbb{E}_{q(X^{(L)}, A^{(L)}) | X^{(l)}, A^{(l)})}[\log p(X'^{(L-l+2)}, A'^{(L-l+2)} \mid X^{(L)}, A^{(L)})],$$

$$\mathcal{L}_{\text{HC-GAE}} = \mathcal{L}_{\text{local}} + \mathcal{L}_{\text{global}},$$

NEURAL INFORMATION PROCESSING SYSTEMS

Experiments

Table 3: Node classification performance based on accuracy. A.R. is the average rank.

Datasets	Cora	CiteSeer	PubMed	Computers	CS	A.R.
DGI	85.41±0.34	74.51±0.51	85.95±0.66	84.68±0.39	91.33±0.30	4.0
VGAE	83.60±0.52	63.37±1.21	78.23±1.63	87.21±0.26	89.79±0.09	5.2
SSL-GCN	57.29±0.13	59.57±1.77	75.06±0.37	80.49±0.10	84.71±0.95	6.8
GraphSage	74.30±1.84	60.20±2.15	81.96±0.74	87.05±0.25	89.74±0.19	5.6
GraphMAE	85.45±0.40	72.48±0.77	85.74±0.14	88.04±0.61	**93.47±0.04**	3.0
S2GAE	86.15±0.25	74.60±0.06	86.91±0.28	90.94±0.08	91.70±0.08	2.2
HC-GAE	**87.97±0.10**	**75.29±0.09**	**87.56±0.35**	**91.07±0.14**	92.28±0.07	**1.2**

Table 4: Graph classification performance based on accuracy. A.R. is the average rank.

Datasets	IMDB-B	IMDB-M	PROTEINS	COLLAB	MUTAG	A. R.
WLSK	64.48±0.90	43.38±0.75	71.70±0.67	N/A	80.72±3.00	7.75
DGCNN	67.45±0.83	46.33±0.73	73.21±0.34	N/A	85.83±1.66	6.25
DiffPool	72.6±3.9	47.2±1.8	75.1±3.5	78.9±2.3	85.0±10.3	5.20
Graph2Vec	71.10±0.54	50.44±0.87	73.30±2.05	N/A	83.15±9.25	5.75
InfoGCL	75.10±0.90	51.40±0.80	N/A	80.00±1.30	91.20±1.30	3.50
GraphMAE	75.52±0.66	51.63±0.52	75.30±0.39	80.32±0.46	88.19±1.26	3.20
S2GAE	75.76±0.62	51.79±0.36	76.37±0.43	**81.02±0.53**	88.26±0.76	2.00
HC-GAE	**76.72±0.60**	**51.90±1.47**	**78.13±1.37**	80.41±0.02	**92.38±1.17**	**1.20**

Table 1: Datasets for node classification

Datasets	Cora	CiteSeer	PubMed	Computers	CS
Nodes	2708	3312	19717	13752	18333
Edges	5429	4660	44338	245861	81894
Features	1433	3703	500	767	6805
Classes	7	6	3	10	15

Table 2: Datasets for graph classification

Datasets	IMDB-B	IMDB-M	PROTEINS	COLLAB	MUTAG
Graphs	1000	1500	1113	5000	188
Nodes(mean)	19.77	13	39.06	74.49	17.93
Edges(mean)	96.53	65.94	72.82	2457.78	19.79
Classes	2	3	2	3	1

NEURAL INFORMATION PROCESSING SYSTEMS

THANKS!

E-mail: cuilixin@cufe.edu.cn

OpenReview Link: