WWW-24 Research Track

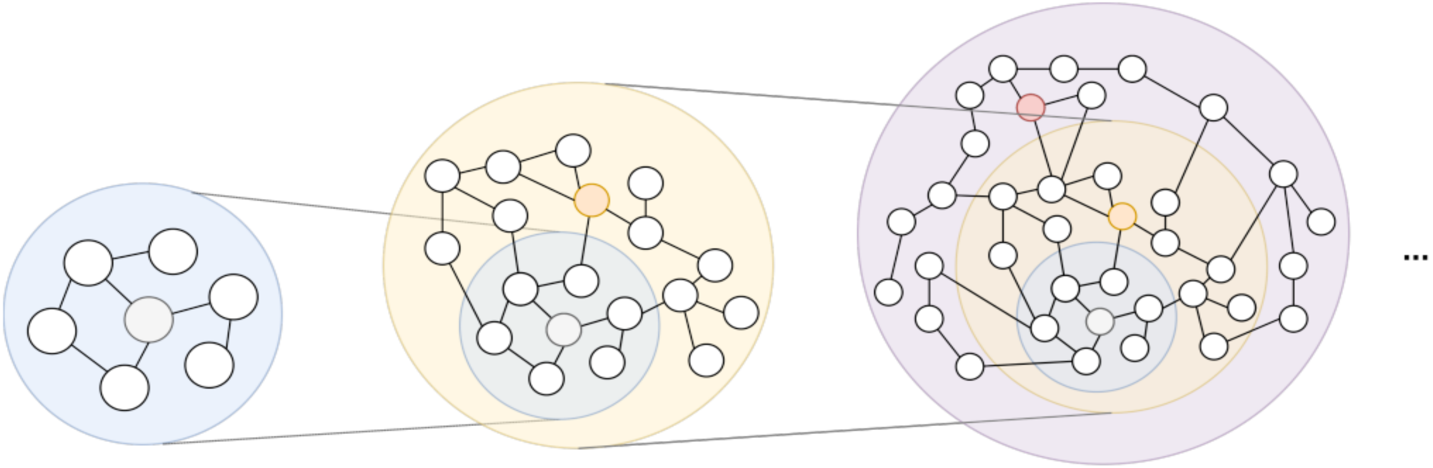
DSLR: Diversity Enhancement and Structure Learning for Rehearsal-based Graph Continual Learning

Seungyoon Choi*, Wonjoong Kim*, Sungwon Kim,
Yeonjun In, Sein Kim, Chanyoung Park

Korean Advanced Institute of Science and Technology (KAIST)

DSAIL Data Science & Artificial Intelligence

Introduction Continual Learning



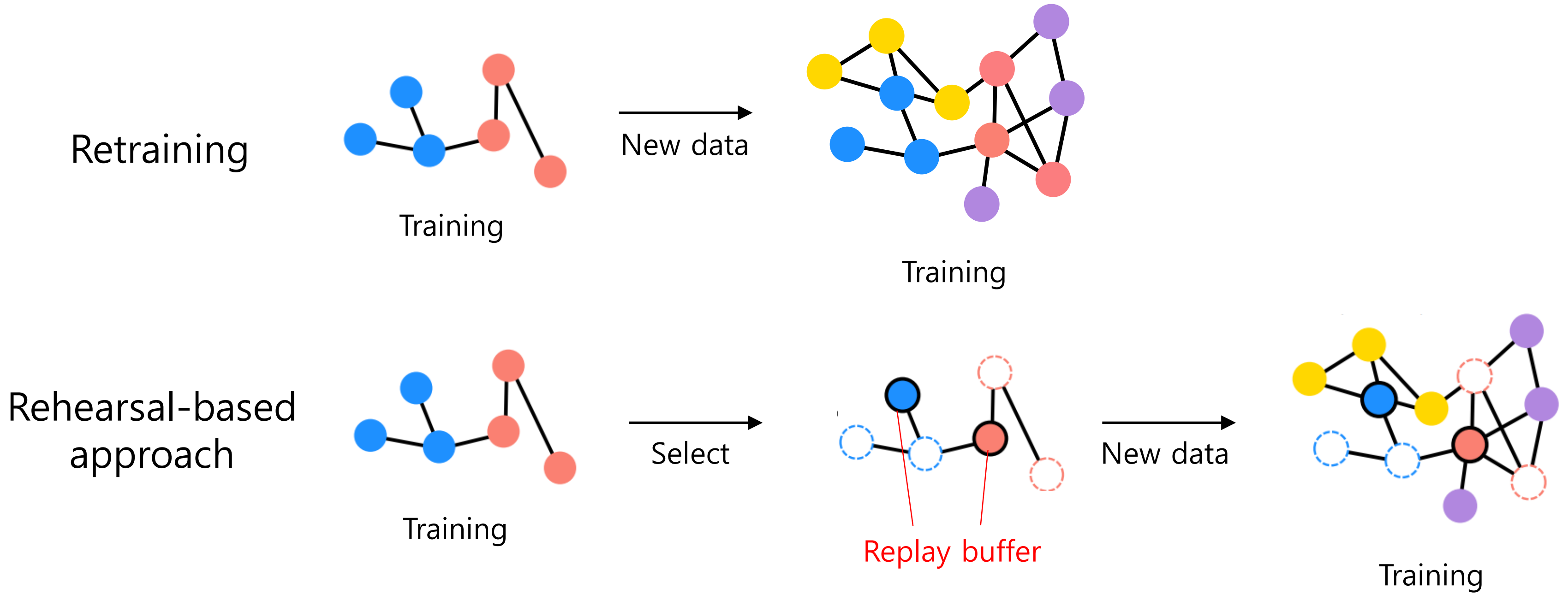
- Efficient learning from **newly introduced data** without retraining the model on the entire dataset, enabling the **preservation of previously acquired knowledge**.

- **Challenge ➔** avoid catastrophic forgetting!

Febrinanto, Falih Gozi, et al. "Graph lifelong learning: A survey." *IEEE Computational Intelligence Magazine* 18.1 (2023): 32-51.

Introduction Continual Learning

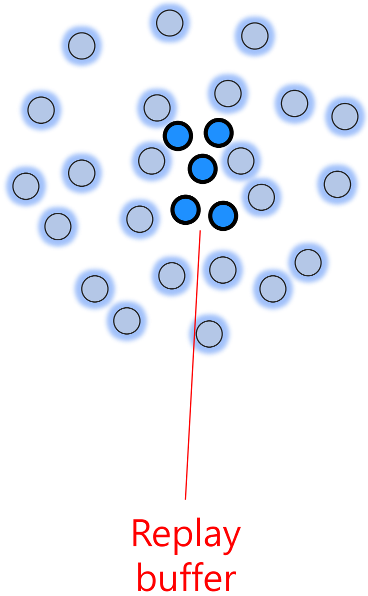
- Continual learning approaches
  - Regularization-based approach: Regularize important parameters to be not changed.
  - Architectural approach: Modify the model's architecture based on the task.
  - Rehearsal-based approach: Store and use important data that effectively represents the entire class from past tasks.



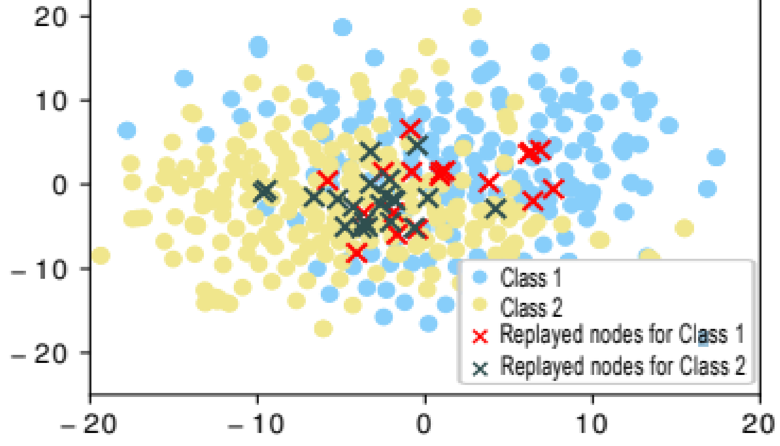
Motivation

- Existing method can cause overfitting to replay buffer

Mean Feature (MF)
Select nodes nearest the center of feature space



※ Dataset: Citeseer



(a) Mean Feature (MF)

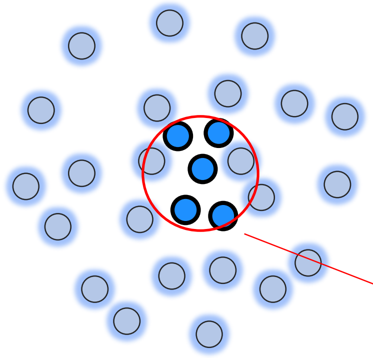
Embeddings of nodes & replayed nodes selected using MF

Zhou, Fan, and Chengtai Cao. "Overcoming catastrophic forgetting in graph neural networks with experience replay." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 5. 2021.

Motivation

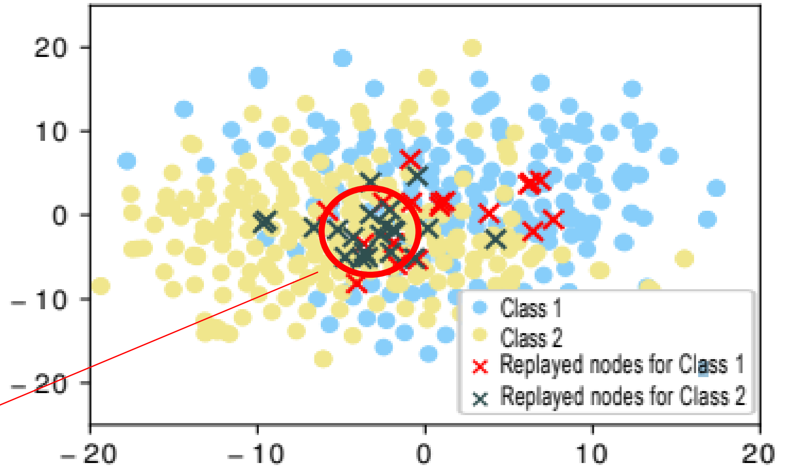
- Existing method can cause overfitting to replay buffer

Mean Feature (MF)
Select nodes nearest the center of feature space



Risk of overfitting

※ Dataset: Citeseer

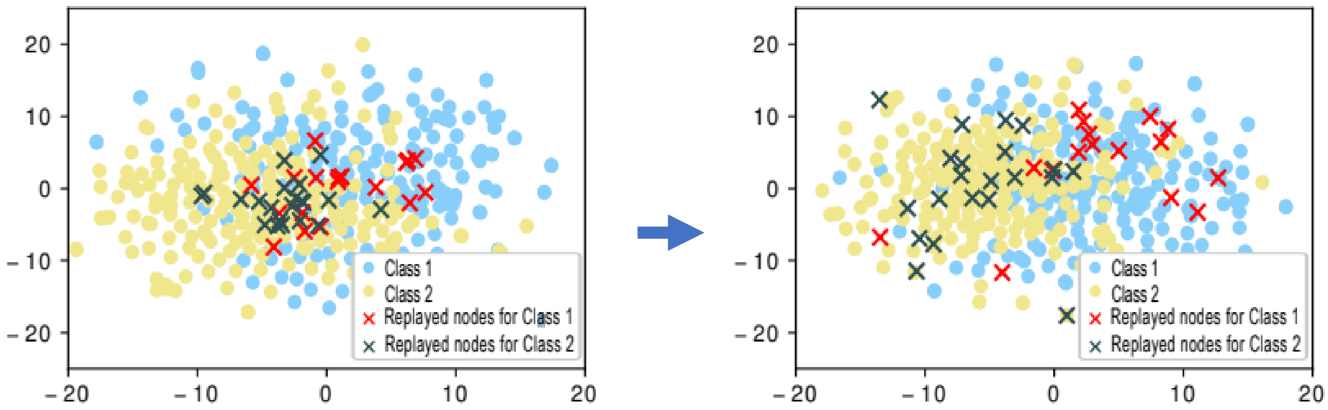


(a) Mean Feature (MF)

Zhou, Fan, and Chengtai Cao. "Overcoming catastrophic forgetting in graph neural networks with experience replay." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 5. 2021.

Motivation

※ Dataset: Citeseer



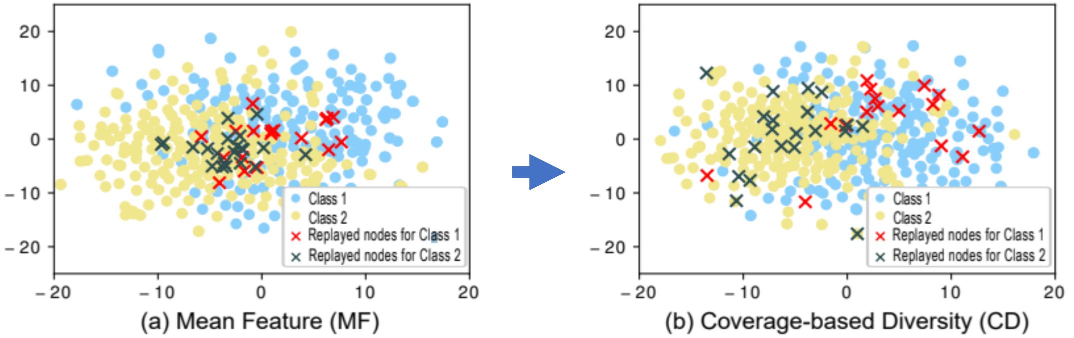
(a) Mean Feature (MF)

(b) Coverage-based Diversity (CD)

Coverage-based Diversity (CD) : Considering both **Representativeness & Diversity**

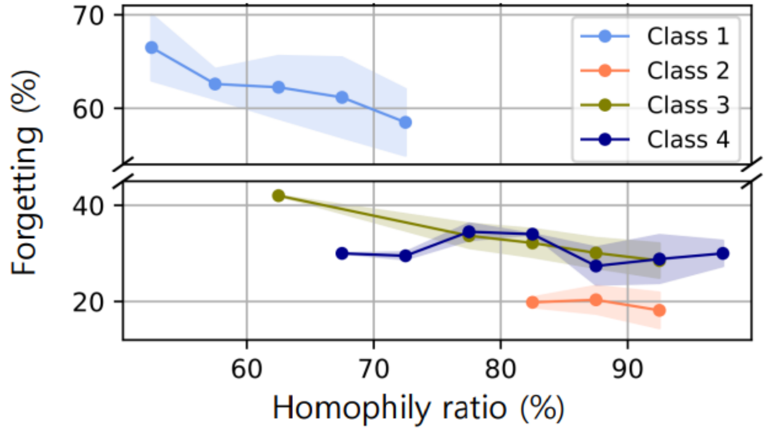
Motivation

- Using CD can lead to another issue → **Homophily Ratio**



	MF	CD
Class 1	0.68 ± 0.43	0.57 ± 0.45
Class 2	0.91 ± 0.24	0.92 ± 0.22
Class 3	0.82 ± 0.28	0.76 ± 0.40
Class 4	0.88 ± 0.26	0.82 ± 0.36

Homophily ratio of replayed nodes using MF & CD

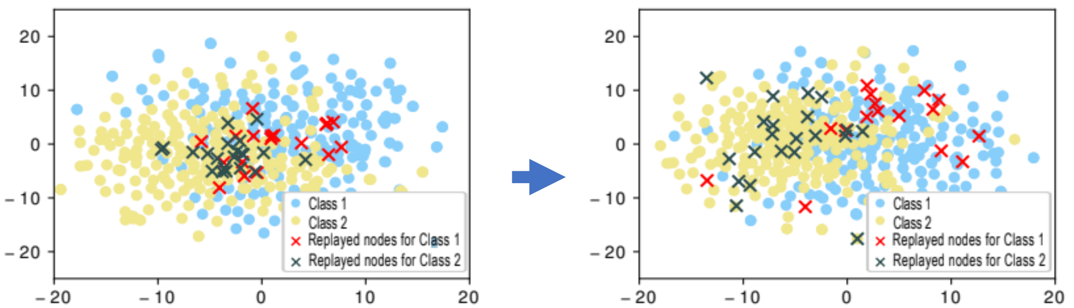


Forgetting over various homophily ratio of the replayed nodes

Can we **just enhance the homophily ratio** of replay nodes?

Motivation

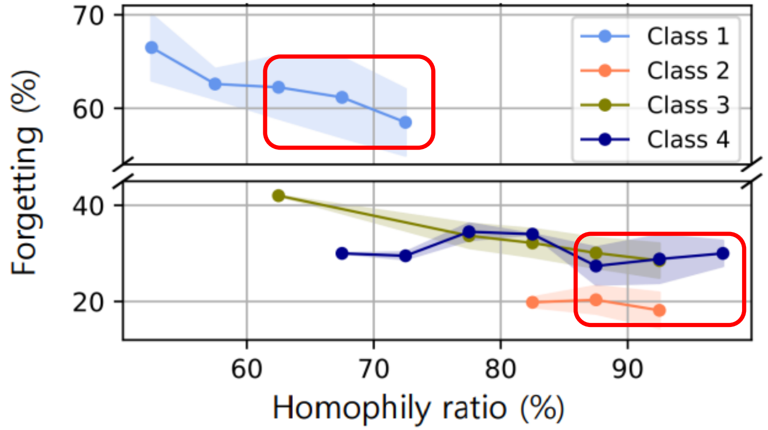
- Using CD can lead to another issue → **Homophily Ratio**



(a) Mean Feature (MF) (b) Coverage-based Diversity (CD)

	MF	CD
Class 1	0.68 ± 0.43	0.57 ± 0.45
Class 2	0.91 ± 0.24	0.92 ± 0.22
Class 3	0.82 ± 0.28	0.76 ± 0.40
Class 4	0.88 ± 0.26	0.82 ± 0.36

Homophily ratio of replayed nodes
using MF & CD



Simply increasing the homophily ratio is not effective!

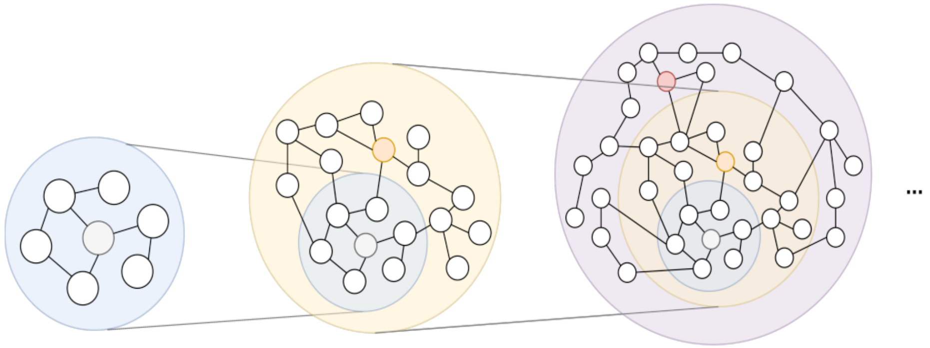
Forgetting over various homophily ratio
of the replayed nodes

Structure Learning for replay buffer!

→ Formulating the structure of replayed nodes to be connected **to truly informative neighbors**

Proposed Method Preliminaries

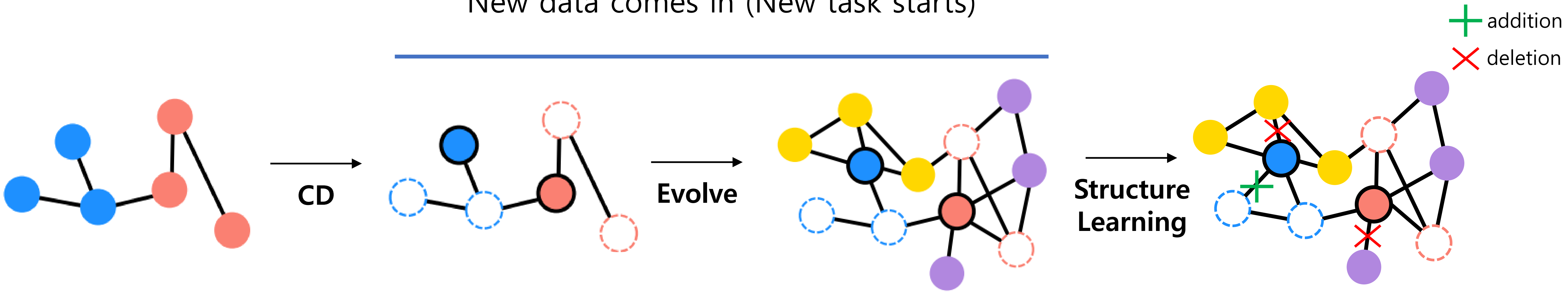
- Continual Learning Scenario
  - Sequential of tasks $\mathcal{T} = \{T_1, T_2, \cdots, T_M\}$
  - Graph at task *t* $\mathcal{G}^t = (A^t, X^t)$
  - Incremental graph $\mathcal{G} = \left\{\mathcal{G}^1, \mathcal{G}^2, ..., \mathcal{G}^M\right\},$ where $\mathcal{G}^t = \mathcal{G}^{t-1} + \Delta\mathcal{G}^t$
  - Goal $\text{GNN}_{\theta^1}, \text{GNN}_{\theta^2}, ..., \text{GNN}_{\theta^M}$



Febrinanto, Falih Gozi, et al. "Graph lifelong learning: A survey." *IEEE Computational Intelligence Magazine* 18.1 (2023): 32-51.

Proposed Method Simplified Framework

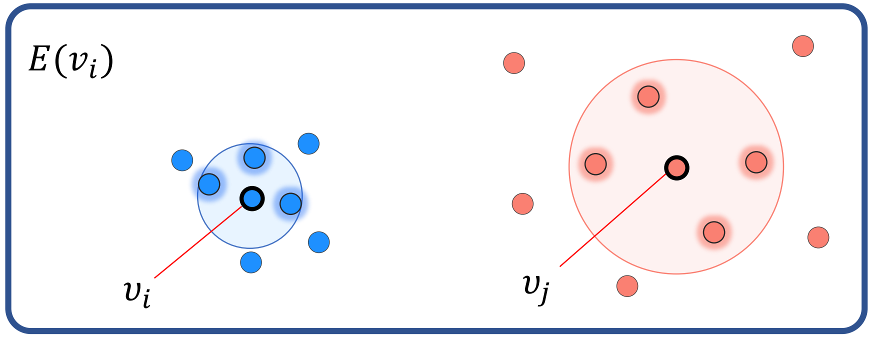
New data comes in (New task starts)



Replay buffer selection considering both representativeness & diversity

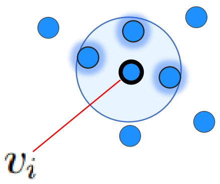
Reformulating the structure of replay buffer to be connected to truly informative neighbors

Proposed Method Coverage-based Diversity (CD)



- Cover of node v_i

$$\mathcal{C}(v_i) = \{v_j \mid dist(\underbrace{h_i}_{\text{Embedding of } v_i}, h_j) < d,\ \underbrace{y_i}_{\text{class of } v_i} = y_j\},\ \text{where } d = r \cdot \underbrace{E(v_i)}_{\text{Average of pairwise distance in same class}}$$



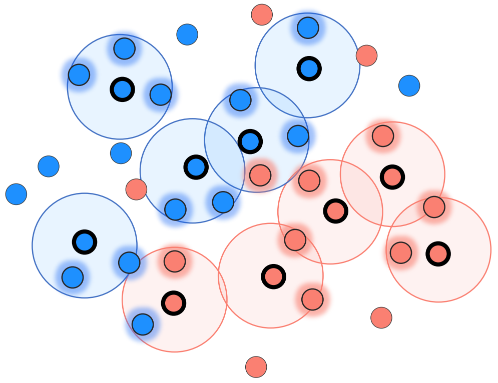
- Set of replayed nodes of class C_l

$$\mathcal{B}_{C_l} = \text{argmax}_{\{v_{b_1},\cdots,v_{b_{e_l}} \mid v_{b_1},\cdots,v_{b_{e_l}} \in train_{C_l}\}} \left| Cover(\{v_{b_1},\cdots,v_{b_{e_l}}\}) \right|$$

$$\text{where } \ Cover(\{v_1,\cdots,v_n\}) = \mathcal{C}(v_1) \cup \cdots \cup \mathcal{C}(v_n)$$

- Size of replay buffer assigned for class C_l

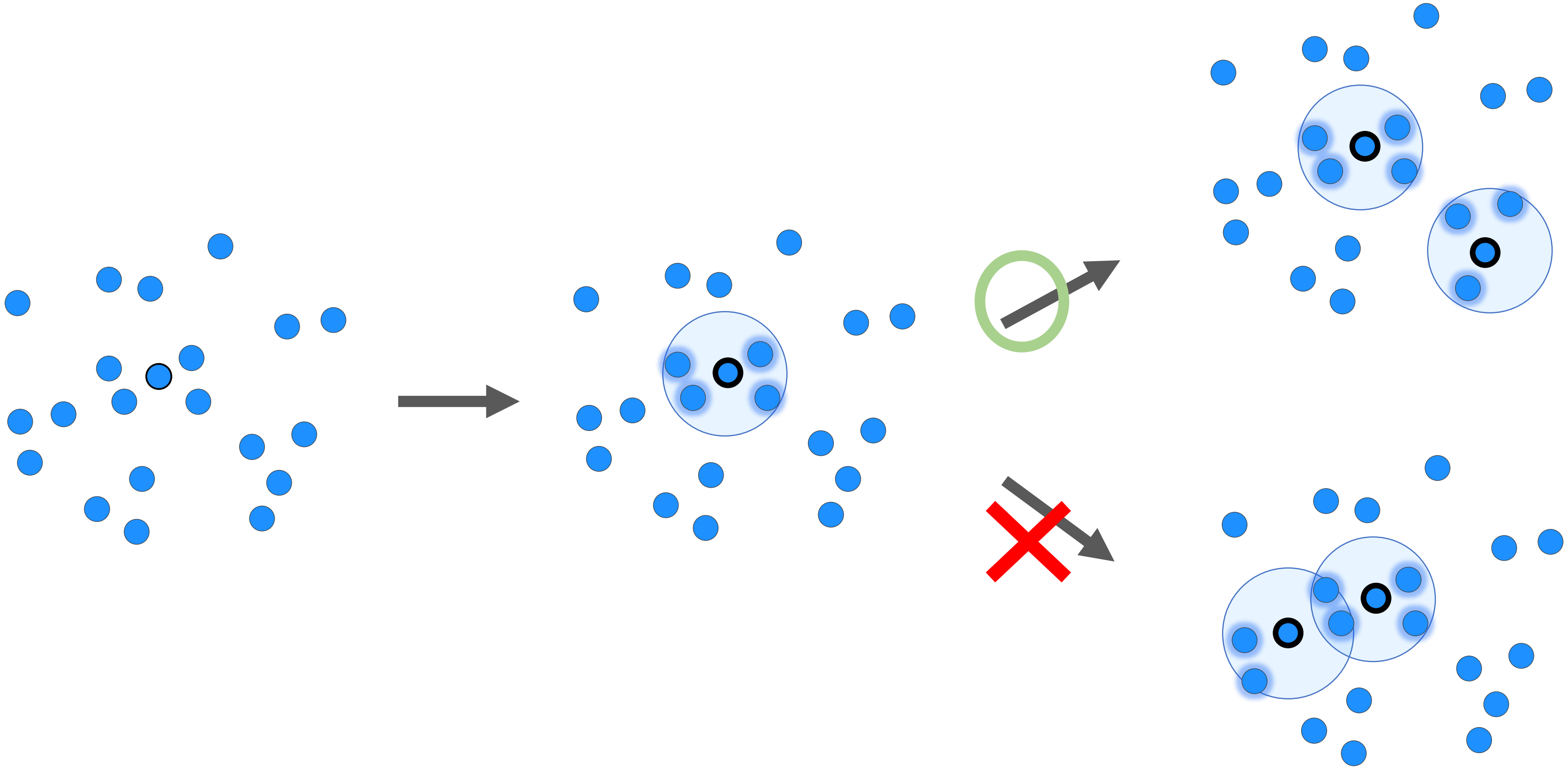
$$e_l = \frac{\text{\# of training nodes for class } C_l}{\text{\# of training nodes for all seen classes}} \text{ X Replay buffer size}$$



Maximize the number of nodes covered by Covers of replayed nodes

Proposed Method Coverage-based Diversity (CD)

$$\mathcal{B}_{C_l} = \underbrace{\text{argmax}}_{\text{Representativeness}}{}_{\{v_{b_1},\cdots,v_{b_{e_l}} \mid v_{b_1},\cdots,v_{b_{e_l}} \in train_{C_l}\}} \left| Cover(\{v_{b_1},\cdots,v_{b_{e_l}}\}) \right| \quad \text{where} \quad Cover(\{v_1,\cdots,v_n\}) = \mathcal{C}(v_1) \underbrace{\cup \cdots \cup}_{\text{Diversity}} \mathcal{C}(v_n)$$



CD (maximize the union of cover)

➔ 7 nodes are covered

Find the largest cover (not consider union)

➔ 6 nodes are covered

Proposed Method Structure Learning for Replay Buffer

- Training link prediction module LP_ϕ

- Link prediction loss

$$\mathcal{L}_{link} = -\Big(\sum_{e_{ij} \in \underbrace{\mathcal{D}_t^{link}}_{\text{Training link set at } T_t}} (A_{ij}^t \log(S_{ij}) + (1 - A_{ij}^t)\log(1 - \underbrace{S_{ij}}_{\text{Similarity based score}})\Big)$$

Capture **structural proximity**

- Node classification loss

$$\mathcal{L}_{node} = \beta \mathcal{L}_{\underbrace{\mathcal{D}_t{}^{tr}}_{\text{Training node set at } T_t}}(\theta^t; A^t, X^t) + (1 - \beta)\mathcal{L}_{\underbrace{\mathcal{B}}_{\text{Replay buffer}}}(\theta^t; A^t, X^t)$$

Capture **homophily ratio**

- Final loss function

$$\mathcal{L}_{LP} = \lambda \mathcal{L}_{link} + (1 - \lambda)\mathcal{L}_{node}$$

Discover **truly informative neighbors**

Proposed Method Structure Learning for Replay Buffer

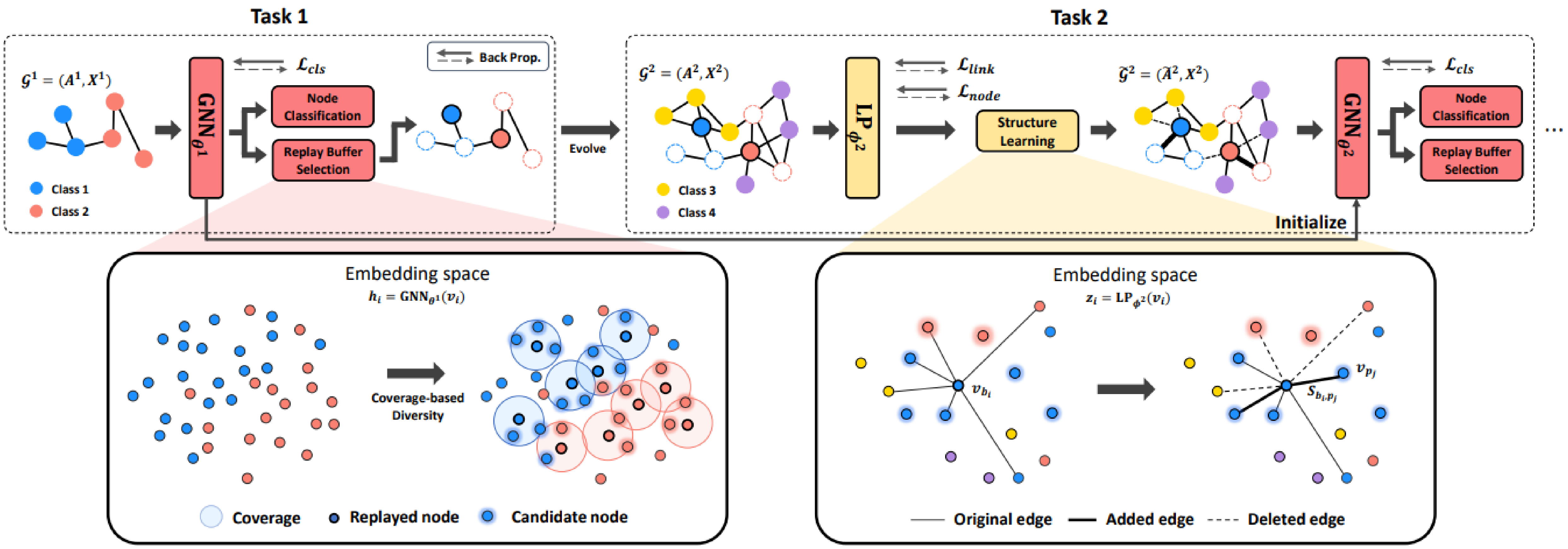
- Structure inference
  - Edge addition : Connect N nodes with highest score, maintaining the original neighbors

$$\tilde{A}_{b_i j} = \begin{cases} 1, & \text{if } v_j \in \mathcal{K}_{b_i} \cup \mathcal{N}(v_{b_i}) \\ 0, & \text{otherwise} \end{cases} \qquad \mathcal{K}_{b_i} = \{\text{argmax}_{v_j}^{(N)} S_{b_i j}\}$$

  - Edge deletion : Remove edges whose score is smaller than the threshold

$$\tilde{A}_{b_i j} = \begin{cases} 1, & \text{if } S_{b_i j} > \tau \\ 0, & \text{otherwise} \end{cases}$$

Proposed Method Overall Architecture



Select replayed nodes using CD

Reformulate the structure of replay buffer

Experiments Dataset

- Cora : citation networks
- Citeseer : citation networks
- Amazon Computer : co-purchase graph
- OGB-arxiv : large citation networks
- Reddit : large social networks

Dataset	# Nodes	# Edges	# Features	# Classes per task	# Tasks
Cora	2,708	5,429	1,433	2	3
Citeseer	3,312	4,732	3,703	2	3
Amazon Computer	13,752	245,778	767	2	4
OGB-arxiv	169,343	1,166,243	128	3	5
Reddit	232,965	114,615,892	602	5	8

Experiments Baselines

- LWF
- EWC
- GEM
- MAS
- TWP
- RCLG
- ContinualGNN
- ER-GNN

Continual learning on vision domain

Continual learning on graph domain

Rehearsal-based continual learning on graph domain

Experiments Experimental Setting

- Hyperparameters

Dataset	β	λ	N	K	τ	r	Buffer size	Learning rate
Cora	0.1	0.5	5	50	0.8	0.3	100	0.005
Citeseer	0.1	0.5	5	50	0.8	0.25	100	0.005
Amazon Computer	0.1	0.5	5	50	0.8	0.2	200	0.005
OGB-arxiv	0.05	0.5	5	50	0.8	0.15	3,000	0.005
Reddit	0.05	0.5	5	50	0.8	0.15	3,000	0.005

- Evaluation protocal

  - **PM** (Performance Mean) $= \frac{1}{T}\sum_{i=1}^{T} A_{T,i}$
  - **FM** (Forgetting Mean) $= \frac{1}{T-1}\sum_{i=1}^{T-1} A_{T,i} - A_{i,i}$

	Performance of task1	Performance of task2	Performance of task3
After Task 1	96.77		
After Task 2	91.7	86.17	
After Task 3	62.21	79.25	76.5

FM FM PM

Ex. PM = (62.21+79.25+76.5) / 3

FM = {(96.77-62.21)+(86.17-79.25)} / 2

Experiments Results

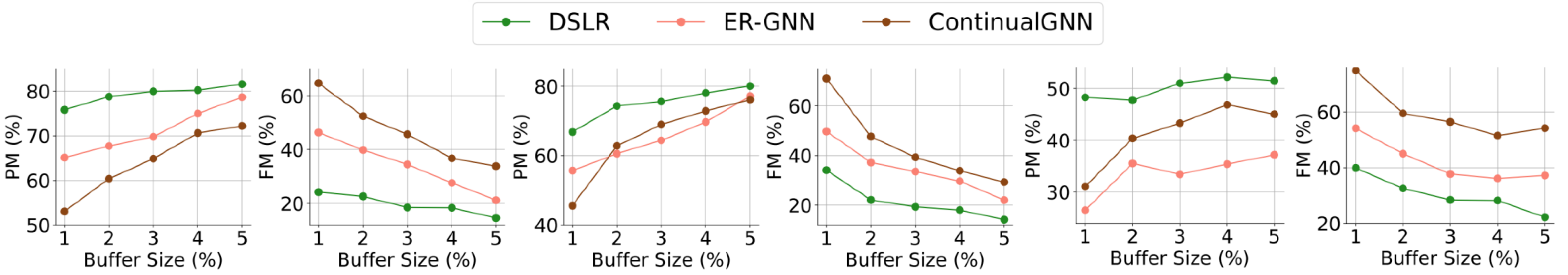
※ Performance with 10 runs

Datasets	Cora		Citeseer		Amazon Computer		OGB-arxiv		Reddit	
Methods \ Metrics	PM ↑	FM ↓	PM ↑	FM ↓	PM ↑	FM ↓	PM ↑	FM ↓	PM ↑	FM ↓
LWF [18]	61.00 ± 4.47	25.73 ± 9.26	50.38 ± 2.02	21.37 ± 4.33	30.28 ± 1.11	80.71 ± 1.68	24.18 ± 2.69	48.56 ± 8.07	23.68 ± 8.74	63.33 ± 10.08
EWC [16]	70.56 ± 3.13	31.90 ± 4.38	60.98 ± 3.45	21.56 ± 4.39	49.63 ± 4.27	49.62 ± 5.73	45.71 ± 6.50	30.91 ± 2.73	20.57 ± 6.25	28.09 ± 6.93
GEM [20]	65.44 ± 5.16	32.97 ± 3.94	60.14 ± 1.72	21.89 ± 2.82	40.74 ± 3.03	42.19 ± 4.52	40.58 ± 4.26	29.28 ± 7.56	36.28 ± 4.77	17.94 ± 2.84
MAS [1]	72.10 ± 5.25	17.21 ± 5.35	60.62 ± 3.32	23.44 ± 3.73	63.37 ± 1.80	23.17 ± 8.18	39.29 ± 2.91	30.36 ± 3.74	10.27 ± 2.84	**13.85 ± 1.42**
ContinualGNN [34]	72.21 ± 1.83	33.84 ± 2.74	60.58 ± 0.86	34.89 ± 1.50	76.12 ± 0.75	29.33 ± 1.03	48.91 ± 4.15	52.83 ± 1.09	OOM	OOM
TWP [19]	71.87 ± 8.45	25.77 ± 4.38	61.80 ± 1.31	24.76 ± 3.93	71.28 ± 3.26	26.55 ± 3.28	39.20 ± 5.92	25.65 ± 4.26	22.56 ± 7.57	21.70 ± 5.51
ER-GNN [46]	78.68 ± 2.10	21.16 ± 3.52	65.49 ± 1.00	30.04 ± 1.19	77.20 ± 2.11	22.00 ± 2.13	37.19 ± 2.50	37.26 ± 1.55	33.62 ± 6.61	19.35 ± 6.08
RCLG [24]	70.77 ± 4.74	15.71 ± 4.01	66.60 ± 3.33	22.67 ± 5.49	51.91 ± 6.57	16.71 ± 9.74	50.04 ± 6.44	41.00 ± 8.16	OOM	OOM
DSLR	**81.59 ± 1.65**	**14.59 ± 2.61**	**69.54 ± 0.74**	**18.21 ± 0.96**	**80.08 ± 0.98**	**14.18 ± 3.15**	**51.46 ± 1.50**	**22.21 ± 3.82**	**38.12 ± 5.91**	16.78 ± 8.12

- DSLR outperforms in terms of both PM and FM over all baselines, demonstrating low variance
- Rehearsal-based approaches (ContinualGNN, ER-GNN) outperforms other baselines in PM, but show worse FM

Experiments Results

- **Memory efficiency of DSLR**

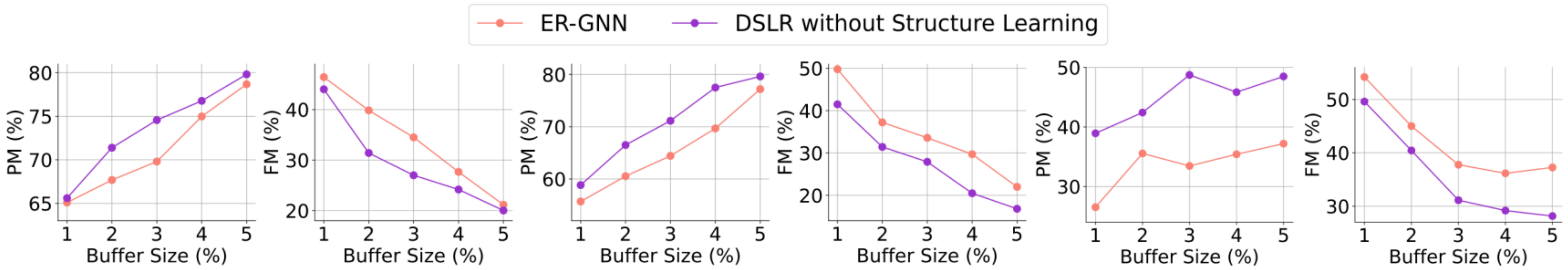


(a) PM on Cora **(b) FM on Cora** **(c) PM on Amazon** **(d) FM on Amazon** **(e) PM on OGB-arxiv** **(f) FM on OGB-arxiv**

- Mild decrease of the performance when the buffer size decreases
- DSLR can achieve comparable performance with a much smaller buffer size

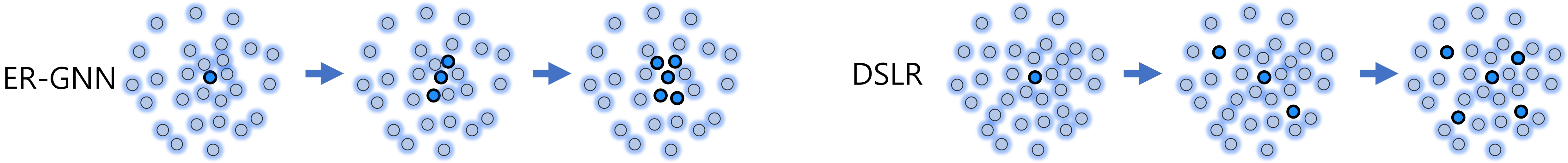
Experiments Results

- **Effectiveness of Coverage-based Diversity (1)**



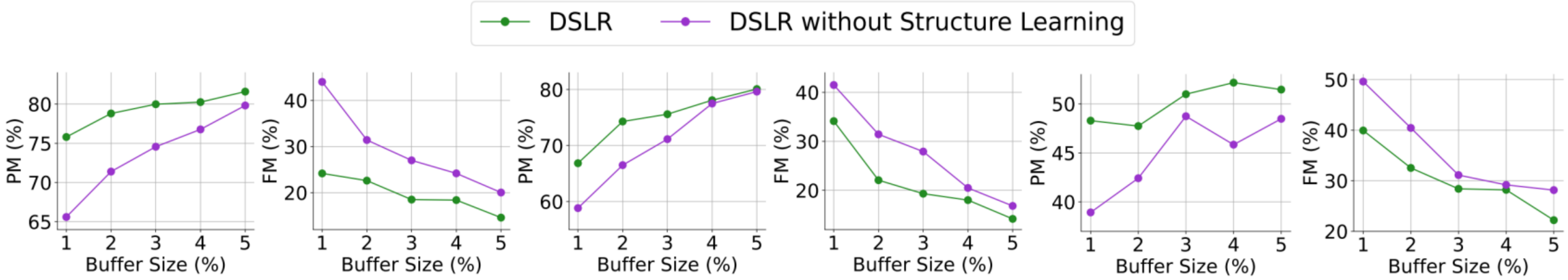
(a) PM on Cora **(b) FM on Cora** **(c) PM on Amazon** **(d) FM on Amazon** **(e) PM on OGB-arxiv** **(f) FM on OGB-arxiv**

- DSLR outperforms ER-GNN regardless of the buffer size, in both PM and FM
- When buffer size increases from 1% to 3% (small to mid-size), the gain of performance of DSLR is more significant



Experiments Results

- **Effectiveness of Structure Learning (1)**



(a) PM on Cora (b) FM on Cora (c) PM on Amazon (d) FM on Amazon (e) PM on OGB-arxiv (f) FM on OGB-arxiv

- Structure learning component not only benefits the performance, but also memory efficiency

Conclusion

- Summary
  - Graph Continual Learning with diverse, representative replayed nodes and structure learning for them

- Contribution
  - Emphasize the consideration of diversity when selecting the replayed nodes
  - Discover the substantial influence of the quality of neighbors surrounding the replayed nodes
  - Extensive experiments demonstrate the effectiveness and efficiency of DSLR

Thank you!

[Full Paper] https://www.arxiv.org/abs/2402.13711

[Source Code] https://github.com/seungyoon-Choi/DSLR_official

[Lab Homepage] http://dsail.kaist.ac.kr

[Email] csyoon08@kaist.ac.kr