Geometer: Graph Few-Shot Class-Incremental Learning via Prototype Representation

Bin Lu, Xiaoying Gan, Lina Yang, Weinan Zhang, Luoyi Fu, Xinbing Wang

Shanghai Jiao Tong University

Shanghai, China

Table of Contents

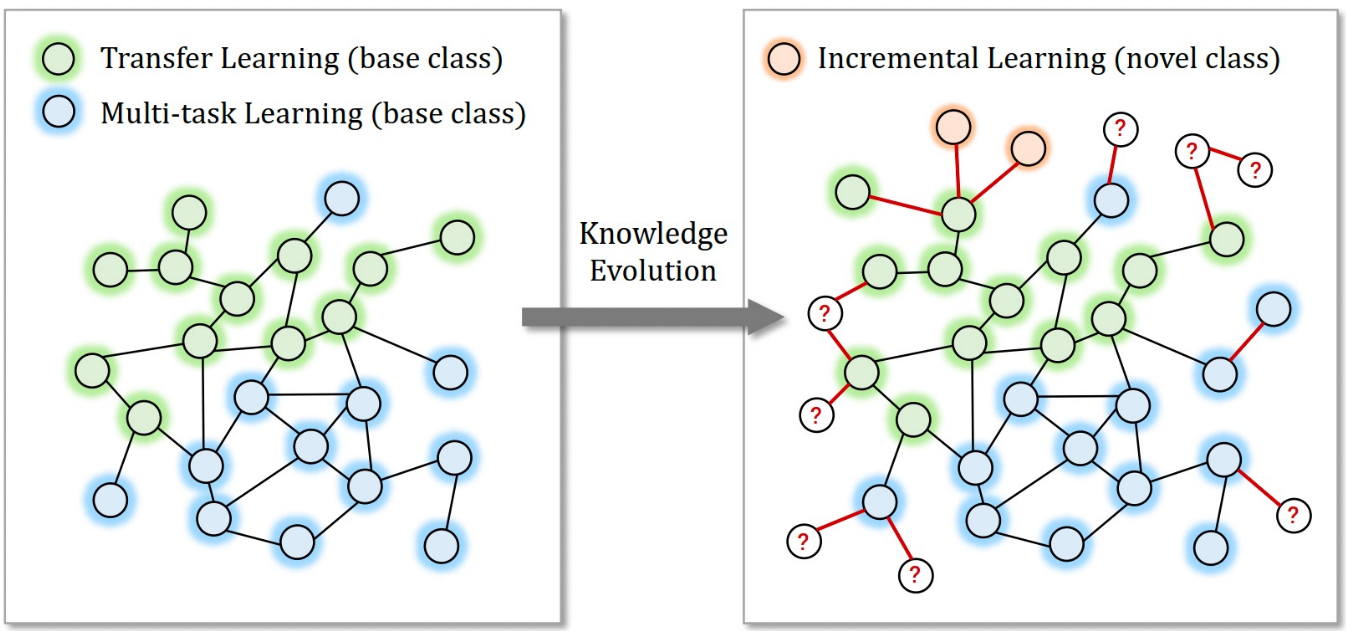
Table of Contents

Background

- Real-world networks evolve with the emergence of new nodes and edges, thereby generating novel classes.
  - Academic citation network: new papers, new interdisciplines
  - Social network: new users, new social groups
  - E-commerce network: new commodities, new types

Background

- Classes of nodes are expanding incrementally and usually accompanied by few labeling due to its newly emergence or lack of exploration.
- A Toy Example of Graph Few-Shot Class-Incremental Learning
  - A novel research field appears with few-shot labeling.



Prior Works

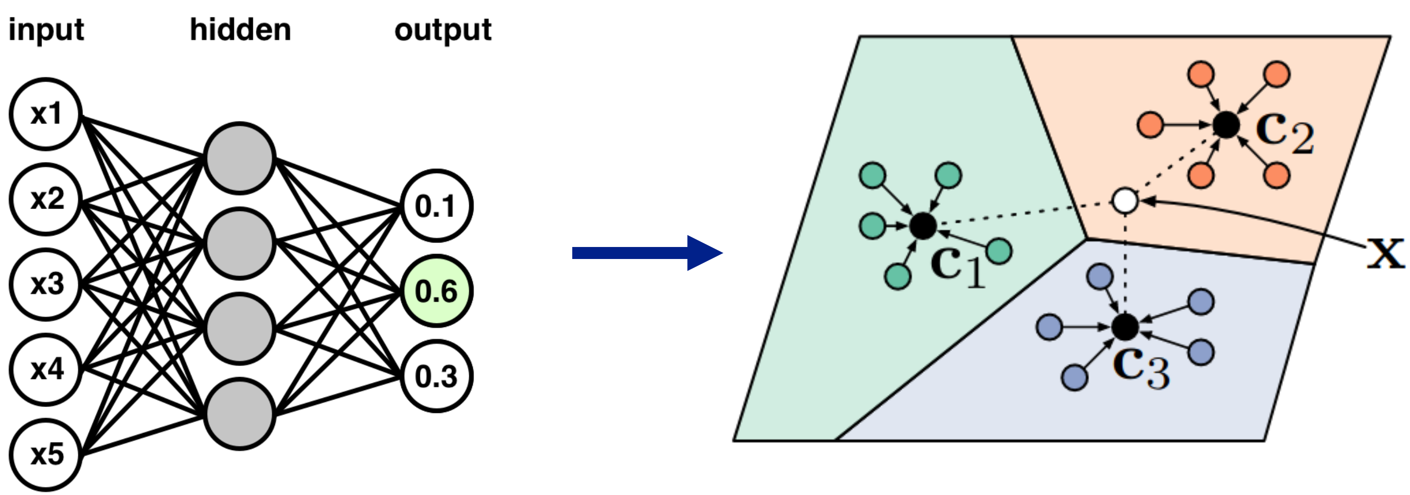
- **#1 Classical GNN methods**
  - Abundant labeling
  - Fixed classes
- **#2 Graph few-shot learning**[1,2,3]
  - A prior N-way K-shot assumption
  - Two separate models for *base* classes and *novel* classes
- **#3 Class-incremental learning**[4,5]
  - Little work focus on non-Euclidean data (e.g., graphs)
  - Network structure evolves dynamically

[1] Zhou et al. "Meta-GNN: On few-shot node classification in graph meta-learning." CIKM, 2019.
[2] Huang et al. "Graph meta learning via local subgraphs" NeurIPS, 2020.
[3] Liu et al. "Relative and absolute location embedding for few-shot node classification on graph" AAAI, 2021.
[4] Li et al. "Learning without forgetting" TPAMI, 2017
[5] Hou et al. "Learning a unified classifier incrementally via rebalancing." CVPR, 2019.

Motivation

Q1: How to cope with the incremental categories for node classification?

- Instead of replacing and retraining the fully connected neural network classifier, we propose to predict the ever-expanding class of a node by finding the nearest *prototype representation*[6].



[6] Snell et al. "Prototypical Networks for Few-shot Learning", NeurIPS, 2017.

Motivation

Q2: How to accurately learn the *prototype representation* ?

- Two Challenges
  - Dynamic graph structure due to evolution attribute
  - Unbalanced labeling between base classes and novel classes
- Our Methods
  - Attention-based Prototype Representation
    - Node-level Graph Attention Network
    - Class-level Multi-head Attention
  - Biased Sampling Strategy
    - Episode Meta Learning
    - Mimic the circumstances encountered when novel classes pop up

Motivation

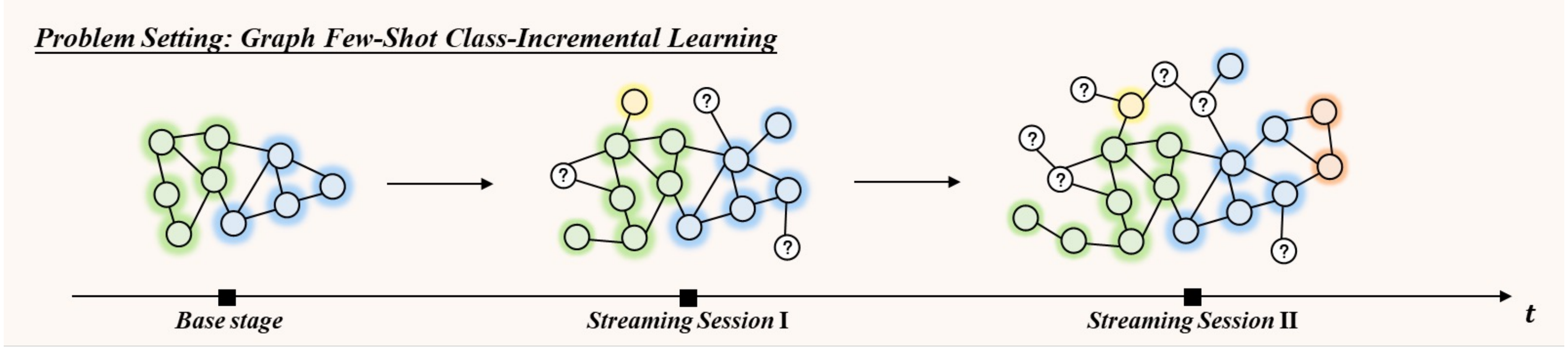
Q3: How to find a way out of "forgetting old" and "overfitting new"?

- We propose to adjust the *prototype representation* delicately.
  - Geometric relations in feature space
    - Inter-class
    - Intra-class
  - Inheritance between previous and revised *prototype representation*

Table of Contents

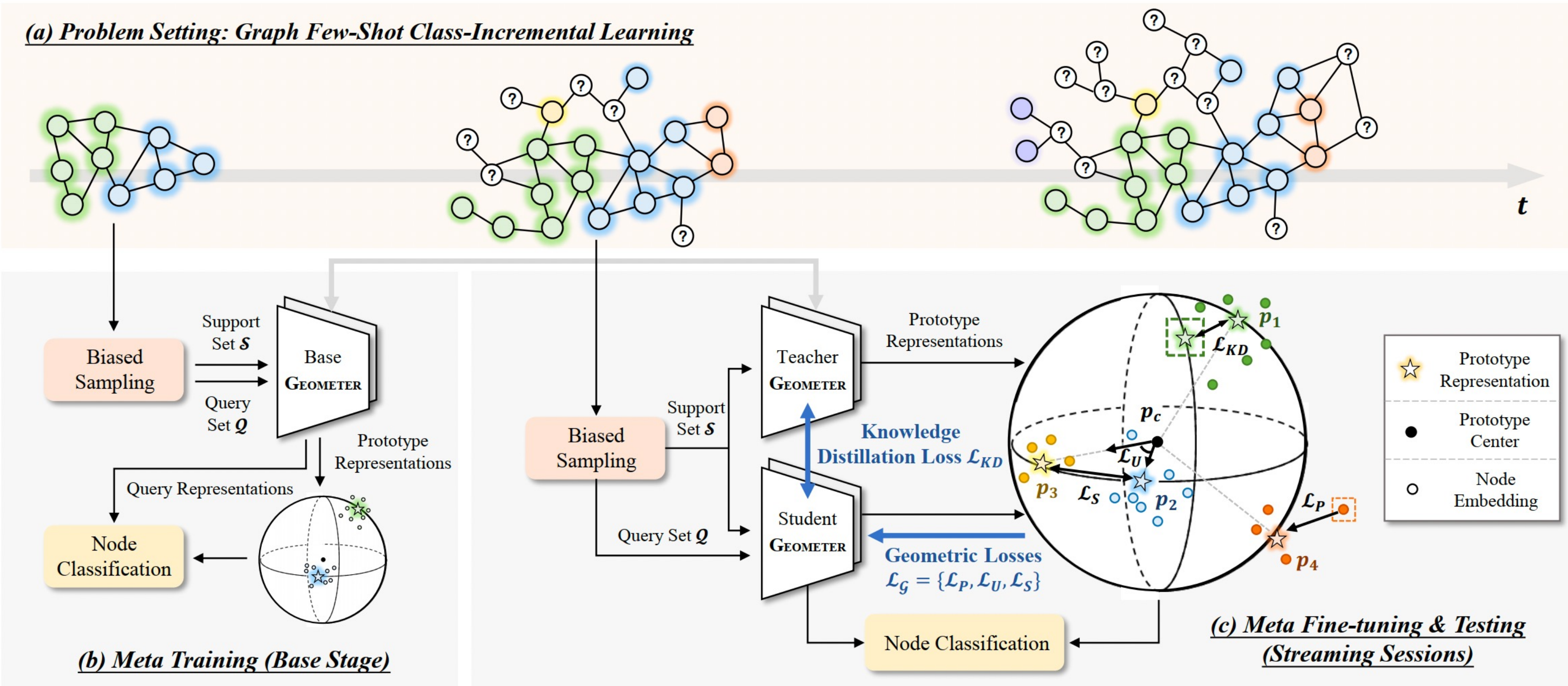
Problem Definition

- Graph Few-Shot Class-Incremental Learning
  - Streaming Sessions
    - Base stage $\mathcal{G}^{\text{base}}$ and T snapshots of $\mathcal{G}^{\text{stream}} = \{\mathcal{G}^1, \cdots, \mathcal{G}^T\}$
    - Sets of classes $\{C^{\text{base}}, C^1, \cdots, C^T\}$, $C^t = C^{\text{base}} + \sum_i \Delta C^i$
  - ΔC^t-way K-shot GFSCIL problem
    - ΔC^t novel classes with K labeled nodes
    - GFSCIL problem is tested to classify unlabeled nodes into all encountered classes C^t



Methodology

- Overview of proposed model: Geometer



(a) Problem setting of GFSCIL, (b) and (c) show the episode meta learning process with biased sampling strategy at base stage and streaming sessions.

Methodology

- **Attention-based Prototype Representation**
  - Node-level Graph Attention Network

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

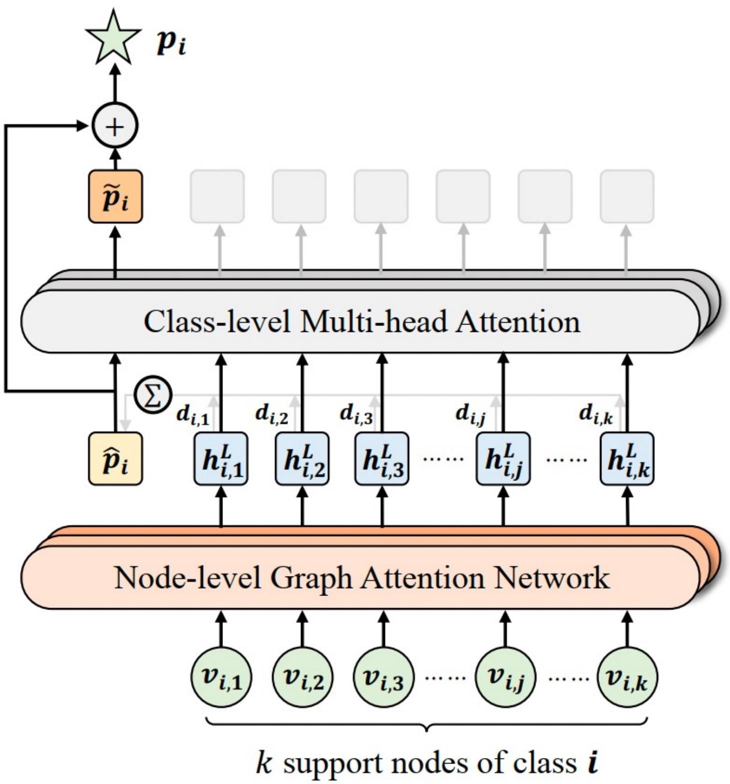
  - Class-level Multi-head Attention
    - Degree-based weighted-sum of support node embedding $\hat{p}_i$

$$\hat{\boldsymbol{p}}_i = \sum_{j\in\mathcal{S}_i} \frac{\text{degree}(v_j)}{\sum_{j'\in\mathcal{S}_i}\text{degree}(v_{j'})} \cdot f_{\mathcal{G}}(\boldsymbol{x}_j)$$

    - Class-level attention score

$$\boldsymbol{p}_i = \hat{\boldsymbol{p}}_i + Attention\left(\hat{\boldsymbol{p}}_i, \boldsymbol{h}_i^{spt}, \boldsymbol{h}_i^{spt}\right)$$

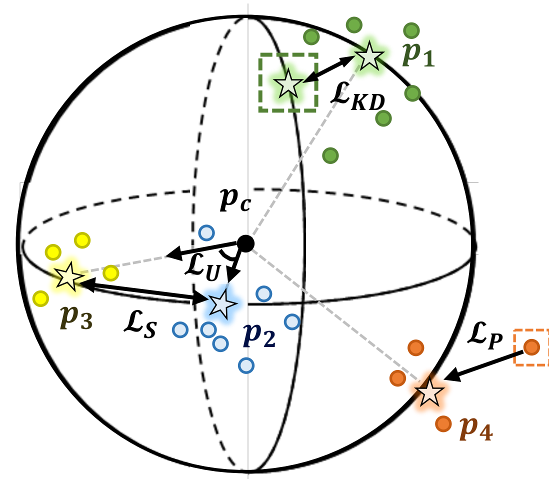
$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$



Methodology

- **Geometric Metric Learning**
  - Intra-Class Proximity $\boldsymbol{\mathcal{L}_P}$
    - Nodes of same classes should be closely clustered

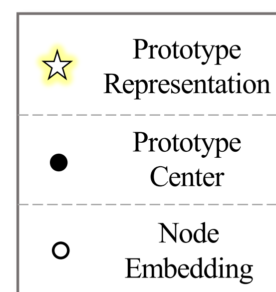
$$\mathcal{L}_P = \sum_{k=1}^{\|C^k\|} \frac{\alpha_k}{n_k} \sum_{i=1}^{n_k} -\log \frac{\exp\left(-d\left(f_{\mathcal{G}}\left(\boldsymbol{x}_i\right), \boldsymbol{p}_k\right)\right)}{\sum_{k' \in C^k} \exp\left(-d\left(f_{\mathcal{G}}\left(\boldsymbol{x}_i\right), \boldsymbol{p}_{k'}\right)\right)}$$



  - Inter-Class Uniformity $\boldsymbol{\mathcal{L}_U}$
    - Uniformity of different prototypes in metric space

$$\mathcal{L}_U = \frac{1}{\|C^k\|} \sum_{i=1}^{\|C^k\|} \left\{1 + \max_{j \in \{C^k\} \setminus i} \left[\frac{(\boldsymbol{p}_i - \boldsymbol{p}_c)}{\|\boldsymbol{p}_i - \boldsymbol{p}_c\|} \cdot \frac{(\boldsymbol{p}_j - \boldsymbol{p}_c)}{\|\boldsymbol{p}_j - \boldsymbol{p}_c\|}\right]\right\}$$

  - Inter-Class Separability $\boldsymbol{\mathcal{L}_S}$
    - Prototypes of novel classes and old classes should keep a distance



$$\mathcal{L}_S = \frac{1}{\Delta C^k} \sum_{i \in \Delta C^k} \min_{j \in C^{k-1}} \exp\left(-d\left(\boldsymbol{p}_i, \boldsymbol{p}_j\right)\right)$$

Methodology

- **Teacher-Student Knowledge Distillation**
  - Mitigate “forgetting old” problem in GFSCIL
  - Temperature-scaled softmax (τ is the temperature factor)

$$y'^{(i)} = \frac{\exp\left(d\left(f_{\mathcal{G}}\left(\boldsymbol{x}^{(i)}\right), \boldsymbol{p}_i\right)/\tau\right)}{\sum_j \exp\left(d\left(f_{\mathcal{G}}\left(x^{(i)}\right), \boldsymbol{p}_j\right)/\tau\right)}$$

  - KL-divergence of the softened logits to make the student model gain the experience of classifying old classes from teacher model

$$\mathcal{L}_{KD} = \frac{1}{\left\|C^{k-1}\right\|} \sum_{i=1}^{\left\|C^{k-1}\right\|} y_S'^{(i)} \cdot \log\left(\frac{y_S'^{(i)}}{y_T'^{(i)}}\right)$$

Methodology

- **Episode Meta Learning**
  - Episode paradigm in learning process
    - Generate a set of tasks $\boldsymbol{\mathcal{T}} = \{\mathcal{T}_1, \mathcal{T}_2, \cdots, \mathcal{T}_N\}$
    - Each task $\mathcal{T}_i$ includes support set $\mathcal{S}_i$ and query set $\mathcal{Q}_i$
  - Biased sampling strategy
    - Pretraining stage
      - Generate class-imbalanced support sets, $\|\mathcal{S}_i\| \sim \boldsymbol{U}[1, K_{max}]$
    - Finetuning stage
      - A higher proportion of old samples will be sampled

Table of Contents

Experiment

- Dataset Description

Dataset	Field	Nodes	Edges	Features	Class
Cora-ML	Academic	2,995	16,316	2,879	7
Flickr	Social Network	7,575	479,476	12,047	9
Amazon	E-commerce	13,752	491,722	767	10
Cora-Full	Academic	19,793	126,842	8,710	70

- Experiment Setting
  - <u>Cora-ML, Flickr, Amazon</u>: 1-way 5-shot GFSCIL setting, 6 streaming sessions
  - <u>Cora-Full</u>: 5-way 5-shot GFSCIL setting, 10 streaming sessions

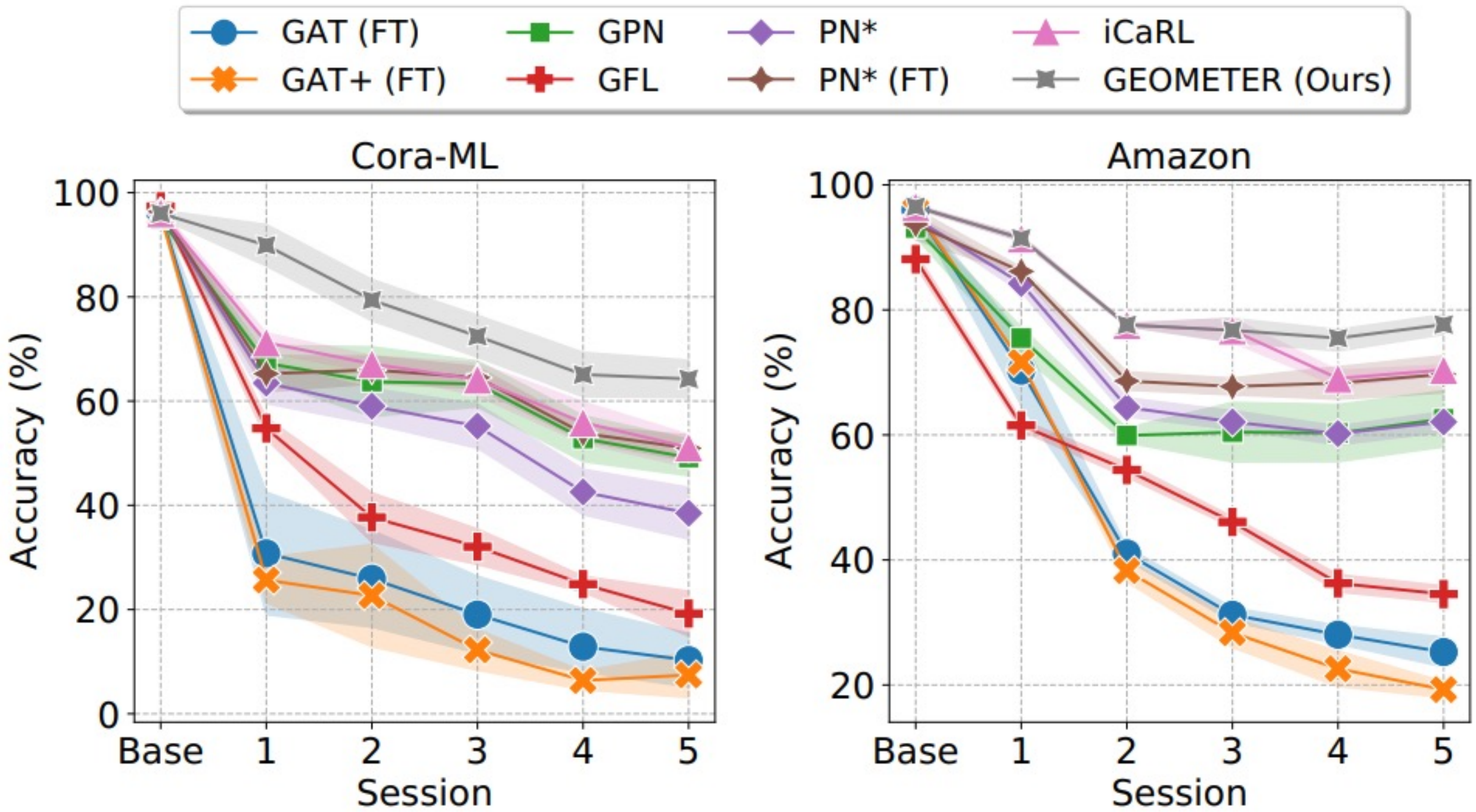
Experiment

- **Baseline Methods**
  - GNN models
    - GAT (FT): fine-tune parameters of MLP predictor only
    - GAT+ (FT): fine-tune all parameters
  - Graph Few-shot learning models
    - GPN: Graph Prototype Network [7]
    - GFL: Graph few-shot learning via knowledge transfer [8]
  - Prototype network methods
    - PN*: Prototype Network for node classification
    - PN*(FT): Prototype Network with fine-tuning strategy
  - iCaRL*: Class-incremental learning methods[9]

[7] Ding et al. “Graph prototypical networks for few-shot learning on attributed networks.” CIKM, 2020.
[8] Yao et al. “Graph few-shot learning via knowledge transfer.” AAAI, 2020.
[9] Rebuffi et al. “iCaRL: Incremental classifier and representation learning” CVPR, 2017.

Experiment

- Performance Comparison



Experiment

- Performance Comparison

Session	Cora-Full (5-way 5-shot GFSCIL setting)								impr.
	GAT (FT)	GAT+ (FT)	GPN	GFL	PN*	PN* (FT)	iCaRL*	Geometer (Ours)	
Base	**80.53±1.32%**	81.11±0.79%	73.82±1.94%	76.02±0.94%	74.88±0.89%	74.18±0.72%	73.92±1.06%	79.88±0.96%	-1.52%
Session 1	33.13±2.51%	37.10±1.51%	55.95±1.52%	60.50±0.74%	56.60±1.11%	58.07±0.92%	59.33±1.79%	**69.48±1.66%**	+14.84%
Session 2	25.39±1.59%	26.34±0.98%	49.49±1.57%	52.85±1.88%	48.59±0.66%	53.97±0.97%	54.05±0.70%	**61.34±0.92%**	+13.49%
Session 3	17.48±1.59%	17.41±1.43%	43.41±1.66%	43.88±2.84%	39.70±1.25%	43.76±0.92%	44.65±0.55%	**53.61±0.81%**	+20.07%
Session 4	12.09±1.35%	12.12±0.78%	39.03±1.29%	38.22±1.81%	37.33±2.07%	41.83±0.91%	40.52±1.56%	**48.24±1.46%**	+15.30%
Session 5	10.04±1.56%	8.54±0.39%	35.12±1.98%	38.69±2.50%	32.66±2.01%	37.35±0.73%	36.25±1.06%	**44.97±1.03%**	+16.23%
Session 6	8.63±0.88%	7.01±0.80%	33.34±1.35%	33.94±3.53%	30.83±2.09%	36.56±0.84%	33.46±1.16%	**42.93±0.88%**	+17.42%
Session 7	7.76±0.62%	5.79±0.42%	31.98±1.03%	32.60±1.65%	29.52±1.92%	34.70±0.20%	32.68±1.44%	**42.82±1.14%**	+23.40%
Session 8	6.99±0.72%	5.38±0.49%	30.63±1.64%	28.32±1.78%	28.39±1.97%	33.97±1.24%	31.02±1.48%	**41.01±0.96%**	+20.72%
Session 9	5.95±0.75%	4.49±0.40%	30.53±1.80%	21.95±1.71%	27.65±2.19%	33.71±0.75%	30.37±1.76%	**40.49±0.97%**	+20.11%
Session 10	5.51±0.95%	3.92±0.61%	28.33±1.48%	21.77±1.50%	26.07±1.89%	31.67±0.55%	29.21±1.71%	**39.32±0.78%**	+24.15%

Session	Flickr (1-way 5-shot GFSCIL setting)								impr.
	GAT (FT)	GAT+ (FT)	GPN	GFL	PN*	PN* (FT)	iCaRL*	Geometer (Ours)	
Base	62.16±1.30%	61.41±1.55%	72.80±1.13%	**84.82±3.13%**	59.86±1.81%	59.43±2.68%	60.81±1.87%	64.75±1.76%	-23.66%
Session 1	24.24±3.91%	26.13±7.62%	51.02±1.70%	**61.09±1.41%**	34.29±1.97%	40.96±2.38%	40.54±1.28%	57.57±2.80%	-5.76%
Session 2	17.54±4.98%	14.83±0.92%	37.77±4.29%	45.53±3.08%	28.30±4.16%	38.78±1.97%	37.03±5.06%	**50.11±2.03%**	+10.05%
Session 3	16.13±5.55%	8.55±0.71%	32.79±2.65%	33.73±1.49%	23.37±3.63%	35.43±5.63%	32.93±7.96%	**45.21±1.04%**	+27.60%
Session 4	8.41±2.11%	6.30±2.31%	24.39±2.54%	31.63±2.35%	23.77±5.02%	38.16±5.31%	31.27±6.79%	**41.77±0.79%**	+9.46%
Session 5	9.04±5.46%	4.94±2.48%	22.01±1.28%	28.15±1.17%	20.23±4.00%	32.74±4.57%	26.57±5.83%	**36.26±2.79%**	+10.75%

Experiment

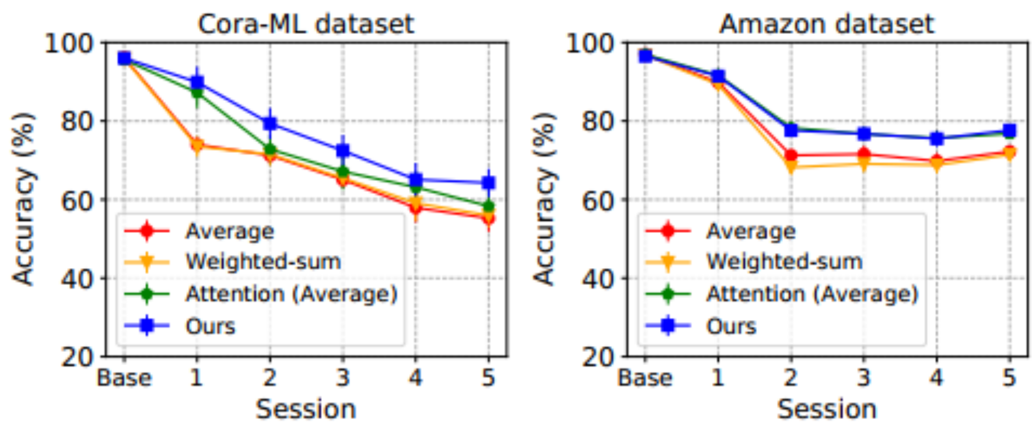
- Ablation Study: Loss functions

Loss functions				Cora-ML (1-way 5-shot GFSCIL setting)			
$\mathcal{L}_P$	$\mathcal{L}_U$	$\mathcal{L}_S$	$\mathcal{L}_{KD}$	**Base Classes**	**Session 1**	**Session 3**	**Session 5**
✓			✓	**96.21±0.67%**	88.25±3.99%	64.89±2.53%	56.21±5.55%
✓	✓		✓	95.85±0.56%	**90.41±3.86%**	68.26±3.65%	58.72±4.66%
✓		✓	✓	95.71±0.55%	89.21±2.88%	69.57±2.71%	54.28±3.90%
✓	✓	✓		95.74±0.61%	90.40±4.82%	68.16±1.45%	62.41±2.37%
✓	✓	✓	✓	96.01±0.92%	89.89±3.97%	**72.45±4.01%**	**64.25±3.60%**

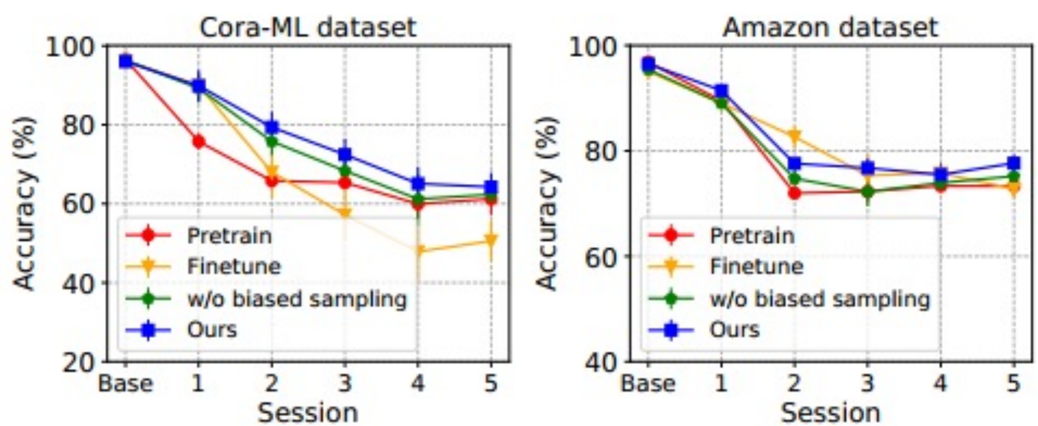
Loss functions				Amazon (1-way 5-shot GFSCIL setting)			
$\mathcal{L}_P$	$\mathcal{L}_U$	$\mathcal{L}_S$	$\mathcal{L}_{KD}$	**Base Classes**	**Session 1**	**Session 3**	**Session 5**
✓			✓	96.72±0.28%	90.91±0.59%	74.74±2.33%	73.73±3.01%
✓	✓		✓	96.72±0.22%	91.39±0.56%	76.55±1.94%	73.97±1.90%
✓		✓	✓	96.83±0.32%	91.15±0.35%	75.08±2.61%	73.92±2.60%
✓	✓	✓		**96.86±0.35%**	91.17±0.38%	75.36±1.28%	74.51±2.53%
✓	✓	✓	✓	96.50±0.29%	**91.44±0.46%**	**76.74±1.89%**	**77.66±1.58%**

Experiment

- Ablation Study: Prototype representation method



- Ablation Study: Biased sampling strategy



Experiment

- Case Study: t-SNE visualization of the query node embeddings and prototypes

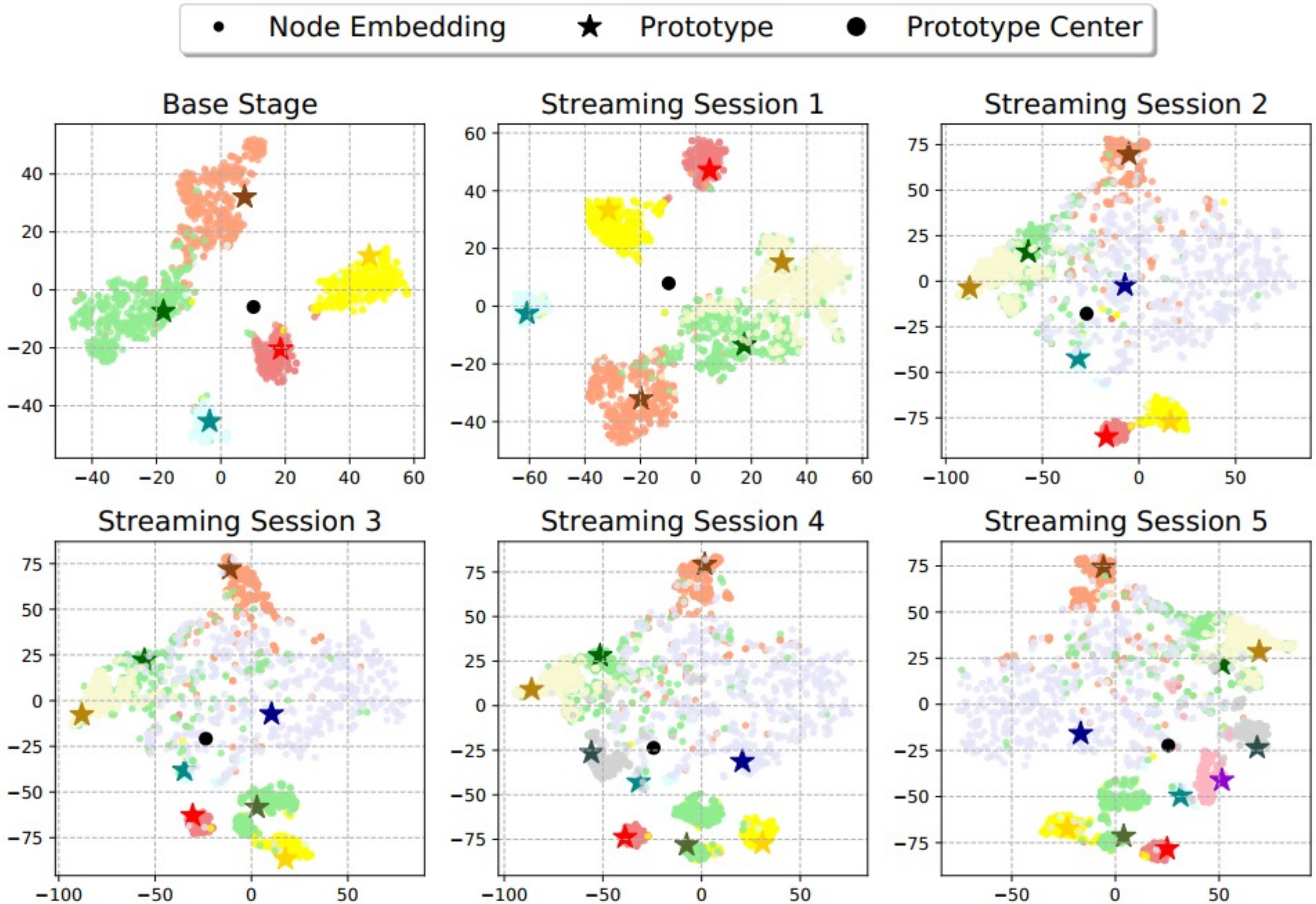


Table of Contents

Conclusion

- We are the first to investigate this novel problem: graph few-shot class-incremental learning (GFSCIL).
- With the novel classes popping up, Geometer learns and adjusts the attention-based prototypes based on the geometric relationships of proximity, uniformity and separability of representations.
- Geometer proposes teacher-student knowledge distillation and biased sampling strategy to further mitigate the catastrophic forgetting and unbalanced labeling in GFSCIL.

Thanks

Bin Lu
robinlu1209@sjtu.edu.cn