Graph Contrastive Learning with Augmentations (GraphCL)

School of Industrial and Management Engineering, Korea University

Jae Hoon Kim

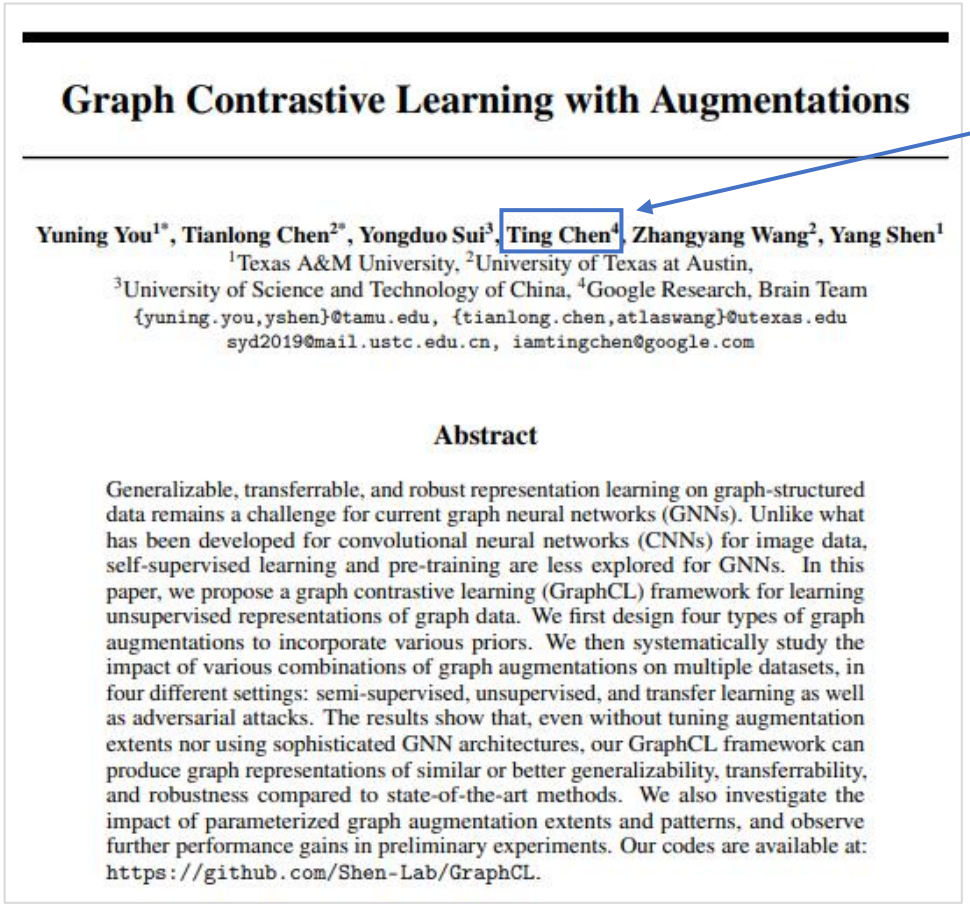
KOREA UNIVERSITY

DMQA

Contents

Research Purpose

- ❖ Graph Contrastive Learning with Augmentations (NeurIPS 2020)
  - 2022년 3월 13일 기준으로 229회 인용됨



Graph Contrastive Learning with Augmentations

Yuning You[1*], Tianlong Chen[2*], Yongduo Sui[3], Ting Chen[4], Zhangyang Wang[2], Yang Shen[1]

[1]Texas A&M University, [2]University of Texas at Austin,
[3]University of Science and Technology of China, [4]Google Research, Brain Team
{yuning.you,yshen}@tamu.edu, {tianlong.chen,atlaswang}@utexas.edu
syd2019@mail.ustc.edu.cn, iamtingchen@google.com

Abstract

Generalizable, transferrable, and robust representation learning on graph-structured data remains a challenge for current graph neural networks (GNNs). Unlike what has been developed for convolutional neural networks (CNNs) for image data, self-supervised learning and pre-training are less explored for GNNs. In this paper, we propose a graph contrastive learning (GraphCL) framework for learning unsupervised representations of graph data. We first design four types of graph augmentations to incorporate various priors. We then systematically study the impact of various combinations of graph augmentations on multiple datasets, in four different settings: semi-supervised, unsupervised, and transfer learning as well as adversarial attacks. The results show that, even without tuning augmentation extents nor using sophisticated GNN architectures, our GraphCL framework can produce graph representations of similar or better generalizability, transferrability, and robustness compared to state-of-the-art methods. We also investigate the impact of parameterized graph augmentation extents and patterns, and observe further performance gains in preliminary experiments. Our codes are available at: https://github.com/Shen-Lab/GraphCL.

SimCLR 저자...

DMQA

Research Purpose

❖ Graph Contrastive Learning with Augmentations (NeurIPS 2020)

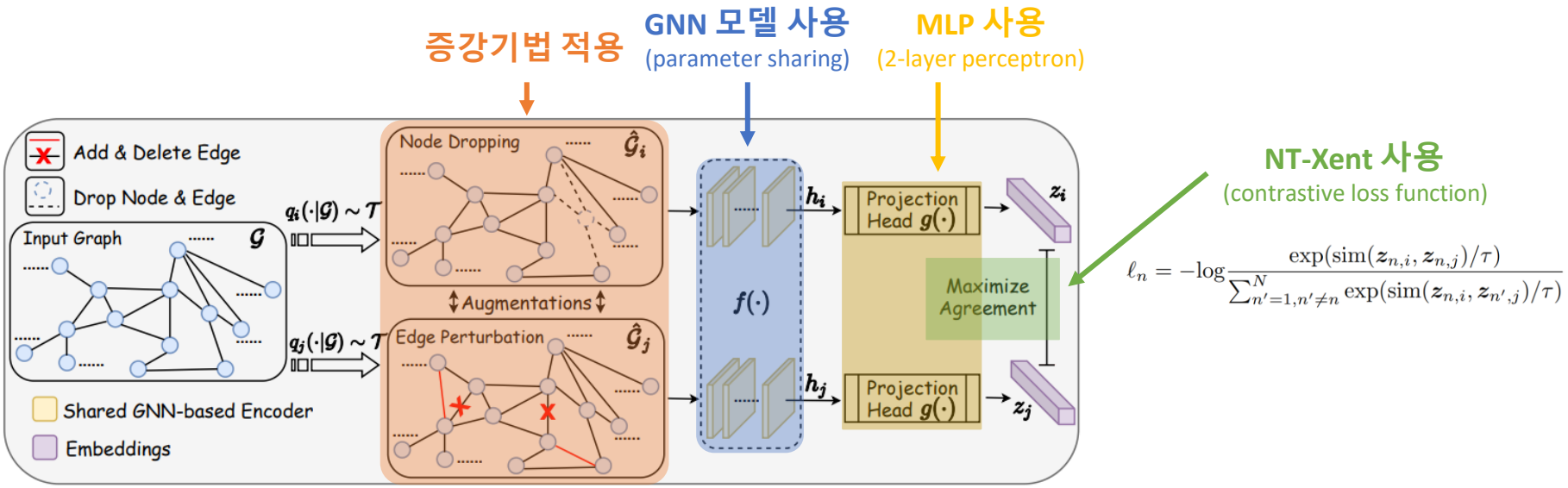
- 그래프 형태의 데이터셋 역시 특정 작업을 위한 레이블이 매우 부족한 경우가 많음
- 이미지 형태의 데이터셋은 위의 관련된 문제를 자기지도학습을 통한 사전학습으로 해결함
- 따라서 이번 논문에서는 그래프 모델에 자기지도학습을 활용한 연구를 진행하였음

- 논문의 기여점은 아래와 같음
  - ✓ 그래프 모델을 활용한 대조학습에서 활용할 수 있는 증강기법 탐색
  - ✓ 그래프 모델을 활용한 대조학습 프레임워크 제안
  - ✓ 데이터 도메인 별 적합한 증강기법 조합 탐색
  - ✓ 제안 방법론을 통해 SOTA 달성

Graph Contrastive Learning (GraphCL)

❖ Graph Contrastive Learning (GraphCL)

- SimCLR 저자가 참여한 논문 답게 기본적인 프레임워크는 SimCLR와 동일함
- GNN에 적합한 증강기법 네 가지를 적용함
  - ✓ Node dropping
  - ✓ Attribute masking
  - ✓ Edge perturbation
  - ✓ Subgraph
- 인코더에는 GNN 기반의 모든 모델을 사용할 수 있음
  - ✓ 논문에서는 Graph Convolution Networks (GCN)을 활용



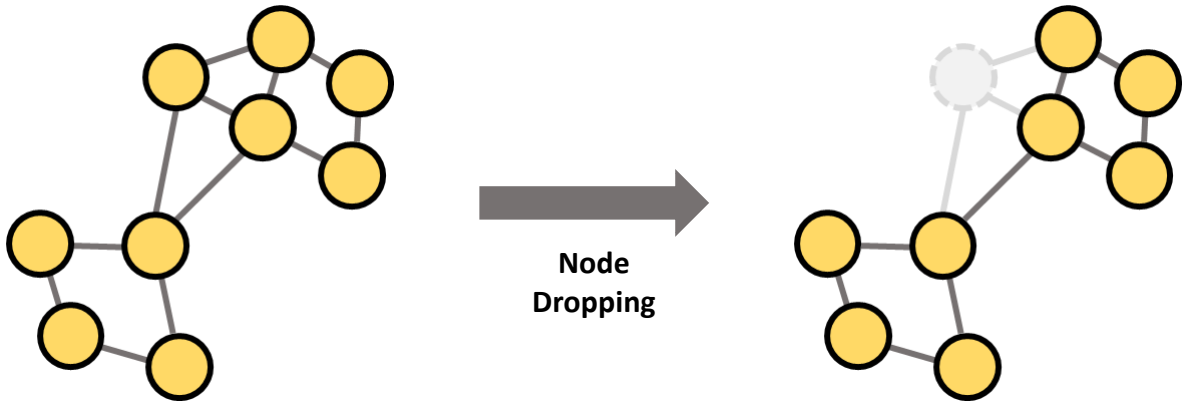
$$\ell_n = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_{n,i}, \boldsymbol{z}_{n,j})/\tau)}{\sum_{n'=1,n'\neq n}^{N} \exp(\text{sim}(\boldsymbol{z}_{n,i}, \boldsymbol{z}_{n',j})/\tau)}$$

DMQA

Graph Contrastive Learning (GraphCL)

❖ Augmentations for GNN (Node dropping)

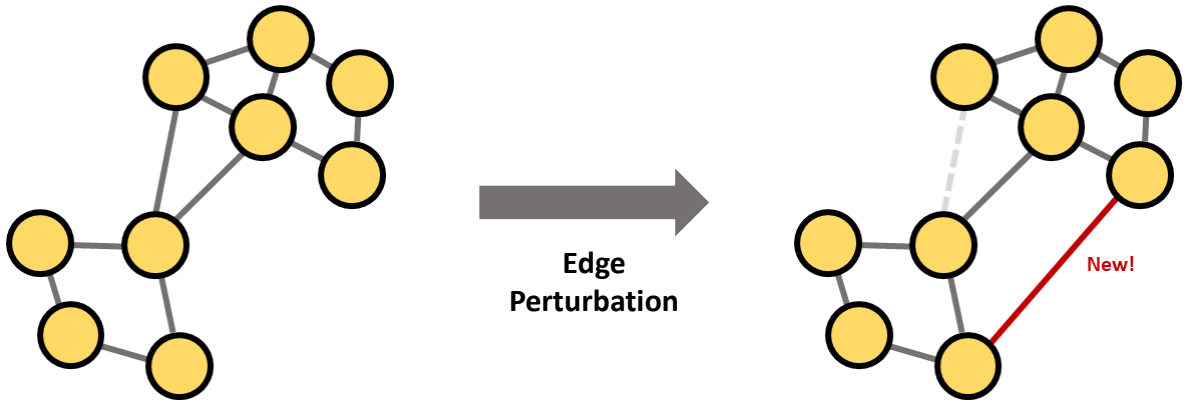
- 노드를 제거하더라도 그래프의 의미(semantic)가 바뀌지 않는다는 전제를 가짐
- 각 노드가 제거될 확률은 기본적으로 균등 분포를 따르며 분포의 종류는 변경될 수 있음



Graph Contrastive Learning (GraphCL)

❖ Augmentations for GNN (Edge perturbation)

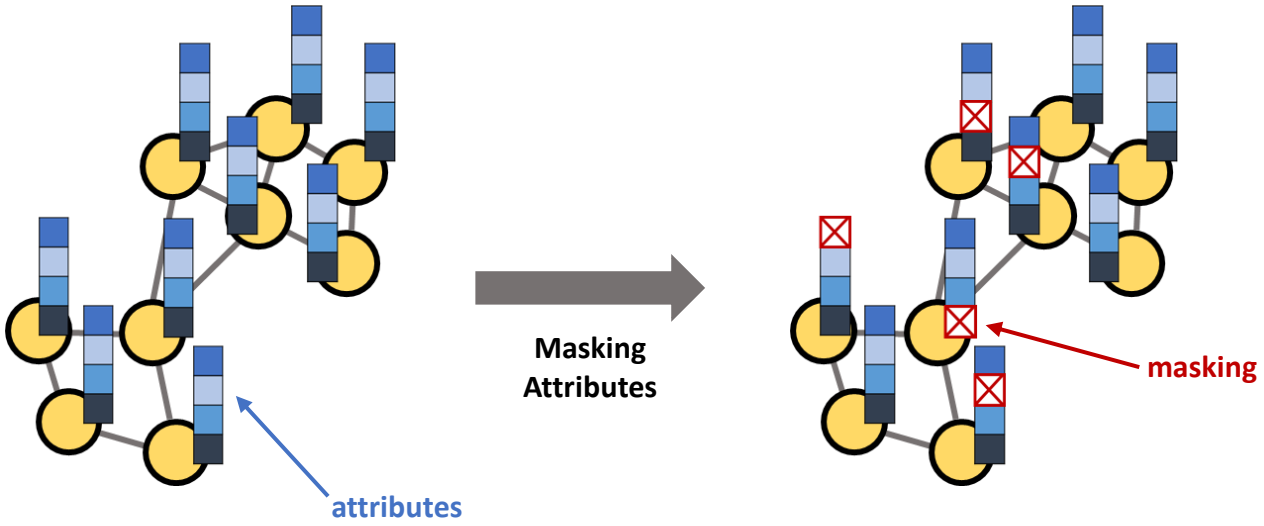
- 특정 비율의 엣지에 대하여 제거 혹은 노드 간의 연결을 추가함
- 연결 구조가 조금 달라지더라도 그래프의 의미가 바뀌지 않는다는 전제를 가짐
- 각 엣지가 추가 혹은 제거될 확률은 기본적으로 균등 분포를 따름



Graph Contrastive Learning (GraphCL)

❖ Augmentations for GNN (Attribute masking)

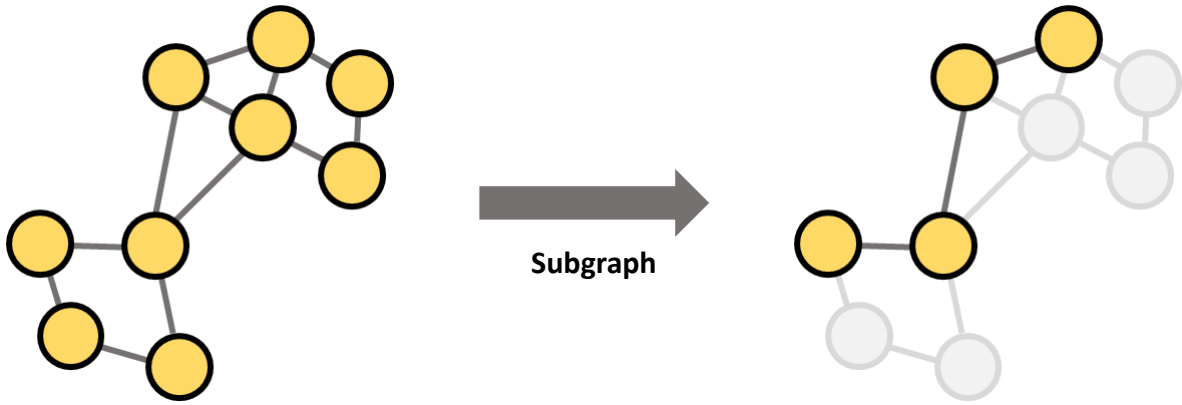
- 각 노드 별 특징 일부에 마스킹을 적용함
- 특징의 일부를 숨기더라도 그래프의 의미가 바뀌지 않는다는 전제를 가짐



Graph Contrastive Learning (GraphCL)

❖ Augmentations for GNN (Subgraph)

- 랜덤워크 알고리즘으로 전체 그래프의 일부를 샘플링함
- 서브 그래프가 전체 그래프 구조의 의미를 추론함에 있어 힌트를 제공한다는 전제가 있음



Experiments

❖ Datasets

- 실험 데이터셋은 크게 자연과학(분자구조)과 인문사회(소셜네트워크) 분야로 구분됨
- 실험 방법은 semi-supervised 방식이며 사전학습 후 전이학습 순서로 진행됨

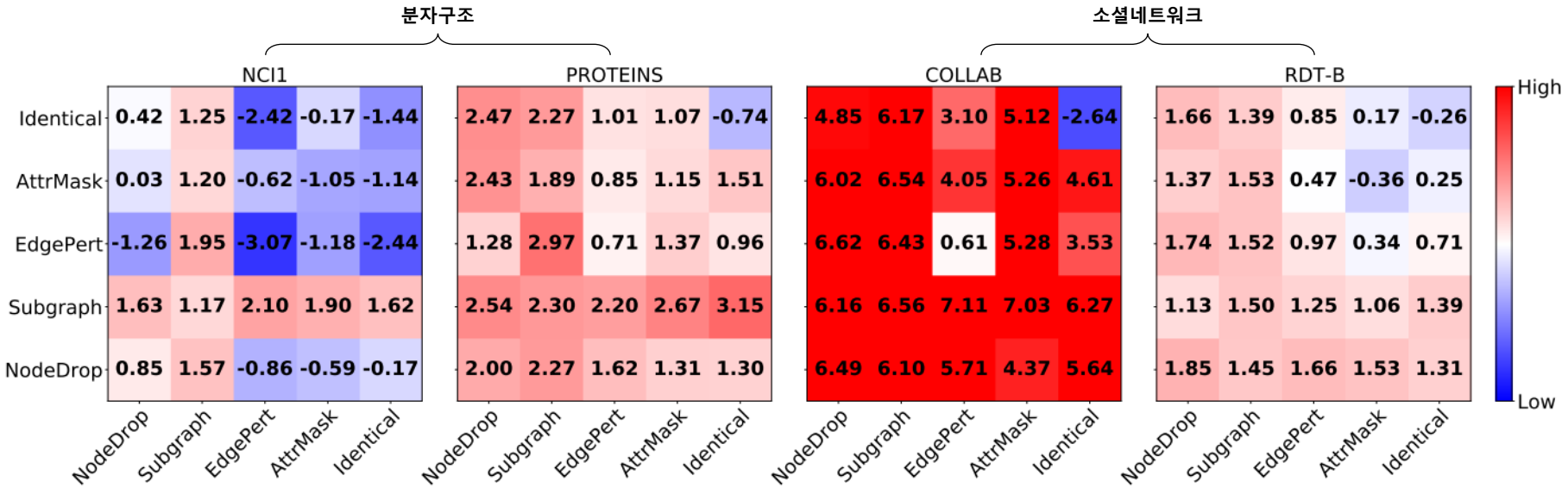
Datasets	Category	Graph Num.	Avg. Node	Avg. Degree
NCI1	Biochemical Molecules	4,110	29.87	1.08
PROTEINS	Biochemical Molecules	1,113	39.06	1.86
COLLAB	Social Networks	5,000	74.49	32.99
RDT-B	Social Networks	2,000	429.63	1.15

Table: Datasets Statistics

Experiments

❖ Augmentations

- 그래프 기반 대조학습(GraphCL)에서 **가장 효과적인 증강기법 조합을 탐색**함
- Node dropping과 subgraph 방식은 모든 데이터셋에서 일반적으로 좋은 성능을 보임
- Attribute masking은 평균 연결성(degree)이 높은 데이터셋(dense graph)에서 더 좋은 성능을 보임
- Edge perturbation은 분자구조 데이터에서는 오히려 성능에 악영향을 미침
  - ✓ 분자구조 데이터에서는 edge 구조의 변경이 그래프의 의미를 훼손하기 때문으로 추정됨



*** Identical은 증강기법을 적용하지 않은 경우를 의미함**

DMQA

Experiments

❖ Semi-supervised learning

- 사전학습 후에 일정 비율의 레이블 데이터만을 가지고 전이학습을 한 비교실험 결과
- 단, baseline과 Aug. 는 사전학습 없이 지도학습을 진행한 결과임
- GCN을 인코더로 사용한 GraphCL이 전반적으로 더 나은 성능을 보여주고 있음

Table 3: Semi-supervised learning with pre-training & finetuning. Red numbers indicate the best performance and the number that overlap with the standard deviation of the best performance (comparable ones). 1% or 10% is label rate; baseline and Aug. represents training from scratch without and with augmentations, respectively.

Dataset	NCI1	PROTEINS	DD	COLLAB	RDT-B	RDT-M5K	GITHUB	MNIST	CIFAR10
1% baseline	60.72±0.45	-	-	57.46±0.25	-	-	54.25±0.22	60.39±1.95	27.36±0.75
1% Aug.	60.49±0.46	-	-	58.40±0.97	-	-	56.36±0.42	67.43±0.36	27.39±0.44
1% GAE	61.63±0.84	-	-	63.20±0.67	-	-	59.44±0.44	57.58±2.07	21.09±0.53
1% Infomax	62.72±0.65	-	-	61.70±0.77	-	-	58.99±0.50	63.24±0.78	27.86±0.43
1% GraphCL	62.55±0.86	-	-	64.57±1.15	-	-	58.56±0.59	83.41±0.33	30.01±0.84
10% baseline	73.72±0.24	70.40±1.54	73.56±0.41	73.71±0.27	86.63±0.27	51.33±0.44	60.87±0.17	79.71±0.65	35.78±0.81
10% Aug.	73.59±0.32	70.29±0.64	74.30±0.81	74.19±0.13	87.74±0.39	52.01±0.20	60.91±0.32	83.99±2.19	34.24±2.62
10% GAE	74.36±0.24	70.51±0.17	74.54±0.68	75.09±0.19	87.69±0.40	53.58±0.13	63.89±0.52	86.67±0.93	36.35±1.04
10% Infomax	74.86±0.26	72.27±0.40	75.78±0.34	73.76±0.29	88.66±0.95	53.61±0.31	65.21±0.88	83.34±0.24	41.07±0.48
10% GraphCL	74.63±0.25	74.17±0.34	76.17±1.37	74.23±0.21	89.11±0.19	52.55±0.45	65.81±0.79	93.11±0.17	43.87±0.77

Experiments

❖ Unsupervised learning

- 사전학습을 한 뒤 전이학습이 아닌 SVM Classifier로 모델의 임베딩 성능을 측정함
- GIN을 인코더로 사용한 GraphCL이 전반적으로 더 나은 성능을 보여주고 있음
- 다만, 매우 작은 그래프로 구성된 데이터셋에서는 비교적 떨어지는 성능을 보여줌

Table 4: Comparing classification accuracy on top of graph representations learned from graph kernels, SOTA representation learning methods, and GIN pre-trained with GraphCL. The compared numbers are from the corresponding papers under the same experiment setting.

Dataset	NCI1	PROTEINS	DD	MUTAG	COLLAB	RDT-B	RDT-M5K	IMDB-B
GL	-	-	-	81.66±2.11	-	77.34±0.18	41.01±0.17	65.87±0.98
WL	80.01±0.50	72.92±0.56	-	80.72±3.00	-	68.82±0.41	46.06±0.21	72.30±3.44
DGK	80.31±0.46	73.30±0.82	-	87.44±2.72	-	78.04±0.39	41.27±0.18	66.96±0.56
node2vec	54.89±1.61	57.49±3.57	-	72.63±10.20	-	-	-	-
sub2vec	52.84±1.47	53.03±5.55	-	61.05±15.80	-	71.48±0.41	36.68±0.42	55.26±1.54
graph2vec	73.22±1.81	73.30±2.05	-	83.15±9.25	-	75.78±1.03	47.86±0.26	71.10±0.54
InfoGraph	76.20±1.06	74.44±0.31	72.85±1.78	89.01±1.13	70.65±1.13	82.50±1.42	53.46±1.03	73.03±0.87
GraphCL	77.87±0.41	74.39±0.45	78.62±0.40	86.80±1.34	71.36±1.15	89.53±0.84	55.99±0.28	71.14±0.44

평균 노드 수가 20개 이하인 매우 작은 그래프로 구성됨

Experiments

❖ Transfer learning

- Strategies for pretraining graph neural networks*에서 제시된 사전학습 방법론과 전이학습 성능을 비교
- 해당 실험은 다양한 데이터셋에서 서로 다른 사전학습 방식의 성능 차이를 비교하려는 것임
- GraphCL에서 제안한 대조학습 방식의 사전학습이 더 좋은 성능을 보여줌

Table 5: Transfer learning comparison with different manually designed pre-training schemes, where the compared numbers are from [9].

Dataset	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BACE	PPI
No Pre-Train	65.8±4.5	74.0±0.8	63.4±0.6	57.3±1.6	58.0±4.4	71.8±2.5	75.3±1.9	70.1±5.4	64.8±1.0
Infomax	68.8±0.8	75.3±0.5	62.7±0.4	58.4±0.8	69.9±3.0	75.3±2.5	76.0±0.7	75.9±1.6	64.1±1.5
EdgePred	67.3±2.4	76.0±0.6	64.1±0.6	60.4±0.7	64.1±3.7	74.1±2.1	76.3±1.0	79.9±0.9	65.7±1.3
AttrMasking	64.3±2.8	76.7±0.4	64.2±0.5	61.0±0.7	71.8±4.1	74.7±1.4	77.2±1.1	79.3±1.6	65.2±1.6
ContextPred	68.0±2.0	75.7±0.7	63.9±0.6	60.9±0.6	65.9±3.8	75.8±1.7	77.3±1.0	79.6±1.2	64.4±1.3
GraphCL	69.68±0.67	73.87±0.66	62.40±0.57	60.53±0.88	75.99±2.65	69.80±2.66	78.47±1.22	75.38±1.44	67.88±0.85

* 배진수 연구원의 발표를 참고
https://github.com/dudwojae/NeverMind_DMQA/blob/main/GraphNeuralNetworks/20220211/%5B20220211%5DSTRATEGIES%20FOR%20PRE-TRAINING%20GRAPH%20NEURAL%20NETWORKS.pdf

Conclusion

❖ Conclusion

- 그래프 데이터셋 역시 레이블이 풍부하지 않다는 문제점을 가지고 있음
- 최근에 유행하고 있는 방법론인 대조 학습을 적용하여 이 문제를 해결하고자 함
- 이미지와는 달리 그래프 분야에서는 적절한 증강기법에 대한 탐색이 이루어지지 않았음
- 도메인에 따라서 증강기법이 의미를 훼손하거나 성능에 민감하게 작용하는 경우가 있었음
- 결론적으로 대조학습은 그래프 문제에서도 충분한 성능을 발휘할 수 있음
- 개인적으로 좋은 논문 하나 써 두면 어디든 참여할 곳이 많다는 생각도 들었음
  (SimCLR를 쓰는 곳에 언제나 등장하는 그 이름 Ting Chen…)

❖ Reference

- You, Yuning, et al. "Graph contrastive learning with augmentations." Advances in Neural Information Processing Systems 33 (2020): 5812-5823.
- Hu, Weihua, et al. "Strategies for pre-training graph neural networks." arXiv preprint arXiv:1905.12265 (2019).


Thank You