Shift-Robust GNNs: Overcoming the Limitations of Localized Graph Training Data

School of Industrial and Management Engineering, Korea University

Sangmin Kim

KOREA UNIVERSITY

DMQA

Contents

Research Purpose

❖ **Shift-Robust GNNs**: Overcoming the Limitations of Localized Graph Training Data(2021, NeurIPS)

- Google Research에서 연구되었으며, 2022년 4월 1일 기준으로 3회 인용됨(Pytorch 기반 코드 공개)

Shift-Robust GNNs: Overcoming the Limitations of Localized Graph Training Data

Qi Zhu* **Natalia Ponomareva†** **Jiawei Han*** **Bryan Perozzi†**

*: University of Illinois Urbana-Champaign
†: Google Research
*{qiz3,hanj}@illinois.edu,
†{nponomareva,bperozzi}@google.com

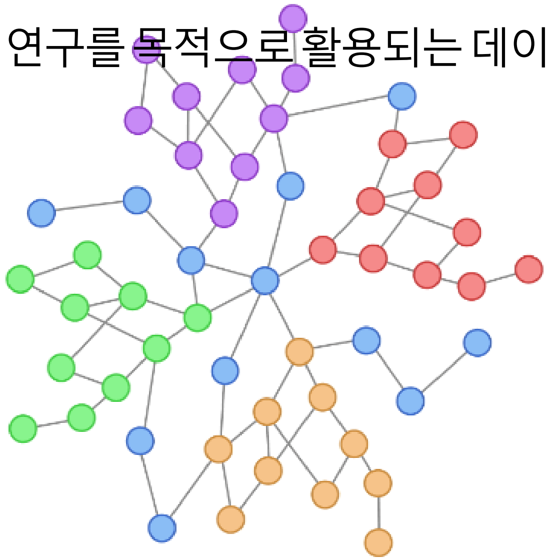
Abstract

There has been a recent surge of interest in designing Graph Neural Networks (GNNs) for semi-supervised learning tasks. Unfortunately this work has assumed that the nodes labeled for use in training were selected uniformly at random (i.e. are an IID sample). However in many real world scenarios gathering labels for graph nodes is both expensive and inherently biased – so this assumption can not be met. GNNs can suffer poor generalization when this occurs, by overfitting to superfluous regularities present in the training data. In this work we present a method, Shift-Robust GNN (SR-GNN), designed to account for distributional differences between biased training data and a graph's true inference distribution. SR-GNN adapts GNN models to the presence of distributional shift between the nodes labeled for training and the rest of the dataset. We illustrate the effectiveness of SR-GNN in a variety of experiments with biased training datasets on common GNN benchmark datasets for semi-supervised learning, where we see that SR-GNN outperforms other GNN baselines in accuracy, addressing at least ∼40% of the negative effects introduced by biased training data. On the largest dataset we consider, `ogb-arxiv`, we observe a 2% absolute improvement over the baseline and are able to mitigate 30% of the negative effects from training data bias [1].

Research Purpose

❖ Introduction

- 일반적인 ML처럼 GNN은 training 을 구성할 때, "IID" 조건 아래 sample을 구성(Unbiased sampling, 영상 자료우측)
    - ✓ https://ai.googleblog.com/2022/03/robust-graph-neural-networks.html(영상 자료)
    - ✓ Open Graph Benchmark: Datasets for Machine Learning on Graphs(2020, NeurIPS)
- 또한, 연구를 목적으로 활용되는 데이터셋에서는 모든 node가 labeled 되어있음



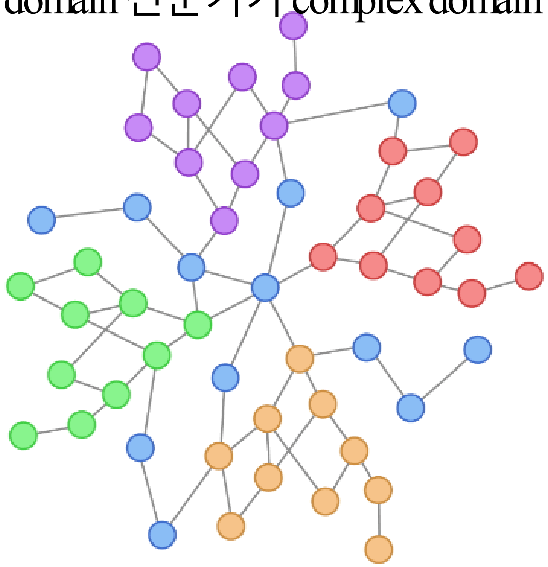
Google AI blog

DMQA

Research Purpose

❖ **Introduction - Biased training data**

- 그러나 real world scenarios 에서는 label 이 없는 데이터가 많으며, 모든 node 에 labeling 하는 것이 어렵고, labeling 을 위해 node를 뽑는 과정 역시 IID 하게 이뤄지지 않기 때문에 biased training data 를 생성하게 됨(Biased sampling, 영상 자료 좌측)
- 또한, domain 전문가가 complex domain knowledge 를 통해 labeling 하면서 biased 되어질 수 있음

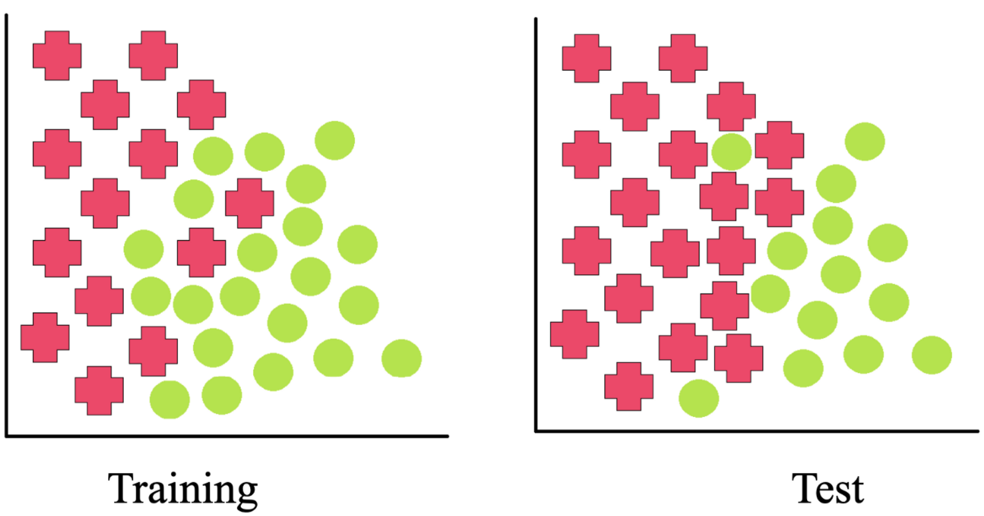


Google AI blog

Research Purpose

❖ Distribution shift

- Biased training data 는 test data 와 다른 distribution 가지므로 "Dataset shift" 문제를 초래함
  - ✓ Data shift : $Pr_{train}(X, Y) \neq Pr_{test}(X, Y)$
- Distribution shift 는 representation shift 를 유발하여 모델 성능 저하
  - ✓ $Pr_{train}(Z, Y) \neq Pr_{test}(Z, Y) \rightarrow Pr_{train}(Z) \neq Pr_{test}(Z) \Longrightarrow$ Representation shift
  - ✓ Z(the output of last activated hidden layer)

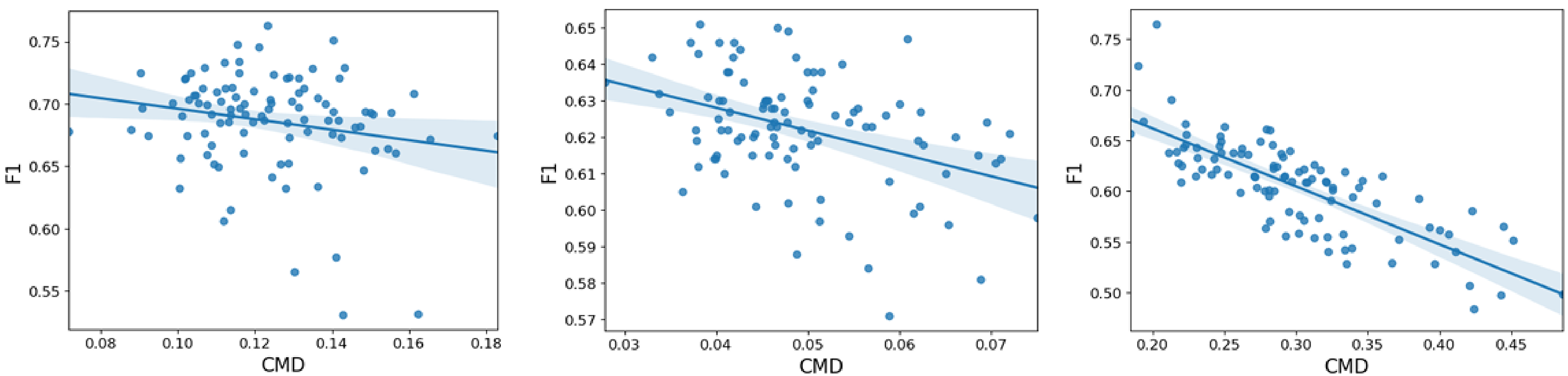


Data shift

Research Purpose

❖ **Distribution shift**

- 따라서, 본 논문은 Biased training data 에 따른 distribution shift 가 GNN에 주는 영향을 파악하여 solution(SR-GNN)을 제시
  - ✓ CMD(Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning, 2017, ICLR) 을 통해 train data 와 test data 차이를 측정
  - ✓ Distribution 차이가 클수록, 즉 training data가 biased 될수록 모델 성능(F1)은 하락함

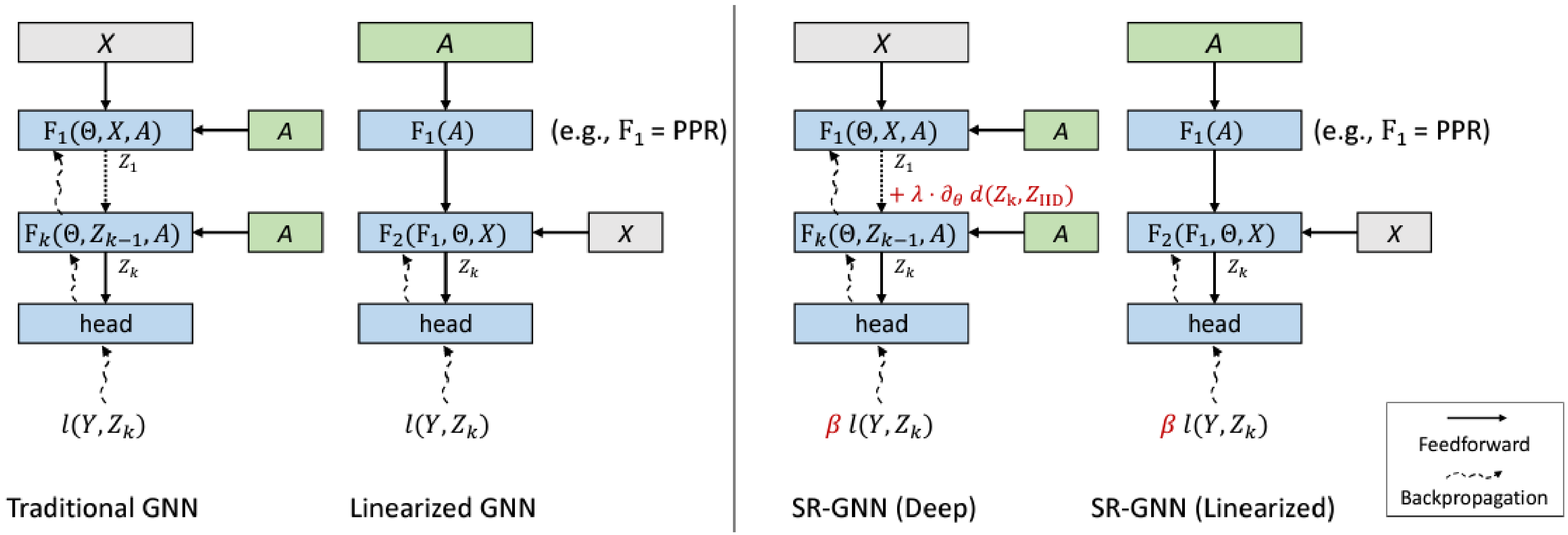


The effect of distribution shift on the Cora, Citeseer, PubMed dataset

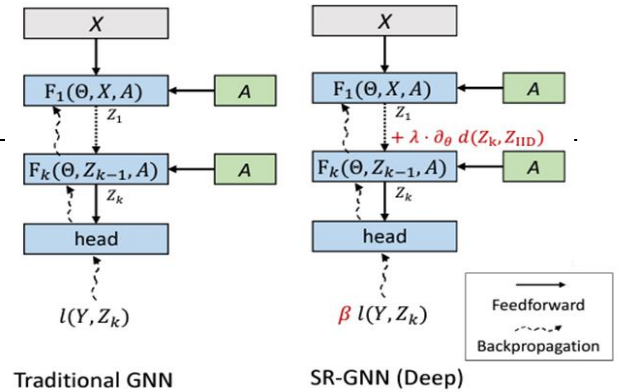
Shift-Robust Graph Neural Networks

❖ Framework

- 본 논문은 두 가지 다른 GNN 모델 계열에서 나타나는 shift 현상을 다룸
  - ✓ Traditional GNN: Basic GCN(2017, Kipf, Thomas N), GAT(2018, Johannes Klicpera)
  - ✓ Linearized GNN: SimpleGCN(2019, Felix Wu), APPNP(2018, Johannes Klicpera)
- Input(X(node feature), A(adjacency matrix)), Output(Z) 은 동일하며, loss term만 변경된 구조



Shift-Robust Graph Neural Networks



❖ **Scenario1: Traditional GNN models(GCN, GAT)**

- ***L*: *cross entropy loss*** 에 regularizer term을 추가
  - ✓ 분포간 shift 된 차이를 측정하기 위해 CMD(Central moment discrepancy (cmd) for domain-invariant representation learning, 2017) 차용, 이때 k는 5를 사용
  - ✓ Z_{train}: **training sample** 에서 biased 하게 추출
  - ✓ Z_{IID}: **training sample + testing samples** iid 하게 추출

$$\mathcal{L} = \boxed{\frac{1}{M}\sum_i l(y_i, z_i)} + \boxed{\lambda \cdot d(Z_{\text{train}}, Z_{\text{IID}})}$$

Cross entropy loss

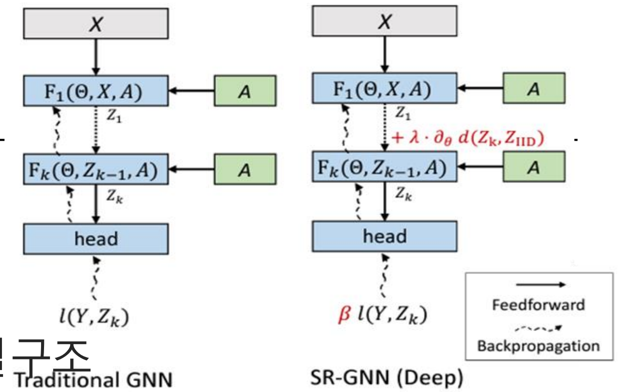
Discrepancy between a biased and unbiased IID sample

$$\boxed{d_{\text{CMD}}(Z_{\text{train}}, Z_{\text{IID}}) = \frac{1}{b-a}\|\mathbf{E}(Z_{\text{train}}) - \mathbf{E}(Z_{\text{IID}})\| + \sum_{k=2}^{\infty}\frac{1}{|b-a|^k}\|c_k(Z_{\text{train}}) - c_k(Z_{\text{IID}})\|}$$

where $\mathbf{E}(Z) = \frac{1}{M}\sum z_i$ and $c_k(Z) = \mathbf{E}(Z - \mathbf{E}(Z))^k$ is the k-order moment (k =5, in paper)

DMQA

Shift-Robust Graph Neural Networks



❖ **Scenario2: Linearized GNN models(SimpleGCN, APPNP)**

- 앞선 전통적인 GCN 계열 모델에서 nonlinearity를 제거한 모델 구조
- ***L*: *cross entropy loss***에 가중치 β를 곱한 손실 함수 형태
  - ✓ kernel mean matching (KMM)을 통해 최적의 β 계산
  - ✓ h_i: **biased training sample**
  - ✓ h_i': **iid sample in training sample + testing sample**

The weight for each training instance

$$\mathcal{L} = \frac{1}{M} \boxed{\beta_i} \boxed{l(y_i, \Phi(h_i))}$$

Cross entropy loss

$$\min_{\beta_i} \left\| \frac{1}{M} \sum_{i=1}^{M} \beta_i \psi(h_i) - \frac{1}{M'} \sum_{i=1}^{M'} \psi(h_i') \right\|^2, \ \textbf{s.t. } B_l \leq \beta < B_u$$

DMQA

Experiments

❖ Biased Training Set Creation Dataset

- Biased training sample을 만들기 위해 Personalized PageRank(PPR) vector를 차용
  - ✓ 특정 seed node 중심으로 인접(nearby) node를 찾는 method
  - ✓ http://dsba.korea.ac.kr/seminar/?mod=document&uid=446 (p42-p57)

Experiments

❖ Experimental settings

- 총 5가지 benckmark datasets 활용하여 모델 성능 확인
  - ✓ Datasets: Cora, Citeseer, Pubmed, ogb-arxiv, Reddit
- 총 6가지 baseline model을 distributional shift가 존재할 때 성능 비교
  - ✓ Traditional GNN Models: GCN, GAT
  - ✓ Linearized GNNs: SGC, APPNP
  - ✓ Unsupervised learning: Deepwalk, DGI
  - ✓ 제안 방법론인 SR-GNN은 APPNP 모델에 scenario1 term: d_{cmd}과 scenario2 term : β 이 포함된 모델

Experiments

❖ Experimental results

- 총 5가지 benckmark datasets 활용하여 모델 성능 확인
- Biased 된 training sample을 통해 각각의 모델을 학습 시킨 후 iid testing sample에 대한 성능
- SR-GNN(Ours) 제안 방법론은 타 방법론 대비 모든 데이터셋에서 우수한 결과를 보임
- IR(scenario1 term: d_{cmd}) 과 Reg.(scenario2 term: β) 에 대한 ablation 을 통해 두가지 term 모두 사용 했을 때 가장 우수한 성능을 보임

Method	Cora			Citeseer			PubMed		
	Micro-F1↑	Macro-F1↑	ΔF1↓	Micro-F1↑	Macro-F1↑	ΔF1↓	Micro-F1↑	Macro-F1↑	ΔF1↓
GCN (IID)	80.8 ± 1.6	80.1 ± 1.3	0	70.3 ± 1.9	66.8 ± 1.3	0	79.8 ± 1.4	78.8 ± 1.4	0
Feat.+MLP	49.7 ± 2.5	48.3 ± 2.2	31.1	55.1 ± 1.3	52.7 ± 1.3	25.2	51.3 ± 2.8	41.8 ± 6.2	28.5
Emb.+MLP	57.6 ± 3.0	56.2 ± 3.0	23.2	38.5 ± 1.2	38.6 ± 1.1	31.8	60.4 ± 2.1	56.6 ± 2.0	19.4
DGI	71.7 ± 4.2	69.2 ± 3.7	9.1	62.6 ± 1.6	60.0 ± 1.6	7.6	58.0 ± 5.3	52.4 ± 8.3	21.8
GCN	67.6 ± 3.5	66.4 ± 3.0	13.2	62.7 ± 1.8	60.4 ± 1.6	7.6	60.6 ± 3.8	56.0 ± 6.0	19.2
GAT	58.4 ± 5.7	58.5 ± 5.0	22.4	58.0 ± 3.5	55.0 ± 2.7	12.3	55.2 ± 3.7	46.0 ± 6.4	14.6
SGC	70.2 ± 3.0	68.0 ± 3.8	10.6	65.4 ± 0.8	62.5 ± 0.8	4.9	61.8 ± 4.5	57.4 ± 7.2	18.0
APPNP	71.3 ± 4.1	69.2 ± 3.4	9.5	63.4 ± 1.8	61.2 ± 1.6	6.9	63.4 ± 4.2	58.7 ± 7.0	16.4
SR-GNN **w.o.** IR	72.1 ± 4.4	69.8 ± 3.7	8.7	63.9 ± 0.7	61.8 ± 0.6	6.4	69.4± 3.4	67.6 ± 4.0	10.4
SR-GNN **w.o.** Reg.	72.0 ± 3.2	69.5 ± 3.7	8.8	66.1 ± 0.9	63.4 ± 0.9	4.2	66.4 ± 4.0	64.0 ± 5.5	13.4
SR-GNN (Ours)	**73.5** ± **3.3**	**71.4**± **3.5**	**7.3**	**67.1** ± **0.9**	**64.0** ± **0.9**	**3.2**	**71.3** ± **2.2**	**70.2** ± **2.4**	**8.5**

Dataset

DMQA

Experiments

❖ Experimental results

- SR-GNN(Ours) 제안 방법론 외 타 방법론을 기준으로 성능 평가
- Biased 된 training sample로 학습 시킬 때, regularizer(scenario1 term: d_{cmd}과 scenario2 term: β) 효과가 큰 것을 확인할 수 있음

Method	Cora Micro-F1↑	Cora Macro-F1↑	Cora Δ(%)	Citeseer Micro-F1↑	Citeseer Macro-F1↑	Citeseer Δ(%)	PubMed Micro-F1↑	PubMed Macro-F1↑	PubMed Δ(%)
GCN (IID)	80.8	80.1	0%	70.3	66.8	0%	79.8	78.8	0%
GCN	67.6	66.4	-12%	62.7	60.4	-8%	60.6	56.0	-19%
SR-GCN	**69.6**	**68.2**	-10%	**64.7**	**62.0**	-6%	**67.0**	**65.2**	-13%
DGI (IID)	80.6	79.3	0%	70.8	66.7	0%	77.6	77.0	0%
DGI	71.7	69.2	-9%	62.6	60.0	-8%	58.0	52.4	-20%
SR-DGI	**74.3**	**72.6**	-6%	**65.8**	**62.6**	-6%	**62.0**	**57.8**	-16%

Comparison of baseline and SR(Shift-Robust) version

Conclusion

❖ Conclusion

- Real world 에서 graph data 형태는 biased 되어 있는 경우가 일반적.
- 따라서, biased 된 data 기준으로 학습한 GNN 모델 성능이 우수해야하나, 그렇지 못함
- Training sample이 biased 될수록 모델 성능은 하락함
- 본 논문은 loss function인 cross entropy에 두가지 regularizer term을 추가하여 문제 상황을 해결함

Reference

❖ Reference

- Zhu, Q., Ponomareva, N., Han, J., & Perozzi, B. (2021). Shift-robust gnns: Overcoming the limitations of localized graph training data. Advances in Neural Information Processing Systems, 34.
- https://ai.googleblog.com/2022/03/robust-graph-neural-networks.html
- https://towardsdatascience.com/understanding-dataset-shift-f2a5a262a766


Thank You