GRADIENT GATING FOR DEEP MULTI-RATE LEARNING ON GRAPHS

김민형@DMLab

Index

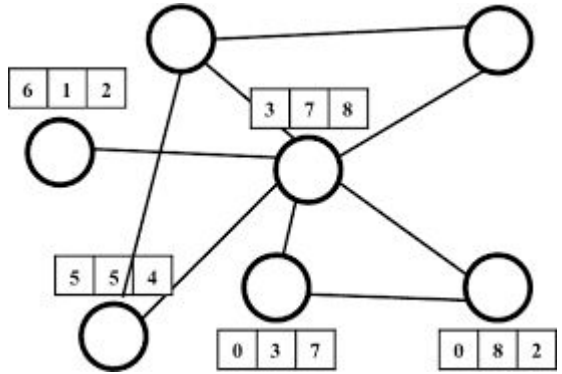
- Problem
- Preliminaries
- Gradient Gating
- Experiments

Problem

- GNN은 그래프 구조 데이터를 처리하는데 좋은 성능을 내는 머신러닝 모델
- 하지만 bottleneck, oversquashing, oversmoothing등 성능 저하를 일으키는 현상이 있음
- 이 논문에서는 oversmoothing을 완화하여 모델의 성능을 높이고자 함

Preliminaries - Graph

- 그래프란 연결 구조를 표현하는데 적합한 자료구조
- 개체를 나타내는 정점과 연결을 나타내는 에지로 구성
- 여러가지 속성 존재 - 정점의 경우 개체의 속성을 나타내는 여러가지 값(문자열, 숫자 등)이 될 수 있으며, 에지의 경우 연결의 성격을 나타내는 값(방향성, 가중치, 거리 등)이 있을 수 있음
- 본 논문은 비방향성 연결 그래프에서 노드에게만 속성이 존재하는 그래프를 다룸

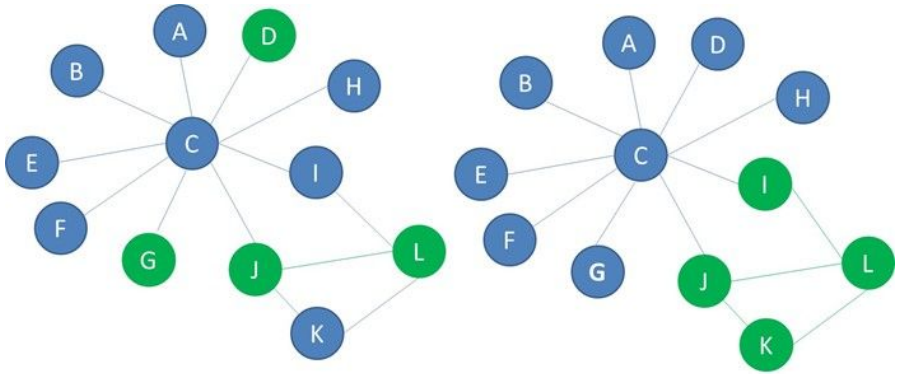


Preliminaries - Graph

- 그래프의 특성을 가지고 여러가지 문제를 만들 수 있음
- Node Classification - 새로운 노드가 들어올때 (혹은 기존 노드의 클래스를 마스킹하고서) 그 노드의 클래스를 그래프의 정보를 이용해 추정함
- Node Regression - 노드의 Feature를 Regression 할 수 있는 모델을 학습함
- Edge Prediction - 그래프의 분포를 바탕으로 잘못된 에지를 제거하거나 누락된 에지를 추가함
- Graph Classification - 객체를 그래프로 표현한 데이터의 경우 그래프가 어떤 객체인지 분류하는 문제가 됨

Preliminaries - Homophilic vs Heterophilic

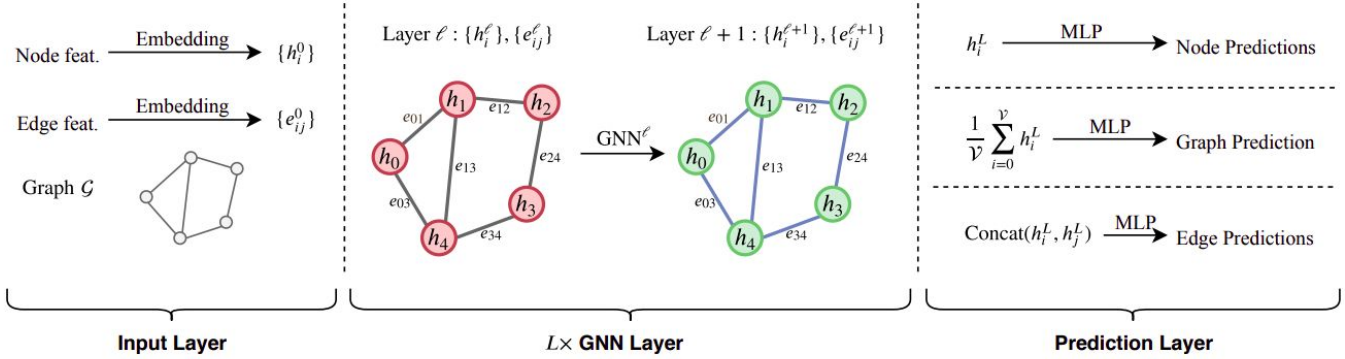
- 그래프의 에지가 다른 클래스의 노드를 연결하는 경우가 많으면 그래프는 Heterophilic임
- 반대로, 에지가 같은 클래스의 노드를 연결하는 경우가 많으면 그래프는 Homophilic임
- 이를 측정하는 다양한 measure가 존재함



<heterophilic graph and homophilic graph>

Preliminaries - GNN

- 그래프의 문제를 해결하는데 여러 방법이 있고, 다른 분야에서 딥러닝이 강력한 성능을 내자 그래프 문제 또한 딥러닝 모델로 해결될 것이라 기대함
- 그래프의 특성 - non euclidean data 이기에 기존 모델을 직접 적용하지 못함
- 초기에는 그래프 스펙트럼 분석 개념을 이용하는 Spectral 모델이 만들어짐
- 최근은 그래프의 연결을 통해 직접 feature를 전달하는 Spacial 모델로 전환
- Spacial 모델은 특징 전파의 특징 때문에 MPNN이라고도 불림



Preliminaries - GCN

- 대표적인 GNN/MPNN 모델
- 전파의 가중치가 Degree와 연결을 기반으로 정해짐
- 주변 노드와 자신의 노드의 값을 이용해 다음 레이어로 출력하는 Feature를 계산하는 것이 Convolution과 유사함

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

Preliminaries - Oversmoothing

- Oversmoothing은 GNN 레이어가 깊어질 수록 출력되는 노드들의 feature가 하나의 값으로 모이는 현상
- 이를 정량적으로 평가하기 위해 저자는 Dirichlet Energy를 사용
- Dirichlet Energy는 이웃과의 평균 거리의 합으로 정의함
- Dirichlet Energy가 지수적으로 감소하거나 지수보다 빠르게 감소하면 Oversmoothing이 발생한다고 할 수 있음
- heterophilic 그래프는 homophilic 그래프 대비 oversmoothing이 더 잘 발생함

$$\mathcal{E}(\mathbf{X}) = \frac{1}{v} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{N}_i} \|\mathbf{X}_i - \mathbf{X}_j\|^2.$$

Gradient Gating

- Gradient Gating의 직관은 다음 생각으로부터 출발했다.
- Residual MPNN의 일반적인 모델은 다음과 같다.
- $\mathbf{X}^n = \mathbf{X}^{n-1} + \sigma(\mathbf{F}_\theta(\mathbf{X}^{n-1}, \mathcal{G}))$
- 이 모델에서 모는 노드는 농일한 속도로 가중치를 갱신한다.
- 하지만 노드에 따라 받는 정보량이 다르기 때문에 갱신 속도 또한 달라질 필요가 있을 것이다.
- 그래서 갱신할 때 각 노드마다 다른 rate를 적용하는 모델을 생각함
- $\mathbf{X}^n = (1 - \boldsymbol{\tau}^n) \odot \mathbf{X}^{n-1} + \boldsymbol{\tau}^n \odot \sigma(\mathbf{F}_\theta(\mathbf{X}^{n-1}, \mathcal{G}))$
- T는 각 노드 피쳐를 얼마나 업데이트할 지 결정하는 행렬
- X와 같은 차원을 가지며 [0,1]의 값을 가지며, 요소별 곱을 적용함

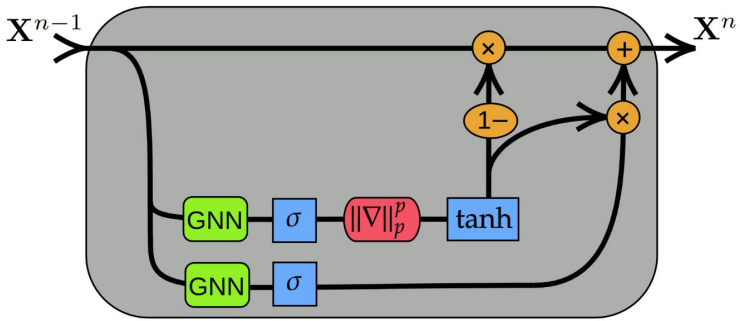
Gradient Gating

- 이때 T를 고정하지 않고 노드 feature와 연결로부터 T를 학습하도록 구성함
- $\boldsymbol{\tau}^n(\mathbf{X}^{n-1}, \mathcal{G}) = \bar{\sigma}(\hat{\mathbf{F}}_{\hat{\theta}}(\mathbf{X}^{n-1}, \mathcal{G}))$
- 이 모델이 반드시 oversmoothing을 방지하지 않으므로 제약을 추가함
- 핵심은 T의 값을 설정할 때 Dirichlet Energy를 이용함
- 만약 해당 노드의 Dirichlet Energy가 0에 가까워지면 이를 더 이상 업데이트 하면 안됨

$$\hat{\boldsymbol{\tau}}^n = \sigma(\hat{\mathbf{F}}_\theta(\mathbf{X}^{n-1}, \mathcal{G})),$$

$$\boldsymbol{\tau}_{ik}^n = \tanh\left(\sum_{j \in \mathcal{N}_i} |\hat{\boldsymbol{\tau}}_{jk}^n - \hat{\boldsymbol{\tau}}_{ik}^n|^p\right),$$

$$\mathbf{X}^n = (1 - \boldsymbol{\tau}^n) \odot \mathbf{X}^{n-1} + \boldsymbol{\tau}^n \odot \sigma(\mathbf{F}_\theta(\mathbf{X}^{n-1}, \mathcal{G}))$$



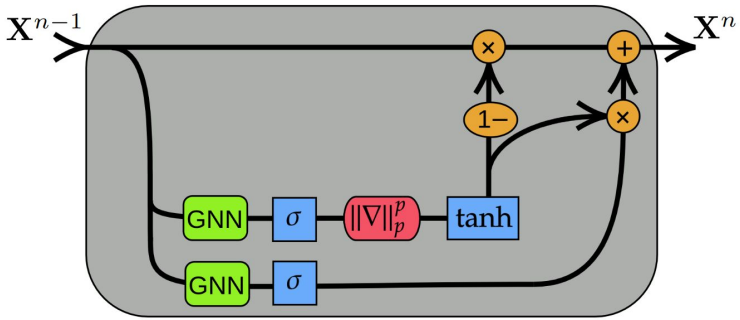
Gradient Gating

- 이때 그래프의 Dirichlet Energy는 그래디언트의 합으로도 해석할 수 있음
- $(\nabla \mathbf{y})_{ij} = \mathbf{y}_j - \mathbf{y}_i$
- Gradient Gating - 그래디언트를 통해 정보의 전파를 차단한다는 의미

$$\hat{\boldsymbol{\tau}}^n = \sigma(\hat{\mathbf{F}}_\theta(\mathbf{X}^{n-1}, \mathcal{G})),$$

$$\boldsymbol{\tau}_{ik}^n = \tanh\left(\sum_{j \in \mathcal{N}_i} |\hat{\boldsymbol{\tau}}_{jk}^n - \hat{\boldsymbol{\tau}}_{ik}^n|^p\right),$$

$$\mathbf{X}^n = (1 - \boldsymbol{\tau}^n) \odot \mathbf{X}^{n-1} + \boldsymbol{\tau}^n \odot \sigma(\mathbf{F}_\theta(\mathbf{X}^{n-1}, \mathcal{G}))$$

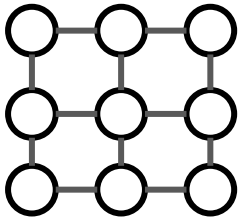


Experiments

- 다음 질문들에 대해 실험을 함
- G2가 Oversmoothing을 완화하는가?
- G2가 Node Classification을 할때 Oversmoothing을 완화하는가?
- G2가 Node Regression을 할때 Oversmoothing을 완화하는가?
- G2가 Homophily가 다를 때 어떠한 영향을 끼치는가?

Experiments

- G2가 Oversmoothing을 완화하는가?
- 10x10 regular grid, X are ramdomly sampled from uniform(0~1) (1dim)의 그래프에서 실험함
- G2를 적용하는 경우 Dirichlet Energy가 (거의) 상수로 유지됨
- 적용하지 않은 경우 급격하게 감소함



<3x3 regular grid graph>

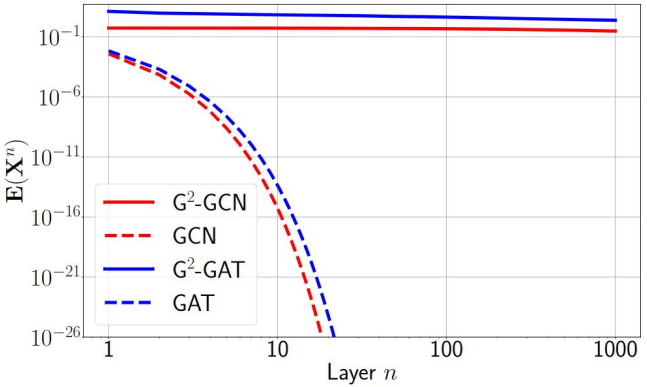


Figure 2: Dirichlet energy $\mathcal{E}(\mathbf{X}^n)$ of layer-wise node features $\mathbf{X}^n$ propagated through a GAT, GCN and their gradient gated versions ($\mathbf{G}^2$-GAT, $\mathbf{G}^2$-GCN).

Experiments

- G2가 Node Classification을 할때 Oversmoothing을 완화하는가?
- DropEdge와 GraphCON은 G2처럼 Oversmoothing을 완화하는 목적의 모델
- Cora Citation Graph에서 Node Classification
- GCN을 그냥 사용하는 경우 레이어가 증가하면 정확도가 떨어짐
- DropEdge, GraphCON은 레이어의 증가에도 정확도가 덜 감소함
- G2는 레이어의 증가에 따라 **정확도 증가**

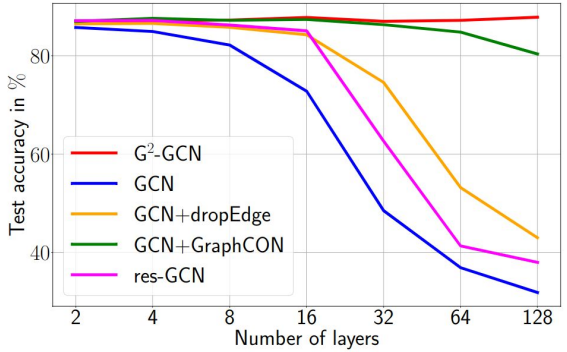


Figure 3: Test accuracies of GCN with gradient gating ($\mathbf{G}^2$-GCN) as well as plain GCN and GCN combined with other methods on the Cora dataset for increasing number of layers.

Dataset	Type	Classes	Features	Nodes	Edges
Cora	citation	7	1433	2485	5069

Experiments

- G2가 Node Regression을 할때 Oversmoothing을 완화하는가?
- Chameleon and Squirrel 데이터셋은 웹페이지 간 연결과 이진 클래스를 가지는 Heterophilic 데이터 세트이지만, 원본 데이터는 트래픽 데이터를 가지고 있음
- 이를 Normalize하여 Regression 데이터셋으로 활용함
- MSE를 비교한 결과 G2를 적용하지 않았을 때 보다 적용했을때 MSE가 감소하는 것을 볼 수 있음

Table 1: Normalized test MSE on multi-scale node-level regression tasks.

	Chameleon	**Squirrel**
#Nodes	2,277	5,201
#Edges	31,421	198,493
GCNII	$\mathbf{0.170 \pm 0.034}$	$\mathbf{0.093 \pm 0.031}$
PairNorm	0.207 ± 0.038	0.140 ± 0.040
GCN	0.207 ± 0.039	0.143 ± 0.039
GAT	0.207 ± 0.038	0.143 ± 0.039
$\mathbf{G^2}$-GCN	$\mathbf{0.137 \pm 0.033}$	$\mathbf{0.070 \pm 0.028}$
$\mathbf{G^2}$-GAT	$\mathbf{0.136 \pm 0.029}$	$\mathbf{0.069 \pm 0.029}$

Experiments

- G2가 Homophily가 다를 때 어떠한 영향을 끼치는가?
- 실험은 의사 Cora 데이터셋으로 진행됨
- Homophily가 1일때 모든 모델이 좋은 정확도를 냄
- Homophily가 감소할 수록 일반 모델은 정확도가 비례하여 떨어짐
- 하지만 G2를 적용한 경우 80% 근처에서 정확도를 유지함

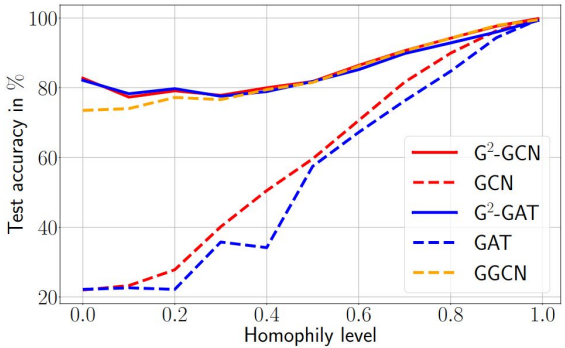


Figure 5: Test accuracy of GCN and GAT with / without gradient gating ($\mathbf{G}^2$) on synthetic Cora with a varying level of true label homophily.

Questions?