All-hands Meeting II
2024 Winter

WalkLM: A Uniform Language Model Fine-tuning Framework for Attributed Graph Embedding

Presenter: Minseo Yoon
(cooki0615@korea.ac.kr)

Introduction

WalkLM: A Uniform Language Model Fine-tuning Framework for Attributed Graph Embedding

Yanchao Tan
College of Computer and Data Science
Fuzhou University
Fuzhou, China
yctan@fzu.edu.cn

Zihao Zhou
College of Computer and Data Science
Fuzhou University
Fuzhou, China
reviverkey@gmail.com

Hang Lv
College of Computer and Data Science
Fuzhou University
Fuzhou, China
lvhangkenn@gmail.com

Weiming Liu
College of Computer Science
Zhejiang University
Hangzhou, China
21831010@zju.edu.cn

Carl Yang*
Department of Computer Science
Emory University
Atlanta, United States
j.carlyang@emory.edu

Introduction

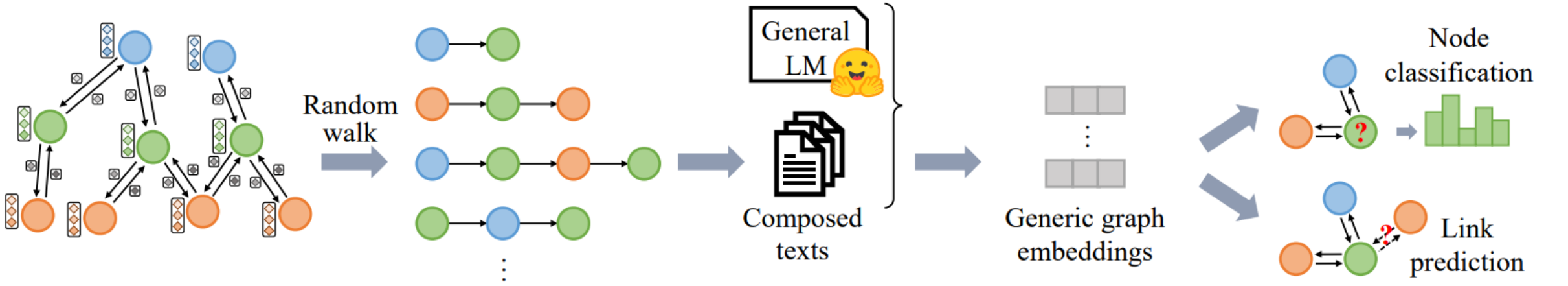
- Graph Neural Networks를 downstream task에 활용할 때 충분한 학습이 없으면 강력한 성능을 달성하기 어려움
- 다양한 downstream task에서 GNN의 unsupervised training은 만족스럽지 못한 성능을 보임
- 복잡한 node/link attribute를 이해하고 유연한 그래프 구조를 비지도 방식으로 통합하는 general-purpose graph embedding method가 필요

Introduction

- WalkLM은 특정 downstream task에 제한되지 않고 비지도적으로 일반적인 그래프 표현을 얻고자 함

- Language Models
  - Node와 link의 complex heterogeneous attribute의 미묘한 의미를 포착하는 모델
- Random Walks
  - 그래프의 유연한 topological structure를 포착하는 방법론

Methodology



(a) Attributed RW-based textualization program (b) Graph-aware LM fine-tuning (c) Downstream tasks

Methodology

- Entity-level Textualization
  - Rule-based program function $\mathcal{P}(\cdot)$
  - Node: $\Phi(v_i) = \{age : 35,\ sex : \text{female},\ pid : \text{P246}\}$

    $\mathcal{P}(v_i) = < \text{A 35-year-old female patient P246} >$
  - Edge: simple relation edges (has, including, ⋯)

Methodology

- Walk-level Textualization
  - Standard RW with a uniform probability
  - $W = \{\mathcal{P}(v_0), \mathcal{P}(e_1), \mathcal{P}(v_1), \ldots, \mathcal{P}(v_{L-1}), \mathcal{P}(e_L), \mathcal{P}(v_L)\}$

- 예시 - PubMed

```
proton_pump causing sphincteric_dysfunction and SIV-infected and schizotypal_symptoms
and SIV-infected and impairment and SIV-infected and increased_oxygen_demand and SIV-
infected and Scale-Depression and DBD and Scale-Depression and psychological_pain
```

Methodology

- Discussion
  - RW는 그래프의 특징적인 특성을 요약하고 node 간의 network proximity를 재구성할 수 있음
    - 다양한 node에서 충분한 양의 attributed RW가 그래프의 topological structures를 잘 반영함
  - RW-based textualization은 node와 edge의 attribute를 완전히 보존
    - 이는 LM이 이해하기 자연스러운 의미 있는 텍스트의 형태

Methodology

- Graph-Aware LM Fine-Tuning
  - Masked Language Modeling
  - LM: DistillRoBERTa

$$\mathcal{T} = \{t_1, t_2, \ldots, t_K, < \text{MASK} >_1, < \text{MASK} >_2, \ldots, < \text{MASK} >_{|\mathcal{M}|}\}$$

$$\mathcal{L}_{FT}(\Theta) = -\frac{1}{|\mathbb{X}|} \sum_{X_i \in \mathbb{X}} \left[\sum_{t_k^* \in \mathcal{M}} \log \frac{\exp(Sim(t_k, t_k^*))}{\sum_{t \in \mathcal{V}} \exp(Sim(t_k, t))} \right]$$

Experiments

- Task
  - Node Classification
  - Link Prediction

Dataset	#attribute type	#node type	#node	#link type	# link	#label	#label node
PubMed	8	4	63,109	10	244,986	8	454
MIMIC-III	10	3	32,267	4	559,290	19	4880

Experiments

- Node Classification
  - GNN의 사용은 의미가 거의 없음
  - LM의 지식이 가장 중요한 요소
  - SimKGC
    - BERT의 지식을 이용한 contrastive learning
  - WalkLM
    - SimKGC에 비해 semantic한 정보와 graph structure를 더 잘 결합함

Task	Node Classification				Link Prediction			
Dataset	PubMed		MIMIC-III		PubMed		MIMIC-III	
Metric	Macro-F1	Micro-F1	Macro-F1	Micro-F1	AUC	MRR	AUC	MRR
M2V	15.35 (±1.27)	20.27 (±3.01)	19.69 (±0.62)	29.24 (±1.57)	74.53 (±3.79)	89.58 (±2.05)	75.05 (±0.41)	88.32 (±0.23)
HIN2Vec	11.57 (±1.23)	18.92 (±2.78)	19.12 (±1.32)	28.05 (±1.44)	74.21 (±5.49)	90.56 (±1.06)	73.46 (±0.41)	88.10 (±0.14)
ConvE	16.06 (±3.69)	19.16 (±4.00)	24.44 (±1.28)	32.89 (±0.86)	76.48 (±4.31)	92.27 (±0.57)	69.56 (±0.36)	84.88 (±0.25)
ComplEx	13.93 (±2.59)	18.27 (±4.12)	9.82 (±0.56)	21.39 (±3.12)	79.81 (±0.97)	91.79 (±0.48)	63.86 (±0.42)	81.40 (±0.40)
SimKGC	21.97 (±3.51)	30.83 (±3.10)	51.62 (±1.81)	58.50 (±1.52)	79.62 (±2.72)	91.43 (±0.48)	67.73 (±1.69)	84.86 (±0.54)
RGCN	12.50 (±2.36)	18.50 (±1.41)	7.19 (±0.77)	14.55 (±3.25)	72.08 (±1.13)	88.20 (±0.47)	57.31 (±0.71)	73.91 (±0.57)
HAN	15.29 (±2.87)	16.95 (±2.71)	6.98 (±0.58)	14.73 (±1.69)	70.57 (±1.58)	87.89 (±0.62)	- -	- -
HGT	11.98 (±2.23)	20.12 (±3.89)	8.03 (±0.87)	17.79 (±0.83)	77.24 (±3.50)	89.63 (±0.84)	64.01 (±0.36)	81.54 (±0.56)
HeCo	10.32 (±1.12)	18.01 (±0.87)	10.78 (±0.41)	15.26 (±1.52)	65.04 (±1.26)	83.29 (±0.72)	53.13 (±0.47)	71.81 (±0.35)
SHGP	10.80 (±3.03)	19.28 (±0.91)	11.34 (±1.29)	17.44 (±1.49)	68.22 (±2.71)	85.34 (±0.48)	54.49 (±0.33)	72.58 (±0.24)
LM (XRoBERTa)	40.10 (±4.62)	44.71 (±3.68)	54.51 (±1.50)	61.27 (±1.22)	60.20 (±2.78)	84.23 (±1.71)	51.21 (±0.17)	74.22 (±0.26)
LM (GPT-2)	59.43 (±4.73)	61.53 (±3.43)	70.26 (±1.43)	72.67 (±0.90)	51.71 (±3.67)	80.54 (±2.49)	50.66 (±0.74)	72.36 (±0.86)
LM (DRoBERTa)	58.29 (±2.44)	60.57 (±2.11)	66.25 (±1.60)	70.14 (±1.52)	60.97 (±2.98)	83.00 (±0.40)	51.44 (±0.14)	75.09 (±0.29)
LM +RGCN	13.83 (±0.73)	22.70 (±3.25)	14.32 (±0.87)	24.59 (±1.17)	72.35 (±4.34)	88.86 (±1.46)	58.62 (±0.50)	78.78 (±0.10)
LM +HGT	12.81 (±1.22)	21.79 (±3.54)	10.49 (±0.41)	20.57 (±0.97)	82.97 (±3.91)	89.98 (±0.88)	65.01 (±0.20)	82.28 (±0.30)
WalkLM	**60.42*** (±2.62)	**62.33*** (±3.13)	**75.16*** (±0.93)	**77.89*** (±0.70)	**85.65*** (±3.28)	**94.16*** (±0.37)	**82.15*** (±0.67)	**92.78*** (±0.68)

Experiments

- Link Prediction
  - HeCo, SHGP
    - 비지도 학습 기반이기 때문에 link prediction과 mechanism이 일치하지 않아 낮은 성능을 보임
  - metapath2vec, ConvE, ComplEx
    - 복잡한 관계를 학습할 수 있는 방법론을 채택한 덕분에 높은 성능을 보임
  - WalkLM
    - RW을 통해 node 간의 proximity를 재구성하여 복잡환 관계를 효과적으로 학습

Task	Node Classification				Link Prediction			
Dataset	PubMed		MIMIC-III		PubMed		MIMIC-III	
Metric	Macro-F1	Micro-F1	Macro-F1	Micro-F1	AUC	MRR	AUC	MRR
M2V	15.35 (±1.27)	20.27 (±3.01)	19.69 (±0.62)	29.24 (±1.57)	74.53 (±3.79)	89.58 (±2.05)	75.05 (±0.41)	88.32 (±0.23)
HIN2Vec	11.57 (±1.23)	18.92 (±2.78)	19.12 (±1.32)	28.05 (±1.44)	74.21 (±5.49)	90.56 (±1.06)	73.46 (±0.41)	88.10 (±0.14)
ConvE	16.06 (±3.69)	19.16 (±4.00)	24.44 (±1.28)	32.89 (±0.86)	76.48 (±4.31)	92.27 (±0.57)	69.56 (±0.36)	84.88 (±0.25)
ComplEx	13.93 (±2.59)	18.27 (±4.12)	9.82 (±0.56)	21.39 (±3.12)	79.81 (±0.97)	91.79 (±0.48)	63.86 (±0.42)	81.40 (±0.40)
SimKGC	21.97 (±3.51)	30.83 (±3.10)	51.62 (±1.81)	58.50 (±1.52)	79.62 (±2.72)	91.43 (±0.48)	67.73 (±1.69)	84.86 (±0.54)
RGCN	12.50 (±2.36)	18.50 (±1.41)	7.19 (±0.77)	14.55 (±3.25)	72.08 (±1.13)	88.20 (±0.47)	57.31 (±0.71)	73.91 (±0.57)
HAN	15.29 (±2.87)	16.95 (±2.71)	6.98 (±0.58)	14.73 (±1.69)	70.57 (±1.58)	87.89 (±0.62)	- -	- -
HGT	11.98 (±2.23)	20.12 (±3.89)	8.03 (±0.87)	17.79 (±0.83)	77.24 (±3.50)	89.63 (±0.84)	64.01 (±0.36)	81.54 (±0.56)
HeCo	10.32 (±1.12)	18.01 (±0.87)	10.78 (±0.41)	15.26 (±1.52)	65.04 (±1.26)	83.29 (±0.72)	53.13 (±0.47)	71.81 (±0.35)
SHGP	10.80 (±3.03)	19.28 (±0.91)	11.34 (±1.29)	17.44 (±1.49)	68.22 (±2.71)	85.34 (±0.48)	54.49 (±0.33)	72.58 (±0.24)
LM (XRoBERTa)	40.10 (±4.62)	44.71 (±3.68)	54.51 (±1.50)	61.27 (±1.22)	60.20 (±2.78)	84.23 (±1.71)	51.21 (±0.17)	74.22 (±0.26)
LM (GPT-2)	59.43 (±4.73)	61.53 (±3.43)	70.26 (±1.43)	72.67 (±0.90)	51.71 (±3.67)	80.54 (±2.49)	50.66 (±0.74)	72.36 (±0.86)
LM (DRoBERTa)	58.29 (±2.44)	60.57 (±2.11)	66.25 (±1.60)	70.14 (±1.52)	60.97 (±2.98)	83.00 (±0.40)	51.44 (±0.14)	75.09 (±0.29)
LM +RGCN	13.83 (±0.73)	22.70 (±3.25)	14.32 (±0.87)	24.59 (±1.17)	72.35 (±4.34)	88.86 (±1.46)	58.62 (±0.50)	78.78 (±0.10)
LM +HGT	12.81 (±1.22)	21.79 (±3.54)	10.49 (±0.41)	20.57 (±0.97)	82.97 (±3.91)	89.98 (±0.88)	65.01 (±0.20)	82.28 (±0.30)
WalkLM	**60.42*** (±2.62)	**62.33*** (±3.13)	**75.16*** (±0.93)	**77.89*** (±0.70)	**85.65*** (±3.28)	**94.16*** (±0.37)	**82.15*** (±0.67)	**92.78*** (±0.68)

Experiments

- Few-shot setting

Dataset	PubMed						MIMIC-III					
Setting	1 shot		3 shot		5 shot		1 shot		3 shot		5 shot	
Metric	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1	Ma-F1	Mi-F1
ComplEx	9.31	12.51	10.32	13.26	10.12	15.94	2.82	5.29	2.00	3.03	3.87	9.26
M2V	9.86	13.42	10.27	12.56	12.97	14.98	5.83	8.72	3.91	5.11	3.40	4.49
ConvE	13.23	13.45	8.84	10.93	11.25	13.53	5.85	6.75	5.61	7.24	6.31	7.69
RGCN	9.34	11.02	8.57	10.58	10.84	13.43	4.97	5.82	5.43	6.28	5.22	5.73
HIN2Vec	8.46	10.54	9.04	12.79	10.96	17.39	5.72	10.33	4.90	5.72	3.57	4.91
SHGP	8.94	12.79	9.12	11.73	10.53	15.14	4.12	6.35	5.36	6.58	4.47	5.34
WalkLM	**28.09***	**30.94***	**32.11***	**35.35***	**35.41***	**37.68***	**23.33***	**27.96***	**34.19***	**40.49***	**41.12***	**46.83***

  - RW 기반의 textualization이 LM의 지식을 활용한 graph representation learning을 가능하게 함

Progress Report

- 재구현 결과
  - Node Classification: Macro-F1: 61.62, Micro-F1: 63.88
  - Link Prediction: AUC: 84.48, MRR: 93.38

WalkLM	**60.42*** (±2.62)	**62.33*** (±3.13)	**75.16*** (±0.93)	**77.89*** (±0.70)	**85.65*** (±3.28)	**94.16*** (±0.37)	**82.15*** (±0.67)	**92.78*** (±0.68)

Progress Report

- 실험 계획

1. Textualization
   - Entity-level: Rule-based → NLG (in-context or PEFT)
     - MLM Loss를 사용하는 부분과 textualization을 위한 NLG를 한 모델에서 학습
       → Encoder-Decoder based LM (T5, BART, …)
       → DistillRoBERTa + Decoder based LM (LLaMA, Solar, …) 2개의 모델을 따로 사용
   - Walk-level: Uniform RW → RW 학습
     - DeepWalk (contrastive loss), JK-Net, …

2. Off-the-shelf LLM의 지식을 활용하는 방법

Questions?