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

A Appendix

A.1 Experiments based on KG datasets

Datasets. We conduct extensive experiments on two new real-world KG datasets, i.e., Freebase [2] and FB15K-237 [3]. Freebase contains a graph of books, films, sports, and locations. The nodes and edges are extracted according to [5]. A large portion of books are labeled into eight genres of literature. Each labeled book has only one label. FB15K-237 is a standard dataset in the knowledge graph community, which contains 310,116 triples with 14,541 entities and 237 relation types. Since we did not manually label the nodes, we only predicted whether a triple is correct or not on this dataset. We matched the entities with Wikidata entities and obtained metadata from Wikidata, and constructed a rough attribute graph dataset by using the names and descriptions of the nodes as textualized features of the nodes, and directly applying their original relationship text as the edge textualized attributes.

Node Classification. As shown in Table 1, our proposed WalkLM has superior performance, indicating the importance of leveraging both semantic and structural information in attributed graphs. WalkLM achieves 40.24% performance gains on Freebase over the second-best performance on average. Specifically, as a text-based Knowledge graph completion method, SimKGC can effectively employ text-based contrastive learning to capture a rich set of semantic information. Compared with SimKGC, WalkLM can effectively combine the complex semantic and graph structure information of attributed graphs, so as to accurately model the complex attributes of nodes.

Link Prediction. We evaluate WalkLM with AUC and MRR. As shown in Table 1, our fine-tuned WalkLM demonstrates outstanding performance in uncovering latent associations among nodes in attributed graphs. In general, WalkLM outperforms all ten baselines with an average of 2.05% performance gain over the second-best performance, showing that our proposed framework can learn accurate edge representation for link prediction. Complex and ConvE consistently demonstrate promising performance by effectively capturing generic node representations. However, as a text-based Knowledge graph completion method, SimKGC can sometimes outperform others in terms of the MRR metric, where SimKGC can enhance semantic similarity between nodes through contrastive

*Corresponding author

[2] http://www.freebase.com/

[3] https://paperswithcode.com/dataset/fb15k-237

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table 1: Different downstream task results (%) with the corresponding std (±) on two KG datasets. The best performances are in bold and the second runners are shaded in gray, where * denotes a significant improvement according to the Wilcoxon signed-rank significance test.

Task	Node Classification		Link Prediction			
Dataset	Freebase		Freebase		FB15K-237	
Metric	Macro-F1	Micro-F1	AUC	MRR	AUC	MRR
M2V	25.74±1.12	50.25±2.57	80.68±1.81	88.97±0.93	90.35±0.50	96.78±0.19
HIN2Vec	15.56±1.07	43.67±2.12	80.04±3.01	90.90 ±0.87	79.68±0.83	92.85±0.40
ConvE	25.13±1.83	49.31±3.45	88.14±1.03	93.57±0.42	92.88±0.42	97.57±0.15
ComplEx	20.25±1.62	49.43±3.57	84.01±1.43	91.46±0.56	95.03±0.35	97.88±0.22
SimKGC	35.88±0.87	56.12±0.45	87.33±1.51	94.21±0.34	93.80±0.31	97.62±0.30
RGCN	15.37±1.54	45.86±1.03	82.75±0.89	91.52±0.64	85.88±0.35	89.84±0.19
HAN	14.25±1.77	39.30±2.18	80.73±1.37	91.61±0.34	82.06±0.53	89.31±0.89
HGT	19.97±1.34	47.99±2.56	81.94±1.84	89.65±0.43	87.41±0.69	94.62±0.34
HeCo	23.95±1.45	48.62±1.13	79.32±0.86	87.40±0.32	78.82±0.37	90.41±0.23
SHGP	13.83±1.27	39.07±1.39	78.37±1.77	85.52±0.69	78.56±0.33	89.84±0.21
XRoBERTa	48.10±2.01	67.95±0.97	73.94±1.62	88.17±0.91	75.62±0.72	91.10±0.71
GPT-2	49.24±2.12	68.28±1.37	60.45±2.43	83.29±1.87	68.87±1.21	85.23±1.73
DRoBERTa	51.76±1.24	69.51±0.73	79.22±1.85	91.21±1.17	84.15±0.63	93.39±0.39
LM+RGCN	28.38±0.63	53.37±2.27	83.63±1.81	96.38±0.67	87.72±0.50	94.47±0.46
LM+HGT	20.79±0.67	48.73±3.13	83.09±1.23	89.79±0.35	88.18±0.61	94.85±0.27
`WalkLM`	**55.01±2.67***	**71.36±1.53***	**92.11±2.24***	**96.54±0.56***	**95.65±0.18***	**98.45±0.33***

Table 2: Accuracy results (%) of graph-level classification on MUTAG.

Dataset	MUTAG					
Model	HIN2Vec	ConvE	ComplEx	LM (DRoBERTa)	`WalkLM` w/o. graph-ID	`WalkLM`
Accuracy	78.72	77.64	78.69	79.23	79.77	**81.39***

learning based on bi-encoder architecture and three types of negatives. Compared with ConvE, ComplEX, and SimKGC, `WalkLM` can effectively capture the complex relations by providing text-based semantic traits of characteristic graph and reconstructing network proximity of nodes that inherit from RWs.

A.2 Graph-level Classification

Compared with node or edge classification, aggregating node embeddings for graph-level classification needs more context information. Furthermore, graph-level classification presents its own set of challenges, which require holistic capturing of graph structures and often do not rely much on attributes. Therefore, it is difficult to find a universal representation learning approach that solves all different levels of graph mining tasks. Technically, adapting our method to graph-level classification necessitates some subtle decisions to make (such as whether to include graph ID as a virtual node). We've conducted a preliminary analysis on aggregating our learned node embeddings for graph-level tasks. Specifically, we adopt a widely-used MUTAG [4] dataset and use mean accuracy as the metric [3, 6]. The results on the popular MUTAG dataset are listed in Table 2. Although the findings are encouraging and show the potential of `WalkLM`, further studies are still needed to establish a clear advantage of our approach over SOTA graph classification baselines.

A.3 Detailed Ablation Studies

From Table 3, we have the following observations: (1) Compared with the graph-based baselines, the LM-based models (e.g., LM (XRoBERTa), LM (GPT-2), and LM (DRoBERTa)) are able to learn accurate and rich node attributes, leading to superior performance in node classification. For the PubMed dataset distributed on 8 classes, LM (XRoBERTa), LM (GPT-2), and LM (DRoBERTa) achieve 63.78%, 135.04%, and 130.89% performance gains over the second-best performance on average, respectively. For the MIMIC-III dataset on the total 19 classes, LM (XRoBERTa), LM

[4] https://ls11-www.cs.tu-dortmund.de/staff/morris/graphkerneldatasets

Table 3: The detailed ablation results (%) with the corresponding std (±) on two datasets. The best performances are in bold and the second runners are shaded in gray, where * denotes a significant improvement according to the Wilcoxon signed-rank significance test.

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)

(GPT-2), and LM (DRoBERTa) achieve 5.17%, 30.17%, and 20.12% average performance gains compared to the second-best performance, respectively.

(2) The choice of LMs can affect the performance of fine-tuning. Due to different pre-training corpora, LM (XRoBERTa) performs worse than LM (DRoBERTa) in most cases. Moreover, LM (GPT-2) achieves an average of 3.30% improvement over LM (DRoBERTa) in node classification, while LM (DRoBERTa) achieves an average of 7.08% improvement over LM (GPT-2) in link prediction. Considering the overall performance of the above three LMs on two different tasks and the goal of learning graph embedding, we choose LM (DRoBERTa) as our starting point for fine-tuning.

(3) Furthermore, LM can further effectively integrate with existing heterogeneous graph algorithms, resulting in a notable performance enhancement over their individual methods. Specifically, compared with RGCN, LM + RGCN achieves an average of 50.38% improvement in node classification, and achieves up to 6.59% improvements in link prediction. Compared with HGT, LM + HGT achieves up to 30.64% improvements in node classification and 7.42% improvements in link prediction.

(4) Compared with LM + RGCN and LM + HGT, our proposed graph-aware LM fine-tuning can achieve the largest improvement gains based on the chosen LM (DRoBERTa) in both node classification and link prediction tasks, showing the effectiveness of capturing topological information together with semantics in modeling attributed graphs. Specifically, our `WalkLM` outperforms the chosen LM (DRoBERTa) by up to 13.45% in node classification. In Particular, our `WalkLM` achieves

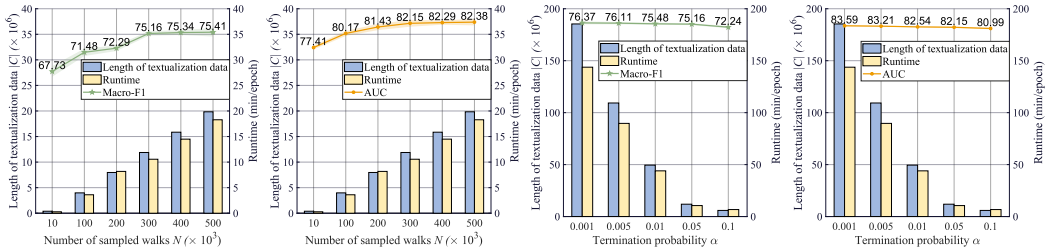
Table 4: Different downstream task results (%) with ratio of masked samples **m** on PubMed.

Task	Node Classification		Link Prediction	
Metric	Macro-F1	Micro-F1	AUC	MRR
m = 0.05	52.97	56.33	83.16	93.47
m = 0.15	**60.42***	**62.33***	**85.65***	**94.16***
m = 0.25	53.80	56.09	82.92	93.75
m = 0.35	52.22	55.61	82.38	92.72

up to 59.70% improvements in link prediction, which demonstrates our `WalkLM` can better learn accurate edge representation for link prediction by the graph-aware LM fine-tuning.

A.4 Detailed Hyper-parameter Studies

We show the results of the model sensitivity on the number of sampled walks N and the termination probability α on MIMIC-III in Figure 1. Overall, `WalkLM` is not sensitive to the two hyper-parameters, where its performance increases slowly with N and α. Note that, setting N around 3×10^5 and α around 0.05 seems appropriate to generate sufficient textual sequences and limit computational costs for fine-tuning, which can achieve a good balance of performance and efficiency. Furthermore, for the ratio of masked samples **m**, the specific results are listed in Table 4. Overall, `WalkLM` is sensitive to **m**, where the optimal value across different tasks is 0.15, which is consistent with the empirical selection in our paper and the previous work [1, 2, 4]



(a) N on node classification (b) N on link prediction (c) α on node classification (d) α on link prediction

Figure 1: Analysis of the number of sampled walks N and the termination probability α.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[3] Qiaoyu Tan, Ninghao Liu, Xiao Huang, Soo-Hyun Choi, Li Li, Rui Chen, and Xia Hu. S2gae: Self-supervised graph autoencoders are generalizable learners with graph masking. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 787–795, 2023.

[4] Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2982–2990, 2022.

[5] Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. Heterogeneous network representation learning: A unified framework with survey and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4854–4873, 2020.

[6] Muhan Zhang, Zhicheng Cui, Marion Neumann, and Yixin Chen. An end-to-end deep learning architecture for graph classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.