Exploiting Mutual Information for Substructure-aware Graph Representation Learning

Pengyang Wang[1] , **Yanjie Fu**[1] , **Yuanchun Zhou**[2*] , **Kunpeng Liu**[1] , **Xiaolin Li**[3] and **Kien Hua**[1]
[1]University of Central Florida
[2]Computer Network Information Center, Chinese Academy of Sciences
[3]Nanjing University
pengyang.wang@knights.ucf.edu, yanjie.fu@ucf.edu, zyc@cnic.cn, kunpengliu@knights.ucf.edu, lixl@nju.edu.cn, kienhua@cs.ucf.edu

Abstract

In this paper, we design and evaluate a new substructure-aware Graph Representation Learning (GRL) approach. GRL aims to map graph structure information into low-dimensional representations. While extensive efforts have been made for modeling global and/or local structure information, GRL can be improved by substructure information. Some recent studies exploit adversarial learning to incorporate substructure awareness, but hindered by unstable convergence. This study will address the major research question: is there a better way to integrate substructure awareness into GRL? As subsets of the graph structure, interested substructures (i.e., subgraph) are unique and representative for differentiating graphs, leading to the high correlation between the representation of the graph-level structure and substructures. Since mutual information (MI) is to evaluate the mutual dependence between two variables, we develop a MI inducted substructure-aware GRL method. We decompose the GRL pipeline into two stages: (1) node-level, where we introduce to maximize MI between the original and learned representation by the intuition that the original and learned representation should be highly correlated; (2) graph-level, where we preserve substructures by maximizing MI between the graph-level structure and substructure representation. Finally, we present extensive experimental results to demonstrate the improved performances of our method with real-world data.

1 Introduction

In this paper, we aim to design, implement, and evaluate a new substructure-aware Graph Representation Learning (GRL) approach. GRL aims to quantify graph by encoding structural information into low dimensional vectors. Due to the impressive effectiveness and robustness of GRL, GRL has drawn attentions in many application domains, such as biomedical sciences, human behavior modeling, social networks, computer vision, etc [Cai *et al.*, 2018].

Unlike global or neighbor (local) connectivity structures, substructures are a set of subgraphs (e.g., sub-circles, high-degree sub-vertexes, etc.) that are represented by a subset of vertexes and edges. Substructures usually demonstrate unique patterns and semantics of graphs that can be used to significantly improve GRL. For example, a mobile user can be described by an activity graph, where nodes are Point of Interests (POIs)[1] and an edge is the transition frequency among two POIs, a sub-circle (a type of substructure) represents a periodical activity sequence of the user, indicating the job occupation and preference of the user. In this case, substructure information can be used to enhance the quality of mobile user profiling [Wang *et al.*, 2019].

Extensive efforts have been made for preserving global and local structures in GRL. For example, Graph Convolutional Networks (GCNs) learn node representations by aggregating neighbors; random walk-based methods decomposed graph structures as a set of random walk paths sampled from a graph [Cai *et al.*, 2018]. However, there is limited studies about preserving substructure information in GRL. In a recent study [Wang *et al.*, 2019], *Wang et al.* propose an adversarial learning based framework to integrate substructures into GRL with a CNN approximated substructure detector. But the performances are significantly hindered by the low convergence of the adversarial learning and the precision uncertainty of the approximated detector.

Therefore, there is a compelling need to develop more effective method to model substructure awareness in GRL. Two unique challenges arise in achieving this goal.

First, how should we guarantee the accountability for the node-level representation? Despite a compressed and effective quantification of graph, the learned node representation should be coherent with the original representation in depicting graph. The coherence can be quantified by correlation that the learned representations should be highly correlated to the original ones. This motivates us to leverage Mutual Information (*MI*), a powerful correlation measurement, to investigate the correlation between the representations of the original and learned node representations. However, due to the untractable exact computation on continuous variables, *MI* is difficult to estimate and maximize, and cannot directly collab-

*Contact Author

[1]In urban setting, Point of Interest (POI) is somewhere that users show interests.

orate with well-studied Graph Convolutional Network (*GCN*) techniques for learning node representations. Fortunately, recent studies on neural *MI* estimation [Belghazi *et al.*, 2018; Hjelm *et al.*, 2018; Veličković *et al.*, 2018] propose to effectively approximate *MI* by designing a neural estimator, which gives a great chance to elaborate *MI* estimation with learning graph representations. Therefore, besides to learn node-level representations through the encode-decode paradigm, we propose to guarantee the accountability of the learned node representations by maximizing the mutual information between the learned and original node representation.

Second, after we obtain the accountable graph-level representation, how should we preserve substructure information for the graph-level representation? As subsets of the graph structure, interested substructures are unique and representative for differentiating graphs. Thus, for the common goal of quantifying graph, the representation of substructures should be highly correlated to graph-level representation. In another word, maximizing the correlation between the graph-level and substructure representations would provide a direction to enforce the graph-level representation to preserve substructure information. To quantify such correlation, similarly, we also exploit to maximize *MI* between the representations of the graph-level and substructures. We first obtain the graph-level representation and substructure representations by graph pooling operation over the entire graph and substructure respectively. We then maximize the *MI* between graph-level and substructure representations to enforce the graph-level representation preserve substructures.

Along these lines, in this paper, we develop a substructure-aware graph representation learning framework. The proposed learning framework includes two stages, (1) node-level representation and (2) graph-level representation. The graph-level representation is the final output that is expected to preserve substructure information of graph. Specifically, we propose to guarantee the accountability in the node-level representation by maximizing *MI* between the learned and original representations to guide the encoding step. And then, we propose to preserve the substructure information in the graph-level representation by maximizing *MI* between the graph-level and substructure representations. We apply the proposed framework to the application of next activity type prediction for mobile user profiling to evaluate the effectiveness with real-world mobile checkin data.

2 Preliminaries

2.1 Mutual Information and Estimation

Mutual Information (*MI*) is a measurement to evaluate the dependency between two random variables. Due to the promising capability of capturing non-linear dependencies, *MI* has been applied in various disciplines, such as cosmology, biomedical sciences, computer vision, feature selection, and information bottleneck [Belghazi *et al.*, 2018].

However, untractable exact computation and limited known distribution families hinder the further adaption of *MI* for continuous variables [Belghazi *et al.*, 2018]. To estimate *MI*, non-parametric methods and approximate gaussianity of data distribution are proposed, which perform poor in expanding data scales (*i.e.*, sample size or dimensions) [Belghazi *et al.*, 2018]. A neural estimator *MINE* is then proposed to exploit gradient descent over neural networks to estimate *MI*, which is linearly scalable in dimensionality and sample size [Belghazi *et al.*, 2018]. Moreover, *DIM* [Hjelm *et al.*, 2018] combine the *MI* with deep representation learning for learning better global and local representations for images. Following *DIM*, *DGI* [Veličković *et al.*, 2018] further extends *MI* neural estimation on learning graph representations. In this paper, we follow the methodology of *MINE*, *DIM* and *DGI* to exploit *MI* for learning graph representations with incorporating substructures.

2.2 Problem Statement

In this paper, we study the problem of learning graph representations with preserving interested substructures. Formally, given a list of graph $\mathcal{G} = \{\mathcal{G}^{(1)}, \mathcal{G}^{(2)}, \cdots, \mathcal{G}^{(K)}\}$, where $\mathcal{G}^{(k)} = \{V^{(k)}, E^{(k)}\}$, we aim to learn a mapping function $f : \mathcal{G} \rightarrow \mathbf{h}_g$ that takes the graph $\mathcal{G}$ as input, and outputs the vectorized graph-level representation $\mathbf{h}_g = \{\mathbf{h}_g^{(1)}, \mathbf{h}_g^{(2)}, \cdots, \mathbf{h}_g^{(K)}\}$ for each graph in $\mathcal{G}$, while subject to the special attention on preserving interested substructures.

3 Methodology

3.1 Model Intuition

We learn the substructure-aware graph representations on the following intuitions.

Intuition 1: Accountability of Representations. For the node-level representation, the learned representation should be a summarized vector that is highly correlated to the original representation. Therefore, we need to guarantee the accountability of the learned representation.

Intuition 2: Substructure Preservation. For the graph-level representation, substructures demonstrates specific patterns of graphs, which would enrich the semantics of graph representations. Therefore, we need to preserve substructures in the graph-level representation.

3.2 Framework Overview

Figure 1 shows the overview of the proposed two-stage framework. In the first stage, we first propose to learn node-level representation with exploiting GCN through the encode-decode paradigm by minimizing the reconstruction loss that follows a contrastive learning-style convention. Then, we maximize *MI* between the learned and original node-level representation for guaranteeing the accountability. In the second stage, we obtain the graph-level representation and substructure representation with the graph pooling operation over the node-level representation of the entire graph and interested subgraph respectively. Then, we propose to preserve the substructure information by maximizing *MI* between the graph-level and substructure representations. Details will be introduced in the following.

Figure 1 shows an overview of the proposed framework. The proposed framework includes three key component. First, the graph is fed into an encode-decode paradigm to learn the node representations by minimizing the contrastive

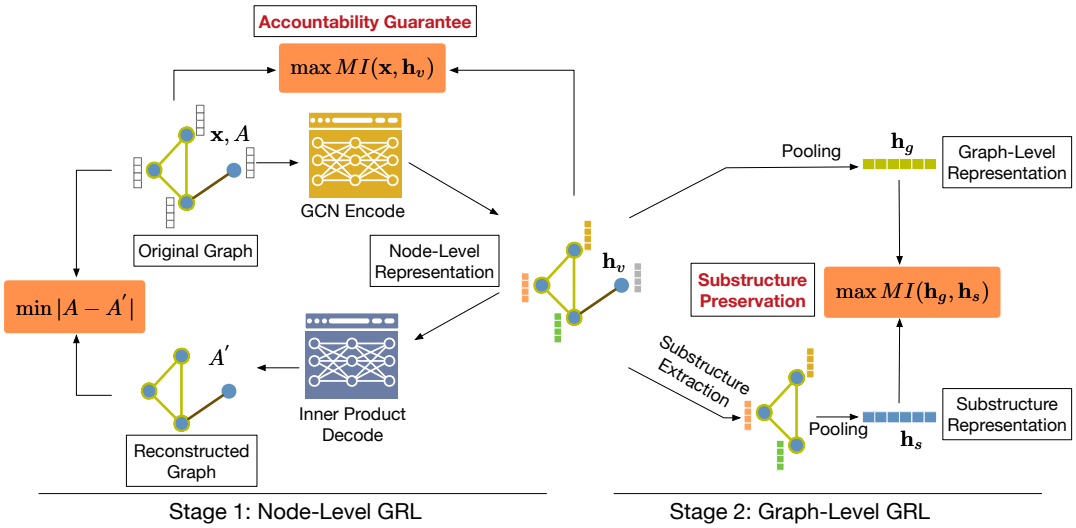


Figure 1: Framework overview.

loss between positive and negative samples. Second, to guarantee the accountability of the learned representation, we aim to maximize the mutual information between the original and the learned node representations. Third, we obtain the graph-level representation and substructured graph representation by the pooling operation. We maximize the mutual information between them to preserve the substructures in the graph-level graph representation. Details will be introduced in rest of the section.

3.3 Learning Node-Level Representation

For better generality, we learn node-level representation with Graph Convolutional Network (GCN) in the unsupervised fashion. We follow the idea of GAE [Kipf and Welling, 2016b] to learn representation in a encode-decode paradigm. Specifically, the encoder is a o-layer GCN. At the l-th layer, the node representation can be denoted as

$$\mathbf{h}^l = \sigma(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{h}^{l-1}\mathbf{W}^{l-1}), \tag{1}$$

where $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, $\mathbf{A}$ is the adjacency matrix. $\mathbf{I}$ is the identity matrix, $\hat{\mathbf{D}}_{ii} = \sum_i \hat{\mathbf{A}}_{ij}$, and $\mathbf{W}^{l-1}$ is the weight. Then, the learned the representation is $\mathbf{h}_v = \mathbf{h}^l$. The decoder is a inner product of the learned representation to recover the adjacency matrix

$$\hat{\mathbf{A}}^{'} = \sigma(\mathbf{h}_v \mathbf{h}_v^T). \tag{2}$$

The objective is to minimize the reconstruction loss between the original adjacency matrix $\hat{\mathbf{A}}$ and reconstructed adjacency matrix $\hat{\mathbf{A}}^{'}$.

We follow the convention of contrastive learning approach to minimize the reconstruction loss. We first sample positive nodes from neighbors, and negative nodes randomly from non-neighbors. Then, we minimize the cross-entropy loss of positive and negative node pairs

$$\mathcal{L}_r = -\log \hat{\mathbf{A}}^{'}_{pos} - \log(1 - \hat{\mathbf{A}}^{'}_{neg}) \tag{3}$$

where $\hat{\mathbf{A}}^{'}_{pos}$ and $\hat{\mathbf{A}}^{'}_{neg}$ are derived from positive nodes pairs and negative node pairs respevtively, based on Equation 3.

3.4 Representation Accountability Guarantee

Intuitively, the learned low dimensional representation should be highly correlated to the original representations. To guarantee such accountability, we exploit mutual information to investigate the correlation between the learned representations $\mathbf{h}_v$ and the original representation $\mathbf{x}$.

Following the idea from *DIM* [Hjelm *et al.*, 2018] and *DGI* [Veličković *et al.*, 2018], we define a Jensen Shannon *MI* estimator to estimate and maximize the *MI* between $\mathbf{x}$ and $\mathbf{h}_v$ as

$$\begin{aligned} MI(\mathbf{x}; \mathbf{h}_x) :=& \mathbb{E}_{\mathbf{X}}[-\text{sp}(-D(\mathbf{x}, \mathbf{h}_v))] + \\ & \mathbb{E}_{\tilde{\mathbf{X}}}[\text{sp}(D(\mathbf{x}, \mathbf{h}_v)] \end{aligned} \tag{4}$$

where sp is the softplus function that $\text{sp}(c) = \log(1 + e^c)$, $\mathbf{X}$ is the positive samples and $\tilde{\mathbf{X}}$ is the negative sample. We will present how we generate positive and negative samples later.

Since the noise-contrastive type objective with a standard binary cross-entropy (BCE) can effectively maximize mutual information [Veličković *et al.*, 2018], we define the loss function as:

$$\begin{aligned} \mathcal{L}_j = -\sum_k \sum_i & \mathbb{E}_{\mathbf{x}}[\log \mathcal{D}(\mathbf{x}_i^{(k)}, \mathbf{h}_{v_i}^{(k)})] - \\ & \mathbb{E}_{\tilde{\mathbf{x}}}[\log(1 - \mathcal{D}(\mathbf{x}_i^{(\tilde{k})}, \mathbf{h}_{v_i}^{(k)})] \end{aligned} \tag{5}$$

where $\mathcal{D}$ denotes a discriminator to provide probability scores for sampled pairs. For the k-th graph, we regard the positive samples as the pairs of $(\mathbf{x}_i^{(k)}, \mathbf{h}_{v_i})$, and the negative samples as the pairs of $(\mathbf{x}_i^{(\tilde{k})}, \mathbf{h}_{v_i})$, where $\mathbf{x}_i^{(\tilde{k})}$ is a randomly picked node from another graph. The objective is to minimize $\mathcal{L}_j$, which is equivalent to maximize the mutual information.

3.5 Preserving Substructures in the Graph-Level Representation

Substructures, which are pivotal for learning complete representations for graphs [Cai *et al.*, 2018; Wang *et al.*, 2019],

are subset of global structures. In another word, the representations of substructures should be highly correlated to the graph-level representations for a given graph. Therefore, this motivates us to exploit to maximize *MI* between the graph-level and substructure representations to provide a direction to enforce the graph-level representation to preserve substructure information.

First, we obtain the graph-level representation and substructure information from node representations generated by the base model. Specifically, as shown in Figure 1, for the k-th graph, on one hand, we exploit the pooling operator over all the nodes representations to obtain the graph-level representation $\mathbf{h}_g^{(k)}$; on the other hand, the pooling operator is also applied over the nodes associated with substructures to generate the representation of substructures $\mathbf{h}_s^{(k)}$.

Then, we exploit to use neural network for estimating and maximizing mutual information $MI(\mathbf{h}_g; \mathbf{h}_s)$ between the graph-level representations $\mathbf{h}_g^{(k)}$ and the representation of substructures $\mathbf{h}_s^{(k)}$ to guarantee the highly correlated relationship. We have the similar noise-contrastive type loss function:

$$\mathcal{L}_s = -\sum_k \sum_i \mathbb{E}_{\mathbf{x}}[\log \mathcal{D}(\mathbf{h}_s^{(k)}, \mathbf{h}_g^{(k)})] - \mathbb{E}_{\tilde{\mathbf{x}}}[\log(1 - \mathcal{D}(\mathbf{h}_s^{(\tilde{k})}, \mathbf{h}_g^{(k)})], \tag{6}$$

where $\mathcal{D}$ denotes a discriminator to provide probability scores for sampled pairs, $\mathbf{h}_s^{(\tilde{k})}$ is the graph-level representations that are not of k-th graph. We design the positive samples as the pairs of $(\mathbf{h}_s^{(k)}, \mathbf{h}_g^{(k)})$, and the negative samples as the pairs of $(\mathbf{h}_s^{(\tilde{k})}, \mathbf{h}_g^{(k)})$. The objective is to minimize $\mathcal{L}_s$, which is equivalent to maximize $MI(\mathbf{h}_g; \mathbf{h}_s)$. In this way, the optimal graph-level representation $\mathbf{h}_g$ would preserve substructures.

3.6 Optimization

The loss of the model includes: (i) the contrastive learning loss for graph reconstruction (Equation 3); (ii) the representation accountability learning loss (Equation 5), and (iii) the substructure preservation loss (Equation 6). The objective is to minimize the overall loss $\mathcal{L}$ as follows:

$$\mathcal{L} = \lambda_r \mathcal{L}_r + \lambda_j \mathcal{L}_j + \lambda_s \mathcal{L}_s \tag{7}$$

where λ_r, λ_j, and λ_s are the weights of $\mathcal{L}_r$, $\mathcal{L}_j$, and $\mathcal{L}_s$ respectively. We employ gradient descent to minimize $\mathcal{L}$.

3.7 Comparison with Related Work

We discuss the differences between our proposed method and recent studies from two perspectives: (1) structure preserving and (2) representation accountability. On one hand, GCN variants (e.g., GCN [Kipf and Welling, 2016a], GAE [Kipf and Welling, 2016b], DGI [Veličković *et al.*, 2018]), and random walk-based approaches (e.g., DeepWalk [Perozzi *et al.*, 2014]) are two main streams for graph representation learning that focus on preserving global and/or local structures of graph, but substructures are not considered in these approaches, while StructRL [Wang *et al.*, 2019] models the substructures through adversarial learning that is limited by the bad convergence and accuracy of the pre-trained approximated detector. On the other hand, DGI first exploits mutual information to guarantee the correlation between learned output and original input to endow the accountability for representations, while others ignores such important property. Different to the literature, in this paper, we simultaneously model substructures and provide the accountability-guarantee to enhance the graph representation. The experiment later will show the superior performance of our proposed framework.

4 Application: Mobile User Profiling For Next Activity Type Prediction

As mentioned in Introduction, preserving substructures of user activity graph can benefit mobile user profiling. Therefore, to validate the effectiveness of preserving substructures, we apply the proposed method to conduct mobile user profiling with the task of predicting next activity type.

First, for each user, we follow the formulation from the work [Wang *et al.*, 2019] to construct a user activity graph, where the nodes are POI categories and edges are the normalized visiting frequencies among POI categories. POI categories are regarded as activity types. User activity graph demonstrates the preference and patterns of participating different activities. For simplicity, we only consider the "circle" substructure [Wang *et al.*, 2019] in this paper.

Then, we exploit the proposed method over the constructed user activity graph to learn user representations as the features, which is further fed into a classifier to predict the next activity type. The more accurate the prediction, the better the user profiling, and then, the better the proposed method preserves substructures of user activity graph.

5 Experiment

City	# Check-ins	# POI Categories	Time Period
New York	227428	400	12 April 2012 to 16 February 2013
Tokyo	573703	385	12 April 2012 to 16 February 2013

Table 1: Statistics of the experimental data.

5.1 Data Description

We evaluate the performance over two real-world check-in datasets [Yang *et al.*, 2014] of New York and Tokyo. Table 1 shows the statistics of the dataset. The format of each dataset is <User ID, Venue ID, Venue Category ID, Venue Category Name, Latitude, Logitudem, Time>.

In the experiment, we chronologically extract POI category visit sequence for each user. We reserve the last visit POI category as the prediction target, and use all of the previous ones to construct user activity graph.

5.2 Evaluation Metrics

We use the prediction accuracy to evaluate the performance. The evaluation metric Accuracy@N is defined as: let T_i denote the target POI category that the user actually visited, P_i^N denote the topN predicted POI category list ranked in a descending order based on the predicted visit probabilities

	@2	Outperform	@3	Outperform	@4	Outperform	@5	Outperform
MI-StrutRL	0.0646	-	0.1200	-	0.1477	-	0.1570	-
DGI	0.0462	+39.9%	0.0554	+116.6%	0.0923	+60.0%	0.1108	+41.7%
GAE	0.0646	+0%	0.0923	+30.0%	0.1293	+14.2%	0.1477	+6.3%
StructRL	0.0462	+39.9%	0.0646	+85.8%	0.0831	+77.7%	0.0923	+70.1%
DeepWalk	0.0369	+75.1%	0.0462	159.7%	0.0646	+128.6%	0.0739	+112.4%

Table 2: Overall comparison of Accuracy@N (%) on the New York dataset.

	@2	Outperform	@3	Outperform	@4	Outperform	@5	Outperform
MI-StrutRL	0.1526	-	0.1614	-	0.1701	-	0.1788	-
DGI	0.1352	+12.9%	0.1439	+12.2%	0.1614	+5.4%	0.1701	+5.1%
GAE	0.1439	+6.0%	0.1570	+2.8%	0.1701	+0%	0.1744	+2.5%
StructRL	0.0392	+39.9%	0.0436	+270.2%	0.0567	+200.0%	0.0785	+127.8%
DeepWalk	0.1265	+289.3%	0.1396	15.6%	0.1483	+14.7%	0.1614	+10.8%

Table 3: Overall comparison of Accuracy@N (%) on the Tokyo dataset.

for user i, we consider the prediction is a success once the $T_i \in P_i^N$. Then

$$Accuracy@N = \frac{1}{|U|}\mathcal{I}(T_i \in P_i^N), \tag{8}$$

where $|U|$ denote the user numbers. We report Accuracy@2, Accuracy@3, Accuracy@4. Accuracy@5 in this paper.

5.3 Baseline Algorithms

(1) GAE. The Graph Autoencoder [Kipf and Welling, 2016b] learned node representations in the encode-decode paradigm with GCN as the encoder and inner production to recover adjancacy matrix as the decoder. In the experiment, we set the number of GCN layer = 2, the input feature size=100, the output feature size = 40, learning rate = 0.001.

(2) DeepWalk. The DeepWalk model [Perozzi *et al.*, 2014] extends the word2vec model [Mikolov *et al.*, 2013] to the scenario of network embedding by truncated random walks. We set the number of walks = 50, the size of representation = 40, the walk length = 40, and the window size = 10.

(3) DGI. Deep Graph Infomax [Veličković *et al.*, 2018] extends learning representations with MI maximation to graph embedding by modeling global and local structures. We set the input feature size=100, the output feature size = 40, learning rate = 0.001.

(4) StructRL. StructRL [Wang *et al.*, 2019] learns graph representations with attention on substructures without considering the accountability of learned representations. We set the input feature size=100, the output feature size is 40, learning rate = 0.001.

All the baseline algorithms and our proposed method are unsupervised learning approach. In the experiment, we first exploit each algorithm to learn graph features for each user. Then, we train a fully connected neural network to predict the visiting probability for each POI category. We conduct 10-fold cross validation and report the average Accuracy@N.

5.4 Overall Performance

We compare our method with the baseline methods in terms of Accuracy@N. In general, Figure 2 and 3 show our model outperforms other baseline methods for both the New York and Tokyo dataset. One interesting observation is that with N increasing, the improvement of our proposed model is less significant.

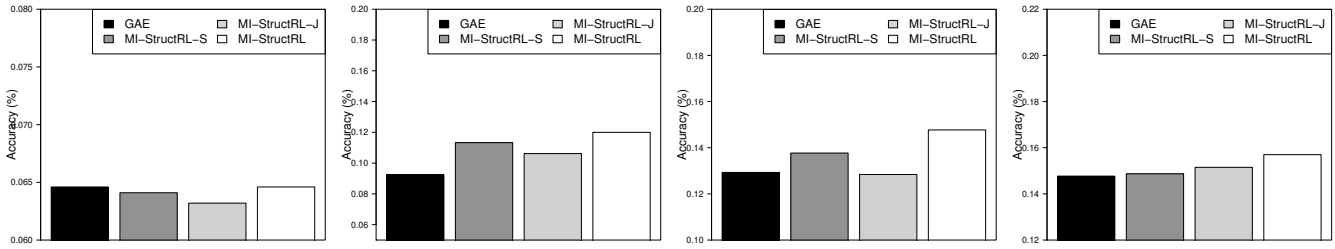
Comparing to DeepWalk, which is the representative of random walk-based methods, graph convolution-based methods (GAE, DGI, our proposed method) and autoencoder-based methods (GAE, StructRL) perform better in modeling structure information of user activity graph. Comparing to GAE, which is the base model of our proposed framework, our proposed framework additionally considers substructures and the accountability of the learned representation, thus, enhance the quality of learned representations. Comparing to DGI, which only models global and local structures, the incorporation of substructures improving the graph representation is validated by the better performance of our proposed method. Comparing to StructRL, which only incorporates substructures, accountability-guarantee provided by our proposed method further elevates the reasonability of the learned representations; and also, mutual information quantifies the non-linear relationship among structures, while only linear relations between countersamples are considered in StructRL.

In summary, the results validate that incorporating substructures and accountability can improve the quality of graph representations.

5.5 Analysis of $\mathcal{L}_j$, $\mathcal{L}_s$

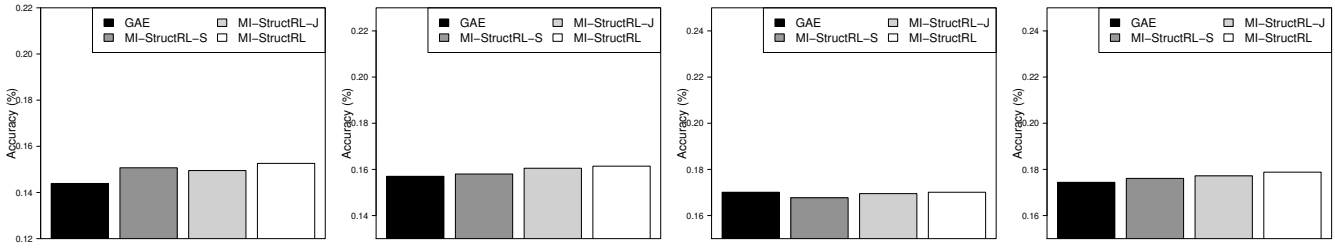
To analyze the contribution of representation accountability and substructures preserving, we define two variants of our proposed model: (1) MI-StructRL-J, which only adds $\mathcal{L}_j$ to the base model for incorporating accountability of the learned representations; (2) MI-StructRL-S, which only adds $\mathcal{L}_s$ to the base model for incorporating substructures. We compare the base model, MI-StructRL-J, MI-StructRL-S, and MI-StructRL in the experiment.

As shown in Figure 2 and 3, when N is relatively small, the improvement by incorporating substructures is more significant than the representation accountability. When N is getting larger, the improvement by the representation accountability is more significant than incorporating substructures. The re-



(a) Accuracy@2 (b) Accuracy@3 (c) Accuracy@4 (d) Accuracy@5

Figure 2: Analysis of $\mathcal{L}_j$, $\mathcal{L}_s$ for New York.



(a) Accuracy@2 (b) Accuracy@3 (c) Accuracy@4 (d) Accuracy@5

Figure 3: Analysis of $\mathcal{L}_j$, $\mathcal{L}_s$ for Tokyo.

sults indicate that incorporating substructures is more important for top short range prediction.

6 Related Work

Graph Representation Learning. Graph representation learning aims to learn representations of graph structures, which can be categorized into matrix factorization based, deep learning based , edge reconstruction based, graph kernel based and generative models [Cai *et al.*, 2018]. Specifically, matrix factorization based methods mainly exploit graph laplacian eigenmaps [Cai *et al.*, 2007] and node proximity matrix factorization [Cao *et al.*, 2015] to learn graph representations; deep learning based methods are a huge group in which many works on random walk [Perozzi *et al.*, 2014] Edge reconstruction based methods carry the insight that the edge connectivity constructed based on embedding and original features should be as similar as possible [Zhang and Wang, 2016]; graph kernel based method aim to model graph structures from the perspectives of graphlet [Yanardag and Vishwanathan, 2015] and substree patterns [Shervashidze *et al.*, 2011]; generative models aim to learn representations by maximizing the joint distribution of the input features and the target labels [Bernardo *et al.*, 2007].

Human mobility modeling. Our work has connection with human mobility modeling. Human mobility modeling aims to learn human patterns from the human mobility data, which has been applied into various applications. For example, *Wang et al.* propose to learn the representation of urban residential communities by modeling human mobility patterns [Wang *et al.*, 2018b; Fu *et al.*, 2019; Zhang *et al.*, 2019]. *Wang et al.* propose to analyze the driving behavior by modeling the human mobility from the perspectives of peer and temporal dependencies [Wang *et al.*, 2018a]. *Liu et al.* propose to predict the travel destination by modeling the patterns of Mobike users with coupling among multi-view spatio-temporal contexts[Liu *et al.*, 2018].

7 Conclusion

Substructures are pivotal for improving graph representations. While recent studies on graph representation learning mainly focus on modeling global and/or local structures of graph, fewer efforts have been made on preserving substructures. Therefore, in this paper, we decompose the GRL pipeline into two stages, (1) node-level and (2) graph-level. In the node-level stage, to further guarantee the accountability of representation, we propose to maximize the mutual information between the learned and original node representations. In the graph-level stage, motivated by the intuition that the representation of substructures should be highly correlated to the graph-level representation, we preserve the substructures by maximizing the mutual information between the substructures and the graph-level structures. We simultaneously optimize the learning procedure of node representations, accountability and substructures. The experimental results show that preserving substructures via maximizing mutual information between substructures and graph-level structures effectively enhance the performance of graph representations.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (NSFC) under Grant 61836013.

References

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[Bernardo *et al.*, 2007] JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Generative or discriminative? getting the best of both worlds. *Bayesian statistics*, 8(3):3–24, 2007.

[Cai *et al.*, 2007] Deng Cai, Xiaofei He, and Jiawei Han. Spectral regression: A unified subspace learning framework for content-based image retrieval. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 403–412. ACM, 2007.

[Cai *et al.*, 2018] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1616–1637, 2018.

[Cao *et al.*, 2015] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 891–900. ACM, 2015.

[Fu *et al.*, 2019] Yanjie Fu, Pengyang Wang, Jiadi Du, Le Wu, and Li Xiaolin. Efficient region embedding with multi-view spatial networks: A perspective of locality-constrained spatial autocorrelations. In *Proceedings of the 33th AAAI Conference on Artificial Intelligence*, page to appear. AAAI, 2019.

[Hjelm *et al.*, 2018] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[Kipf and Welling, 2016a] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Kipf and Welling, 2016b] Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[Liu *et al.*, 2018] Kunpeng Liu, Pengyang Wang, Jiawei Zhang, Yanjie Fu, and Sajal K Das. Modeling the interaction coupling of multi-view spatiotemporal contexts for destination prediction. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 171–179. SIAM, 2018.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[Perozzi *et al.*, 2014] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.

[Shervashidze *et al.*, 2011] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M Borgwardt. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(Sep):2539–2561, 2011.

[Veličković *et al.*, 2018] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.

[Wang *et al.*, 2018a] Pengyang Wang, Yanjie Fu, Jiawei Zhang, Pengfei Wang, Yu Zheng, and Charu Aggarwal. You are how you drive: Peer and temporal-aware representation learning for driving behavior analysis. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2457–2466. ACM, 2018.

[Wang *et al.*, 2018b] Pengyang Wang, Jiawei Zhang, Guannan Liu, Yanjie Fu, and Charu Aggarwal. Ensemble-spotting: Ranking urban vibrancy via poi embedding with multi-view spatial graphs. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 351–359. SIAM, 2018.

[Wang *et al.*, 2019] Pengyang Wang, Yanjie Fu, Hui Xiong, and Xiaolin Li. Adversarial substructured representation learning for mobile user profiling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 130–138. ACM, 2019.

[Yanardag and Vishwanathan, 2015] Pinar Yanardag and SVN Vishwanathan. Deep graph kernels. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1365–1374. ACM, 2015.

[Yang *et al.*, 2014] Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2014.

[Zhang and Wang, 2016] Qing Zhang and Houfeng Wang. Not all links are created equal: An adaptive embedding approach for social personalized ranking. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 917–920. ACM, 2016.

[Zhang *et al.*, 2019] Yunchao Zhang, Pengyang Wang, Xiaolin Li, Yu Zheng, and Yanjie Fu. Unifying inter-region autocorrelation and intra-region structures for spatial embedding via collective adversarial learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019.