SimGNN: A Neural Network Approachto Fast Graph Similarity Computation

作者：Yunsheng Bai, Hao Ding, Song Bian, Ting Chen, Yizhou Sun, Wei Wang

University of California, Los Angeles, CA, USA

期刊： WSDM 2019

Abstract

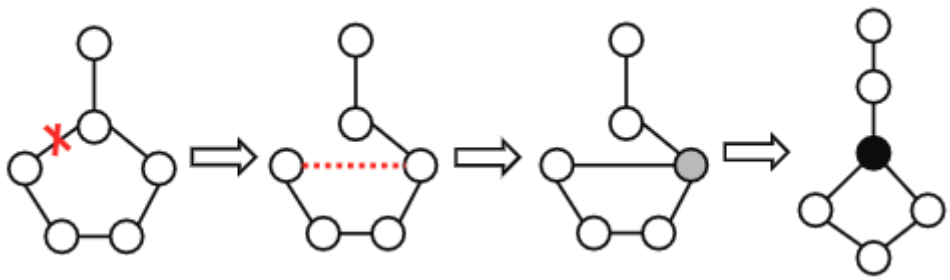
图的相似性搜索是重要的基于图的应用（例如化合物相似性等），但是GNN的使用一般就是用在点的分类上，而没有用在整个图的计算上

本文提出 SimGNN 的网络框架进行图相似性计算

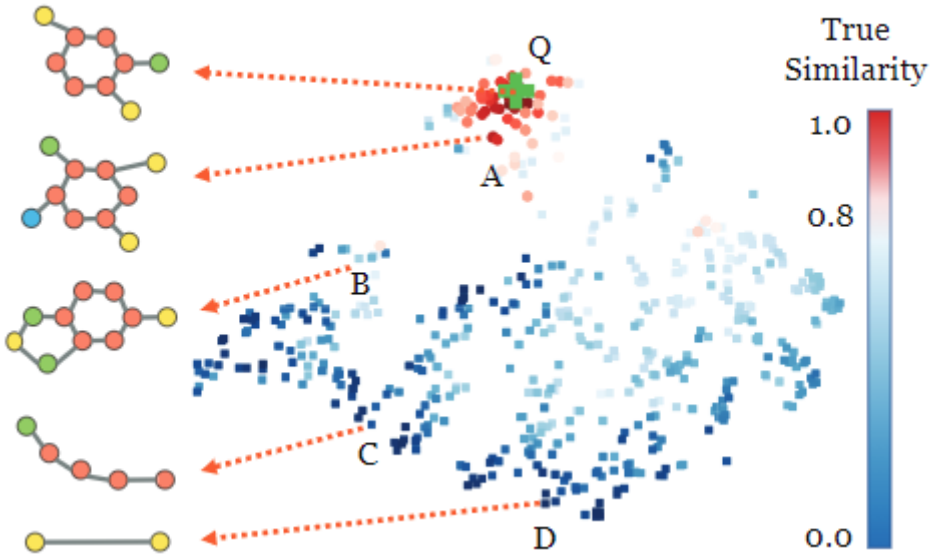
创新点/贡献/优势：

- 使用基于神经网络的方法，将图相似性计算视为学习问题。
- 提出了两种新颖的策略。
  - 提出一种的注意力机制，以选择图的最相关部分以生成图级别的嵌入，从而保留图之间的相似性。
  - 提出了一种成对节点比较方法来补充图级嵌入，以便更有效地建模两个图之间的相似度。
- 基于三个真实的网络数据集，使用GED作为实验的评价参数，证明该方法的有效性和效率

Introduction

- **GED**（图编辑距离）
  - 
    这样需要3步变化，所以最左边和最右边图的GED为3
  - 使用GED做为一个相似性的评价指标
- **图的相似性查询**就是解决，从一个图数据库种获取与查询图相似的图
- 相似性评价标准：GED，MCS（最大公共子图）。
- 目前的相似性查询方法
  - 使用传统算法，通过剪枝优化（不能处理超过16个节点的图）
  - 使用启发式优化和组合搜索的方法查找近似解（A*算法）
- 本文将图相似问题变成一个学习问题，设计一个神经网络，将两张图映射到一个相似性分数上面
  - **特征不变性**：节点的顺序和特征无关
  - **归纳性**：可以直接用于未见过的图
  - **可学习性**：通过训练调整参数
- 为了满足上面的条件，设计了两个策略

- 设计一个可学习的嵌入，将图直接映射到一个向量

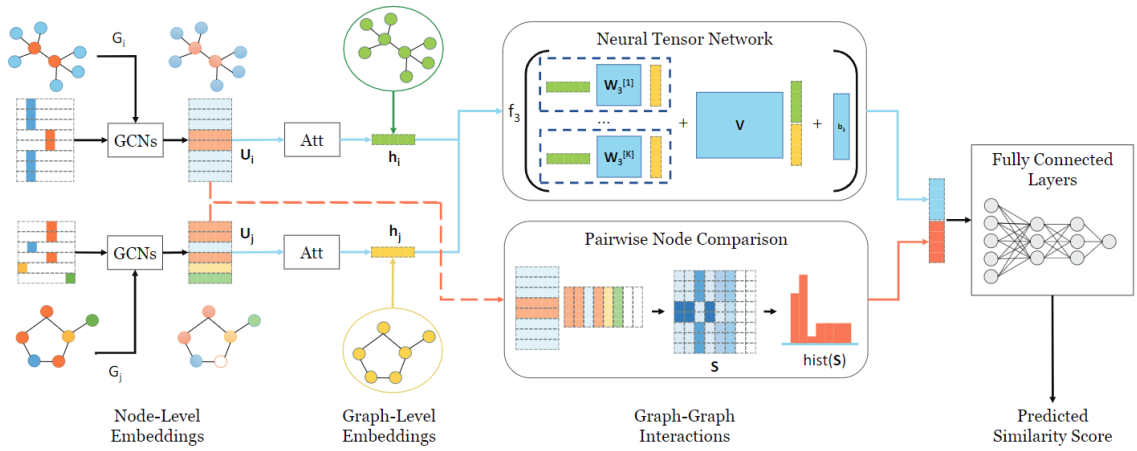


图从可以看出，对于相似的图对应到向量上的距离确实很接近

- 两张图之间点对的比较，做为对上面图特征的一个补充

Model

SimGNN



第一种策略：（蓝线）

- 使用GCN学习点的特征。
- 使用注意力机制

$$\boldsymbol{h} = \sum_{n=1}^{N} f_2(\boldsymbol{u}_n^T \boldsymbol{c})\boldsymbol{u}_n = \sum_{n=1}^{N} f_2(\boldsymbol{u}_n^T \tanh((\frac{1}{N}\sum_{m=1}^{N} \boldsymbol{u}_m)\boldsymbol{W}_2))\boldsymbol{u}_n \qquad (2)$$

其中f2函数是一个激活函数，$u_n^T c$ 就是一个注意力系数

- 图和图之间的交互，使用NTN（Neural Tensor Networks是一种用于解决文本实体之间关系的方法）

$$g(h_i, h_j) = f_3(h_i^T W_3^{[1:K]} h_j + V \begin{bmatrix} h_i \\ h_j \end{bmatrix} + b_3)$$

k是一个超参数，f3也是一个激活函数，W，V，b是权重矩阵（可学习的）

- 图相似性分数计算

  将一个向量通过全连接层再激活，得到一个相似性分数

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{(i,j) \in \mathcal{D}} (\hat{s_{ij}} - s(\mathcal{G}_i, \mathcal{G}_j))^2$$

第二种策略：（红线）

- 由于策略一的节点信息再图信息嵌入后可能会出现丢失的情况，需要策略二进行补充
- 将经过GCN后的点特征，进行矩阵运算，得到一个二维矩阵（需要使用空向量进行补齐）
- 由于需要考虑特征不变性，所以再使用直方图来表示特征，但是这样就不能使用反向传播了

时间复杂度： 第一种策略是基础，第二种策略是辅助，但是耗时，最坏情况下是 $O(n^2)$ ，点数的平方

Experiments

Datasets

Dataset	Graph Meaning	#Graphs	#Pairs
AIDS	Chemical Compounds	700	490K
LINUX	Program Dependency Graphs	1000	1M
IMDB	Actor/Actress Ego-Networks	1500	2.25M

pairs：代表点对的数量

- AIDS：每个图都小于等于10个点，点的标签一个29种
- LINUX：点无标签，每一个图都小于等于10个点
- IMDB：点无标签，并且不限制每一个图的点数量

Preprocessing

- 对于IMDB数据集没有一个算法可以在可接受时间内计算出正确的GED，故处理时，对于IMDB数据集，使用3个最有名的近似算法Beam、Hungarian、VJ，并且取其中的最小值作为答案GED
- 对于GED需要将其转化为相似性分数在[0,1]之间，$\mathrm{nGED}(\mathcal{G}_1, \mathcal{G}_2) = \frac{\mathrm{GED}(\mathcal{G}_1, \mathcal{G}_2)}{(|\mathcal{G}_1| + |\mathcal{G}_2|)/2}$，在将其通过激活函数就转化为相似性分数了

Result

斯皮尔曼等级相关系数 (ρ)　Spearman’s Rank Correlation Coefficient

肯德尔等级相关系数 (τ)　　Kendall’sRank Correlation Coefficient

用来评价预测结果核正确结果的相关性

Table 2: Results on AIDS.

Method	mse(10^{-3})	ρ	τ	p@10	p@20
Beam	12.090	0.609	0.463	**0.481**	0.493
Hungarian	25.296	0.510	0.378	0.360	0.392
VJ	29.157	0.517	0.383	0.310	0.345
SimpleMean	3.115	0.633	0.480	0.269	0.279
HierarchicalMean	3.046	0.681	0.629	0.246	0.340
HierarchicalMax	3.396	0.655	0.505	0.222	0.295
AttDegree	3.338	0.628	0.478	0.209	0.279
AttGlobalContext	1.472	0.813	0.653	0.376	0.473
AttLearnableGC	1.340	0.825	0.667	0.400	0.488
SimGNN	**1.189**	**0.843**	**0.690**	**0.421**	**0.514**

Table 3: Results on LINUX.

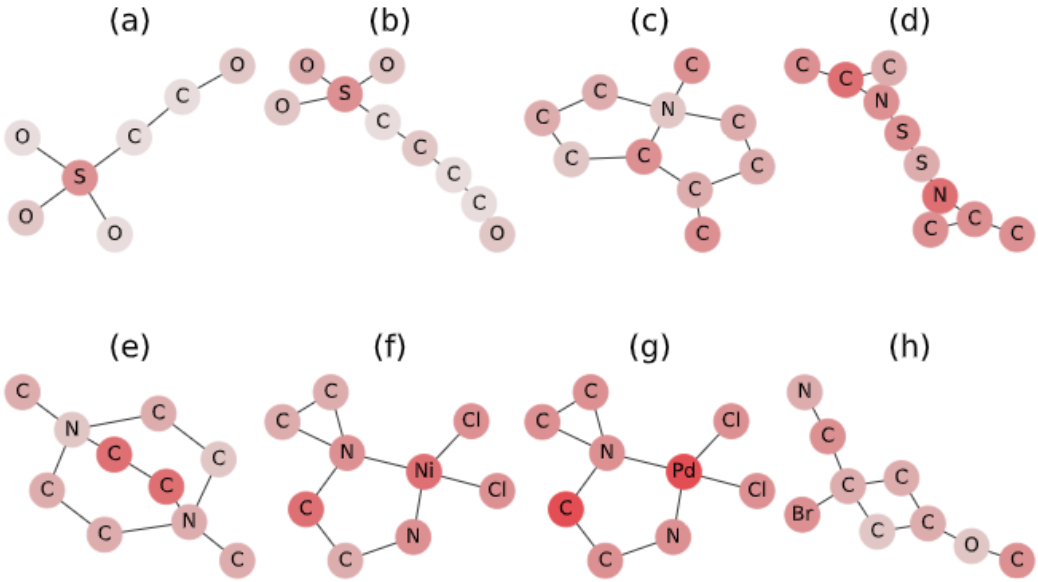
Method	mse(10^{-3})	ρ	τ	p@10	p@20
Beam	9.268	0.827	0.714	**0.973**	0.924
Hungarian	29.805	0.638	0.517	0.913	0.836
VJ	63.863	0.581	0.450	0.287	0.251
SimpleMean	16.950	0.020	0.016	0.432	0.465
HierarchicalMean	6.431	0.430	0.525	0.750	0.618
HierarchicalMax	6.575	0.879	0.740	0.551	0.575
AttDegree	8.064	0.742	0.609	0.427	0.460
AttGlobalContext	3.125	0.904	0.781	0.874	0.864
AttLearnableGC	2.055	0.916	0.804	0.903	0.887
SimGNN	**1.509**	**0.939**	**0.830**	**0.942**	**0.933**

Table 4: Results on IMDB. Beam, Hungarian, and VJ together are used to determine the ground-truth results.

Method	mse(10^{-3})	ρ	τ	p@10	p@20
SimpleMean	3.749	0.774	0.644	0.547	0.588
HierarchicalMean	5.019	0.456	0.378	0.567	0.553
HierarchicalMax	6.993	0.455	0.354	0.572	0.570
AttDegree	2.144	0.828	0.695	0.700	0.695
AttGlobalContext	3.555	0.684	0.553	0.657	0.656
AttLearnableGC	1.455	0.835	0.700	0.732	0.742
SimGNN	**1.264**	**0.878**	**0.770**	**0.759**	**0.777**

注意力训练的效果，对于两种节点的注意力系数高

- 度数大的节点
- 仅仅出现一次的节点或者结构



Conclusion and Future Works

该论文是图深度学习和图搜索问题的结合，并通过基于神经网络的方法解决图相似性计算的核心操作。

核心思想是学习一种基于神经网络的函数，该函数具有表征不变性，归纳性和对特定相似性度量的适应性。

输入任意两个图并输出其相似性得分。 与现有的经典算法相比，模型在近似的Graph Edit Distance计算上运行速度非常快，并且具有高准确性。

未来方向

- 考虑处理带有边特征信息的图。
- 使用不同的技术使前k的结果更佳。
- 考虑如何将小图中计算的GED应用到大型图之中。