LinkPrediction 实验

程智镝、陈凌

2023 年 9 月 17 日

作业任务

- 从代码（自己实现 or 复现）、数据集（直接获取或自己处理得到）两个角度权衡是否选择某个 link prediction 的工作。。
- 论文摘要 abstract 和 introduction 翻译
- 问题描述。
- 输入、输出、模型算法描述（附框架图；有多个的挑 1 个主要实现）
- 评价指标及其计算公式
- 对比方法及这些对比方法的引用论文出处
- 结果
- 打包提交 code、运行配置说明（数据集太大的可以是开放链接，需描述）

实验难点：

- 论文为全英文描述，阅读难度提升
- 论文实验复现环境搭配
- 相关神经网络、机器学习、图论的知识暂且未知

论文:Sampling Enclosing Subgraphs for Link Prediction

abstract 和 introduction 翻译

ABSTRACT Link prediction is a fundamental problem for graph-structured data (e.g., social networks, drug side-effect networks, etc.). Graph neural networks have offered robust solutions for this problem, specifically by learning the representation of the subgraph enclosing the target link (i.e., pair of nodes). However, these solutions do not scale well to large graphs as extraction and operation on enclosing subgraphs are computationally expensive, especially for large graphs. This paper presents a scalable link prediction solution, that we call ScaLed, which utilizes sparse enclosing subgraphs to make predictions. To extract sparse enclosing subgraphs, ScaLed takes multiple random walks from a target pair of nodes, then operates on the sampled enclosing subgraph induced by all visited nodes. By leveraging the smaller sampled enclosing subgraph, ScaLed can scale to larger graphs with much less overhead while maintaining high accuracy. ScaLed further provides the flexibility to control the trade-off between computation overhead and accuracy. Through comprehensive experiments, we have shown that ScaLed can produce comparable accuracy to those reported by the existing subgraph representation learning frameworks while being less computationally demanding.

摘要：链接预测是图结构数据（例如社交网络、药物副作用网络等）的一个基本问题。图神经网络为这个问题提供了鲁棒的解决方案，尤其是通过对包含目标链接（即节点对）的子图的表示的学习。然而，这些解决方案不能很好地扩展到大型图，因为对封闭子图的提取和操作在计算上是昂贵的。本文提出了一种可扩展的链接预测解决方案，我们称之为 ScaLed，它利用稀疏封闭子图进行预测。为了提取稀疏封闭子图，ScaLed 从目标节点对开始进行多次随机地游走，然后对所有访问节点产生的采样封闭子图进行操

作。通过利用较小的采样封闭子图，ScaLed 可以在保持高精度的同时以更少的开销扩展到更大的图。通过全面的实验，我们表明 ScaLed 可以达到那些现有的子图表示学习框架所报告的精度，同时拥有较低的计算要求。

INTRODUCTION Graph-structured data such as user interactions, collaborations, protein-protein interactions, drug-drug interactions are prevalent in natural and social sciences. Link prediction —a fundamental problem on graph-structured data—intends to quantify the likelihood of a link (or interaction) occurring between a pair of nodes (e.g., proteins, drugs, etc.). Link prediction has many diverse applications such as predicting drug side effects, drug-repurposing [14], understanding molecule interactions [18], friendship recommendation [9], and recommender systems [39]. Many solutions to link prediction problem [24, 26–28, 35] has been proposed ranging from simple heuristics (e.g., common neighbors, Adamic-Adar [1], Katz [19]) to graph neural networks (GNNs) [5, 6, 17, 21, 30, 45]. Among these solutions, GNNs [15, 36, 48] have emerged as the widely-accepted and successful solution for learning rich latent representations of graph data to tackle link prediction problems. The early GNNs focused on shallow encoders [13, 32] in which the latent nodes' representations was first learnt through a sequence of random walks, and then a likelihood of a link is determined by combining its two-end nodes' latent representations. However, these shallow encoders were limited by not incorporating nodal features and their incompatibility with inductive settings as they require that all nodes are present for training. These two challenges were (partially) addressed with the emergence of message-passing graph neural networks [16, 22, 37]. These advancements motivate the research on determining and extending the expressive power of GNNs [3, 12, 40-42, 46] for all downstream tasks of link prediction, node classification, and graph classification. For link prediction, subgraph-based representation learning (SGRL) methods [5, 6, 25, 30, 45]—by learning the enclosing subgraphs

around the two-end nodes rather than independently learning two end-node's embedding—have improved GNNs expressive power, and offered state-of-the-art solutions. However, these solutions suffer from the lack of scalability, thus preventing them to be applied to large-scale graphs. This is primarily due to the computation overhead in extracting, preprocessing, and learning (large) enclosing subgraphs for any pair of nodes. We focus on addressing this scalability issue. Contribution. We introduce Sampling Enclosing Subgraph for Link Prediction (ScaLed) to extend SGRL methods and enhance their scalability. The crux of ScaLed is to sample enclosing subgraphs using a sequence of random walks. This sampling reduces the computational overhead of large subgraphs while maintaining their key structural information. can be integrated into any GNN, and also offers parallelizability and model compression that can be exploited for large-scale graphs. Furthermore, the two hyperparameters, walk length and number of walks, in ScaLed provides a way to control the trade-off between scalability and accuracy, if needed. Our extensive experiments on real-world datasets demonstrate that ScaLed produces comparable results to the state-of-the-art methods (e.g, SEAL [45]) in link prediction, but requiring magnitudes less training data, time, and memory. ScaLed combines the benefits of SGRL framework and random walks for link prediction.

引言：用户交互协作、蛋白质间相互作用、药物间相互作用等图结构数据在自然科学和社会科学中普遍存在。链接预测——图结构数据的一个基本问题——旨在量化一对节点（例如蛋白质、药物等）之间产生链接（或交互）的可能性。链接预测有许多不同的应用，例如预测药物副作用、药物重新利用、理解分子间的相互作用和推荐系统。人们已经提出了许多链接预测问题的解决方案，从简单的启发式方法（例如共同邻居、Adamic-Adar、Katz）到图神经网络（GNNs）。在这些解决方案中，GNN 已经成为了学习图数据的丰富潜在表示以解决链接预测问题的前景良好的解决方案。早期的 GNN 专注于浅层编码器，其中潜在节点的表示首先通过一系列随机游走来获取，然后通过组合其两端节点的潜在表示来确定链接的可能性。然而，这些浅层编

码器因未结合节点特征且与感应设置不兼容而受到限制。这两个问题已（部分）通过消息传递图神经网络得到解决。这些进步激发了关于确定和扩展 GNN 对于链路预测、节点分类和图分类等所有下游任务的表达能力的研究。对于链接预测，基于子图的表示学习（SGRL）方法——通过学习两端节点周围的封闭子图，而不是独立学习两端节点的嵌入——提高了 GNN 的表达能力，并提供了最先进的解决方案。然而，这些解决方案缺乏大规模图的可扩展性。这主要是由于提取、预处理和学习（大）封闭子图的计算开销。我们引入了用于链路预测的采样封闭子图（ScaLed）来扩展 SGRL 方法并增强其可扩展性。ScaLed 使用一系列随机游走对封闭子图进行采样。这种采样减少了大型子图的计算开销，同时保留了关键的结构信息。ScaLed 可以集成到任何 GNN 中，并且还提供了可用于大规模图的并行性和模型压缩。如果需要，ScaLed 中的两个超参数（步行长度和步行次数）提供了一种控制可扩展性和准确性之间的权衡的方法。我们对真实世界数据集的广泛实验表明，ScaLed 在链路预测中产生的结果与最先进的方法（例如 SEAL）相当，但需要的训练数据、时间和内存要少得多。

问题描述

- 链接预测是图形结构化数据的基本问题。
- 图神经网络通过学习包围目标链接的子图的表示，为链接预测提供了强大的解决方案。
- 但是，由于计算开销，这些解决方案无法扩展到大型图形。
- ScaLed 是一种可扩展的链路预测解决方案，它利用稀疏封闭子图。
- ScaLed 通过从目标节点对中随机游走来提取稀疏封闭子图，并对所有访问节点诱导的采样子图进行操作。
- ScaLed 可以扩展到更大的图形，计算开销更少，同时保持高精度。
- 实验结果表明，ScaLed 的精度与现有方法相当，同时需要更少的训练数据、时间和内存。

链路预测 - Link Prediction

图论前导知识

假设有图 $G=(V,E,A)$，其中 V 为图的节点集合，E 为图的边集合，而张量 $A\in\mathbb{R}^{n\times n\times d}$ 包含了所有节点的属性（例如，用户资料）和边的属性（例如，交互作用的强度或类型）。对于每个节点 $v\in V$，它的属性（如果有的话）存储在对角组件 A_{vv} 中，而非对角线组件 A_{uv} 则可以包含边 (u,v) 的属性，如果 $(u,v)\in E$；否则 $A_{uv}=0$。

链路预测问题

在链路预测中的目标是根据观察到的张量 A 推断目标节点对之间是否存在边。学习问题是找到一个似然（或评分）函数 f，它为每对目标节点 $(u,v)\notin E$ 分配交互似然性（分数）$\hat{A}_{uv}$，这些节点之间的关系未被观察到。较大的 $\hat{A}_{uv}$ 表示 (u,v) 形成链接或缺失链接的可能性较高。函数 f 可以表示为 $\hat{A}_{uv}=f(u,v,A|\Theta)$，其中 Θ 表示模型参数。

ScaLed 模型

定义 1（包围子图 [43]）：给定一个图 G，目标节点对 (u,v) 周围的 h-跳包围子图是从 G 中导出的子图 G_{huv}，其节点集合是由满足条件 $d(j,x)\leq h$ 或 $d(j,y)\leq h$ 的节点 j 组成，其中 $d(i,j)$ 表示节点 i 和节点 j 之间的测地距离。

这里，G_{huv} 表示目标节点 (u,v) 周围的 h-跳包围子图，该子图是从图 G 中选取的，其中包括那些满足条件 $d(j,x)\leq h$ 或 $d(j,y)\leq h$ 的节点 j。

在 SEAL 模型中，对于每对目标节点 (u,v)，它们的包围子图 G_{huv} 是通过两个 h-跳广度优先搜索（BFS）找到的，其中每个 BFS 都从 u 和 v 开始。包围子图中的节点还使用 Double-Radius Node Labeling（DRNL）哈希函数 [43] 增强，以指示它们与目标节点的距离：

$$DRNL_L(x,G_{huv})=1+\min(d_{xu},d_{xv})+\left\lfloor\frac{d'}{2}\right\rfloor\left\lceil\frac{d'}{2}-1\right\rceil$$

其中，x 代表子图 G_{huv} 中的节点，d_{xu} 是节点 x 到 u 的测地距离

（当移除节点 v 时），而 $d' = d_{xu} + d_{xv}$。请注意，节点 x 到目标节点 u 的距离是通过从子图中移除另一个目标节点 v 来计算的。目标节点被赋予标签 1，而到至少一个目标节点的距离为无穷大的节点被赋予标签 0。然后，每个节点标签都用其一热编码表示，并扩展了初始节点特征（如果有的话）。子图 G_{huv} 以及增强的节点特征被输入到图神经网络中，用于预测边的存在或不存在。在 SEAL 中，链路预测被视为通过确定包围子图是否会由目标节点对之间的链接关闭而进行的二进制分类。因此，SEAL 使用了图池化机制来计算用于分类任务的包围子图表示。

ScaLed。通过观察，SEAL 及其变体的计算瓶颈源于包围子图的指数增长和大小。我们提出了采样包围子图，其规模更可控：

定义 2（随机游走采样包围子图）。给定图 G，目标节点对 (u,v) 周围的 h-跳采样包围子图是从 G 中导出的子图 $\hat{G}_{hkuv}$，其节点集合为 $\hat{V}_{hkuv} \in W_{uhk} \cup W_{vhk}$，其中 W_{ihk} 是从节点 i 开始的 k 个长度为 h 的随机游走访问的节点集合。

图 1(b) 展示了在图 1(a) 的原始图中针对目标对 (u,v) 的采样包围子图，其中 $h=2$ 和 $k=2$。在这里，$W_{vhk} = \{v,d,e,f,g\}$，$W_{uhk} = \{u,a,b,c\}$，结果是 $\hat{V}_{hkuv} = \{a,b,c,d,e,f,g,u,v\}$。图中包含的子图包含了 $\hat{V}_{hkuv}$ 中节点之间的所有节点和边。

与定义 1 相比，可以得出几个重要观察：(i) 采样包围子图 $\hat{G}_{hkuv}$ 是包围子图 G_{huv} 的子图，因为 h-长度的随机游走无法到达距起始节点 h 跳远的节点；(ii) 采样子图的大小受到 $O(hk)$ 的限制，并由这两个参数控制，相对于定义 1 中 h 的指数增长。这两个观察突显了 ScaLed 通过用稀疏的（子）子图替代密集的包围子图，提供了可扩展性。ScaLed 还具有通过其采样参数 h 和 k 控制稀疏性和可扩展性程度的灵活性。