## Evaluating Graph Neural Networks for Link Prediction: Current Pitfalls and New Benchmarking

## 摘要

Link prediction attempts to predict whether an unseen edge exists based on only a portion of edges of a graph. A flurry of methods have been introduced in recent years that attempt to make use of graph neural networks (GNNs) for this task. Furthermore, new and diverse datasets have also been created to better evaluate the effectiveness of these new models. However, multiple pitfalls currently exist that hinder our ability to properly evaluate these new methods. These pitfalls mainly include: (1) Lower than actual performance on multiple baselines, (2) A lack of a unified data split and evaluation metric on some datasets, and (3) An unrealistic evaluation setting that uses easy negative samples. To overcome these challenges, we first conduct a fair comparison across prominent methods and datasets, utilizing the same dataset and hyperparameter search settings. We then create a more practical evaluation setting based on a Heuristic Related Sampling Technique (HeaRT), which samples hard negative samples via multiple heuristics. The new evaluation setting helps promote new challenges and opportunities in link prediction by aligning the evaluation with real-world situations. Our implementation and data are available at https://github.com/Juanhui28/HeaRT

链接预测试图仅基于图的一部分边来预测是否存在看不见的边。近年来，人们引入了一系列方法，试图利用图神经网络（GNN）来完成这项任务。此外，还创建了新的多样化数据集，以更好地评估这些新模型的有效性。然而，目前存在多个陷阱，阻碍了我们正确评估这些新方法的能力。这些陷阱主要包括：（1）在多个基线上的性能低于实际，（2）在一些数据集上缺乏统一的数据分割和评估指标，以及（3）使用简单的负样本的不切实际的评估设置。为了克服这些挑战，我们首先利用相同的数据集和超参数搜索设置，在突出的方法和数据集之间进行公平的比较。然后，我们基于启发式相关采样技术（HeaRT）创建了一个更实用的评估设置，该技术通过多种启发式方法对硬阴性样本进行采样。新的评估设置通过使评估与现实世界的情况相一致，有助于促进链接预测方面的新挑战和机遇。我们的实施和数据可在https://github.com/Juanhui28/HeaRT

文章主页 : https://paperswithcode.com/paper/evaluating-graph-neural-networks-for-link

PDF : https://arxiv.org/pdf/2306.10453v1.pdf

github : https://github.com/juanhui28/heart

## Knowledge Graph Embeddings in the Biomedical Domain: Are They Useful? A Look at Link Prediction, Rule Learning, and Downstream Polypharmacy Tasks

## 摘要

Knowledge graphs are powerful tools for representing and organising complex biomedical data. Several knowledge graph embedding algorithms have been proposed to learn from and complete knowledge graphs. However, a recent study demonstrates the limited efficacy of these embedding algorithms when applied to biomedical knowledge graphs, raising the question of whether knowledge graph embeddings have limitations in biomedical settings. This study aims to apply state-of-the-art knowledge graph embedding models in the context of a recent biomedical knowledge graph, BioKG, and evaluate their performance and potential downstream uses. We achieve a three-fold improvement in terms of performance based on the HITS@10 score over previous work on the same biomedical knowledge graph. Additionally, we provide interpretable predictions through a rule-based method. We demonstrate that knowledge graph embedding models are applicable in practice by evaluating the best-performing model on four tasks that represent real-life polypharmacy situations. Results suggest that knowledge learnt from large biomedical knowledge graphs can be transferred to such downstream use cases. Our code is available at https://github.com/aryopg/biokge.

知识图是表示和组织复杂生物医学数据的强大工具。已经提出了几种知识图嵌入算法来学习和完善知识图。然而，最近的一项研究表明，这些嵌入算法在应用于生物医学知识图时效果有限，这引发了知识图嵌入在生物医学环境中是否有局限性的问题。本研究旨在将最先进的知识图嵌入模型应用于最近的生物医学知识图BioKG，并评估其性能和潜在的下游用途。基于HITS@10在同一生物医学知识图谱上的先前工作得分。此外，我们通过基于规则的方法提供可解释的预测。我们通过评估四个任务中表现最好的模型来证明知识图嵌入模型在实践中是适用的，这四个任务代表了现实生活中的多药物情况。结果表明，从大型生物医学知识图中学习的知识可以转移到此类下游用例中。我们的代码可在https://github.com/aryopg/biokge.

文章主页 : https://paperswithcode.com/paper/knowledge-graph-embeddings-in-the-biomedical

PDF : https://arxiv.org/pdf/2305.19979v1.pdf

github : https://github.com/aryopg/biokge

## DotHash: Estimating Set Similarity Metrics for Link Prediction and Document Deduplication

## 摘要

Metrics for set similarity are a core aspect of several data mining tasks. To remove duplicate results in a Web search, for example, a common approach looks at the Jaccard index between all pairs of pages. In social network analysis, a much-celebrated metric is the Adamic-Adar index, widely used to compare node neighborhood sets in the important problem of predicting links. However, with the increasing amount of data to be processed, calculating the exact similarity between all pairs can be intractable. The challenge of working at this scale has motivated research into efficient estimators for set similarity metrics. The two most popular estimators, MinHash and SimHash, are indeed used in applications such as document deduplication and recommender systems where large volumes of data need to be processed. Given the importance of these tasks, the demand for advancing estimators is evident. We propose DotHash, an unbiased estimator for the intersection size of two sets. DotHash can be used to estimate the Jaccard index and, to the best of our knowledge, is the first method that can also estimate the Adamic-Adar index and a family of related metrics. We formally define this family of metrics, provide theoretical bounds on the probability of estimate errors, and analyze its empirical performance. Our experimental results indicate that DotHash is more accurate than the other estimators in link prediction and detecting duplicate documents with the same complexity and similar comparison time.

集合相似性度量是几个数据挖掘任务的核心方面。例如，为了删除Web搜索中的重复结果，一种常见的方法是查看所有页面对之间的Jaccard索引。在社交网络分析中，一个非常著名的指标是Adamic-Adar指数，在预测链接的重要问题中，它被广泛用于比较节点邻域集。然而，随着要处理的数据量的增加，计算所有对之间的确切相似性可能很困难。在这种规模下工作的挑战促使人们研究集合相似性度量的有效估计量。两种最流行的估计器MinHash和SimHash确实用于需要处理大量数据的应用程序，如文档重复数据消除和推荐系统。考虑到这些任务的重要性，对高级估计器的需求是显而易见的。我们提出了DotHash，一个关于两个集合的交集大小的无偏估计量。DotHash可以用于估计Jaccard指数，据我们所知，它是第一种也可以估计Adamic-Adar指数和一系列相关指标的方法。我们正式定义了这类度量，提供了估计误差概率的理论界限，并分析了其实证性能。我们的实验结果表明，在相同复杂度和相似比较时间的情况下，DotHash在链接预测和检测重复文档方面比其他估计器更准确。

文章主页 : https://paperswithcode.com/paper/dothash-estimating-set-similarity-metrics-for

PDF : https://arxiv.org/pdf/2305.17310v1.pdf

github : https://github.com/mikeheddes/dothash

## Link Prediction without Graph Neural Networks

## 摘要

Link prediction, which consists of predicting edges based on graph features, is a fundamental task in many graph applications. As for several related problems, Graph Neural Networks (GNNs), which are based on an attribute-centric message-passing paradigm, have become the predominant framework for link prediction. GNNs have consistently outperformed traditional topology-based heuristics, but what contributes to their performance? Are there simpler approaches that achieve comparable or better results? To answer these questions, we first identify important limitations in how GNN-based link prediction methods handle the intrinsic class imbalance of the problem -- due to the graph sparsity -- in their training and evaluation. Moreover, we propose Gelato, a novel topology-centric framework that applies a topological heuristic to a graph enhanced by attribute information via graph learning. Our model is trained end-to-end with an N-pair loss on an unbiased training set to address class imbalance. Experiments show that Gelato is 145% more accurate, trains 11 times faster, infers 6,000 times faster, and has less than half of the trainable parameters compared to state-of-the-art GNNs for link prediction.

链接预测是许多图应用中的一项基本任务，它包括基于图特征的边缘预测。对于几个相关的问题，基于以属性为中心的消息传递范式的图神经网络（GNN）已经成为链路预测的主要框架。GNN一直优于传统的基于拓扑的启发式算法，但是什么对它们的性能有贡献？有没有更简单的方法可以达到类似或更好的结果？为了回答这些问题，我们首先确定了基于GNN的链接预测方法在训练和评估中如何处理问题的内在类不平衡（由于图的稀疏性）的重要局限性。此外，我们提出了Gelato，这是一种新的以拓扑为中心的框架，它通过图学习将拓扑启发式应用于由属性信息增强的图。我们的模型是在无偏训练集上用N对损失进行端到端训练的，以解决类不平衡问题。实验表明，与最先进的链路预测GNN相比，Gelato的准确率高145%，训练速度快11倍，推断速度快6000倍，并且具有不到一半的可训练参数。

文章主页 : https://paperswithcode.com/paper/link-prediction-without-graph-neural-networks

PDF : https://arxiv.org/pdf/2305.13656v1.pdf

github : https://github.com/facebookresearch/SEAL\_OGB

## Friendly Neighbors: Contextualized Sequence-to-Sequence Link Prediction

## 摘要

We propose KGT5-context, a simple sequence-to-sequence model for link prediction (LP) in knowledge graphs (KG). Our work expands on KGT5, a recent LP model that exploits textual features of the KG, has small model size, and is scalable. To reach good predictive performance, however, KGT5 relies on an ensemble with a knowledge graph embedding model, which itself is excessively large and costly to use. In this short paper, we show empirically that adding contextual information - i.e., information about the direct neighborhood of the query entity - alleviates the need for a separate KGE model to obtain good performance. The resulting KGT5-context model is simple, reduces model size significantly, and obtains state-of-the-art performance in our experimental study.

我们提出了KGT5上下文，这是一个用于知识图（KG）中链接预测（LP）的简单序列到序列模型。我们的工作扩展了KGT5，这是一个最近的LP模型，它利用了KG的文本特征，具有较小的模型大小，并且是可扩展的。然而，为了达到良好的预测性能，KGT5依赖于与知识图嵌入模型的集成，该模型本身过大且使用成本高昂。在这篇简短的论文中，我们实证地表明，添加上下文信息——即关于查询实体的直接邻域的信息——减轻了对单独的KGE模型的需求，以获得良好的性能。由此产生的KGT5上下文模型简单，显著减小了模型大小，并在我们的实验研究中获得了最先进的性能。

文章主页 : https://paperswithcode.com/paper/friendly-neighbors-contextualized-sequence-to

PDF : https://arxiv.org/pdf/2305.13059v2.pdf

github : https://github.com/uma-pi1/kgt5-context

## Musical Voice Separation as Link Prediction: Modeling a Musical Perception Task as a Multi-Trajectory Tracking Problem

## 摘要

This paper targets the perceptual task of separating the different interacting voices, i.e., monophonic melodic streams, in a polyphonic musical piece. We target symbolic music, where notes are explicitly encoded, and model this task as a Multi-Trajectory Tracking (MTT) problem from discrete observations, i.e., notes in a pitch-time space. Our approach builds a graph from a musical piece, by creating one node for every note, and separates the melodic trajectories by predicting a link between two notes if they are consecutive in the same voice/stream. This kind of local, greedy prediction is made possible by node embeddings created by a heterogeneous graph neural network that can capture inter- and intra-trajectory information. Furthermore, we propose a new regularization loss that encourages the output to respect the MTT premise of at most one incoming and one outgoing link for every node, favouring monophonic (voice) trajectories; this loss function might also be useful in other general MTT scenarios. Our approach does not use domain-specific heuristics, is scalable to longer sequences and a higher number of voices, and can handle complex cases such as voice inversions and overlaps. We reach new state-of-the-art results for the voice separation task in classical music of different styles.

本文的目标是在复调音乐作品中分离不同的相互作用的声音，即单声道旋律流。我们以符号音乐为目标，其中音符被明确编码，并根据离散观测（即音高时间空间中的音符）将该任务建模为多轨迹跟踪（MTT）问题。我们的方法通过为每个音符创建一个节点，从音乐作品中构建一个图，并通过预测两个音符之间的链接来分离旋律轨迹，如果它们在同一语音/流中是连续的。这种局部贪婪预测是通过异构图神经网络创建的节点嵌入实现的，该网络可以捕获轨迹间和轨迹内信息。此外，我们提出了一种新的正则化损失，该损失鼓励输出尊重每个节点最多一个传入和一个传出链路的MTT前提，有利于单声道（语音）轨迹；这种损失函数在其他一般MTT场景中也可能是有用的。我们的方法不使用特定领域的启发式方法，可扩展到更长的序列和更高数量的语音，并且可以处理复杂的情况，如语音反转和重叠。对于不同风格的古典音乐中的声音分离任务，我们得出了最先进的新结果。

文章主页 : https://paperswithcode.com/paper/musical-voice-separation-as-link-prediction

PDF : https://arxiv.org/pdf/2304.14848v1.pdf

github : https://github.com/manoskary/vocsep\_ijcai2023

## IMF: Interactive Multimodal Fusion Model for Link Prediction

## 摘要

Link prediction aims to identify potential missing triples in knowledge graphs. To get better results, some recent studies have introduced multimodal information to link prediction. However, these methods utilize multimodal information separately and neglect the complicated interaction between different modalities. In this paper, we aim at better modeling the inter-modality information and thus introduce a novel Interactive Multimodal Fusion (IMF) model to integrate knowledge from different modalities. To this end, we propose a two-stage multimodal fusion framework to preserve modality-specific knowledge as well as take advantage of the complementarity between different modalities. Instead of directly projecting different modalities into a unified space, our multimodal fusion module limits the representations of different modalities independent while leverages bilinear pooling for fusion and incorporates contrastive learning as additional constraints. Furthermore, the decision fusion module delivers the learned weighted average over the predictions of all modalities to better incorporate the complementarity of different modalities. Our approach has been demonstrated to be effective through empirical evaluations on several real-world datasets. The implementation code is available online at https://github.com/HestiaSky/IMF-Pytorch.

链接预测旨在识别知识图中潜在的缺失三元组。为了获得更好的结果，最近的一些研究将多模态信息引入到链路预测中。然而，这些方法分别利用多模态信息，忽略了不同模态之间的复杂交互。在本文中，我们旨在更好地对模态间信息建模，从而引入一种新的交互式多模态融合（IMF）模型来集成来自不同模态的知识。为此，我们提出了一个两阶段的多模态融合框架，以保留模态特定知识并利用不同模态之间的互补性。我们的多模态融合模块不是直接将不同模态投影到统一的空间中，而是限制不同模态的独立表示，同时利用双线性池进行融合，并将对比学习作为附加约束。此外，决策融合模块在所有模态的预测上提供学习的加权平均值，以更好地结合不同模态的互补性。通过对几个真实世界数据集的实证评估，我们的方法已被证明是有效的。实现代码可在线获取，网址为https://github.com/HestiaSky/IMF-Pytorch.

文章主页 : https://paperswithcode.com/paper/imf-interactive-multimodal-fusion-model-for

PDF : https://arxiv.org/pdf/2303.10816v1.pdf

github : https://github.com/hestiasky/imf-pytorch

## You Only Transfer What You Share: Intersection-Induced Graph Transfer Learning for Link Prediction

## 摘要

Link prediction is central to many real-world applications, but its performance may be hampered when the graph of interest is sparse. To alleviate issues caused by sparsity, we investigate a previously overlooked phenomenon: in many cases, a densely connected, complementary graph can be found for the original graph. The denser graph may share nodes with the original graph, which offers a natural bridge for transferring selective, meaningful knowledge. We identify this setting as Graph Intersection-induced Transfer Learning (GITL), which is motivated by practical applications in e-commerce or academic co-authorship predictions. We develop a framework to effectively leverage the structural prior in this setting. We first create an intersection subgraph using the shared nodes between the two graphs, then transfer knowledge from the source-enriched intersection subgraph to the full target graph. In the second step, we consider two approaches: a modified label propagation, and a multi-layer perceptron (MLP) model in a teacher-student regime. Experimental results on proprietary e-commerce datasets and open-source citation graphs show that the proposed workflow outperforms existing transfer learning baselines that do not explicitly utilize the intersection structure.

链接预测是许多现实世界应用程序的核心，但当感兴趣的图稀疏时，其性能可能会受到阻碍。为了缓解稀疏性引起的问题，我们研究了一个以前被忽视的现象：在许多情况下，可以为原始图找到一个密连接的互补图。密集图可以与原始图共享节点，这为传递选择性的、有意义的知识提供了一个自然的桥梁。我们将这种设置确定为图交叉诱导迁移学习（GITL），其动机是在电子商务或学术合作预测中的实际应用。我们制定了一个框架，以便在这种情况下有效地利用结构性优势。我们首先使用两个图之间的共享节点创建一个交集子图，然后将知识从源丰富的交集子图转移到完整的目标图。在第二步中，我们考虑两种方法：改进的标签传播和师生制度中的多层感知器（MLP）模型。在专有电子商务数据集和开源引文图上的实验结果表明，所提出的工作流优于未明确利用交叉结构的现有迁移学习基线。

文章主页 : https://paperswithcode.com/paper/you-only-transfer-what-you-share-intersection

PDF : https://arxiv.org/pdf/2302.14189v2.pdf

github : https://github.com/amazon-science/gnn-tail-generalization

## PaGE-Link: Path-based Graph Neural Network Explanation for Heterogeneous Link Prediction

## 摘要

Transparency and accountability have become major concerns for black-box machine learning (ML) models. Proper explanations for the model behavior increase model transparency and help researchers develop more accountable models. Graph neural networks (GNN) have recently shown superior performance in many graph ML problems than traditional methods, and explaining them has attracted increased interest. However, GNN explanation for link prediction (LP) is lacking in the literature. LP is an essential GNN task and corresponds to web applications like recommendation and sponsored search on web. Given existing GNN explanation methods only address node/graph-level tasks, we propose Path-based GNN Explanation for heterogeneous Link prediction (PaGE-Link) that generates explanations with connection interpretability, enjoys model scalability, and handles graph heterogeneity. Qualitatively, PaGE-Link can generate explanations as paths connecting a node pair, which naturally captures connections between the two nodes and easily transfer to human-interpretable explanations. Quantitatively, explanations generated by PaGE-Link improve AUC for recommendation on citation and user-item graphs by 9 - 35% and are chosen as better by 78.79% of responses in human evaluation.

透明度和问责制已经成为黑匣子机器学习（ML）模型的主要关注点。对模型行为的适当解释增加了模型的透明度，并帮助研究人员开发出更负责任的模型。图神经网络（GNN）最近在许多图ML问题上显示出比传统方法更优越的性能，并且解释它们引起了越来越多的兴趣。然而，文献中缺乏对链接预测（LP）的GNN解释。LP是一项重要的GNN任务，与推荐和赞助搜索等网络应用程序相对应。鉴于现有的GNN解释方法仅针对节点/图级任务，我们提出了基于路径的异构链路预测GNN解释（PaGE-Link），该方法生成具有连接可解释性的解释，具有模型可扩展性，并处理图的异构性。定性地说，PaGE-Link可以将解释生成为连接节点对的路径，从而自然地捕捉两个节点之间的连接，并容易地转换为人类可解释的解释。从数量上讲，PaGE-Link生成的解释将引文和用户项目图上推荐的AUC提高了9-35%，在人类评估中，78.79%的回答被选为更好。

文章主页 : https://paperswithcode.com/paper/page-link-path-based-graph-neural-network

PDF : https://arxiv.org/pdf/2302.12465v3.pdf

github : https://github.com/amazon-science/page-link-path-based-gnn-explanation

## Learning Representations of Bi-level Knowledge Graphs for Reasoning beyond Link Prediction

## 摘要

Knowledge graphs represent known facts using triplets. While existing knowledge graph embedding methods only consider the connections between entities, we propose considering the relationships between triplets. For example, let us consider two triplets $T\_1$ and $T\_2$ where $T\_1$ is (Academy\_Awards, Nominates, Avatar) and $T\_2$ is (Avatar, Wins, Academy\_Awards). Given these two base-level triplets, we see that $T\_1$ is a prerequisite for $T\_2$. In this paper, we define a higher-level triplet to represent a relationship between triplets, e.g., $\langle T\_1$, PrerequisiteFor, $T\_2\rangle$ where PrerequisiteFor is a higher-level relation. We define a bi-level knowledge graph that consists of the base-level and the higher-level triplets. We also propose a data augmentation strategy based on the random walks on the bi-level knowledge graph to augment plausible triplets. Our model called BiVE learns embeddings by taking into account the structures of the base-level and the higher-level triplets, with additional consideration of the augmented triplets. We propose two new tasks: triplet prediction and conditional link prediction. Given a triplet $T\_1$ and a higher-level relation, the triplet prediction predicts a triplet that is likely to be connected to $T\_1$ by the higher-level relation, e.g., $\langle T\_1$, PrerequisiteFor, ?$\rangle$. The conditional link prediction predicts a missing entity in a triplet conditioned on another triplet, e.g., $\langle T\_1$, PrerequisiteFor, (Avatar, Wins, ?)$\rangle$. Experimental results show that BiVE significantly outperforms all other methods in the two new tasks and the typical base-level link prediction in real-world bi-level knowledge graphs.

知识图使用三元组表示已知事实。虽然现有的知识图嵌入方法只考虑实体之间的连接，但我们建议考虑三元组之间的关系。例如，让我们考虑两个三元组$T\_1>和$T\_2$，其中$T\_1$是（Academy\_Awards，Nominates，Avatar），$T\_2美元是（Avatar，Wins，Academy\_Awards）。给定这两个基本级别的三元组，我们看到$T\_1$是$T\_2$的先决条件。在本文中，我们定义了一个更高级别的三元组来表示三元组之间的关系，例如$\langle T\_1$、PrerequisiteFor、$T\_2\rangle$，其中PrerequisteFor是一个更高级的关系。我们定义了一个双层知识图，它由基本层和更高层的三元组组成。我们还提出了一种基于双层知识图上随机行走的数据扩充策略，以扩充看似合理的三元组。我们称为BiVE的模型通过考虑基本层和更高层三元组的结构来学习嵌入，并额外考虑增强三元组。我们提出了两个新任务：三元组预测和条件链接预测。给定三元组$T\_1$和更高级别的关系，三元组预测预测一个可能通过更高级别关系连接到$T\_1$$的三元组，例如$\langle T\_1$，PrerequisiteFor$\激怒了$。条件链接预测预测以另一个三元组为条件的三元组中缺失的实体，例如$\langle T\_1$，PrerequisiteFor，（Avatar，Wins，？）$\langle$。实验结果表明，在这两个新任务和真实世界的双层知识图中的典型基层链接预测中，BiVE显著优于所有其他方法。

文章主页 : https://paperswithcode.com/paper/learning-representations-of-bi-level

PDF : https://arxiv.org/pdf/2302.02601v3.pdf

github : https://github.com/bdi-lab/bive

## Neural Common Neighbor with Completion for Link Prediction

## 摘要

Despite its outstanding performance in various graph tasks, vanilla Message Passing Neural Network (MPNN) usually fails in link prediction tasks, as it only uses representations of two individual target nodes and ignores the pairwise relation between them. To capture the pairwise relations, some models add manual features to the input graph and use the output of MPNN to produce pairwise representations. In contrast, others directly use manual features as pairwise representations. Though this simplification avoids applying a GNN to each link individually and thus improves scalability, these models still have much room for performance improvement due to the hand-crafted and unlearnable pairwise features. To upgrade performance while maintaining scalability, we propose Neural Common Neighbor (NCN), which uses learnable pairwise representations. To further boost NCN, we study the unobserved link problem. The incompleteness of the graph is ubiquitous and leads to distribution shifts between the training and test set, loss of common neighbor information, and performance degradation of models. Therefore, we propose two intervention methods: common neighbor completion and target link removal. Combining the two methods with NCN, we propose Neural Common Neighbor with Completion (NCNC). NCN and NCNC outperform recent strong baselines by large margins. NCNC achieves state-of-the-art performance in link prediction tasks. Our code is available at https://github.com/GraphPKU/NeuralCommonNeighbor.

尽管普通消息传递神经网络在各种图任务中表现出色，但它通常在链路预测任务中失败，因为它只使用两个单独目标节点的表示，而忽略了它们之间的成对关系。为了捕捉成对关系，一些模型将手动特征添加到输入图中，并使用MPNN的输出来生成成对表示。相反，其他人直接使用手动特征作为成对表示。尽管这种简化避免了对每个链路单独应用GNN，从而提高了可扩展性，但由于手工制作和不可编程的成对功能，这些模型仍有很大的性能改进空间。为了在保持可扩展性的同时提高性能，我们提出了使用可学习成对表示的神经公共邻居（NCN）。为了进一步提高NCN，我们研究了未观测到的链路问题。图的不完全性是普遍存在的，并导致训练集和测试集之间的分布偏移、公共邻居信息的丢失以及模型的性能退化。因此，我们提出了两种干预方法：公共邻居完成和目标链路去除。将这两种方法与NCN相结合，提出了具有补全的神经公共邻居（NCNC）。NCN和NCNC在很大程度上优于最近的强劲基线。NCNC在链路预测任务中实现了最先进的性能。我们的代码可在https://github.com/GraphPKU/NeuralCommonNeighbor.

文章主页 : https://paperswithcode.com/paper/neural-common-neighbor-with-completion-for

PDF : https://arxiv.org/pdf/2302.00890v2.pdf

github : https://github.com/GraphPKU/NeuralCommonNeighbor

## Simplifying Subgraph Representation Learning for Scalable Link Prediction

## 摘要

Link prediction on graphs is a fundamental problem. Subgraph representation learning approaches (SGRLs), by transforming link prediction to graph classification on the subgraphs around the links, have achieved state-of-the-art performance in link prediction. However, SGRLs are computationally expensive, and not scalable to large-scale graphs due to expensive subgraph-level operations. To unlock the scalability of SGRLs, we propose a new class of SGRLs, that we call Scalable Simplified SGRL (S3GRL). Aimed at faster training and inference, S3GRL simplifies the message passing and aggregation operations in each link's subgraph. S3GRL, as a scalability framework, accommodates various subgraph sampling strategies and diffusion operators to emulate computationally-expensive SGRLs. We propose multiple instances of S3GRL and empirically study them on small to large-scale graphs. Our extensive experiments demonstrate that the proposed S3GRL models scale up SGRLs without significant performance compromise (even with considerable gains in some cases), while offering substantially lower computational footprints (e.g., multi-fold inference and training speedup).

图上的链接预测是一个基本问题。子图表示学习方法（SGRL）通过将链接预测转换为对链接周围的子图进行图分类，在链接预测方面取得了最先进的性能。然而，SGRL在计算上是昂贵的，并且由于昂贵的子图级操作而不能扩展到大规模图。为了解锁SGRL的可扩展性，我们提出了一类新的SGRL，称为可扩展简化SGRL（S3GRL）。为了更快地进行训练和推理，S3GRL简化了每个链接子图中的消息传递和聚合操作。S3GRL作为一个可伸缩性框架，提供了各种子图采样策略和扩散算子，以模拟计算成本高昂的SGRL。我们提出了S3GRL的多个实例，并在小到大图上对它们进行了实证研究。我们的大量实验表明，所提出的S3GRL模型在没有显著性能折衷的情况下扩大了SGRL（在某些情况下甚至有相当大的增益），同时提供了显著更低的计算足迹（例如，多重推理和训练加速）。

文章主页 : https://paperswithcode.com/paper/simplifying-subgraph-representation-learning

PDF : https://arxiv.org/pdf/2301.12562v2.pdf

github : https://github.com/venomouscyanide/s3grl

## There is No Big Brother or Small Brother: Knowledge Infusion in Language Models for Link Prediction and Question Answering

## 摘要

The integration of knowledge graphs with deep learning is thriving in improving the performance of various natural language processing (NLP) tasks. In this paper, we focus on knowledge-infused link prediction and question answering using language models, T5, and BLOOM across three domains: Aviation, Movie, and Web. In this context, we infuse knowledge in large and small language models and study their performance, and find the performance to be similar. For the link prediction task on the Aviation Knowledge Graph, we obtain a 0.2 hits@1 score using T5-small, T5-base, T5-large, and BLOOM. Using template-based scripts, we create a set of 1 million synthetic factoid QA pairs in the aviation domain from National Transportation Safety Board (NTSB) reports. On our curated QA pairs, the three models of T5 achieve a 0.7 hits@1 score. We validate out findings with the paired student t-test and Cohen's kappa scores. For link prediction on Aviation Knowledge Graph using T5-small and T5-large, we obtain a Cohen's kappa score of 0.76, showing substantial agreement between the models. Thus, we infer that small language models perform similar to large language models with the infusion of knowledge.

知识图与深度学习的集成在提高各种自然语言处理（NLP）任务的性能方面蓬勃发展。在本文中，我们重点讨论了在航空、电影和网络三个领域中使用语言模型T5和BLOOM进行知识注入的链接预测和问题回答。在这种情况下，我们在大型和小型语言模型中注入知识，并研究它们的性能，发现它们的性能是相似的。对于航空知识图上的链接预测任务，我们获得了0.2hits@1使用T5小、T5基础、T5大和BLOOM得分。使用基于模板的脚本，我们根据美国国家运输安全委员会（NTSB）的报告在航空领域创建了一组100万个合成事实QA对。在我们精心策划的QA配对中，T5的三款车型获得了0.7hits@1分数我们用配对学生t检验和Cohen的kappa评分来验证研究结果。对于使用T5小和T5大的航空知识图上的链接预测，我们获得了0.76的Cohen’s kappa分数，显示了模型之间的基本一致性。因此，我们推断，随着知识的注入，小语言模型的表现与大语言模型相似。

文章主页 : https://paperswithcode.com/paper/there-is-no-big-brother-or-small-brother

PDF : https://arxiv.org/pdf/2301.04013v1.pdf

github : https://github.com/ankush9812/knowledge-infusion-in-lm-for-qa

## Generative Graph Neural Networks for Link Prediction

## 摘要

Inferring missing links or detecting spurious ones based on observed graphs, known as link prediction, is a long-standing challenge in graph data analysis. With the recent advances in deep learning, graph neural networks have been used for link prediction and have achieved state-of-the-art performance. Nevertheless, existing methods developed for this purpose are typically discriminative, computing features of local subgraphs around two neighboring nodes and predicting potential links between them from the perspective of subgraph classification. In this formalism, the selection of enclosing subgraphs and heuristic structural features for subgraph classification significantly affects the performance of the methods. To overcome this limitation, this paper proposes a novel and radically different link prediction algorithm based on the network reconstruction theory, called GraphLP. Instead of sampling positive and negative links and heuristically computing the features of their enclosing subgraphs, GraphLP utilizes the feature learning ability of deep-learning models to automatically extract the structural patterns of graphs for link prediction under the assumption that real-world graphs are not locally isolated. Moreover, GraphLP explores high-order connectivity patterns to utilize the hierarchical organizational structures of graphs for link prediction. Our experimental results on all common benchmark datasets from different applications demonstrate that the proposed method consistently outperforms other state-of-the-art methods. Unlike the discriminative neural network models used for link prediction, GraphLP is generative, which provides a new paradigm for neural-network-based link prediction.

根据观察到的图推断缺失的链接或检测虚假链接，称为链接预测，是图数据分析中的一个长期挑战。随着深度学习的最新进展，图神经网络已被用于链路预测，并取得了最先进的性能。然而，为此目的开发的现有方法通常是有区别的，计算两个相邻节点周围的局部子图的特征，并从子图分类的角度预测它们之间的潜在链接。在这种形式中，用于子图分类的封闭子图和启发式结构特征的选择显著影响了方法的性能。为了克服这一限制，本文基于网络重构理论，提出了一种新的、完全不同的链路预测算法，称为GraphLP。GraphLP没有对正链接和负链接进行采样并启发式计算其封闭子图的特征，而是利用深度学习模型的特征学习能力，在假设真实世界的图不是局部孤立的情况下，自动提取图的结构模式用于链接预测。此外，GraphLP探索了高阶连接模式，以利用图的层次组织结构进行链接预测。我们在来自不同应用程序的所有常见基准数据集上的实验结果表明，所提出的方法始终优于其他最先进的方法。与用于链路预测的判别式神经网络模型不同，GraphLP是生成的，它为基于神经网络的链路预测提供了一种新的范式。

文章主页 : https://paperswithcode.com/paper/generative-graph-neural-networks-for-link

PDF : https://arxiv.org/pdf/2301.00169v1.pdf

github : https://github.com/star4455/graphlp

## Bring Your Own View: Graph Neural Networks for Link Prediction with Personalized Subgraph Selection

## 摘要

Graph neural networks (GNNs) have received remarkable success in link prediction (GNNLP) tasks. Existing efforts first predefine the subgraph for the whole dataset and then apply GNNs to encode edge representations by leveraging the neighborhood structure induced by the fixed subgraph. The prominence of GNNLP methods significantly relies on the adhoc subgraph. Since node connectivity in real-world graphs is complex, one shared subgraph is limited for all edges. Thus, the choices of subgraphs should be personalized to different edges. However, performing personalized subgraph selection is nontrivial since the potential selection space grows exponentially to the scale of edges. Besides, the inference edges are not available during training in link prediction scenarios, so the selection process needs to be inductive. To bridge the gap, we introduce a Personalized Subgraph Selector (PS2) as a plug-and-play framework to automatically, personally, and inductively identify optimal subgraphs for different edges when performing GNNLP. PS2 is instantiated as a bi-level optimization problem that can be efficiently solved differently. Coupling GNNLP models with PS2, we suggest a brand-new angle towards GNNLP training: by first identifying the optimal subgraphs for edges; and then focusing on training the inference model by using the sampled subgraphs. Comprehensive experiments endorse the effectiveness of our proposed method across various GNNLP backbones (GCN, GraphSage, NGCF, LightGCN, and SEAL) and diverse benchmarks (Planetoid, OGB, and Recommendation datasets). Our code is publicly available at \url{https://github.com/qiaoyu-tan/PS2}

图神经网络（GNNs）在链路预测（GNNLP）任务中取得了显著的成功。现有的工作首先预先定义整个数据集的子图，然后通过利用固定子图引起的邻域结构来应用GNN来编码边缘表示。GNNLP方法的突出性在很大程度上依赖于adhoc子图。由于真实世界图中的节点连通性很复杂，一个共享子图对所有边都是有限的。因此，子图的选择应该针对不同的边进行个性化。然而，执行个性化的子图选择是不平凡的，因为潜在的选择空间以指数形式增长到边的尺度。此外，在链路预测场景的训练过程中，推理边缘不可用，因此选择过程需要是归纳的。为了弥补这一差距，我们引入了一种个性化子图选择器（PS2）作为即插即用框架，以在执行GNNLP时自动、亲自和归纳地识别不同边缘的最优子图。PS2被实例化为可以以不同方式有效地解决的双层优化问题。将GNNLP模型与PS2相结合，我们提出了一个全新的GNNLP训练角度：首先识别边缘的最优子图；然后重点利用采样子图训练推理模型。综合实验证实了我们提出的方法在各种GNNLP主干（GCN、GraphSage、NGCF、LightGCN和SEAL）和各种基准（Planetoid、OGB和Recommendation数据集）上的有效性。我们的代码可在\url上公开获取{https://github.com/qiaoyu-tan/PS2}

文章主页 : https://paperswithcode.com/paper/bring-your-own-view-graph-neural-networks-for

PDF : https://arxiv.org/pdf/2212.12488v1.pdf

github : https://github.com/qiaoyu-tan/ps2

## FakeEdge: Alleviate Dataset Shift in Link Prediction

## 摘要

Link prediction is a crucial problem in graph-structured data. Due to the recent success of graph neural networks (GNNs), a variety of GNN-based models were proposed to tackle the link prediction task. Specifically, GNNs leverage the message passing paradigm to obtain node representation, which relies on link connectivity. However, in a link prediction task, links in the training set are always present while ones in the testing set are not yet formed, resulting in a discrepancy of the connectivity pattern and bias of the learned representation. It leads to a problem of dataset shift which degrades the model performance. In this paper, we first identify the dataset shift problem in the link prediction task and provide theoretical analyses on how existing link prediction methods are vulnerable to it. We then propose FakeEdge, a model-agnostic technique, to address the problem by mitigating the graph topological gap between training and testing sets. Extensive experiments demonstrate the applicability and superiority of FakeEdge on multiple datasets across various domains.

链接预测是图结构数据中的一个关键问题。由于图神经网络（GNN）最近的成功，提出了各种基于GNN的模型来处理链路预测任务。具体来说，GNN利用消息传递范式来获得节点表示，这依赖于链路连接。然而，在链路预测任务中，训练集中的链路总是存在，而测试集中的链路尚未形成，这导致了连接模式的差异和学习表示的偏差。这导致了数据集偏移的问题，从而降低了模型的性能。在本文中，我们首先识别了链接预测任务中的数据集偏移问题，并对现有的链接预测方法如何容易受到该问题的影响进行了理论分析。然后，我们提出了一种模型不可知技术FakeEdge，通过缓解训练集和测试集之间的图拓扑间隙来解决该问题。大量实验证明了FakeEdge在不同领域的多个数据集上的适用性和优越性。

文章主页 : https://paperswithcode.com/paper/fakeedge-alleviate-dataset-shift-in-link

PDF : https://arxiv.org/pdf/2211.15899v2.pdf

github : https://github.com/Barcavin/FakeEdge

## Link Prediction with Non-Contrastive Learning

## 摘要

A recent focal area in the space of graph neural networks (GNNs) is graph self-supervised learning (SSL), which aims to derive useful node representations without labeled data. Notably, many state-of-the-art graph SSL methods are contrastive methods, which use a combination of positive and negative samples to learn node representations. Owing to challenges in negative sampling (slowness and model sensitivity), recent literature introduced non-contrastive methods, which instead only use positive samples. Though such methods have shown promising performance in node-level tasks, their suitability for link prediction tasks, which are concerned with predicting link existence between pairs of nodes (and have broad applicability to recommendation systems contexts) is yet unexplored. In this work, we extensively evaluate the performance of existing non-contrastive methods for link prediction in both transductive and inductive settings. While most existing non-contrastive methods perform poorly overall, we find that, surprisingly, BGRL generally performs well in transductive settings. However, it performs poorly in the more realistic inductive settings where the model has to generalize to links to/from unseen nodes. We find that non-contrastive models tend to overfit to the training graph and use this analysis to propose T-BGRL, a novel non-contrastive framework that incorporates cheap corruptions to improve the generalization ability of the model. This simple modification strongly improves inductive performance in 5/6 of our datasets, with up to a 120% improvement in Hits@50--all with comparable speed to other non-contrastive baselines and up to 14x faster than the best-performing contrastive baseline. Our work imparts interesting findings about non-contrastive learning for link prediction and paves the way for future researchers to further expand upon this area.

图神经网络（GNN）领域最近的一个焦点领域是图自监督学习（SSL），其目的是在没有标记数据的情况下导出有用的节点表示。值得注意的是，许多最先进的图SSL方法都是对比方法，它们使用正样本和负样本的组合来学习节点表示。由于负采样的挑战（缓慢和模型敏感性），最近的文献引入了非对比方法，只使用正样本。尽管这些方法在节点级任务中表现出了良好的性能，但它们对链路预测任务的适用性尚未得到探索，链路预测任务涉及预测节点对之间的链路存在（并对推荐系统上下文具有广泛的适用性）。在这项工作中，我们广泛评估了现有的非对比方法在转导和归纳环境中的链接预测性能。虽然大多数现有的非对比方法总体表现不佳，但我们发现，令人惊讶的是，BGRL在转导环境中通常表现良好。然而，它在更现实的归纳设置中表现不佳，在这种情况下，模型必须推广到去往/来自看不见的节点的链接。我们发现非对比模型倾向于过度拟合训练图，并利用这一分析提出了T-BGRL，这是一种新的非对比框架，它结合了廉价的腐蚀来提高模型的泛化能力。这种简单的修改大大提高了我们5/6个数据集的归纳性能，在Hits@50--all其速度与其他非对比基线相当，并且比表现最好的对比基线快14倍。我们的工作提供了关于链接预测的非对比学习的有趣发现，并为未来的研究人员进一步扩展这一领域铺平了道路。

文章主页 : https://paperswithcode.com/paper/link-prediction-with-non-contrastive-learning

PDF : https://arxiv.org/pdf/2211.14394v2.pdf

github : https://github.com/snap-research/non-contrastive-link-prediction

## RAILD: Towards Leveraging Relation Features for Inductive Link Prediction In Knowledge Graphs

## 摘要

Due to the open world assumption, Knowledge Graphs (KGs) are never complete. In order to address this issue, various Link Prediction (LP) methods are proposed so far. Some of these methods are inductive LP models which are capable of learning representations for entities not seen during training. However, to the best of our knowledge, none of the existing inductive LP models focus on learning representations for unseen relations. In this work, a novel Relation Aware Inductive Link preDiction (RAILD) is proposed for KG completion which learns representations for both unseen entities and unseen relations. In addition to leveraging textual literals associated with both entities and relations by employing language models, RAILD also introduces a novel graph-based approach to generate features for relations. Experiments are conducted with different existing and newly created challenging benchmark datasets and the results indicate that RAILD leads to performance improvement over the state-of-the-art models. Moreover, since there are no existing inductive LP models which learn representations for unseen relations, we have created our own baselines and the results obtained with RAILD also outperform these baselines.

由于开放世界的假设，知识图是不完整的。为了解决这个问题，到目前为止，已经提出了各种链路预测（LP）方法。这些方法中的一些是归纳LP模型，其能够学习训练期间未看到的实体的表示。然而，据我们所知，现有的归纳LP模型中没有一个专注于学习不可见关系的表示。在这项工作中，为KG完成提出了一种新的关系感知归纳链接预判（RAILD），它学习看不见实体和看不见关系的表示。除了通过使用语言模型来利用与实体和关系相关的文本文本外，RAILD还引入了一种新的基于图的方法来生成关系的特征。使用不同的现有和新创建的具有挑战性的基准数据集进行了实验，结果表明，RAILD比最先进的模型提高了性能。此外，由于目前还没有学习不可见关系表示的归纳LP模型，我们创建了自己的基线，使用RAILD获得的结果也优于这些基线。

文章主页 : https://paperswithcode.com/paper/raild-towards-leveraging-relation-features

PDF : https://arxiv.org/pdf/2211.11407v1.pdf

github : https://github.com/genetasefa/raild

## ReInform: Selecting paths with reinforcement learning for contextualized link prediction

## 摘要

We propose to use reinforcement learning to inform transformer-based contextualized link prediction models by providing paths that are most useful for predicting the correct answer. This is in contrast to previous approaches, that either used reinforcement learning (RL) to directly search for the answer, or based their prediction on limited or randomly selected context. Our experiments on WN18RR and FB15k-237 show that contextualized link prediction models consistently outperform RL-based answer search, and that additional improvements (of up to 13.5% MRR) can be gained by combining RL with a link prediction model. The PyTorch implementation of the RL agent is available at https://github.com/marina-sp/reinform

我们建议使用强化学习，通过提供对预测正确答案最有用的路径，为基于变换器的上下文链接预测模型提供信息。这与以前的方法形成了对比，以前的方法要么使用强化学习（RL）直接搜索答案，要么基于有限或随机选择的上下文进行预测。我们在WN18RR和FB15k-237上的实验表明，上下文链接预测模型始终优于基于RL的答案搜索，并且通过将RL与链接预测模型相结合可以获得额外的改进（高达13.5%MRR）。RL代理的PyTorch实现可在https://github.com/marina-sp/reinform

文章主页 : https://paperswithcode.com/paper/reinform-selecting-paths-with-reinforcement

PDF : https://arxiv.org/pdf/2211.10688v2.pdf

github : https://github.com/marina-sp/reinform

## New Frontiers in Graph Autoencoders: Joint Community Detection and Link Prediction

## 摘要

Graph autoencoders (GAE) and variational graph autoencoders (VGAE) emerged as powerful methods for link prediction (LP). Their performances are less impressive on community detection (CD), where they are often outperformed by simpler alternatives such as the Louvain method. It is still unclear to what extent one can improve CD with GAE and VGAE, especially in the absence of node features. It is moreover uncertain whether one could do so while simultaneously preserving good performances on LP in a multi-task setting. In this workshop paper, summarizing results from our journal publication (Salha-Galvan et al. 2022), we show that jointly addressing these two tasks with high accuracy is possible. For this purpose, we introduce a community-preserving message passing scheme, doping our GAE and VGAE encoders by considering both the initial graph and Louvain-based prior communities when computing embedding spaces. Inspired by modularity-based clustering, we further propose novel training and optimization strategies specifically designed for joint LP and CD. We demonstrate the empirical effectiveness of our approach, referred to as Modularity-Aware GAE and VGAE, on various real-world graphs.

图形自动编码器（GAE）和变分图自动编码器（VGAE）作为链路预测（LP）的强大方法出现。它们在社区检测（CD）方面的表现不那么令人印象深刻，在社区检测中，它们的表现往往优于更简单的替代方案，如Louvain方法。目前尚不清楚使用GAE和VGAE可以在多大程度上改善CD，尤其是在没有节点特征的情况下。此外，还不确定是否可以在多任务设置中同时保持LP上的良好性能。在这篇研讨会论文中，总结了我们期刊发表的结果（Salha Galvan等人，2022），我们表明，以高精度联合解决这两项任务是可能的。为此，我们引入了一种保留社区的消息传递方案，在计算嵌入空间时，通过考虑初始图和基于Louvain的先验社区来掺杂我们的GAE和VGAE编码器。受基于模块化的聚类的启发，我们进一步提出了专门为联合LP和CD设计的新的训练和优化策略。我们在各种真实世界的图上证明了我们的方法（称为模块化感知GAE和VGAE）的经验有效性。

文章主页 : https://paperswithcode.com/paper/new-frontiers-in-graph-autoencoders-joint

PDF : https://arxiv.org/pdf/2211.08972v1.pdf

github : https://github.com/guillaumesalhagalvan/modularity\_aware\_gae

## Graph Sequential Neural ODE Process for Link Prediction on Dynamic and Sparse Graphs

## 摘要

Link prediction on dynamic graphs is an important task in graph mining. Existing approaches based on dynamic graph neural networks (DGNNs) typically require a significant amount of historical data (interactions over time), which is not always available in practice. The missing links over time, which is a common phenomenon in graph data, further aggravates the issue and thus creates extremely sparse and dynamic graphs. To address this problem, we propose a novel method based on the neural process, called Graph Sequential Neural ODE Process (GSNOP). Specifically, GSNOP combines the advantage of the neural process and neural ordinary differential equation that models the link prediction on dynamic graphs as a dynamic-changing stochastic process. By defining a distribution over functions, GSNOP introduces the uncertainty into the predictions, making it generalize to more situations instead of overfitting to the sparse data. GSNOP is also agnostic to model structures that can be integrated with any DGNN to consider the chronological and geometrical information for link prediction. Extensive experiments on three dynamic graph datasets show that GSNOP can significantly improve the performance of existing DGNNs and outperform other neural process variants.

动态图上的链接预测是图挖掘中的一项重要任务。基于动态图神经网络（DGNN）的现有方法通常需要大量的历史数据（随时间的交互），而这在实践中并不总是可用的。随着时间的推移，缺失的链接是图形数据中的一种常见现象，这进一步加剧了问题，从而创建了极其稀疏和动态的图形。为了解决这个问题，我们提出了一种基于神经过程的新方法，称为图序列神经ODE过程（GSNOP）。具体而言，GSNOP结合了神经过程和神经常微分方程的优点，将动态图上的链接预测建模为动态变化的随机过程。通过定义函数上的分布，GSNOP将不确定性引入到预测中，使其推广到更多的情况，而不是过度拟合稀疏数据。GSNOP对可以与任何DGNN集成的模型结构也是不可知的，以考虑链路预测的时间和几何信息。在三个动态图数据集上的大量实验表明，GSNOP可以显著提高现有DGNN的性能，并优于其他神经过程变体。

文章主页 : https://paperswithcode.com/paper/graph-sequential-neural-ode-process-for-link

PDF : https://arxiv.org/pdf/2211.08568v1.pdf

github : https://github.com/rmanluo/gsnop

## KGLM: Integrating Knowledge Graph Structure in Language Models for Link Prediction

## 摘要

The ability of knowledge graphs to represent complex relationships at scale has led to their adoption for various needs including knowledge representation, question-answering, and recommendation systems. Knowledge graphs are often incomplete in the information they represent, necessitating the need for knowledge graph completion tasks. Pre-trained and fine-tuned language models have shown promise in these tasks although these models ignore the intrinsic information encoded in the knowledge graph, namely the entity and relation types. In this work, we propose the Knowledge Graph Language Model (KGLM) architecture, where we introduce a new entity/relation embedding layer that learns to differentiate distinctive entity and relation types, therefore allowing the model to learn the structure of the knowledge graph. In this work, we show that further pre-training the language models with this additional embedding layer using the triples extracted from the knowledge graph, followed by the standard fine-tuning phase sets a new state-of-the-art performance for the link prediction task on the benchmark datasets.

知识图在规模上表示复杂关系的能力导致了它们被用于各种需求，包括知识表示、问答和推荐系统。知识图所代表的信息往往是不完整的，这就需要知识图完成任务。经过预训练和微调的语言模型在这些任务中显示出了希望，尽管这些模型忽略了知识图中编码的内在信息，即实体和关系类型。在这项工作中，我们提出了知识图语言模型（KGLM）架构，其中我们引入了一个新的实体/关系嵌入层，该层学习区分不同的实体和关系类型，从而允许模型学习知识图的结构。在这项工作中，我们表明，使用从知识图中提取的三元组，通过这个额外的嵌入层进一步预训练语言模型，然后是标准的微调阶段，为基准数据集上的链接预测任务设置了新的最先进的性能。

文章主页 : https://paperswithcode.com/paper/kglm-integrating-knowledge-graph-structure-in

PDF : https://arxiv.org/pdf/2211.02744v2.pdf

github : https://github.com/ibpa/kglm

## Rethinking the positive role of cluster structure in complex networks for link prediction tasks

## 摘要

Clustering is a fundamental problem in network analysis that finds closely connected groups of nodes and separates them from other nodes in the graph, while link prediction is to predict whether two nodes in a network are likely to have a link. The definition of both naturally determines that clustering must play a positive role in obtaining accurate link prediction tasks. Yet researchers have long ignored or used inappropriate ways to undermine this positive relationship. In this article, We construct a simple but efficient clustering-driven link prediction framework(ClusterLP), with the goal of directly exploiting the cluster structures to obtain connections between nodes as accurately as possible in both undirected graphs and directed graphs. Specifically, we propose that it is easier to establish links between nodes with similar representation vectors and cluster tendencies in undirected graphs, while nodes in a directed graphs can more easily point to nodes similar to their representation vectors and have greater influence in their own cluster. We customized the implementation of ClusterLP for undirected and directed graphs, respectively, and the experimental results using multiple real-world networks on the link prediction task showed that our models is highly competitive with existing baseline models. The code implementation of ClusterLP and baselines we use are available at https://github.com/ZINUX1998/ClusterLP.

聚类是网络分析中的一个基本问题，它可以找到紧密连接的节点组，并将它们与图中的其他节点分离，而链路预测是预测网络中的两个节点是否可能有链路。两者的定义自然决定了聚类必须在获得准确的链接预测任务方面发挥积极作用。然而，研究人员长期以来一直忽视或使用不恰当的方式来破坏这种积极的关系。在本文中，我们构建了一个简单但高效的聚类驱动链接预测框架（ClusterLP），目的是直接利用聚类结构在无向图和有向图中尽可能准确地获得节点之间的连接。具体而言，我们提出在无向图中具有相似表示向量和聚类趋势的节点之间更容易建立链接，而有向图中的节点可以更容易地指向与其表示向量相似的节点，并在自己的聚类中具有更大的影响。我们分别为无向图和有向图定制了ClusterLP的实现，在链路预测任务中使用多个真实世界网络的实验结果表明，我们的模型与现有的基线模型具有很强的竞争力。ClusterLP的代码实现和我们使用的基线可在https://github.com/ZINUX1998/ClusterLP.

文章主页 : https://paperswithcode.com/paper/rethinking-the-positive-role-of-cluster

PDF : https://arxiv.org/pdf/2211.02396v2.pdf

github : https://github.com/zinux1998/clusterlp

## Line Graph Contrastive Learning for Link Prediction

## 摘要

Link prediction tasks focus on predicting possible future connections. Most existing researches measure the likelihood of links by different similarity scores on node pairs and predict links between nodes. However, the similarity-based approaches have some challenges in information loss on nodes and generalization ability on similarity indexes. To address the above issues, we propose a Line Graph Contrastive Learning(LGCL) method to obtain rich information with multiple perspectives. LGCL obtains a subgraph view by h-hop subgraph sampling with target node pairs. After transforming the sampled subgraph into a line graph, the link prediction task is converted into a node classification task, which graph convolution progress can learn edge embeddings from graphs more effectively. Then we design a novel cross-scale contrastive learning framework on the line graph and the subgraph to maximize the mutual information of them, so that fuses the structure and feature information. The experimental results demonstrate that the proposed LGCL outperforms the state-of-the-art methods and has better performance on generalization and robustness.

链接预测任务的重点是预测未来可能的连接。现有的大多数研究都是通过节点对上不同的相似性得分来衡量链接的可能性，并预测节点之间的链接。然而，基于相似性的方法在节点信息丢失和相似性指数泛化能力方面存在一些挑战。为了解决上述问题，我们提出了一种折线图对比学习（LGCL）方法，以从多个角度获得丰富的信息。LGCL通过对目标节点对进行h跳子图采样来获得子图视图。在将采样子图转换为线图后，将链接预测任务转换为节点分类任务，图卷积进程可以更有效地从图中学习边缘嵌入。然后，我们在折线图和子图上设计了一个新的跨尺度对比学习框架，以最大化它们之间的相互信息，从而融合结构和特征信息。实验结果表明，所提出的LGCL优于现有的方法，在泛化和鲁棒性方面具有更好的性能。

文章主页 : https://paperswithcode.com/paper/line-graph-contrastive-learning-for-link

PDF : https://arxiv.org/pdf/2210.13795v2.pdf

github : https://github.com/shilinsun/lgcl

## Parameter-free Dynamic Graph Embedding for Link Prediction

## 摘要

Dynamic interaction graphs have been widely adopted to model the evolution of user-item interactions over time. There are two crucial factors when modelling user preferences for link prediction in dynamic interaction graphs: 1) collaborative relationship among users and 2) user personalized interaction patterns. Existing methods often implicitly consider these two factors together, which may lead to noisy user modelling when the two factors diverge. In addition, they usually require time-consuming parameter learning with back-propagation, which is prohibitive for real-time user preference modelling. To this end, this paper proposes FreeGEM, a parameter-free dynamic graph embedding method for link prediction. Firstly, to take advantage of the collaborative relationships, we propose an incremental graph embedding engine to obtain user/item embeddings, which is an Online-Monitor-Offline architecture consisting of an Online module to approximately embed users/items over time, a Monitor module to estimate the approximation error in real time and an Offline module to calibrate the user/item embeddings when the online approximation errors exceed a threshold. Meanwhile, we integrate attribute information into the model, which enables FreeGEM to better model users belonging to some under represented groups. Secondly, we design a personalized dynamic interaction pattern modeller, which combines dynamic time decay with attention mechanism to model user short-term interests. Experimental results on two link prediction tasks show that FreeGEM can outperform the state-of-the-art methods in accuracy while achieving over 36X improvement in efficiency. All code and datasets can be found in https://github.com/FudanCISL/FreeGEM.

动态交互图已被广泛采用来对用户-项目交互随时间的演变进行建模。在动态交互图中为链接预测建模用户偏好时，有两个关键因素：1）用户之间的协作关系和2）用户个性化交互模式。现有的方法通常隐含地将这两个因素放在一起考虑，这可能会在两个因素出现分歧时导致有噪声的用户建模。此外，它们通常需要耗时的参数学习和反向传播，这对于实时用户偏好建模来说是禁止的。为此，本文提出了一种用于链路预测的无参数动态图嵌入方法FreeGEM。首先，为了利用协作关系，我们提出了一种增量图嵌入引擎来获得用户/项目嵌入，这是一种在线监控离线架构，由在线模块组成，用于随着时间的推移近似嵌入用户/项目，用于实时估计近似误差的监视器模块和用于在在线近似误差超过阈值时校准用户/项目嵌入的离线模块。同时，我们将属性信息集成到模型中，使FreeGEM能够更好地对属于某些代表性不足群体的用户进行建模。其次，我们设计了一个个性化的动态交互模式建模器，将动态时间衰减与注意力机制相结合，对用户的短期兴趣进行建模。在两个链路预测任务上的实验结果表明，FreeGEM在精度上优于最先进的方法，同时在效率上提高了36X以上。所有代码和数据集都可以在中找到https://github.com/FudanCISL/FreeGEM.

文章主页 : https://paperswithcode.com/paper/parameter-free-dynamic-graph-embedding-for

PDF : https://arxiv.org/pdf/2210.08189v2.pdf

github : https://github.com/fudancisl/freegem

## SMiLE: Schema-augmented Multi-level Contrastive Learning for Knowledge Graph Link Prediction

## 摘要

Link prediction is the task of inferring missing links between entities in knowledge graphs. Embedding-based methods have shown effectiveness in addressing this problem by modeling relational patterns in triples. However, the link prediction task often requires contextual information in entity neighborhoods, while most existing embedding-based methods fail to capture it. Additionally, little attention is paid to the diversity of entity representations in different contexts, which often leads to false prediction results. In this situation, we consider that the schema of knowledge graph contains the specific contextual information, and it is beneficial for preserving the consistency of entities across contexts. In this paper, we propose a novel Schema-augmented Multi-level contrastive LEarning framework (SMiLE) to conduct knowledge graph link prediction. Specifically, we first exploit network schema as the prior constraint to sample negatives and pre-train our model by employing a multi-level contrastive learning method to yield both prior schema and contextual information. Then we fine-tune our model under the supervision of individual triples to learn subtler representations for link prediction. Extensive experimental results on four knowledge graph datasets with thorough analysis of each component demonstrate the effectiveness of our proposed framework against state-of-the-art baselines. The implementation of SMiLE is available at https://github.com/GKNL/SMiLE.

链接预测是推断知识图中实体之间缺失链接的任务。基于嵌入的方法通过对三元组中的关系模式进行建模，在解决这个问题方面显示出了有效性。然而，链接预测任务通常需要实体邻域中的上下文信息，而大多数现有的基于嵌入的方法都无法捕捉到这些信息。此外，很少注意不同上下文中实体表示的多样性，这往往会导致错误的预测结果。在这种情况下，我们认为知识图的模式包含特定的上下文信息，这有利于保持实体在上下文之间的一致性。在本文中，我们提出了一种新的模式增强多级对比学习框架（SMiLE）来进行知识图链接预测。具体来说，我们首先利用网络模式作为先验约束来对否定进行采样，并通过采用多级对比学习方法来预训练我们的模型，以产生先验模式和上下文信息。然后，我们在单个三元组的监督下微调我们的模型，以学习链接预测的微妙表示。在四个知识图数据集上的大量实验结果以及对每个组件的彻底分析证明了我们提出的框架相对于最先进的基线的有效性。SMiLE的实施可在https://github.com/GKNL/SMiLE.

文章主页 : https://paperswithcode.com/paper/smile-schema-augmented-multi-level

PDF : https://arxiv.org/pdf/2210.04870v2.pdf

github : https://github.com/gknl/smile

## Graph Neural Networks for Link Prediction with Subgraph Sketching

## 摘要

Many Graph Neural Networks (GNNs) perform poorly compared to simple heuristics on Link Prediction (LP) tasks. This is due to limitations in expressive power such as the inability to count triangles (the backbone of most LP heuristics) and because they can not distinguish automorphic nodes (those having identical structural roles). Both expressiveness issues can be alleviated by learning link (rather than node) representations and incorporating structural features such as triangle counts. Since explicit link representations are often prohibitively expensive, recent works resorted to subgraph-based methods, which have achieved state-of-the-art performance for LP, but suffer from poor efficiency due to high levels of redundancy between subgraphs. We analyze the components of subgraph GNN (SGNN) methods for link prediction. Based on our analysis, we propose a novel full-graph GNN called ELPH (Efficient Link Prediction with Hashing) that passes subgraph sketches as messages to approximate the key components of SGNNs without explicit subgraph construction. ELPH is provably more expressive than Message Passing GNNs (MPNNs). It outperforms existing SGNN models on many standard LP benchmarks while being orders of magnitude faster. However, it shares the common GNN limitation that it is only efficient when the dataset fits in GPU memory. Accordingly, we develop a highly scalable model, called BUDDY, which uses feature precomputation to circumvent this limitation without sacrificing predictive performance. Our experiments show that BUDDY also outperforms SGNNs on standard LP benchmarks while being highly scalable and faster than ELPH.

许多图神经网络（GNN）在链路预测（LP）任务上与简单的启发式算法相比表现不佳。这是由于表达能力的限制，例如无法计算三角形（大多数LP启发式算法的主干），以及因为它们无法区分自同构节点（具有相同结构角色的节点）。这两个表达问题都可以通过学习链接（而不是节点）表示和结合三角形计数等结构特征来缓解。由于显式链接表示通常过于昂贵，最近的工作求助于基于子图的方法，这些方法已经为LP实现了最先进的性能，但由于子图之间的高度冗余，效率较差。我们分析了用于链路预测的子图GNN（SGNN）方法的组成部分。基于我们的分析，我们提出了一种新的全图GNN，称为ELPH（带哈希的高效链接预测），它将子图草图作为消息传递，以在没有显式子图构造的情况下近似SGNN的关键组件。可以证明，ELPH比消息传递GNN（MPNN）更具表达能力。它在许多标准LP基准上优于现有的SGNN模型，同时速度快几个数量级。然而，它有一个共同的GNN限制，即只有当数据集适合GPU内存时，它才有效。因此，我们开发了一个高度可扩展的模型，称为BUDDY，它使用特征预计算来规避这一限制，而不牺牲预测性能。我们的实验表明，BUDDY在标准LP基准测试上也优于SGNN，同时具有高度可扩展性，比ELPH更快。

文章主页 : https://paperswithcode.com/paper/graph-neural-networks-for-link-prediction

PDF : https://arxiv.org/pdf/2209.15486v3.pdf

github : https://github.com/melifluos/subgraph-sketching

## MEIM: Multi-partition Embedding Interaction Beyond Block Term Format for Efficient and Expressive Link Prediction

## 摘要

Knowledge graph embedding aims to predict the missing relations between entities in knowledge graphs. Tensor-decomposition-based models, such as ComplEx, provide a good trade-off between efficiency and expressiveness, that is crucial because of the large size of real world knowledge graphs. The recent multi-partition embedding interaction (MEI) model subsumes these models by using the block term tensor format and provides a systematic solution for the trade-off. However, MEI has several drawbacks, some of which carried from its subsumed tensor-decomposition-based models. In this paper, we address these drawbacks and introduce the Multi-partition Embedding Interaction iMproved beyond block term format (MEIM) model, with independent core tensor for ensemble effects and soft orthogonality for max-rank mapping, in addition to multi-partition embedding. MEIM improves expressiveness while still being highly efficient, helping it to outperform strong baselines and achieve state-of-the-art results on difficult link prediction benchmarks using fairly small embedding sizes. The source code is released at https://github.com/tranhungnghiep/MEIM-KGE.

知识图嵌入旨在预测知识图中实体之间缺失的关系。基于张量分解的模型，如ComplEx，在效率和表达性之间提供了良好的权衡，这一点至关重要，因为现实世界中的知识图很大。最近的多分区嵌入交互（MEI）模型通过使用块项张量格式将这些模型纳入其中，并为权衡提供了系统的解决方案。然而，MEI有几个缺点，其中一些缺点来自于其包含的基于张量分解的模型。在本文中，我们解决了这些缺点，并介绍了改进的超越块项格式（MEIM）的多分区嵌入交互模型，除了多分区嵌入之外，该模型还具有用于集成效应的独立核心张量和用于最大秩映射的软正交性。MEIM提高了表现力，同时仍然非常高效，帮助它超越强基线，并使用相当小的嵌入大小在困难的链接预测基准上实现最先进的结果。源代码发布于https://github.com/tranhungnghiep/MEIM-KGE.

文章主页 : https://paperswithcode.com/paper/meim-multi-partition-embedding-interaction

PDF : https://arxiv.org/pdf/2209.15597v2.pdf

github : https://github.com/tranhungnghiep/meim-kge

## Predicting the Future of AI with AI: High-quality link prediction in an exponentially growing knowledge network

## 摘要

A tool that could suggest new personalized research directions and ideas by taking insights from the scientific literature could significantly accelerate the progress of science. A field that might benefit from such an approach is artificial intelligence (AI) research, where the number of scientific publications has been growing exponentially over the last years, making it challenging for human researchers to keep track of the progress. Here, we use AI techniques to predict the future research directions of AI itself. We develop a new graph-based benchmark based on real-world data -- the Science4Cast benchmark, which aims to predict the future state of an evolving semantic network of AI. For that, we use more than 100,000 research papers and build up a knowledge network with more than 64,000 concept nodes. We then present ten diverse methods to tackle this task, ranging from pure statistical to pure learning methods. Surprisingly, the most powerful methods use a carefully curated set of network features, rather than an end-to-end AI approach. It indicates a great potential that can be unleashed for purely ML approaches without human knowledge. Ultimately, better predictions of new future research directions will be a crucial component of more advanced research suggestion tools.

一个可以通过从科学文献中获得见解来提出新的个性化研究方向和想法的工具可以显著加速科学的进步。人工智能研究是一个可能从这种方法中受益的领域，在过去几年中，该领域的科学出版物数量呈指数级增长，这使得人类研究人员很难跟踪进展。在这里，我们使用人工智能技术来预测人工智能本身的未来研究方向。我们基于真实世界的数据开发了一种新的基于图的基准——Science4Cast基准，旨在预测人工智能语义网络的未来状态。为此，我们使用了超过100000篇研究论文，建立了一个包含64000多个概念节点的知识网络。然后，我们提出了十种不同的方法来解决这一任务，从纯粹的统计方法到纯粹的学习方法。令人惊讶的是，最强大的方法使用了一组精心策划的网络功能，而不是端到端的人工智能方法。它表明，在没有人类知识的情况下，纯ML方法可以释放出巨大的潜力。最终，更好地预测未来新的研究方向将是更先进的研究建议工具的关键组成部分。

文章主页 : https://paperswithcode.com/paper/predicting-the-future-of-ai-with-ai-high

PDF : https://arxiv.org/pdf/2210.00881v1.pdf

github : https://github.com/mariokrenn6240/futureofaiviaai

## Disconnected Emerging Knowledge Graph Oriented Inductive Link Prediction

## 摘要

Inductive link prediction (ILP) is to predict links for unseen entities in emerging knowledge graphs (KGs), considering the evolving nature of KGs. A more challenging scenario is that emerging KGs consist of only unseen entities, called as disconnected emerging KGs (DEKGs). Existing studies for DEKGs only focus on predicting enclosing links, i.e., predicting links inside the emerging KG. The bridging links, which carry the evolutionary information from the original KG to DEKG, have not been investigated by previous work so far. To fill in the gap, we propose a novel model entitled DEKG-ILP (Disconnected Emerging Knowledge Graph Oriented Inductive Link Prediction) that consists of the following two components. (1) The module CLRM (Contrastive Learning-based Relation-specific Feature Modeling) is developed to extract global relation-based semantic features that are shared between original KGs and DEKGs with a novel sampling strategy. (2) The module GSM (GNN-based Subgraph Modeling) is proposed to extract the local subgraph topological information around each link in KGs. The extensive experiments conducted on several benchmark datasets demonstrate that DEKG-ILP has obvious performance improvements compared with state-of-the-art methods for both enclosing and bridging link prediction. The source code is available online.

归纳链接预测（ILP）是考虑新兴知识图（KGs）的进化性质，预测新兴知识图中看不见的实体的链接。一个更具挑战性的场景是，新兴KGs只由看不见的实体组成，称为断开连接的新兴KGs（DEKGs）。现有的DEKG研究仅侧重于预测封闭链接，即预测新兴KG内部的链接。桥接链接承载着从原始KG到DEKG的进化信息，迄今为止，前人尚未对其进行研究。为了填补这一空白，我们提出了一个名为DEKG-ILP（Disconnected Emerging Knowledge Graph Oriented Inductive Link Prediction）的新模型，该模型由以下两个部分组成。（1） 开发了CLRM（基于对比学习的特定关系特征建模）模块，以通过一种新颖的采样策略提取原始KGs和DEKGs之间共享的基于全局关系的语义特征。（2） 提出了基于GNN的子图建模模块GSM来提取KGs中每个链路周围的局部子图拓扑信息。在几个基准数据集上进行的大量实验表明，与最先进的封闭和桥接链路预测方法相比，DEKG-ILP具有明显的性能改进。源代码可在线获取。

文章主页 : https://paperswithcode.com/paper/disconnected-emerging-knowledge-graph

PDF : https://arxiv.org/pdf/2209.01397v1.pdf

github : https://github.com/ninecl/dekg-ilp

## Link-Backdoor: Backdoor Attack on Link Prediction via Node Injection

## 摘要

Link prediction, inferring the undiscovered or potential links of the graph, is widely applied in the real-world. By facilitating labeled links of the graph as the training data, numerous deep learning based link prediction methods have been studied, which have dominant prediction accuracy compared with non-deep methods. However,the threats of maliciously crafted training graph will leave a specific backdoor in the deep model, thus when some specific examples are fed into the model, it will make wrong prediction, defined as backdoor attack. It is an important aspect that has been overlooked in the current literature. In this paper, we prompt the concept of backdoor attack on link prediction, and propose Link-Backdoor to reveal the training vulnerability of the existing link prediction methods. Specifically, the Link-Backdoor combines the fake nodes with the nodes of the target link to form a trigger. Moreover, it optimizes the trigger by the gradient information from the target model. Consequently, the link prediction model trained on the backdoored dataset will predict the link with trigger to the target state. Extensive experiments on five benchmark datasets and five well-performing link prediction models demonstrate that the Link-Backdoor achieves the state-of-the-art attack success rate under both white-box (i.e., available of the target model parameter)and black-box (i.e., unavailable of the target model parameter) scenarios. Additionally, we testify the attack under defensive circumstance, and the results indicate that the Link-Backdoor still can construct successful attack on the well-performing link prediction methods. The code and data are available at https://github.com/Seaocn/Link-Backdoor.

链接预测，即推断图中未发现或潜在的链接，在现实世界中得到了广泛应用。通过将图的标记链接作为训练数据，已经研究了许多基于深度学习的链接预测方法，与非深度方法相比，这些方法具有显著的预测精度。然而，恶意制作的训练图的威胁会在深度模型中留下特定的后门，因此当一些特定的例子被输入到模型中时，它会做出错误的预测，被定义为后门攻击。这是当前文学中被忽视的一个重要方面。在本文中，我们对链路预测提出了后门攻击的概念，并提出了链路后门来揭示现有链路预测方法的训练漏洞。具体来说，链接后门将假节点与目标链接的节点组合起来，形成触发器。此外，它通过来自目标模型的梯度信息来优化触发。因此，在后门数据集上训练的链接预测模型将预测具有到目标状态的触发的链接。在五个基准数据集和五个性能良好的链路预测模型上进行的大量实验表明，链路后门在白盒（即目标模型参数可用）和黑盒（即对象模型参数不可用）场景下都达到了最先进的攻击成功率。此外，我们还验证了防御环境下的攻击，结果表明，在性能良好的链路预测方法上，链路后门仍然可以构建成功的攻击。代码和数据可在https://github.com/Seaocn/Link-Backdoor.

文章主页 : https://paperswithcode.com/paper/link-backdoor-backdoor-attack-on-link

PDF : https://arxiv.org/pdf/2208.06776v1.pdf

github : https://github.com/seaocn/link-backdoor

## Path-aware Siamese Graph Neural Network for Link Prediction

## 摘要

In this paper, we propose a Path-aware Siamese Graph neural network(PSG) for link prediction tasks. First, PSG captures both nodes and edge features for given two nodes, namely the structure information of k-neighborhoods and relay paths information of the nodes. Furthermore, a novel multi-task GNN framework with self-supervised contrastive learning is proposed for differentiation of positive links and negative links while content and behavior of nodes can be captured simultaneously. We evaluate the proposed algorithm PSG on two link property prediction datasets, ogbl-ddi and ogbl-collab. PSG achieves top 1 performance on ogbl-ddi until submission and top 3 performance on ogbl-collab. The experimental results verify the superiority of our proposed PSG

在本文中，我们提出了一种用于链路预测任务的路径感知暹罗图神经网络（PSG）。首先，PSG捕获给定两个节点的节点和边缘特征，即k邻域的结构信息和节点的中继路径信息。此外，提出了一种新的具有自监督对比学习的多任务GNN框架，用于区分正链路和负链路，同时可以捕获节点的内容和行为。我们在两个链路特性预测数据集ogbl-ddi和ogbl-colab上评估了所提出的算法PSG。PSG在提交之前在ogbl-ddi上获得了前1名的性能，在ogbl-colab上获得了后3名的性能。实验结果验证了我们提出的PSG的优越性

文章主页 : https://paperswithcode.com/paper/path-aware-siamese-graph-neural-network-for

PDF : https://arxiv.org/pdf/2208.05781v2.pdf

github : https://github.com/jingsonglv/PSG

## Subgraph Neighboring Relations Infomax for Inductive Link Prediction on Knowledge Graphs

## 摘要

Inductive link prediction for knowledge graph aims at predicting missing links between unseen entities, those not shown in training stage. Most previous works learn entity-specific embeddings of entities, which cannot handle unseen entities. Recent several methods utilize enclosing subgraph to obtain inductive ability. However, all these works only consider the enclosing part of subgraph without complete neighboring relations, which leads to the issue that partial neighboring relations are neglected, and sparse subgraphs are hard to be handled. To address that, we propose Subgraph Neighboring Relations Infomax, SNRI, which sufficiently exploits complete neighboring relations from two aspects: neighboring relational feature for node feature and neighboring relational path for sparse subgraph. To further model neighboring relations in a global way, we innovatively apply mutual information (MI) maximization for knowledge graph. Experiments show that SNRI outperforms existing state-of-art methods by a large margin on inductive link prediction task, and verify the effectiveness of exploring complete neighboring relations in a global way to characterize node features and reason on sparse subgraphs.

知识图的归纳链接预测旨在预测看不见的实体之间的缺失链接，即那些在训练阶段没有显示的实体。大多数先前的作品学习实体的实体特定嵌入，这些嵌入不能处理看不见的实体。最近的几种方法利用封闭子图来获得归纳能力。然而，所有这些工作都只考虑子图的封闭部分，而没有完全的相邻关系，这导致了部分相邻关系被忽略，稀疏子图很难处理的问题。为此，我们提出了子图邻居关系Infomax，SNRI，它从两个方面充分利用了完全的邻居关系：节点特征的邻居关系特征和稀疏子图的邻居关系路径。为了进一步以全局方式对相邻关系进行建模，我们创新性地将互信息（MI）最大化应用于知识图。实验表明，SNRI在归纳链路预测任务上大大优于现有技术，并验证了在稀疏子图上以全局方式探索完全相邻关系来表征节点特征和原因的有效性。

文章主页 : https://paperswithcode.com/paper/subgraph-neighboring-relations-infomax-for

PDF : https://arxiv.org/pdf/2208.00850v3.pdf

github : https://github.com/Tebmer/SNRI

## Towards Better Evaluation for Dynamic Link Prediction

## 摘要

Despite the prevalence of recent success in learning from static graphs, learning from time-evolving graphs remains an open challenge. In this work, we design new, more stringent evaluation procedures for link prediction specific to dynamic graphs, which reflect real-world considerations, to better compare the strengths and weaknesses of methods. First, we create two visualization techniques to understand the reoccurring patterns of edges over time and show that many edges reoccur at later time steps. Based on this observation, we propose a pure memorization baseline called EdgeBank. EdgeBank achieves surprisingly strong performance across multiple settings because easy negative edges are often used in the current evaluation setting. To evaluate against more difficult negative edges, we introduce two more challenging negative sampling strategies that improve robustness and better match real-world applications. Lastly, we introduce six new dynamic graph datasets from a diverse set of domains missing from current benchmarks, providing new challenges and opportunities for future research. Our code repository is accessible at https://github.com/fpour/DGB.git.

尽管最近在从静态图学习方面取得了普遍的成功，但从时间演化图学习仍然是一个悬而未决的挑战。在这项工作中，我们为动态图的链接预测设计了新的、更严格的评估程序，这些程序反映了真实世界的考虑，以更好地比较方法的优势和劣势。首先，我们创建了两种可视化技术来了解边随时间的重复出现模式，并显示许多边在以后的时间步骤中重复出现。基于这一观察，我们提出了一个称为EdgeBank的纯记忆基线。EdgeBank在多个设置中实现了令人惊讶的强大性能，因为在当前评估设置中经常使用简单的负边。为了针对更困难的负边缘进行评估，我们引入了两种更具挑战性的负采样策略，这两种策略可以提高鲁棒性并更好地匹配现实世界的应用。最后，我们介绍了六个新的动态图数据集，它们来自当前基准中缺失的一组不同领域，为未来的研究提供了新的挑战和机遇。我们的代码库可访问https://github.com/fpour/DGB.git.

文章主页 : https://paperswithcode.com/paper/towards-better-evaluation-for-dynamic-link

PDF : https://arxiv.org/pdf/2207.10128v2.pdf

github : https://github.com/fpour/dgb

## DHGE: Dual-View Hyper-Relational Knowledge Graph Embedding for Link Prediction and Entity Typing

## 摘要

In the field of representation learning on knowledge graphs (KGs), a hyper-relational fact consists of a main triple and several auxiliary attribute-value descriptions, which is considered more comprehensive and specific than a triple-based fact. However, currently available hyper-relational KG embedding methods in a single view are limited in application because they weaken the hierarchical structure that represents the affiliation between entities. To overcome this limitation, we propose a dual-view hyper-relational KG structure (DH-KG) that contains a hyper-relational instance view for entities and a hyper-relational ontology view for concepts that are abstracted hierarchically from the entities. This paper defines link prediction and entity typing tasks on DH-KG for the first time and constructs two DH-KG datasets, JW44K-6K, extracted from Wikidata, and HTDM based on medical data. Furthermore, we propose DHGE, a DH-KG embedding model based on GRAN encoders, HGNNs, and joint learning. DHGE outperforms baseline models on DH-KG, according to experimental results. Finally, we provide an example of how this technology can be used to treat hypertension. Our model and new datasets are publicly available.

在知识图的表示学习领域，超关系事实由一个主三元组和几个辅助属性值描述组成，这被认为比基于三元组的事实更全面、更具体。然而，目前在单个视图中可用的超关系KG嵌入方法在应用上受到限制，因为它们削弱了表示实体之间隶属关系的层次结构。为了克服这一限制，我们提出了一种双视图超关系KG结构（DH-KG），该结构包含实体的超关系实例视图和从实体中分层抽象的概念的超关系本体视图。本文首次在DH-KG上定义了链接预测和实体分类任务，并基于医学数据构建了两个DH-KG数据集，即从Wikidata中提取的JW44K-6K和HTDM。此外，我们提出了DHGE，一种基于GRAN编码器、HGNN和联合学习的DH-KG嵌入模型。根据实验结果，DHGE在DH-KG上优于基线模型。最后，我们提供了一个如何将这项技术用于治疗高血压的例子。我们的模型和新数据集是公开的。

文章主页 : https://paperswithcode.com/paper/dhge-dual-view-hyper-relational-knowledge

PDF : https://arxiv.org/pdf/2207.08562v4.pdf

github : https://github.com/lhrlab/dhge

## The maximum capability of a topological feature in link prediction

## 摘要

Link prediction aims to predict links of a network that are not directly visible, with profound applications in biological and social systems. Despite intensive utilization of the topological feature in this task, it is unclear to what extent a particular feature can be leveraged to infer missing links. Here, we show that the maximum capability of a topological feature follows a simple mathematical expression, which is independent of how an index gauges the feature. Hence, a family of indexes associated with one topological feature shares the same performance limit. A feature's capability is lifted in the supervised prediction, which in general gives rise to better results compared with unsupervised prediction. The universality of the pattern uncovered is empirically verified by 550 structurally diverse networks, which can be applied to feature selection and the analysis of network characteristics associated with a topological feature in link prediction.

链接预测旨在预测网络中不直接可见的链接，在生物和社会系统中有着深刻的应用。尽管在这项任务中大量利用了拓扑特征，但尚不清楚在多大程度上可以利用特定特征来推断缺失的链接。在这里，我们展示了拓扑特征的最大能力遵循一个简单的数学表达式，该表达式与索引如何衡量特征无关。因此，与一个拓扑特征相关联的一系列索引共享相同的性能极限。在监督预测中提高了特征的能力，与无监督预测相比，这通常会产生更好的结果。550个结构多样的网络实证验证了所揭示的模式的普遍性，这些网络可以应用于特征选择和分析链路预测中与拓扑特征相关的网络特征。

文章主页 : https://paperswithcode.com/paper/the-maximum-capability-of-a-topological

PDF : https://arxiv.org/pdf/2206.15101v2.pdf

github : https://github.com/YijunRan/Maximum-Capability-Link-Prediction

## Sampling Enclosing Subgraphs for Link Prediction

## 摘要

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.

链接预测是图结构数据（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稳健的解决方案，特别是通过学习包围目标链接的子图（即节点对）的表示。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，尤其是对于大型图。本文提出了一种可扩展的链路预测解决方案，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从目标节点对进行多次随机遍历，然后对所有访问节点诱导的采样封闭子图进行运算。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。ScaLed进一步提供了控制计算开销和准确性之间权衡的灵活性。通过全面的实验，我们已经表明ScaLed可以产生与现有子图表示学习框架所报告的精度相当的精度，同时计算要求更低。

文章主页 : https://paperswithcode.com/paper/sampling-enclosing-subgraphs-for-link

PDF : https://arxiv.org/pdf/2206.12004v1.pdf

github : https://github.com/venomouscyanide/scaled

## Two-Dimensional Weisfeiler-Lehman Graph Neural Networks for Link Prediction

## 摘要

Link prediction is one important application of graph neural networks (GNNs). Most existing GNNs for link prediction are based on one-dimensional Weisfeiler-Lehman (1-WL) test. 1-WL-GNNs first compute node representations by iteratively passing neighboring node features to the center, and then obtain link representations by aggregating the pairwise node representations. As pointed out by previous works, this two-step procedure results in low discriminating power, as 1-WL-GNNs by nature learn node-level representations instead of link-level. In this paper, we study a completely different approach which can directly obtain node pair (link) representations based on \textit{two-dimensional Weisfeiler-Lehman (2-WL) tests}. 2-WL tests directly use links (2-tuples) as message passing units instead of nodes, and thus can directly obtain link representations. We theoretically analyze the expressive power of 2-WL tests to discriminate non-isomorphic links, and prove their superior link discriminating power than 1-WL. Based on different 2-WL variants, we propose a series of novel 2-WL-GNN models for link prediction. Experiments on a wide range of real-world datasets demonstrate their competitive performance to state-of-the-art baselines and superiority over plain 1-WL-GNNs.

链路预测是图神经网络的一个重要应用。大多数现有的用于链路预测的GNN是基于一维Weisfeiler-Lehman（1-WL）检验的。1-WL-GNN首先通过迭代地将相邻节点特征传递到中心来计算节点表示，然后通过聚合成对节点表示来获得链路表示。正如先前的工作所指出的，这种两步过程导致低辨别能力，因为1-WL-GNN本质上学习节点级表示而不是链路级表示。在本文中，我们研究了一种完全不同的方法，该方法可以基于\textit{二维Weisfeiler-Lehman（2-WL）检验}直接获得节点对（链路）表示。2-WL测试直接使用链路（2-元组）作为消息传递单元，而不是节点，因此可以直接获得链路表示。我们从理论上分析了2-WL测试判别非同构链路的表现力，并证明了其优于1-WL的链路判别力。基于不同的2-WL变体，我们提出了一系列新的2-WL-GNN链路预测模型。在广泛的真实世界数据集上进行的实验表明，它们与最先进的基线相比具有竞争力，并且优于普通的1-WL-GNN。

文章主页 : https://paperswithcode.com/paper/two-dimensional-weisfeiler-lehman-graph

PDF : https://arxiv.org/pdf/2206.09567v1.pdf

github : https://github.com/graphpku/2wl\_link\_pred

## Neo-GNNs: Neighborhood Overlap-aware Graph Neural Networks for Link Prediction

## 摘要

Graph Neural Networks (GNNs) have been widely applied to various fields for learning over graph-structured data. They have shown significant improvements over traditional heuristic methods in various tasks such as node classification and graph classification. However, since GNNs heavily rely on smoothed node features rather than graph structure, they often show poor performance than simple heuristic methods in link prediction where the structural information, e.g., overlapped neighborhoods, degrees, and shortest paths, is crucial. To address this limitation, we propose Neighborhood Overlap-aware Graph Neural Networks (Neo-GNNs) that learn useful structural features from an adjacency matrix and estimate overlapped neighborhoods for link prediction. Our Neo-GNNs generalize neighborhood overlap-based heuristic methods and handle overlapped multi-hop neighborhoods. Our extensive experiments on Open Graph Benchmark datasets (OGB) demonstrate that Neo-GNNs consistently achieve state-of-the-art performance in link prediction. Our code is publicly available at https://github.com/seongjunyun/Neo\_GNNs.

图神经网络（GNN）已被广泛应用于对图结构数据进行学习的各个领域。在节点分类和图分类等各种任务中，它们比传统的启发式方法有了显著的改进。然而，由于GNN在很大程度上依赖于平滑的节点特征而不是图结构，因此在链路预测中，它们通常比简单的启发式方法表现出较差的性能，其中结构信息，例如重叠邻域、度和最短路径，是至关重要的。为了解决这一限制，我们提出了邻域重叠感知图神经网络（Neo-GNN），该网络从邻接矩阵中学习有用的结构特征，并估计用于链路预测的重叠邻域。我们的新GNN推广了基于邻域重叠的启发式方法，并处理了重叠的多跳邻域。我们在开放图基准数据集（OGB）上进行的大量实验表明，Neo-GNN在链路预测方面始终实现了最先进的性能。我们的代码可在https://github.com/seongjunyun/Neo\_GNNs.

文章主页 : https://paperswithcode.com/paper/neo-gnns-neighborhood-overlap-aware-graph-1

PDF : https://arxiv.org/pdf/2206.04216v1.pdf

github : https://github.com/seongjunyun/neo\_gnns

## OOD Link Prediction Generalization Capabilities of Message-Passing GNNs in Larger Test Graphs

## 摘要

This work provides the first theoretical study on the ability of graph Message Passing Neural Networks (gMPNNs) -- such as Graph Neural Networks (GNNs) -- to perform inductive out-of-distribution (OOD) link prediction tasks, where deployment (test) graph sizes are larger than training graphs. We first prove non-asymptotic bounds showing that link predictors based on permutation-equivariant (structural) node embeddings obtained by gMPNNs can converge to a random guess as test graphs get larger. We then propose a theoretically-sound gMPNN that outputs structural pairwise (2-node) embeddings and prove non-asymptotic bounds showing that, as test graphs grow, these embeddings converge to embeddings of a continuous function that retains its ability to predict links OOD. Empirical results on random graphs show agreement with our theoretical results.

这项工作首次对图消息传递神经网络（gMPNN）（如图神经网络（GNN））执行归纳分布外（OOD）链路预测任务的能力进行了理论研究，其中部署（测试）图的大小大于训练图。我们首先证明了非渐近界，表明当测试图变得更大时，基于gMPNNs获得的置换等变（结构）节点嵌入的链接预测器可以收敛到随机猜测。然后，我们提出了一个理论上合理的gMPNN，它输出结构成对（2节点）嵌入，并证明了非渐近界，表明随着测试图的增长，这些嵌入收敛于连续函数的嵌入，该函数保持了预测链接OOD的能力。随机图的经验结果与我们的理论结果一致。

文章主页 : https://paperswithcode.com/paper/ood-link-prediction-generalization

PDF : https://arxiv.org/pdf/2205.15117v5.pdf

github : https://github.com/yangzez/ood-link-prediction-generalization-mpnn