DOI:10.7544/issn1000-1239.20220074
59(8):1742-1756,2022

面向知识超图链接预测的生成对抗负采样方法

郭正山　左　劼　段　磊　李仁昊　何承鑫　肖英劼　王培妍
(四川大学计算机学院　成都　610065)
(guozhengshan@stu.scu.edu.cn)

A Generative Adversarial Negative Sampling Method for Knowledge Hypergraph Link Prediction

Guo Zhengshan, Zuo Jie, Duan Lei, Li Renhao, He Chengxin, Xiao Yingjie, and Wang Peiyan
(*School of Computer Science, Sichuan University, Chengdu* 610065)

Abstract　As an extension of the knowledge graph, the knowledge hypergraph has a strong ability to express n-ary relational facts. Using the knowledge hypergraph to model known facts in the real world and discover unknown facts through link prediction has become a current research hotspot. Among existing knowledge hypergraph (or knowledge graph) link prediction methods, constructing the loss function using true labels of samples and their predicted labels is a key step, where negative samples have a great influence on the training of the link prediction model. However, when applying the negative sampling methods for knowledge graph link prediction (e.g., the uniformly random sampling) to the knowledge hypergraph, we may face problems such as low quality of negative samples and high complexity of models. As a result, we design a generative adversarial negative sampling method, named HyperGAN, for knowledge hypergraph link prediction, which generates high-quality negative samples through adversarial training to solve the zero loss problem, thereby improving the accuracy of the link prediction model. Besides, HyperGAN does not require pre-training, which makes it more efficient than previous negative sampling methods in assisting the training of link prediction models. Comparative experiments on multiple real-world datasets show that HyperGAN outperforms the baselines in terms of performance and efficiency. In addition, the case study and quantitative analysis further validate our method in improving the quality of negative samples.

Key words　knowledge hypergraph; link prediction; generative adversarial network; negative sampling; n-ary relation

摘　要　知识超图作为知识图谱的拓展,对多元关系事实具有良好表达能力.利用知识超图对现实世界中已知事实进行建模,并通过链接预测发现未知事实成为当前研究热点.在现有知识超图(知识图谱)链接预测方法中,构建样本真实标签与预测标签间的损失函数是关键步骤,其中负样本对链接预测模型的训练具有极大的影响.将知识图谱链接预测的负采样方法(如均匀随机负采样)用于知识超图链接预测

收稿日期:2022-01-09;**修回日期**:2022-03-22
基金项目:国家重点研发计划项目(2020YFB0704502);国家自然科学基金项目(61972268);四川省科技计划项目(2020YFG0034)
This work was supported by the National Key Research and Development Program of China (2020YFB0704502), the National Natural Science Foundation of China (61972268), and the Sichuan Science and Technology Program (2020YFG0034).
通信作者:左劼(zuojie@scu.edu.cn)

会面临负样本质量低下、复杂度过高等问题.对此,设计了面向知识超图链接预测的生成对抗负采样方法 HyperGAN,通过对抗训练生成高质量负样本以解决“零损失”问题,从而提升链接预测模型的准确度.HyperGAN 方法无需预训练,因此在辅助链接预测模型进行训练时相比现有负采样方法具有更高的效率.在多个真实数据集上的对比实验表明:HyperGAN 在性能与效率方面均优于基线方法.此外,具体案例分析及定量分析亦验证了 HyperGAN 方法在提升负样本质量方面的有效性.

关键词 知识超图;链接预测;生成对抗网络;负采样;多元关系

中图法分类号 TP181

知识图谱是一种以图结构存储真实世界中事实的知识库,其基本构成单位是一个包含头尾实体与描述两者间关系的三元组.然而,现实世界中的多元关系事实通常无法以三元组形式进行表达,如 Freebase[1] 知识库中多元关系的占比超过 61%[2].另一方面,将多元关系事实拆分为多个三元组会导致部分事实的信息丢失.例如,图 1 左部分中的事实“刘翔 2004 年在雅典奥运会的 110 米栏项目上获得金牌”,“人物-时间-赛事-项目-奖项”是一个涉及“刘翔”“2004 年”“雅典奥运会”“110 米栏”和“金牌”的五元关系,无法在保持事实完整性的情况下进行拆分.针对此问题,知识超图对知识图谱的事实表达形式进行了推广,以包含多个实体及多元关系的多元组为基本单位,增强了对多元关系事实的表达能力,在信息检索[3]、推荐系统[4-5]、自然语言处理[6]等场景中均发挥重要作用.

由于现实世界中新知识不断产生及知识获取困难等问题,现有知识超图不可避免地面临不完整性问题.例如,在 Freebase 知识库中,99%的“人物”缺少“种族”信息[7];在 DBpedia[8] 知识库中,58%的“学者”缺少“研究领域”的信息[9].因此,面向知识超图补全的链接预测任务已成为该领域的研究热点.其通过实体间现有的链接关系预测未知链接,以能自动学习和推理知识超图为目标,是完成诸多下游应用的重要前置任务.

针对知识超图链接预测任务,现有工作主要将实体和关系编码到低维向量空间中以捕获两者之间的联系,通过评分函数分别对正、负样本进行合理性打分,再结合样本真实标签计算预测损失以优化模型参数.在上述过程中,事实与非事实分别作为正、负样本为链接预测模型提供训练梯度.然而,现有知识超图出于对空间有效性的考虑只存储事实[10],这使得如何构建代表非事实的负样本成为关键.目前,均匀随机采样方法[11]是链接预测任务中广泛使用的一种负采样方法,其通过将正样本中的任一实体以相同概率随机替换为知识超图中其他实体以生成负样本.然而,该方法大概率会采样到不属于同一实体类型或语义不相关的实体作为替换,生成的低质量负样本可能会使模型在训练阶段面临零损失(zero loss)问题[12].

例 1. 对图 1 左图中的知识超图多元组 P_3 使用均匀随机负采样,生成图 1 右图中的 6 个候选负样本 $N_1,N_2,\cdots,N_6$.其中 N_1,N_2 和 N_3 因采样到与原实体语义不相关的替换实体,在映射至嵌入空间中时与原事实 P_3 距离较远,是无法在链接预测训练阶段有效贡献梯度的低质量负样本.反之,N_4,N_5 和 N_6 是有助于链接预测模型训练的高质量负样本.

为解决上述问题,研究者提出了针对知识图谱链接预测的负采样方法,包括基于缓存和基于生成对抗网络的方法.对于前者,NSCaching[13] 通过 AutoML[14] 技术动态更新缓存中的高质量负样本,然而该模型需针对元组中的不同实体位置分别设定缓存,具有较高的时空开销,因此不适合扩展至知识超图.对于后者,一系列工作[10,12,15]将生成对抗网络用于知识库链接预测任务.具体地说,网络中的生成器倾向于生成使判别器难以判断真伪的样本,而判别器在模型迭代过程中不断提高判别能力.在引入到知识库链接预测任务中后,该结构适用于生成可以解决零损失问题的高质量负样本.然而,这些方法均是针对知识图谱场景设计的,迁移至知识超图场景时,会面临 2 种挑战:

挑战 1.现有负采样方法均针对具有固定实体数目的三元组进行采样,而在知识超图场景下,多元组中实体数目会变化,如:四元组、五元组等.因此无法直接替换其中的链接预测方法以适用于知识超图.

挑战 2.现有基于生成对抗网络的负采样方法均需加入预训练步骤以提高生成对抗网络的稳定性.然而在知识超图场景下,多元组中实体数量的增加会进一步提高预训练步骤的复杂度.

为应对 2 种挑战,本文提出一个面向知识超图

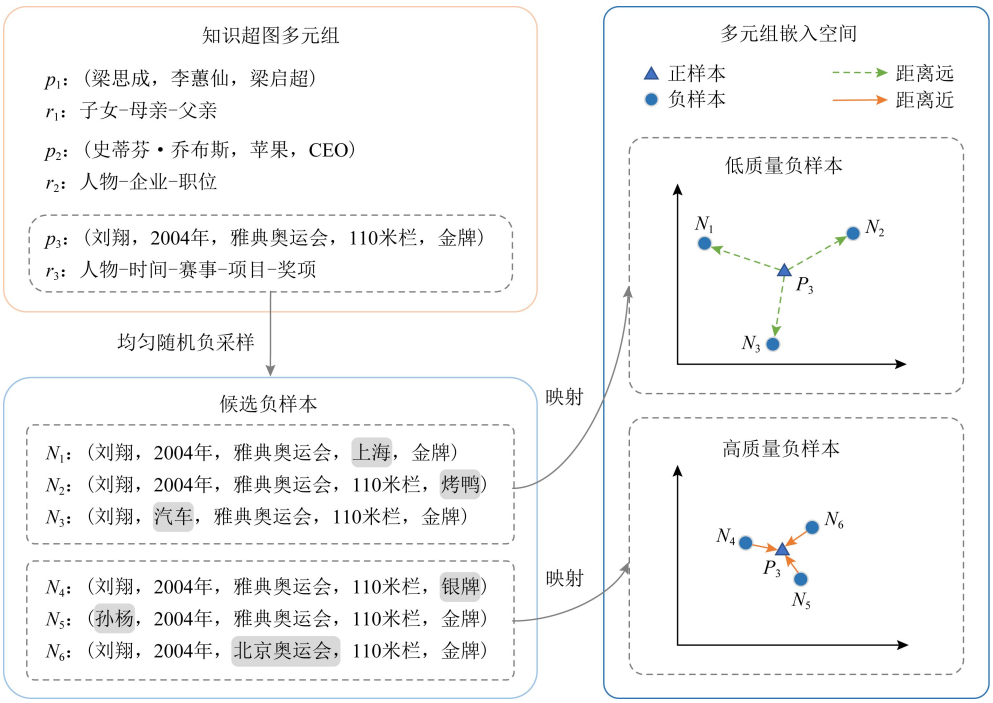


Fig. 1 An example of negative sampling in knowledge hypergraph

图 1 知识超图负采样的简单示例

链接预测的生成对抗负采样方法 HyperGAN，通过引入生成对抗网络提高负样本质量，以对链接预测模型提供指导.具体地，针对挑战 1，对以往负采样方法中共享的链接预测与判别器模块进行了分离.解耦后的链接预测模块可适用于知识超图的链接预测方法，而不影响判别器模块进行样本质量判别.针对挑战 2，简化了生成器与判别器的结构，因此无需像以往负采样方法一样进行预训练，在适用于知识超图链接预测任务的同时提高了对抗训练的稳定性.在多个知识超图数据集上的实验表明，HyperGAN 能生成高质量负样本并提升链接预测模型的准确度.

本文工作的主要贡献有 3 个方面：

1）提出了一个面向知识超图链接预测的生成对抗负采样方法 HyperGAN.其通过提高负样本质量改善链接预测任务的效果.

2）HyperGAN 通过解耦预测模型与判别器提升了方法通用性，简化后的生成对抗网络结构无需预训练也可进行稳定的对抗训练，降低了时间开销.

3）在 5 个公开的真实知识超图数据集上进行了多组对比实验，相比其他先进方法，HyperGAN 方法在性能与效率方面均取得提升.针对负样本的具体案例分析与定量分析进一步验证了 HyperGAN 方法的有效性.

1 相关工作

本节主要介绍知识超图链接预测方法和针对知识库链接预测领域的负采样方法的相关工作.

1.1 知识超图链接预测方法

当前针对知识超图的链接预测研究工作可分为 3 类：基于平移的方法、基于神经网络的方法和基于张量分解的方法[16].

1）基于平移的方法.TransE[11] 是首次提出基于平移的知识图谱链接预测方法，其通过将知识图谱三元组中的关系建模为从头实体到尾实体的平移变换，从而学习实体和关系之间的联系.然而，该方法在关系建模上的限制使其不具有完全表达性[17].此外，该方法只适用于知识图谱，不能应用到知识超图.为解决此问题，Wen 等人[18] 提出的 m-TransH 方法通过将实体映射到知识超图多元关系超平面，用基于平移的方法解决了知识超图链接预测问题.RAE 方法[19] 在此基础上考虑到 2 个实体同时出现在一个多元关系中的重要性，其将该情况下的相关性概率引入到损失函数中并使用一个全连接神经网络对模型进行训练.然而，上述扩展自 TransE 的 m-TransH 及 RAE 方法均不可避免地继承了其不具有完全表达性的缺点.最近，Abboud 等人[7] 提出的

BoxE 方法解决了这一问题,该方法将多元关系视为空间中的一组框(boxes),实体视为空间中的点,通过计算实体点到对应关系框的距离,来对知识超图中的事实进行评分.

2) 基于神经网络的方法.由于神经网络在各类自然语言处理任务上的成功应用,部分研究工作将其引入到知识超图链接预测中,并取得了较好的表现.NaLP[20]方法将知识超图中多元关系事实表示为"角色-实体"对的形式,使用卷积神经网络(convolutional neural network, CNN)和全连接网络(fully connected network, FCN)来衡量所有角色与实体对之间的相关性,从而对事实成立的合理性进行评分.然而,NaLP 并未考虑不同"角色-实体"对间的差异.为改进此不足,HINGE[21]与 NeuInfer[22]方法将多元关系事实表示为主三元组与辅助"角色-实体"对的形式,并分别通过 CNN 与 FCN 来计算事实的合理性得分.Galkin 等人[23]提出的 StarE 方法将信息传递网络应用到知识超图链接预测问题上,通过迭代更新实体和关系的嵌入表示并将其送入 Transformer[24]中,进而对事实合理性进行评估.一般来说,上述基于神经网络的链接预测方法需要经过额外数据处理步骤确定主三元组或"角色-实体"对中的实体和关系,同时也会面临参数过多导致的训练困难和不易推广到大规模知识超图等问题.

3) 基于张量分解的方法.最初,许多基于张量分解的知识超图链接预测方法是对知识图谱链接预测方法的推广,例如 n-CP[2],n-TuckER[25]和 m-DistMult[26]方法分别是对 CP[27],TuckER[28]和 DistMult[2]方法的拓展,但是这些方法不但没有改善原始方法的不足,甚至在知识超图链接预测任务中对缺点进行了放大,于是研究者开始针对知识超图场景设计特定的链接预测方法.Fatemi 等人[2]提出的 HypE 模型考虑了实体在多元组中不同位置的信息差异,其基于位置信息学习实体的嵌入表示并进一步计算事实成立的合理性得分.Liu 等人[25]提出的 GETD 模型通过使用 TuckER 方法[28]将知识超图中事实的高阶张量表示分解为一个核张量加若干个因子矩阵,并使用分解得到的核张量与关系和实体的嵌入表示相乘以获得多元关系事实成立的合理性得分.此外,针对核张量具有较多参数的问题,该模型利用张量环式(tensor ring, TR)分解方法[29]将其分解为多个三阶张量以降低模型复杂度.虽然 GETD 具有完全表达性,但其不能同时处理具有不同元数的事实.因此,当数据集中同时包含不同元数的事实时,必须将事实以元数进行划分,再分别进行模型训练.最近,Yao 等人[30]提出的 S2S 模型,通过将实体和关系的嵌入表示划分为多个块,允许嵌入表示从具有混合元数的事实中学习.此外,该方法采用神经网络架构搜索技术,稀疏化核张量的参数以降低模型复杂度.

综上所述,基于张量分解的链接预测方法能有效捕获实体和关系间的潜在交互,取得了当前最好的预测准确度,并且大多数有严格数学理论支撑其完全表达性,因而在知识超图链接预测任务上发挥着越来越大的作用.

1.2 负采样方法

1.1 节的知识超图链接预测模型一般通过构造事实评分函数,根据输入样本与真实标签的差距计算损失以优化模型参数.其中,高质量负样本应对训练梯度有所贡献以使模型具有更好的泛化能力,因此一些研究人员将不同负采样方法应用于知识库链接预测任务,以获得高质量的负样本.

最初,大多数模型使用 TransE[11]在知识图谱上提出的均匀随机采样(uniformly random sampling)方法,其将正样本中的头实体或尾实体以相同概率随机替换为实体集中的其他实体,以生成对应的负样本.该方法因简单有效的特性,被后续应用到知识超图的链接预测任务中.然而,均匀随机采样方法的采样空间过大,这使得随机采样到高质量负样本的概率过小,对提升模型泛化能力的贡献不大[10].另一方面,该方法还可能采样到错误的负样本(即未出现在数据集中的正样本),不利于模型训练[15].针对该问题,Wang 等人[31]在 TransH 中提出一种基于伯努利采样(Bernoulli sampling)的方法,通过统计知识库中关系的一对多、多对一、多对多映射属性,使模型以不同的概率替换头实体和尾实体,从而减少了生成错误负样本的概率,但该方法也是从固定分布中采样,所以不能随着模型的训练动态改变,造成大多数生成的负样本可信度得分很小,对基于梯度的优化算法几乎没有贡献.因此,上述 2 种方法都难以生成高质量的负样本,这可能导致模型训练时出现零损失问题[12],即在训练一段时间后,生成的大部分负样本对模型损失函数为零,导致梯度消失,对模型训练没有贡献.

Zhang 等人[13]提出针对知识图谱链接预测的负采样模型 NSCaching,其通过设定缓存(cache)来直接存储负样本,并在训练过程中对负样本进行动态更新,以使其中包含更多高质量的负样本.然而,

该模型需针对元组中不同实体位置设定相应缓存，且在每次模型训练后，需及时更新所有缓存中的负样本，具有较高的时空开销，因而不适合应用于知识超图.Ahrabian 等人[32]考虑到知识图谱的空间结构，提出了 SANS 模型.该模型假设头、尾实体的 k 跳邻居实体与其有更高的相关性，于是将这些相关性高的实体作为事实中负采样的候选集，进而从中采样生成高质量的负样本.然而，该方法亦不易于拓展应用至空间结构复杂的知识超图.RotatE[33]采用自对抗负采样方法进行知识图谱链接预测，其使用模型评分函数获得负样本的概率分布，然后将概率作为负样本的权重引入到链接预测模型的损失函数中.该方法通过模型本身衡量负样本的质量，无额外开销且简单有效，因而可灵活应用于知识超图.

近年来受生成对抗网络(generative adversarial network, GAN)的启发，一些工作尝试使用对抗式训练框架生成高质量的负样本，以解决对模型训练的零损失问题.Cai 等人[10]提出的 KBGAN 模型采用 2 个不同的知识图谱链接预测模型分别作为对抗网络的生成器与判别器，其中生成器负责负样本生成，而判别器负责对输入的样本进行正负判别.随着模型的迭代，负样本生成器与样本判别器通过对抗式训练同时提高性能.测试阶段则采用判别器中的嵌入表示对样本进行评估，以获得良好的链接预测效果.KSGAN[15]通过在 KBGAN 模型上增加一个选择器，筛选生成器生成的负样本，进一步提升了模型性能.Wang 等人[12]提出的 IGAN 模型通过采用双层全连接神经网络作为生成器，为正样本生成对应的高质量负样本，辅助基于平移的链接预测模型进行训练，缓解模型零损失问题的同时也降低了生成器的复杂性.尽管上述基于 GAN 的负采样方法取得了较好的表现，但为增强稳定性并缓解退化问题，必须进行预训练，这带来了额外的训练成本[13].

与 1.2 节提到的负采样方法相比，本文工作主要致力于解决知识超图链接预测中的负采样问题.由于知识超图结构复杂、采样空间大，目前工作大多采用均匀随机采样的方式生成负样本，还未有简单通用的有效方法.本文将生成对抗网络引入到知识超图负采样中，提出的 HyperGAN 方法能够应用于不同关系元数及多种关系类型的知识超图，通过采用 GAN 的输出得分作为衡量负样本质量的标准，过滤生成高质量负样本以提升链接预测模型的性能.此外，HyperGAN 的简单结构使得其不需要进行额外预训练，节约训练成本.

2 预备知识

本节给出知识超图链接预测的一些基本概念.

定义 1. 知识超图.一个知识超图可以表示为 $\mathcal{B}=(\mathcal{E},\mathcal{R},\mathcal{F})$，其中 $\mathcal{E}$ 为实体(节点)集，$\mathcal{R}$ 为多元关系集，$\mathcal{F}$ 为知识超图中的事实集合.对于每个多元关系 $r\in\mathcal{R}$，其元数 $|r|$ 为关系 r 涉及到的实体数量.

在知识图谱中，一个事实通常被表示为三元组 (e_h,r,e_t) 的形式，其中 $e_h,e_t\in\mathcal{E}$ 分别表示头实体与尾实体，$r\in\mathcal{R}$ 表示头尾实体间的关系.类似地，知识超图中的事实 $f\in\mathcal{F}$ 可以被表示为多元组 $(r,e_1,e_2,\cdots,e_n)$ 的形式，其中 $r\in\mathcal{R}$，$e_i\in\mathcal{E}$ 表示关系中第 i 个实体，该事实 f 也被称为 n 元关系事实.

这里用 $\mathcal{F}_{\text{all}}$ 表示世界上全部事实的集合，所有在 $\mathcal{F}_{\text{all}}$ 中的多元组 $(r,e_1,e_2,\cdots,e_n)$ 为正样本，否则为负样本.由于知识库高度不完整，并不能存储所有事实[25]，所以当前知识超图仅包含一部分事实 $\mathcal{F}\subset\mathcal{F}_{\text{all}}$.

定义 2. 知识超图链接预测.给定一个候选多元组 $f\notin\mathcal{F}$，链接预测任务的目的就是判断该多元组 f 是否属于 $\mathcal{F}$ 中缺少的事实，即 f 是否属于 $\mathcal{F}_{\text{all}}\backslash\mathcal{F}$.

3 HyperGAN 方法

本文提出的 HyperGAN 由 3 个模块组成：1)知识超图链接预测模块.将负样本置信度引入损失函数以改善模型性能，并将训练好的嵌入表示共享至其余 2 个模块.2)生成器模块.生成高质量负样本与判别器进行对抗训练.3)判别器模块.在链接预测阶段，为链接预测模块提供负样本置信度；在生成对抗负采样阶段，通过判别正样本与生成器提供的负样本提高自身辨别能力.图 2 为 HyperGAN 的整体架构.

3.1 知识超图链接预测模块

在知识超图链接预测模块中，为了学习到多元关系事实中实体和关系的嵌入表示，本文采用常用的对数损失函数[25,33]优化链接预测模块：

$$\mathcal{L}=\sum_{f\in\mathcal{F}_{\text{batch}}}\left(-\phi(f)+\log\left(\mathrm{e}^{\phi(f)}+\sum_{f'\in\mathcal{N}(f)}\mathrm{e}^{\phi(f')}\right)\right),\tag{1}$$

其中，$\mathcal{F}_{\text{batch}}$ 表示训练过程中每次迭代使用的小批量正样本集，f 表示其中的一个正样本多元组 $(r,e_1,\cdots,e_n)$，ϕ 表示链接预测模块中的事实合理性评分函数，$\mathcal{N}(f)$ 表示基于正样本 f 均匀随机采样生成

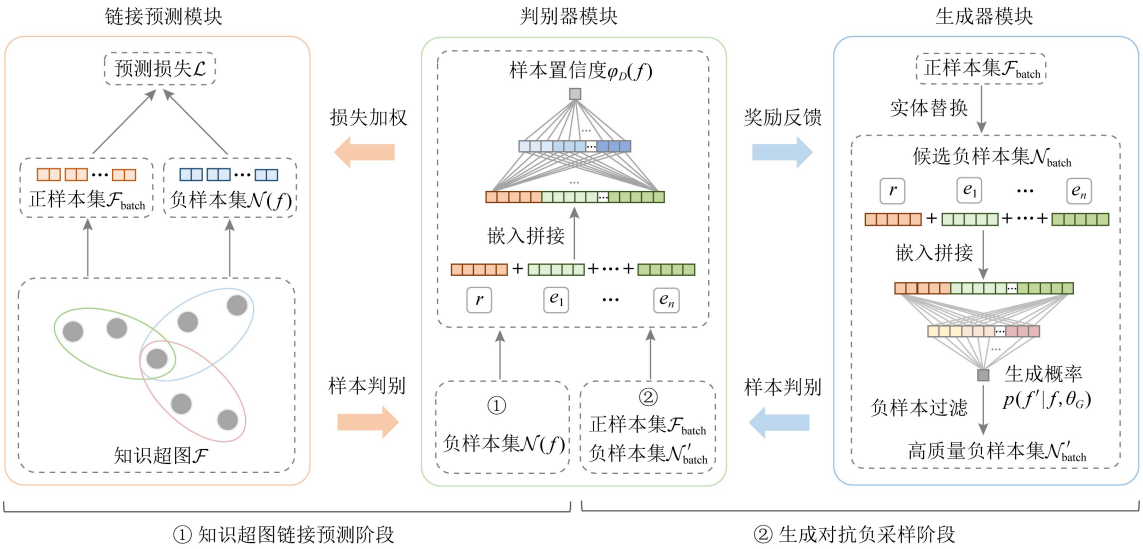


Fig. 2 The architecture of HyperGAN

图 2 HyperGAN 架构图

的负样本集合,$f'\in\mathcal{N}(f)$表示其中的一个负样本.具体地,对于 f 中的任意一个实体 e_i,用实体集 $\mathcal{E}$ 中的其他实体替换该实体以生成不属于 $\mathcal{F}$ 的负样本:

$$\mathcal{N}(f)=\bigcup_{i=1}^{n}\{(r,\cdots,e_i^-,\cdots)\notin\mathcal{F}\mid e_i^-\in\mathcal{E}\},\tag{2}$$

其中,n 表示正样本 f 的元数.根据式(1),损失函数同等对待所有负样本,即评分函数中的权重均为 1,未考虑负样本质量不同对训练梯度贡献的影响.因此,本文在链接预测阶段引入生成对抗网络中的判别器来为每个训练批次中的负样本进行合理性打分,对于 $f'_i\in\mathcal{N}(f)$,计算负样本权重 $\lambda(f'_i)$:

$$\lambda(f'_i\mid\mathcal{N}(f))=\frac{\exp\varphi_D(f'_i)}{\sum\limits_{f'_j\in\mathcal{N}(f)}\exp\varphi_D(f'_j)},\tag{3}$$

其中,φ_D 表示判别器模块的评分函数.为平衡不同负样本之间的影响,将上述负样本权重融入式(1)所述链接预测模块,最终损失函数表示为

$$\mathcal{L}=\sum_{f\in\mathcal{F}_{batch}}\Big(-\phi(f)+\log\big(e^{\phi(f)}+\sum_{f'\in\mathcal{N}(f)}e^{\lambda(f')\phi(f')}\big)\Big).\tag{4}$$

3.2 生成器模块

生成对抗网络中生成器的目标是生成能在对抗生成阶段高效贡献梯度的负样本,即高质量负样本.具体地,HyperGAN 中生成器使用链接预测模块训练好的实体和关系的嵌入表示,将正样本集 $\mathcal{F}_{batch}$ 作为输入,经过实体替换操作,即对于 $\mathcal{F}_{batch}$ 中每个正样本 f,按式(2)替换其中的实体,最终获得候选负样本集 $\mathcal{N}_{batch}$:

$$\mathcal{N}_{batch}=\bigcup_{f\in\mathcal{F}_{batch}}\mathcal{N}(f),\tag{5}$$

之后,生成器使用 3.1 节链接预测模块共享的实体、关系嵌入表示进行训练.对于任一负样本 $f'\in\mathcal{N}_{batch}$,将其关系与实体的嵌入表示按顺序拼接后,作为双层全连接神经网络的输入,经过前向传播,输出负样本 f'生成的概率:

$$p(f'\mid f,\theta_G)=\sigma(\tanh(\oplus(f)\cdot\boldsymbol{W}_1+\boldsymbol{b}_1)\cdot\boldsymbol{W}_2+\boldsymbol{b}_2),\tag{6}$$

其中,σ 表示 sigmoid 函数,$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 分别表示全连接神经网络第 1,2 层的权重矩阵,$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 分别表示偏置向量,$\oplus$表示将事实多元组中关系和实体嵌入表示进行顺序拼接,θ_G 表示生成器模块的参数.则对于候选负样本集中每个负样本,均获得其对应的生成概率.最终,通过将生成概率作为衡量负样本质量的标准,并根据超参数负样本集大小 c,从候选负样本集中过滤概率最大的前 c 个负样本,以此构成高质量负样本集 $\mathcal{N}'_{batch}$.

为优化生成器模块的参数,需要判别器模块对生成的负样本质量进行反馈以计算预期奖励.具体地说,对于生成的高质量负样本 $f'\in\mathcal{N}'_{batch}$,奖励函数被定义为

$$R(f'\mid f)=\tanh(\varphi_D(f')-\varphi_D(f)),\tag{7}$$

其中,φ_D 表示判别器模块的评分函数.对于生成的高质量负样本集,生成器的总奖励为其中所有负样本的奖励期望值:

$$R_G(\theta_G)=\sum_{f\in\mathcal{F}_{batch}}E_{f'\sim p(f'\mid f,\theta_G)}[R(f'\mid f)].\tag{8}$$

由于生成器输出的负样本下标是一个离散值，面临不可微分的问题，因此采用基于策略梯度的强化学习方法[34]计算生成器梯度：

$$\nabla R_G(\theta_G)=R(f'|f)\times\nabla_G\log p(f'|f,\theta_G),\tag{9}$$

根据式(9)中 θ_G 的优化方式，生成器倾向于为高质量负样本输出更高的生成概率，以最大化奖励期望.

3.3 判别器模块

生成对抗网络中判别器的目标是根据监督信号不断提高分辨正负样本的能力.具体地说，HyperGAN方法中的判别器模块在链接预测阶段以均匀随机采样的负样本 $f'_i\in\mathcal{N}(f)$ 作为输入，输出样本置信度 $\varphi_D(f'_i)$ 为预测损失 $\mathcal{L}$ 提供负样本权重.在对抗生成阶段，该模块使用生成器中使用的训练正样本集 $\mathcal{F}_{\text{batch}}$ 与经过过滤得到的高质量负样本集 $\mathcal{N}'_{\text{batch}}$ 作为输入，输出样本置信度得分以判断事实成立的合理性.与生成器类似，本文也采用双层全连接神经网络构建生成对抗网络判别器，由于该模块本质上是一个二元分类器，其目标函数可以表示为最小化二元交叉熵损失的形式：

$$\mathcal{L}_D=-\sum_{f\in\mathcal{F}_{\text{batch}}}\varphi_D(f)-\sum_{f'\in\mathcal{N}'_{\text{batch}}}(1-\varphi_D(f')),\tag{10}$$

其中，φ_D 表示判别器中的评分函数.以正样本 f 为例，经过判别器输出的置信度 $\varphi_D(f|\theta_D)$ 表示为

$$\varphi_D(f|\theta_D)=\sigma(\tanh(\oplus(f)\cdot\boldsymbol{W}_3+\boldsymbol{b}_3)\cdot\boldsymbol{W}_4+\boldsymbol{b}_4),\tag{11}$$

其中，$\boldsymbol{W}_3$ 和 $\boldsymbol{W}_4$ 分别表示全连接神经网络第一、二层的权重矩阵，$\boldsymbol{b}_3$ 和 $\boldsymbol{b}_4$ 分别表示偏置向量，θ_D 表示判别器模块的参数.

3.4 模型训练

HyperGAN方法的训练过程由知识超图链接预测和生成对抗负采样2阶段构成.

知识超图链接预测阶段的主要目的是优化知识超图节点的嵌入表示与链接预测评分函数的参数，因此，链接预测模块与判别器模块共同参与计算预测损失.其中，固定参数的判别器只进行前向传播以输出权重，并不进行参数更新.这一阶段的训练通过最小化式(4)中的损失函数提升高质量负样本对模型的贡献、降低低质量负样本对模型的干扰，以使模型学得更好的实体和关系的嵌入表示，提升模型的链接预测准确度.

生成对抗负采样阶段的主要目的是对抗训练生成器和判别器模块，共同提高生成器生成高质量负样本的能力与判别器分辨正负样本的能力.在这一阶段中，链接预测模块的参数被固定，生成器模块使用链接预测模块训练好的实体和关系的嵌入表示，对训练正样本集生成高质量负样本集合并送至判别器；判别器对输入的正负样本进行事实合理性判别.在对抗训练过程中，分别采用随机梯度下降法与Adam[35]算法优化式(9)的梯度与式(10)中的二元交叉熵损失，2个模块交替进行参数更新，预期在达到平衡时获得各自最好的性能.

值得注意的是，以往方法将结构复杂的链接预测模型作为生成对抗网络判别器，HyperGAN方法则将预测模型与判别器解耦，采用双层全连接神经网络构建生成器与判别器，因而在保持预测能力的同时简化了生成对抗网络结构.同时，考虑到可能面临的生成对抗网络不稳定性[36]，HyperGAN使用结构相同、参数不一致的生成器与判别器缓解两者间建模能力不平衡的问题.此外，HyperGAN方法采用间断式两阶段训练模式，即在每次迭代时均进行链接预测阶段训练，而间隔 C 次迭代才进行一次生成对抗训练，减小每次迭代带来的不稳定性.基于以上原因，HyperGAN方法无需像以往结构复杂[15]或共享链接预测模块与判别器[10,12]的工作一样进行预训练，也可保证其训练过程中的稳定性.设计的间断式两阶段训练相比每次迭代均进行训练，亦提升了训练效率.

最终，使用知识超图链接预测模块学习到的嵌入表示作为对实体和关系的最终表示.算法1给出了HyperGAN方法的伪代码.

算法 1. HyperGAN方法.

输入：事实集 $\mathcal{F}$、链接预测模块损失函数 $\mathcal{L}$、判别器损失函数 $\mathcal{L}_D$、生成器奖励期望 R_G、总训练次数 K、生成器 G、判别器 D、间隔次数 C、对抗网络训练次数 M；

输出：事实集 $\mathcal{F}$ 中所有实体嵌入表示 $\boldsymbol{e}$、关系嵌入表示 $\boldsymbol{r}$.

① for $i\leftarrow 1$ to K do
② if i mod $C=0$ do /*每间隔 C 次*/
③ for $j\leftarrow 1$ to M do
④ for $\mathcal{F}_{\text{batch}}$ of $\mathcal{F}$ do
⑤ 通过 G 生成候选负样本集 $\mathcal{N}_{\text{batch}}$；
⑥ 过滤得到高质量负样本集；
⑦ 计算 $\mathcal{L}_D$ 并更新参数 θ_D；
/*式(10)(11)*/
⑧ 计算 R_G 并更新参数 θ_G；
/*式(8)(9)*/

```
⑨          end for
⑩        end for
⑪      end if
⑫      for F_batch of F do
⑬        for f∈F_batch do
⑭          生成负样本集 N(f);
⑮          通过 D 计算负样本权重;
             /*式(3)*/
⑯        end for
⑰        计算链接预测损失 L; /*式(4)*/
⑱        通过反向传播更新 e,r;
⑲      end for
⑳  end for
```

4 实　　验

本节为验证 HyperGAN 方法在知识超图链接预测任务上的有效性,设计了 4 组实验:1)在 5 个公开数据集上对多个知识超图链接预测模型应用 HyperGAN 负采样方法,并将其与均匀随机采样和自对抗负采样方法进行对比,以评估 HyperGAN 方法的有效性;2)通过效率实验对比不同负采样方法对链接预测模型效率的影响;3)通过参数敏感性实验分析模型参数对实验的鲁棒性影响;4)通过负样本案例分析及定量分析验证 HyperGAN 方法生成了质量较高的负样本.以上实验均基于 PyTorch[37] 1.6.0,CUDA 11.0 和 Python 3.6.6 的实验环境,在使用 RTX3090 显卡的服务器上进行.在详细说明这些实验之前,首先介绍实验中用到的数据集、评价指标、基线模型和实验设置.

4.1 数据集

本文使用 5 个公开的真实大型知识超图数据集对 HyperGAN 方法进行评估,其详细统计信息如表 1 所示.数据集的简要描述:

1) JF17K-3[25]是基于 Freebase① 过滤得到的三元关系数据集.首先将 Freebase 中出现次数较少的实体所对应的事实删除,在删除涉及到字符串、数字及枚举类型的事实之后,从每个多元关系中随机选出 10 000 个事实并进一步删除出现次数少于 5 的实体所对应的事实;然后,利用文献[18]中的逆向化方法生成多元组;最后过滤元数为 3 的多元组来构成数据集.

2) JF17K-4[25]是基于 Freebase 依据类似 1)中流程构建的四元关系数据集.

3) WikiPeople-3[30]是基于 Wikidata② 过滤得到的三元关系数据集.首先抽取出实体类型为"人物"的全部事实,移除其中涉及到图像类型及包含未知实体的事实,以及删除出现次数少于 30 的实体所对应的事实,最终过滤得到元数为 3 的事实来构成数据集.

4) WikiPeople-4[30]是基于 Wikidata 依据类似 3)中流程构建的四元关系数据集.

5) M-FB15K-3 是基于 M-FB15K[2] 过滤得到的三元关系数据集.

Table 1 Statistics of Datasets

表 1 数据集的统计信息

数据集	实体	关系	训练集	验证集	测试集
JF17K-3	11 541	104	27 635	3 454	3 455
JF17K-4	6 536	23	7 607	951	951
WikiPeople-3	12 270	66	20 656	2 582	2 582
WikiPeople-4	9 528	50	12 150	1 519	1 519
M-FB15K-3	4 240	20	336 754	31 897	31 376

4.2 评价指标

本文使用 2 个广泛应用于知识超图链接预测模型的度量指标对模型性能进行评估:平均倒数排名 MRR 和击中率 $Hits@k$,$k\in\{1,3,10\}$.给定知识超图事实集 $\mathcal{F}$,用 $f=(r,e_1,e_2,\cdots,e_n)$ 表示其中任一正样本.对于 f 中位置 m 上的实体 e_m,使用知识超图实体集 $\mathcal{E}$ 中的全部实体将其替换,从而得到 $|\mathcal{E}|$ 个样本,删除其中出现在 $\mathcal{F}$ 中的样本,这样就构造出对应于正样本 f 位置 m 的负样本集 $\mathcal{N}_m(f)$,令 $\mathcal{H}_m(f)=\{f\}\cup\mathcal{N}_m(f)$.通过使用链接预测模型中评分函数对 $\mathcal{H}_m(f)$ 中所有样本打分,将得到的评价分数进行降序排序,得到正样本 f 在 $\mathcal{H}_m(f)$ 中的排名 $rank_m(f)$.MRR 是事实集中所有正样本排名的倒数平均值,$Hits@k$ 是正样本排序在前 k 个的比例.MRR 和 $Hits@k$ 的具体计算:

$$MRR=\frac{1}{\sum\limits_{f\in\mathcal{F}}|r|}\sum_{f\in\mathcal{F}}\sum_{m=1}^{|r|}\frac{1}{rank_m(f)},\tag{12}$$

$$Hits@k=\frac{\sum\limits_{f\in\mathcal{F}}\sum\limits_{m=1}^{|r|}cond(rank_m(f)\leqslant k)}{\sum\limits_{f\in\mathcal{F}}|r|},\tag{13}$$

① http://www.freebase.com/

② https://www.wikidata.org/

其中,r 是正样本 f 中的关系,$cond(\cdot)$是条件函数,当条件成立时值为 1,否则为 0.MRR 和 $Hits@k$ 值越大表明模型性能越好.

4.3 基线模型和实验设置

本文选择 6 个知识超图链接预测模型作为实验基线模型:

1) m-DistMult[26].对 DistMult 算法[2]进行了扩展,通过对多元关系和实体的嵌入表示计算哈达玛积,将其加和得到多元关系事实成立的得分.

2) RAE[19].对 m-TransH 算法[18]进行了扩展,考虑到 2 个实体同时出现在一个多元关系中的相关性,将其引入模型损失函数进行训练.

3) n-CP[2].对 CP 算法[27]进行了扩展,通过赋予一个实体多个嵌入表示,将实体在关系中的位置信息引入模型以进行知识超图链接预测.

4) n-TuckER[25].对 TuckER 算法[28]进行了扩展,基于 TuckER 分解将知识超图中事实的高阶张量表示分解为一个核张量加若干个因子矩阵的形式,并通过将关系和实体的嵌入表示与核张量相乘,获得多元关系事实成立的得分.

5) GETD[25].对 n-TuckER 算法进行了扩展,通过引入 TR 分解,将核张量进一步分解为多个三阶张量,降低了模型的复杂度.

6) S2S[30].对 n-TuckER 算法进行了扩展,通过将实体和关系的嵌入表示划分为多个块,允许模型从具有混合元数的事实中学习嵌入表示.

为验证 HyperGAN 方法对链接预测模型的性能提升效果,本文分别在基线模型中链接预测性能良好的 GETD 与 S2S 模型上应用本文提出的 HyperGAN 方法,并在相同的参数设置下,与模型本身采用的均匀随机采样方法和拓展到知识超图的自对抗负采样(Self-Adv)方法[33]进行对比实验.

HyperGAN 中生成器和判别器模块分别通过 SGD 和 Adam[35]优化器进行训练,实体和关系的嵌入表示维度、批处理大小、训练次数等参数与框架中采用的知识超图链接预测模型相关,与其在原文中的参数保持一致.其他实验默认参数设置如下,生成对抗网络的间隔迭代次数 $C=20$,SGD 和 Adam 优化器的学习率为 0.01,自对抗负采样方法的采样率 α 为 0.5 或 1,生成对抗网络训练次数 M 与全连接神经网络隐藏层的维度 d 在不同数据集中的设置如表 2 所示.HyperGAN 的实验代码和所用数据集将被公开在 https://github.com/GuoZhengshan/HyperGAN.

Table 2 Hyper-Parameter Settings of HyperGAN

表 2 HyperGAN 的超参数设置

数据集	GETD			S2S		
	M	d	α	M	d	α
JF17K-3	16	8	0.5	16	50	0.5
JF17K-4	38	11	0.5	38	50	0.5
WikiPeople-3	16	8	1.0	16	50	0.5
WikiPeople-4	38	11	1.0	38	50	1.0

4.4 有效性分析

表 3 和表 4 分别给出了基于 HyperGAN 方法的多个知识超图链接预测模型在数据集 JF17K-3/4 和 WikiPeople-3/4 上的实验结果.其中涉及 GETD 模型的结果由取自其论文文献[25]中的公开源代码①实验完成.对于 S2S 模型,由于其未在论文中公开源代码,为进行对比分析,本文使用作者公开 GitHub②仓库中的源代码进行实验,并对缺失部分进行了复现补全(该代码未给出元数为 4 情况下的代码).此外,针对 S2S 模型代码中缺失超参数的部分,本文参考 GETD 模型中的超参数进行设置与微调,因此本文报告的部分实验结果与 S2S 原论文报告的结果有一定差距.尽管如此,由于 HyperGAN 方法是面向知识超图链接预测模型的通用负采样优化方法,对比加入 HyperGAN 方法前后的性能实验结果依旧验证了该方法的有效性.除上述基线模型,其余方法的实验结果均采用对应文章中报告的最好结果.

实验结果表明:本文提出的 HyperGAN 方法在上述数据集上都获得了良好的效果.具体地说,就 $Hits@1$ 指标而言,基于 GETD 模型的 HyperGAN 方法相较于其原本均匀随机采样方法在 JF17K-3/4 和 WikiPeople-3/4 数据集上分别提升了 3.29%,1.59%,4.58%和 3.02%.扩展到超图的自对抗负采样(Self-Adv)方法也提升了 GETD 模型的链接预测性能,就 WikiPeople-3 数据集而言,其在 $Hits@1$ 和 $Hits@3$ 指标上分别提升了 2.47%和 2.24%,但 HyperGAN 方法与自对抗负采样方法相比,在各数据集上均表现更优,特别是在 WikiPeople-4 数据集上,$Hits@1$ 和 $Hits@3$ 指标分别提升了 2.25%和

① https://github.com/liuyuaa/GETD/

② https://github.com/AutoML-Research/

2.63%.基于 S2S 模型的 HyperGAN 方法对链接预测性能提升实验结果可以发现,HyperGAN 方法同样在各数据集上都取得了良好的表现,特别是在 WikiPeople-4 数据集上的 *MRR* 和 *Hits*@1 指标分别提升了 1.35%和 3.75%.然而,自对抗负采样方法在 S2S 模型上性能不稳定,在 JF17K-3/4 数据集上的指标稍加提升,而在 WikiPeople-3/4 数据集上反而下降.导致这一点的原因可能是,更复杂的链接预测模型结构会使自对抗负采样方法在计算权重时具有偏差.相比之下,HyperGAN 方法为优化负样本质量而设计的生成对抗网络在结合复杂结构的链接预测模型时依旧能够具有稳定的性能.

Table 3 The Link Prediction Results on the JF17K-3/4 Datasets

表 3 JF17K-3/4 数据集链接预测结果

方法	JF17K-3				JF17K-4			
	MRR	*Hits*@1	*Hits*@3	*Hits*@10	*MRR*	*Hits*@1	*Hits*@3	*Hits*@10
RAE[19]	0.505	0.430	0.532	0.644	0.707	0.636	0.751	0.835
n-CP[2]	0.700	0.635	0.736	0.827	0.787	0.733	0.821	0.890
n-TuckER[25]	0.727	0.664	0.761	0.852	0.804	0.748	0.841	0.902
GETD[25]	0.732	0.669	0.764	0.856	0.810	0.755	0.844	**0.913**
S2S[30]	0.739	0.680	0.769	0.855	**0.816**	0.769	0.848	0.900
GETD+Self-Adv[33]	**0.743**	0.681	**0.778**	**0.864**	0.814	0.760	0.849	**0.910**
S2S+Self-Adv[33]	0.742	0.683	0.769	0.861	**0.818**	**0.772**	0.848	0.903
GETD+HyperGAN (本文)	**0.750**	**0.691**	**0.781**	**0.868**	**0.818**	0.767	**0.850**	**0.910**
S2S+HyperGAN (本文)	**0.743**	**0.684**	0.773	**0.864**	**0.818**	**0.770**	**0.851**	0.908

注:黑体数字表示在 2 种模型中最优链接预测结果.

Table 4 The Link Prediction Results on the WikiPeople-3/4 Datasets

表 4 WikiPeople-3/4 数据集链接预测结果

方法	WikiPeople-3				WikiPeople-4			
	MRR	*Hits*@1	*Hits*@3	*Hits*@10	*MRR*	*Hits*@1	*Hits*@3	*Hits*@10
RAE[19]	0.239	0.168	0.252	0.379	0.150	0.080	0.149	0.273
n-CP[2]	0.330	0.250	0.356	0.496	0.265	0.169	0.315	0.445
n-TuckER[25]	0.365	0.274	0.400	0.548	0.362	0.246	0.432	0.570
GETD[25]	0.373	0.284	0.401	**0.558**	**0.386**	0.265	**0.462**	0.596
S2S[30]	0.371	0.293	0.404	0.522	0.313	0.216	0.357	0.511
GETD+Self-Adv[33]	**0.378**	0.291	**0.410**	**0.558**	**0.386**	**0.267**	0.456	**0.601**
S2S+Self-Adv[33]	0.369	0.292	0.406	0.517	0.304	0.208	0.351	0.490
GETD+HyperGAN (本文)	**0.382**	**0.297**	**0.414**	**0.553**	**0.393**	**0.273**	**0.468**	**0.605**
S2S+HyperGAN (本文)	0.376	**0.304**	**0.410**	0.519	0.315	0.215	0.362	0.514

注:黑体数字表示在 2 种模型中最优链接预测结果.

为进一步研究 HyperGAN 在更大规模数据集上的效果,本文基于 GETD 链接预测模型在数据集 M-FB15K-3 上进行了实验,表 5 给出了相应的实验结果.具体地说,Self-Adv 方法相较于其原本采用的均匀随机采样方法在各性能指标上均表现不佳,而 HyperGAN 方法取得了良好的表现,特别是在 *Hits*@10 指标上提升了 1.21%,从而进一步表明 HyperGAN 在大规模数据集上同样具有稳定的性能.

Table 5 The Link Prediction Results on the M-FB15K-3

表 5 M-FB15K-3 链接预测结果

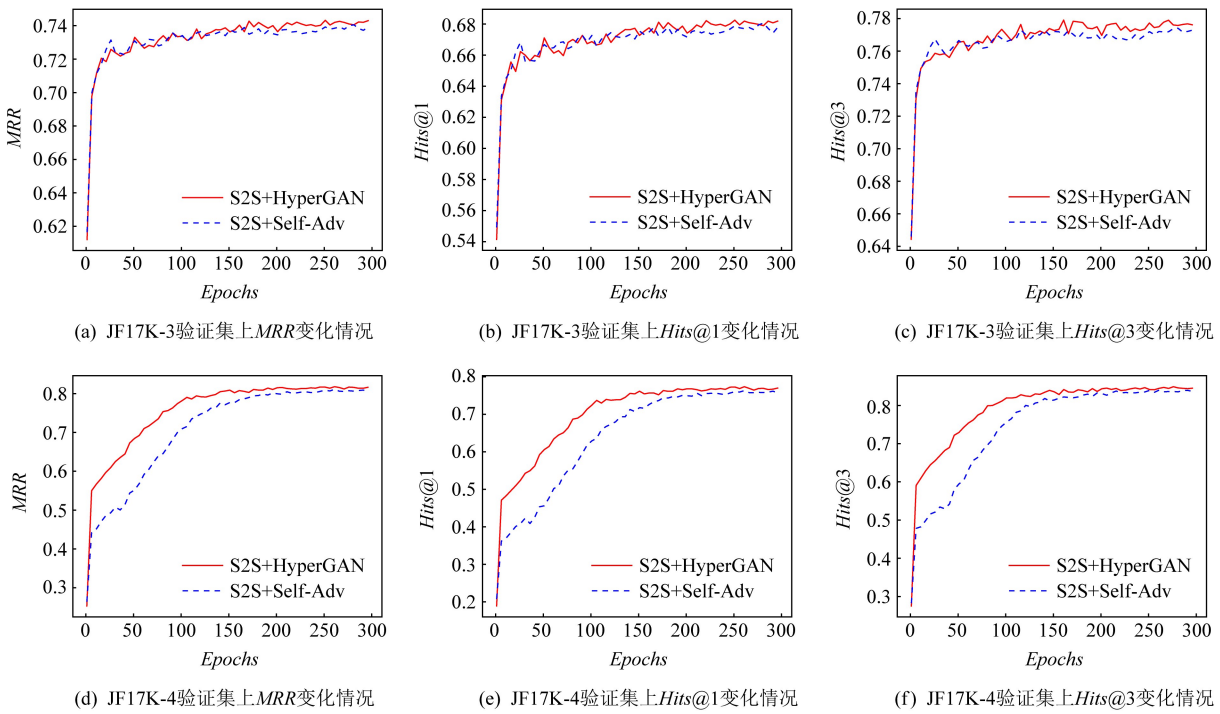
方法	M-FB15K-3		
	MRR	*Hits*@3	*Hits*@10
GETD[25]	0.715	0.737	0.825
GETD+Self-Adv[33]	0.707	0.730	0.819
GETD+HyperGAN (本文)	**0.719**	**0.744**	**0.835**

注:黑体数字表示最优链接预测结果.

4.5 效率分析

本节进一步对比了 HyperGAN 方法和自对抗负采样(Self-Adv)方法[33]对知识超图链接预测模型的效率影响.具体地,在 JF17K-3/4 数据集的验证集上分别训练加入了 HyperGAN 与自对抗负采样方法的 S2S 链接预测模型,并记录性能评价指标随训练迭代次数增加的变化.

如图 3 所示,HyperGAN 方法与自对抗负采样方法在三元关系数据集 JF17K-3 上对链接预测模型的收敛速度影响相差不大,但在四元关系数据集 JF17K-4 上,HyperGAN 方法相对自对抗负采样方法具有更快的收敛速度.这一现象表明 HyperGAN 方法相比以往负采样方法不仅对链接预测模型有更好的性能提升,在元数更大的知识超图数据集上也有更高的效率.



(a) JF17K-3验证集上*MRR*变化情况 (b) JF17K-3验证集上*Hits*@1变化情况 (c) JF17K-3验证集上*Hits*@3变化情况

(d) JF17K-4验证集上*MRR*变化情况 (e) JF17K-4验证集上*Hits*@1变化情况 (f) JF17K-4验证集上*Hits*@3变化情况

Fig. 3 Performance comparison of HyperGAN and Self-Adv on multiple metrics of S2S method

图 3 HyperGAN 方法和 Self-Adv 方法在 S2S 模型各评价指标上的对比

4.6 参数敏感性分析

为测试 HyperGAN 方法的鲁棒性,本文基于知识超图链接预测模型 GETD 在 JF17K-3 数据集上分析了部分超参数对模型性能的影响,包括生成对抗网络与链接预测模型的训练迭代间隔次数 C、生成对抗网络训练次数 M 以及生成器与判别器模块的全连接神经网络隐藏层维度 d.令间隔次数 $C\in\{10,20,30,40,50,60\}$,生成对抗网络训练次数 $M\in\{2,6,12,14,16,18,24\}$,隐藏层的维度 $d\in\{4,5,6,8,10,16,32\}$,其余超参数均与 4.3 节实验设置中的参数保持一致,图 4 给出了相应的实验结果.

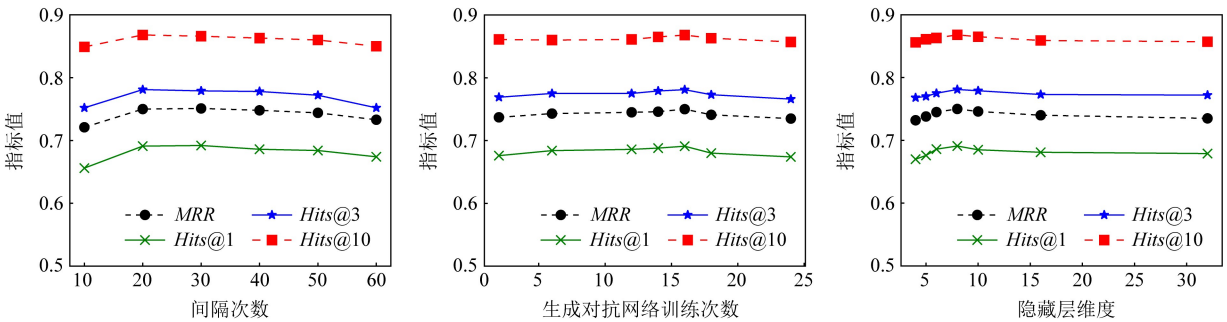
1) 间隔次数.如图 4(a)所示,当训练间隔次数小于 20 时,随着次数的增加,模型的各性能指标均呈上升趋势,表明增大链接预测模型与生成对抗网络的训练间隔,有助于各模块自身训练,从而使整体的性能得以提升.然而,当间隔次数到达 20 时,上述指标逐渐平稳不再增加,表明 HyperGAN 方法性能逐渐稳定,而当次数到达 50 时,性能出现下降,表明间隔次数过多会导致 HyperGAN 各模块间信息交互不及时,对模型训练不利.

2) 生成对抗网络训练次数.如图 4(b)所示,随着训练次数的增加,各项性能指标均随之变好,表明充分训练生成对抗网络有助于提升模型性能.但当训练次数达到 16 时,性能指标开始呈下降趋势,表明过多的训练可能导致过拟合问题.因此,生成对抗网络的训练次数对 HyperGAN 方法而言是一个较为敏感的超参数.

3) 隐藏层维度.如图 4(c)所示,当维度开始增加时,各项性能指标均有所提升,表明加大隐藏层的维度有助于链接预测模型性能的提升,太小的维度可能导致学习过程中信息的丢失.然而,当维度到达 10 时,各项性能指标逐渐平稳不再增加,表明维度到达一定程度时已能够表达所有信息.

为研究链接预测性能的提升与 HyperGAN

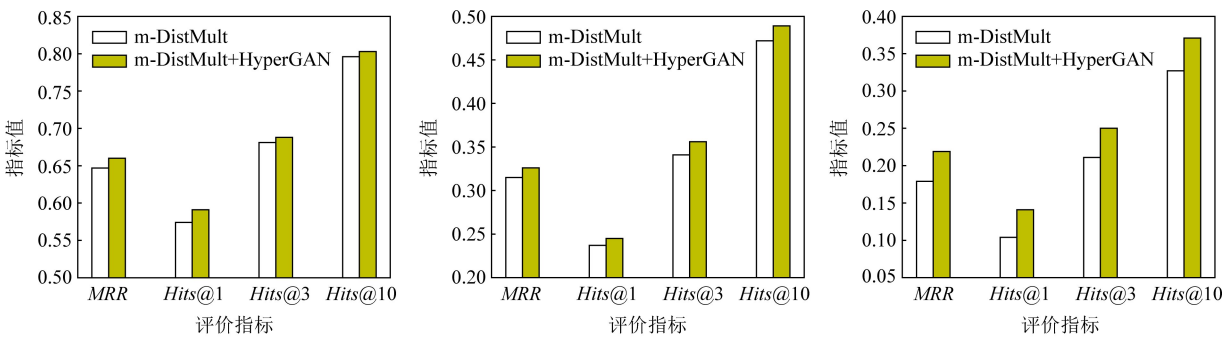
方法的相关性,本文将 HyperGAN 应用于知识超图场景下基础链接预测模型 m-DistMult[26] 并记录性能变化,结果如图 5 所示.多个数据集上的 4 种性能指标显示,HyperGAN 方法即使面对最基础的链接预测模块,也能通过提升负样本质量辅助提升预测精度.



(a) 间隔次数对指标的影响 (b) 生成对抗网络训练次数 (c) 隐藏层维度对指标的影响

Fig. 4 Parameter sensitivity analysis

图 4 参数敏感性分析



(a) JF17K-3数据集 (b) WikiPeople-3数据集 (c) WikiPeople-4数据集

Fig. 5 Performance comparison of m-DistMult and m-DistMult+HyperGAN on multiple datasets

图 5 m-DistMult 和 m-DistMult+HyperGAN 方法在各数据集上的对比

4.7 负样本质量分析

为了验证 HyperGAN 方法能生成高质量的负样本,本文基于知识超图链接预测模型 GETD 在 WikiPeople-3/4 和 JF17K-3/4 数据集上进行了案例分析.如表 6 所示,列 2 正样本以多元组的形式表示,其中黑体单词表示被替换的实体,列 3 为经过 HyperGAN 方法生成的替换后实体.在 WikiPeople-3/4 数据集中,HyperGAN 方法更倾向于选择同一类型的实体进行替换,例如时间实体会被替换为其他时间,影视作品"西城故事"会替换为"指环王:王者归来",奖项名称"图灵奖"被替换为"诺贝尔物理学奖".在 JF17K-3/4 数据集上,HyperGAN 方法更倾向于选择有相似语义关系的实体进行替换.例如对于描述网球比赛关系的事实,"塞雷娜·威廉姆斯"会被替换为同为网球运动员的"贾斯汀·海宁";对于描述大不列颠贵族等级的事实,"公爵"被替换为"子爵".在以上 4 个数据集上的案例分析表明,本文提出的 HyperGAN 方法会选择同类型或是在语义上相似、具有内在联系的实体作为替换,以生成能为链接预测模型训练贡献更大梯度的高质量负样本,而传统的均匀随机采样方法大概率会采样到不属于同一类型或几乎完全不相关的实体作为替换,从而可能会使模型在训练阶段面临零损失问题.

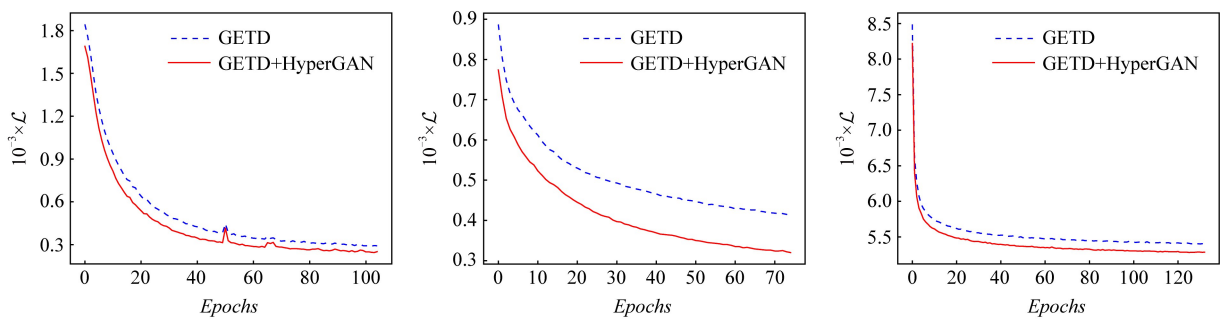
此外,为进一步定量分析负样本质量对链接预测模型的影响,本文在 JF17K,WikiPeople 和 m-FB15K 数据集上基于链接预测模型 GETD 给出了模型损失随训练次数的变化情况.如图 6 所示,在应用 HyperGAN 方法后,模型损失相比原方法下降更快,在正样本一致的情况下,这一现象表明通过 HyperGAN 方法生成的高质量负样本能为链接预测模型带来更大损失,贡献更多梯度,通过解决零损失问题提升了知识超图链接预测的准确度.

Table 6 The Positive Samples and Replaced Entities with HyperGAN on the JF17K-3/4 and WikiPeople-3/4

表 6 JF17K-3/4 和 WikiPeople-3/4 中的正样本与 HyperGAN 替换后的实体

数据集	正样本	替换后的实体
WikiPeople-3	(nominated for, **Neil Gaiman**, Locus Award for Best Short Story, 2003)	Salomon Eberhard Henschen
	(relative, Marcel Mauss, **Émile Durkheim**, maternal uncle)	Vladimir Kryukov
	(nominated for, Robert Burns Woodward, Nobel Prize in Chemistry, **1956**)	1943
WikiPeople-4	(member of sports team, **Andrey Chernyshov**, P.A.O.K. Thessaloniki F.C., 1996, 1997)	Ricardo Costa
	(award received, Allen Newell, **Turing Award**, 1975, Herbert Alexander Simon)	Nobel Prize in Physics
	(nominated for, Robert Wise, Academy Award for Best Picture, **West Side Story**, 34th Academy Awards)	The Lord of the Rings: The Return of the King
	(position held, Egon Jüttner, member of the German Bundestag, 2002, **2005**)	2008
JF17K-3	(tennismatch, **Serena Williams**, US Open, Venus Williams)	Justine Henin
	(royalty precedence, marquess, **duke**, peerage of Great Britain)	baron
	(basketball player stats, 2007-08 NBA season, Kobe Bryant, **Los Angeles Lakers**)	Washington Wizards
JF17K-4	(film dubbing performance, **English**, Shikamaru Nara, Naruto Shippuden the Movie, Tom Gibis)	Chinese
	(regular TV appearance, Lawrence, **Plankton and Karen**, season 4, SpongeBob SquarePants)	Jill Talley
	(sports award, 2008-09 NBA season, Kobe Bryant, Los Angeles Lakers, **Bill Russell NBA Finals Most Valuable Player Award**)	NBA Most Valuable Player Award
	(Olympic medal honor, 2010 Winter Olympics, Finland, Kimmo Timonen, **men's ice hockey tournament**)	women's ice hockey tournament

注:列 2 中的黑体单词表示被替换的实体,列 3 为 HyperGAN 进行替换后的实体.



(a) JF17K-4数据集 (b) WikiPeople-4数据集 (c) M-FB15K-3数据集

Fig. 6 Training loss in one epoch versus number of epochs for GETD and GETD+HyperGAN on datasets

图 6 GETD 和 GETD+HyperGAN 方法在各数据集上模型损失随训练次数的变化情况

5 总 结

本文提出了面向知识超图链接预测的生成对抗负采样方法 HyperGAN,其通过对抗训练生成高质量负样本以提升链接预测模型的性能.此外,HyperGAN 方法无需预训练的特性使其在辅助链接预测模型进行训练时相比以往负采样方法具有更高的效率.在 5 个大型公开知识超图数据集上的对比实验验证了 HyperGAN 方法在性能与效率方面的有效性与先进性.

下一步,计划探究知识超图不同类型信息间的隐式关联,以提高链接预测中的负样本质量,并分析负样本质量在不同衡量准则下对链接预测性能的影响.其次,考虑将变分自编码器用于链接预测负采样,并且将并行计算融入链接预测模型以进一步提升其在大规模数据集上的训练效率.此外,知识超图的其余下游应用也是值得探索的研究方向.

作者贡献声明:郭正山提出了方法思路和实验方案,并完成实验和撰写论文;左劼指导研究方案设计并修改论文;段磊提出指导性意见并修改论文;李仁昊、何承鑫、王培妍共同辅助撰写论文;肖英劼负责辅助完善实验,整理参考文献.

参 考 文 献

[1] Bollacker K, Evans C, Paritosh P, et al. Freebase: A collaboratively created graph database for structuring human knowledge [C] //Proc of the 2008 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2008: 1247-1250

[2] Fatemi B, Taslakian P, Vazquez D, et al. Knowledge hypergraphs: Prediction beyond binary relations [C] //Proc of the 29th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2020: 2191-2197

[3] Lande D, Fu Minglei, Guo Wen, et al. Link prediction of scientific collaboration networks based on information retrieval [J]. World Wide Web-Internet and Web Information Systems, 2020, 23(4): 2239-2257

[4] Zhang Fuzheng, Yuan N, Lian Defu, et al. Collaborative knowledge base embedding for recommender systems [C] //Proc of the 22nd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2016: 353-362

[5] JiShuyi, Feng Yifan, Ji Rongong, et al. Dual channel hypergraph collaborative filtering [C] //Proc of the 26th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2020: 2020-2029

[6] Yang Bishan, Mitchell T. Leveraging knowledge bases in LSTMs for improving machine reading [C] //Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2017: 1436-1446

[7] Abboud R, Ceylan I, Lukasiewicz T, et al. BoxE: A box embedding model for knowledge base completion [C] //Proc of the 34th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2020: 9649-9661

[8] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A nucleus for a Web of open data [C] //Proc of the 6th Int Semantic Web Conf and 2nd Asian Semantic Web Conf. Berlin: Springer, 2007: 722-735

[9] Krompaβ D, Baier S, Tresp V. Type-constrained representation learning in knowledge graphs [C] //Proc of the 14th Int Semantic Web Conf. Berlin: Springer, 2015: 640-655

[10] Cai Liwei, Wang Y W. KBGAN: Adversarial learning for knowledge graph embeddings [C] //Proc of the 2018 Conf of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 1470-1480

[11] Bordes A, Usunier N, Garcia-Durán A, et al. Translating embeddings for modeling multi-relational data [C] //Proc of the 27th Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 2787-2795

[12] Wang Peifeng, Li Shuangyin, Pan Rong. Incorporating GAN for negative sampling in knowledge representation learning [C] //Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 2005-2012

[13] Zhang Yongqi, Yao Quanming, Shao Yingxia, et al. NSCaching: Simple and efficient negative sampling for knowledge graph embedding [C] //Proc of the 35th Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2019: 614-625

[14] Zhang Ziwei, Wang Xin, Zhu Wenwu. Automated machine learning on graphs: A survey [C] //Proc of the 30th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2021: 4704-4712

[15] Hu Kairong, Liu Hai, Hao Tianyong. A knowledge selective adversarial network for link prediction in knowledge graph [C] //Proc of the 8th CCF Int Conf on Natural Language Processing and Chinese Computing. Berlin: Springer, 2019: 171-183

[16] Liu Yu, Yao Quanming, Li Yong. Role-aware modeling for n-ary relational knowledge bases [C] //Proc of the 2021 World Wide Web Conf. New York: ACM, 2021: 2660-2671

[17] Kazemi S, Poole D. SimplE embedding for link prediction in knowledge graphs [C] //Proc of the 32nd Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2018: 4289-4300

[18] Wen Jianfeng, Li Jianxin, Mao Yongyi, et al. On the representation and embedding of knowledge bases beyond binary relations [C] //Proc of the 25th Int Joint Conf on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2016: 1300-1307

[19] Zhang Richong, Li Junpeng, Mei Jiajie, et al. Scalable instance reconstruction in knowledge bases via relatedness affiliated embedding [C] //Proc of the 2018 World Wide Web Conf. New York: ACM, 2018: 1185-1194

[20] Guan Saiping, Jin Xiaolong, Wang Yuanzhuo, et al. Link prediction on n-ary relational data [C] //Proc of the 2019 World Wide Web Conf. New York: ACM, 2019: 583-593

[21] Rosso P, Yang Dingqi, Cudré-Mauroux P. Beyond triplets: Hyper-relational knowledge graph embedding for link prediction [C] //Proc of the 2020 World Wide Web Conf. New York: ACM, 2020: 1885-1896

[22] Guan Saiping, Jin Xiaolong, Guo Jiafeng, et al. NeuInfer: Knowledge inference on n-ary facts [C] //Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 6141-6151

[23] Galkin M, Trivedi P, Maheshwari G, et al. Message passing for hyper-relational knowledge graphs [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 7346-7359

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] //Proc of the 31st Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2017: 5998-6008

[25] Liu Yu, Yao Quanming, Li Yong. Generalizing tensor decomposition for n-ary relational knowledge bases [C] //Proc of the 2020 World Wide Web Conf. New York: ACM, 2020: 1104-1114

[26] Wang Peiyan, Duan Lei, Guo Zhengshan, et al. Knowledge hypergraph link prediction model based on tensor decomposition [J]. Journal of Computer Research and Development, 2021, 58(8): 1599-1611 (in Chinese)
(王培妍, 段磊, 郭正山, 等. 基于张量分解的知识超图链接预测模型[J]. 计算机研究与发展, 2021, 58(8): 1599-1611)
[27] Hitchcock F. The expression of a tensor or a polyadic as a sum of products [J]. Journal of Mathematics and Physics, 1927, 6(1-4): 164-189
[28] Balažević I, Allen C, Hospedales T. TuckER: Tensor factorization for knowledge graph completion [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing and the 9th Int Joint Conf on Natural Language Processing. Stroudsburg, PA: ACL, 2019: 5184-5193
[29] Yuan Longhao, Li Chao, Mandic D, et al. Tensor ring decomposition with rank minimization on latent space: An efficient approach for tensor completion [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 9151-9158
[30] Di Shimin, Yao Quanming, Chen Lei. Searching to sparsify tensor decomposition for n-ary relational data [C] //Proc of the 2021 World Wide Web Conf. New York: ACM, 2021: 4043-4054
[31] Wang Zhen, Zhang Jianwei, Feng Jianlin, et al. Knowledge graph embedding by translating on hyperplanes [C] //Proc of the 28th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2014: 1112-1119
[32] Ahrabian K, Feizi A, Salehi Y, et al. Structure aware negative sampling in knowledge graphs [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 6093-6101
[33] Sun Zhiqing, Deng Zhihong, Nie Jianyun, et al. RotatE: Knowledge graph embedding by relational rotation in complex space [C/OL] //Proc of the 7th Int Conf on Learning Representations. Amherst, MA: UMASS, 2019 [2022-03-10]. https://arxiv.org/abs/1902.10197
[34] Sutton R, McAllester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation [C] //Proc of the 14th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 1999: 1057-1063
[35] Kingma D, Ba J. Adam: A method for stochastic optimization [C/OL] //Proc of the 3rd Int Conf on Learning Representations. Amherst, MA: UMASS, 2015 [2022-03-10]. https://arxiv.org/abs/1412.6980
[36] Ma Xiaohan, Jin Rize, Sohn K, et al. Improving generative adversarial networks with adaptive control learning [C] //Proc of the 2018 IEEE Visual Communications and Image Processing. Piscataway, NJ: IEEE, 2018: 1-4
[37] Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library [C] //Proc of the 33rd Conf on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2019: 8024-8035

Guo Zhengshan, born in 1997. Master candidate. Student member of CCF. His main research interests include knowledge base and data mining.
郭正山, 1997 年生. 硕士研究生. CCF 学生会员. 主要研究方向为知识库和数据挖掘.

Zuo Jie, born in 1977. PhD, associate professor, master supervisor. His main research interests include database, data mining and big data processing.
左 劼, 1977 年生. 博士, 副教授, 硕士生导师. 主要研究方向为数据库、数据挖掘和大数据处理.

Duan Lei, born in 1981. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include bioinformatics, data mining, and database.
段 磊, 1981 年生. 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为生物信息学、数据挖掘和数据库.

Li Renhao, born in 1997. Master candidate. Student member of CCF. His main research interests include knowledge base and natural language inference.
李仁昊, 1997 年生. 硕士研究生. CCF 学生会员. 主要研究方向为知识库和自然语言推理.

He Chengxin, born in 1996. PhD candidate. Student member of CCF. His main research interests include data mining, network analysis and bioinformatics.
何承鑫, 1996 年生. 博士研究生. CCF 学生会员. 主要研究方向为数据挖掘、网络分析和生物信息学.

Xiao Yingjie, born in 2000. Bachelor. Student member of CCF. His main research interests include database and data mining.
肖英劼, 2000 年生. 学士. CCF 学生会员. 主要研究方向为数据库和数据挖掘.

Wang Peiyan, born in 1997. Master. Student member of CCF. Her main research interests include data mining and knowledge base.
王培妍, 1997 年生. 硕士. CCF 学生会员. 主要研究方向为数据挖掘和知识库.