SimKGC:使用简单对比知识图谱完成预训练语言模型

王亮1* 、赵伟2 、魏卓宇2 、刘景明2
1微软亚洲研究院
2猿辅导人工智能实验室,北京,中国
wangliang@microsoft.com
{zhaowei01,weizhuoyu,liujm}@yuanfudao.com

抽象的

知识图谱补全（KGC)旨在
推理已知事实并推断缺失的环节。基于文本的方法（如 KG-
BERT （Yao 等人， 2019）)从自然语言描述中学习实体表示，
并且

具有归纳 KGC 的潜力。然而,基于文本的方法的性能

仍然远远落后于基于图嵌入的
TransE 等方法（Bordes 等， 2013）
和 RotatE (Sun et al., 2019b)。在本文中,我们确定关键问题
是效率
对比学习。为了提高学习
效率,我们引入了三种类型的底片 :批次内底片、批次前底片、

以及作为一种简单形式的自我否定
硬否定。结合 InfoNCE
损失,我们提出的模型 SimKGC 在多个基准数据集上的表现明
显优于基于嵌入的方法。在

平均倒数排名（MRR）,我们前进
WN18RR 比最先进的车型高出 19%，
在 Wikidata5M 传导设置上 +6.8%,在 Wikidata5M 感应设
置上 +22%
设置。进行彻底的分析以
深入了解每个组件。我们的代码
可在https://github.com/上找到
intfloat/SimKGC。

1 简介

大规模知识图谱（KG)很重要
知识密集型应用的组件，
例如问答（Sun et al.， 2019a），
推荐系统（Huang 等人， 2018 年)和智能对话代理（Dinan 等人，
2019 年）
等。知识图谱通常由一组三元组(h, r,
t),其中h是主实体， r是关系,并且
t是尾部实体。流行的公共知识图谱包括
Freebase （Bollacker 等人， 2008）、 Wikidata （Vran-
deci ˇ c 和 Krötzsch ， 2014）、 YAGO （Suchanek 等人，
2007)、 ConceptNet (Speer et al., 2017)和 WordNet
(Miller, 1992)等。尽管它们很有用
在实践中,它们往往是不完整的。知识
图形完成（KGC)技术是必需的
用于自动构建和验证
知识图谱。

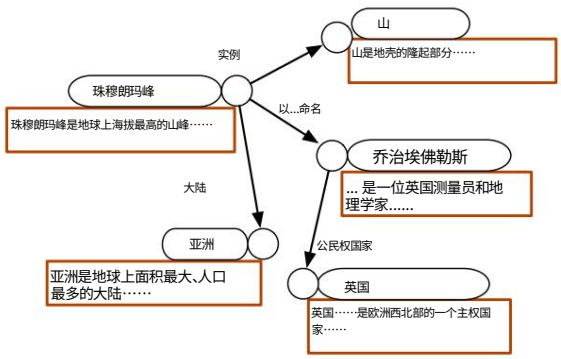

图 1:知识图谱示例。每个实体
有其名称和文字描述。

现有的 KGC 方法可以分为
两个系列:基于嵌入和基于文本
方法。基于嵌入的方法将每个
将实体和关系转化为低维向量，
无需使用任何辅助信息,例如实体
描述。该家族包括TransE (Bor- des et al., 2013)、 TransH
(Wang et al., 2014)、 Ro- tatE (Sun et al., 2019b)和
TuckER (Balaze-vic et al., 2019)等。相比之下,基于文本的

方法（Yao 等， 2019； Xie 等， 2016； Wang
等， 2021c)纳入实体的现有文本
表征学习,如图1所示。直观地看,基于文本的方法应
该优于
嵌入方法相比基于嵌入的方法,因为它们可以访问额
外的输入信号。然而,结果
在流行的基准测试（例如 WN18RR、FB15k- 237、
Wikidata5M)上,情况有所不同:基于文本的
即使使用预先训练的语言模型,方法仍然落后。

我们假设这种性能下降的关键问题是对比学习的效率低下。基
于嵌入的方法

不涉及文本编码的昂贵计算

在猿辅导人工智能实验室期间完成的工作。

编码器,因此在使用大量负样本量进行训练时效率极高。例如， 1 RotatE 的默认配置在Wikidata5M 数据集上训练了1000 个epoch,负样本量为64。而基于文本的方法KEPLER (Wang et al., 2021c)只能训练30 个epoch,负样本量为1,这是因为RoBERTa 的计算成本很高。

在本文中,受对比学习最新进展的启发,我们引入了三种类型的负样本来改进基于文本的 KGC 方法:批次内负样本、批次前负样本和自负样本。通过采用双编码器而不是交叉编码器(Yao et al., 2019)架构,可以使用更大的批次大小来增加批次内负样本的数量。来自先前批次的向量被缓存并充当批次前负样本(Karpukhin et al., 2020)。此外,挖掘困难负样本可能有利于改进对比学习。

我们发现,头部实体本身可以充当硬性负样本,我们称之为“自负样本”。因此,负样本量可以增加到数千个规模。我们还建议将损失函数从基于边距的排名损失更改为 InfoNCE,这可以使模型专注于硬性负样本。

基于文本的方法的一个优点是它们能够进行归纳实体表示学习。在训练期间未见过的实体仍然可以被适当地建模,而基于嵌入的方法（如 TransE)只能在

传导设置。由于每天都会出现新的[2]实体,因此归纳知识图谱补全在现实世界中非常重要。此外,基于文本的方法可以利用最先进的预训练语言模型来学习更好的表示。最近的一系列研究（Shin 等人， 2020 年； Petroni 等人， 2019 年)试图从 BERT 中引出隐式存储的知识。KGC 的任务也可以看作是检索此类知识的一种方式。

如果两个实体在图中通过一条短路径连接,则它们更有可能相关。从经验上讲,我们发现基于文本的模型严重依赖语义匹配,并在一定程度上忽略了这种拓扑偏差。我们提出了一种简单的重新排名策略,即提高头实体的k 跳邻居的得分。

我们通过以下方式评估我们提出的模型 SimKGC 在三个流行的基准测试集上进行实验:WN18RR、FB15k-237 和 Wikidata5M（传导和归纳设置）。根据自动评估指标（MRR、Hits@{1,3,10}） ,SimKGC在 WN18RR（MRR 47.6 → 66.6）、 Wikidata5M 传导设置（MRR 29.0 → 35.8)和归纳设置（MRR 49.3 → 71.4)上的表现都远远优于最先进的方法。在 FB15k-237 数据集上,我们的结果也很有竞争力。为了帮助更好地理解我们提出的方法,我们进行了一系列分析并报告了人工评估结果。希望 SimKGC 能够促进未来更好的 KGC 系统的开发。

2 相关工作

知识图谱补全涉及对多关系数据进行建模,以帮助自动构建大规模知识图谱。在基于翻译的方法（例如 TransE （Bordes 等人， 2013)和 TransH （Wang 等人， 2014）)中,三元组(h, r, t) 是从头实体h到尾实体t的关系特定翻译。Trouillon等人（2016)引入了复数嵌入来提高模型的表达能力。

RotatE (Sun et al., 2019b)将三元组建模为复空间中的关系旋转。Nickel等人(2011)； Balazevic 等人(2019)将 KGC 视为 3-D 二元张量分解问题,并研究了几种分解技术的有效性。一些方法尝试合并实体描述。DKRL (Xie et al., 2016)使用CNN 对文本进行编码,而 KG-BERT (Yao et al., 2019)、 StAR (Wang et al., 2021a)和 BLP (Daza et al., 2021)都采用预训练语言模型来计算实体嵌入。

GraIL (Teru 等人， 2020)和 BERTRL (Zha等人， 2021)利用子图或路径信息进行归纳关系预测。在基准测试性能方面(Wang 等人， 2021c),基于文本的方法仍然不如RotatE 等方法。

预训练语言模型包括 BERT （Devlin 等人， 2019 年）、 GPT （Radford 等人， 2018 年)和 T5 （Raffel 等人， 2019 年） ,它们已经导致 NLP 的学习范式转变。首先使用语言建模目标在大量未标记的文本语料库上对模型进行预训练,然后在下游任务上进行微调。考虑到它们在少样本甚至零样本中的良好表现

1https://github.com/DeepGraphLearning/ graphvite 2测试集中的所有实体也出现在训练集中。

场景（Brown 等人， 2020 年），一个有趣的问题是：“预训练的语言模型可以用作知识库吗？” Petroni 等人（2019 年）提出使用手动设计的提示来探测语言模型。一系列后续工作（Shin 等人， 2020 年； Zhong 等人， 2021 年； Jiang等人， 2020 年）专注于寻找更好的提示来引出隐含在模型参数中的知识。另一项工作（Zhang 等人， 2019 年； Liu 等人， 2020 年； Wang 等人， 2021c）将符号知识注入语言模型预训练中，并在几个知识密集型任务上显示出一些性能提升。

对比学习通过对比正例和负例来学习有用的表征（Le-Khac 等人， 2020 年）。正例和负例的定义是特定于任务的。在自监督视觉表征学习中（Chen 等人， 2020 年； He 等人， 2020 年； Grill 等人， 2020 年），正例对是同一图像的两个增强视图，而负例对是不同图像的两个增强视图。最近，对比学习范式在许多不同领域取得了巨大成功，包括多模态预训练（Radford et al., 2021）、视频文本检索（Liu et al., 2021）和自然语言理解（Gunel et al., 2021）等。在NLP 社区中，通过利用来自自然语言推理数据（Gao et al., 2021）、 QA 对（Ni et al., 2021）和平行语料库（Wang et al., 2021b) 的监督信号，这些方法在语义相似性基准上超越了非对比方法（Reimers and Gurevych, 2019) 。Karpukhin等人（2020）；Qu 等人（2021）； Xiong 等人（2021）采用对比学习来改进开放域问答的密集段落检索，其中正段落是包含正确答案的段落。

3 方法论

3.1 符号

知识图谱G是一个有向图，顶点为实体E，每条边可以表示为三元组(h,r,t)，其中h、 r 和t分别对应于头实体、关系和尾实体。KGC 的链接预测任务是在给定不完整G 的情况下推断缺失的三元组。在广泛采用的实体排名评估协议下，尾实体预测(h, r ,?) 需要对给定h和r 的所有实体进行排名，对于头实体也是如此预测（？， r， t）。在本文中，对于每个三元组(h,r,t)，我们添加一个逆三元组(t,r $^{-1}$,h)，其中是 r 的 r $^{-1}$ 逆关系。基于这样的改写后，我们只需要处理尾部实体预测问题（Malaviya 等人，2020 年）。

3.2 模型架构

我们提出的 SimKGC 模型采用双编码器架构。两个编码器分别初始化具有相同的预训练语言模型，但不共享参数。

给定一个三元组(h,r,t)，第一个编码器 BERThr用于计算头实体 h 的关系感知嵌入。我们首先将实体h和关系r的文本描述用中间的特殊符号 [SEP] 连接起来。应用 BERThr来获取最后一层隐藏状态。我们不直接使用第一个 token 的隐藏状态，而是使用均值池化，然后进行L2归一化来获取关系感知嵌入ehr，因为均值池化已被证明可以产生更好的句子嵌入（Gao et al., 2021; Reimers and Gurevych, 2019）。ehr是关系感知的，因为不同的关系会有不同的输入，从而有不同的嵌入，即使头实体是相同的。

类似地，第二个编码器 BERTt用于计算尾部实体t的L2 归一化嵌入et。BERTt的输入仅包含实体 t 的文本描述。

由于嵌入ehr和et都是L2归一化的，余弦相似度cos(ehr, et)只是两个嵌入之间的点积：

$$\cos(\text{ehr}, \text{et}) = \frac{\text{等}}{\text{埃雷特}} = \text{ehr} \cdot \text{et} \quad (1)$$

对于尾部实体预测（h， r，？ ），我们计算ehr与E中所有实体之间的余弦相似度，并预测得分最高的实体：

$$\underset{\text{量}}{\text{精子最大摄取}} \cos(\text{ehr}, \text{eti}), \text{ti} \in \text{E} \quad (2)$$

3.3 负采样在知识图谱补全中，训练数据仅由正三元组组成。给定一个正三元组(h, r, t)，“负采样”需要采样一个或多个负三元组来训练判别模型。大多数现有方法会随机破坏h或t ，然后过滤掉训练图G中出现的假负例。

不同三元组的负样本不共享,因此是独立的。基于嵌入的方法的典型负样本数量为 64 （Sun et al.， 2019b），基于文本的方法的典型负样本数量为 5 （Wang et al.， 2021a）。我们结合了三种类型的负样本来提高训练效率,而不会产生大量的计算和内存开销。

批次内负样本 (IB)这是视觉表征学习(Chen et al., 2020)和密集段落检索(Karpukhin et al., 2020)等中广泛采用的一种策略。同一批次内的实体可用作负样本。此类批次内负样本允许高效重用双编码器模型的实体嵌入。

批次前负样本 (PB)批次内负样本的缺点是负样本的数量与批次大小相关。批次前负样本(Lee et al., 2021)使用来自先前批次的实体嵌入。由于这些嵌入是使用较早版本的模型参数计算的,因此它们与批次内负样本不一致。通常仅使用1或2 个批次前样本。其他方法（如 MoCo (He et al., 2020))也可以提供更多负样本。我们将 MoCo 的研究留待将来研究。

自我负例（SN)除了增加负例的数量之外,挖掘硬负例（Gao et al.， 2021； Xiong et al.， 2021)对于改进对比表征学习也很重要。

对于尾部实体预测（h、 r、？ ）,基于文本的方法倾向于为头部实体 h 分配高分,这可能是由于文本重叠度高。为了缓解这个问题,我们提出了使用头部实体h作为硬负例的自负例。包括自负例可以使模型更少地依赖虚假文本匹配。

我们用NIB、 NPB和NSN来表示上述三类负样本。在训练过程中,可能会存在一些假负样本。例如,正确的实体恰好出现在同一个批次中的另一个三元组中。我们用二元掩码过滤掉这样的实体。将它们全部组合起来,负样本集合 N (h, r) 为：

$$\{t\,|t\in \mathrm{NIB}\cup \mathrm{NPB}\cup \mathrm{NSN}\,,\ (h,r,t)\in G\}/ \tag{3}$$

[3] 训练数据中未出现的假阴性将不会被过滤。

假设批次大小为1024,并且使用2 个预批次,则我们总共会有|NIB| = 1024 − 1、 |NPB| = 2 × 1024、 |NSN| = 1 和|N (h, r)| = 3072 个负数。

3.4 基于图的重新排序

知识图谱通常表现出空间局部性。
相距较远的实体与相距较近的实体相比,它们之间的关联性更强。基于文本的 KGC 方法擅长捕捉语义关联性,但可能无法完全捕捉这种归纳偏差。我们提出了一种基于图的简单重排序策略:如果根据训练集中的图， ti位于头实体 h 的k 跳邻居Ek(h)中,则将候选尾部实体ti 的得分提高α ⩾ 0：

$$\underset{\text{量}}{\text{精子最大摄取}}\ \cos(e_{hr}, e_{ti}) + \alpha(t_i \in E_k(h)) \tag{4}$$

3.5 训练和推理

在训练过程中,我们使用带有附加边距的InfoNCE 损失（Chen et al.， 2020； Yang et al.， 2019）：

$$L = -\log \frac{e^{(\phi(h,r,t)-\gamma)/\tau}}{e^{(\phi(h,r,t)-\gamma)/\tau} + \sum_{i=1}^{|\mathbb{N}|} e^{\phi(h,r,t_i')/\tau}} \tag{5}$$

附加边际γ > 0鼓励模型提高正确三元组(h,r,t) 的得分。ϕ (h,r,t)是候选三元组的得分函数,这里我们定义ϕ(h,r,t) = cos(ehr, et) ∈[−1,1],如公式1 所示。温度τ可以调整负样本的相对重要性,较小的τ使得损失更加侧重于难样本,但也存在过度拟合标签噪声的风险。为了避免将τ作为超参数进行调整,我们将log 重新参数化为可学习参数。

$\frac{1}{\tau}$ 作为

对于推理,最耗时的部分是O(|E|) BERT 实体嵌入的前向传递计算。假设有|T |个测试三元组。对于每个三元组(h, r, ?) 和(t, r) ,输入关系感知的头实体嵌入并使用点积来获取所有实体的排名分数。总的来说,SimKGC 需要|E| + 2 × |T | BERT $^{-1}$， ？ ）,我们需要前向传递,而像KG-BERT (Yao et al., 2019)这样的跨编码器模型需要|E| × 2 × |T |。

能够扩展到大型数据集对于实际使用非常重要。对于双编码器模型,我们可以预先计算实体嵌入,并借助Faiss 等快速相似性搜索工具高效地检索前 k 个实体（Johnson 等人， 2021 年）。

数据集	#实体	#关系	#训练	#有效的	#测试
WN18RR	40,943	11	86、835	3034	3134
FB15k-237 14, 541 237			272、115	17、535	20、466
Wikidata5M-Trans 4, 594, 485 822			20, 614, 279	5, 163	5, 163
维基数据5M-Ind 4, 579, 609 822			20, 496, 514	6, 699	6, 894

表 1:本文使用的数据集统计。“Wikidata5M-Trans”和“Wikidata5M-Ind”指的是
分别是传导和感应设置。

4 实验

4.1 实验装置

数据集我们使用三个数据集进行评估：
WN18RR、FB15k-237 和 Wikidata5M （王
et al., 2021c).统计数据见表
1. Bordes 等人（2013 年)提出了 WN18 和
FB15k 数据集。后期研究（Toutanova 等人，
2015； Dettmers 等人， 2018)表明,这两个
数据集遭受测试集泄漏并发布
WN18RR 和 FB15k-237 数据集,删除
逆关系。WN18RR 数据集包括
来自 WordNet的 41k 个同义词集和11 个关系
（Miller， 1992）， FB15k-237 数据集包括
来自 Freebase的 15k 个实体和237 个关系。
Wikidata5M 数据集的规模更大
包含约 500万个实体和约2000万个三元组。
它提供两种设置:传导和感应。
对于传导设置,测试中的所有实体
集合也出现在训练集中,而对于归纳设置,实体之间
没有重叠
训练和测试集。我们使用“Wikidata5M-Trans”和
“Wikidata5M-Ind” 来表示这两个设置。
对于文本描述,我们使用数据
由 KG-BERT （ Yao et al.， 2019)提供
WN18RR 和 FB15k-237 数据集。Wiki-data5M 数据集已经包含
以下描述
所有实体和关系。

评估指标根据之前的研究,我们的
所提出的 KGC 模型通过实体排名任务进行评估:对于每
个测试三元组(h, r, t),尾部实体
预测给定h对所有实体进行排序以预测t
和r,对于头部实体预测也类似。我们使用
四个自动评估指标:平均倒数排名（MRR)和 Hits@k(k
∈{1,3,10}) （H@k
简称）。MRR 是
所有测试三元组。H@k计算
正确实体排名在前 k 名。MRR和
H@k在过滤设置下报告（Bor- des 等， 2013）,过滤设置忽略
了
训练中所有已知真实三元组的得分,val-
idation 和测试集。所有指标均通过以下方式计算：
在两个方向上取平均:头实体预测和尾实体预测。

我们还对
Wikidata5M 数据集提供更准确的
模型性能的评估。

超参数编码器已初始化
使用 bert-base-uncased (英语)。使用更好的
预训练语言模型有望改善
进一步提高性能。大多数超参数除了
学习率和训练周期在
所有数据集,以避免针对特定数据集进行调整。我们
对学习率进行范围网格搜索
{10−5 , 3×10−5 ，5×10−5 }。实体描述是
截断为最多50 个标记。温度
τ初始化为0.05,并且附加边际
InfoNCE 损失为0.02。对于重新排序,我们设置
α = 0.05。使用2 个预批次和 logit 权重
0.5.我们使用 AdamW 优化器进行线性学习
衰减率。模型的训练批次大小为1024
在4 个V100 GPU 上。对于 WN18RR,FB15k-237，
和 Wikidata5M（两种设置)数据集,我们训练
分别为50、10和1 个时期。请参阅
附录A了解更多详细信息。

4.2 主要结果

我们重复使用了Wang 等人报告的数字。
（2021c)针对 TransE 和 DKRL,结果
RotatE 的测试数据来自 GraphVite 官方基准。在表2和表 3 [4]
中,我们提出的
SimKGCIB+PB+SN模型在 WN18RR 上的表现远超最先进的方
法，
Wikidata5M-Trans 和 Wikidata5M-Ind 数据集，
但在 FB15k-237 数据集上略有落后
（MRR 33.6% vs 35.8%）。据我们所知,SimKGC 是第一
个基于文本的 KGC 方法
比基于嵌入的
对应者。

4https://graphvite.io/docs/latest/
基准

方法	Wikidata5M-Trans				Wikidata5M-Ind			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
基于嵌入的方法								
TransE （Bordes 等人， 2013 年）	25.3	17.0	31.1	39.2	29.0	23.4	32.2	
RotatE (Sun et al., 2019b)基于文本	39.0				-	-	-	-
的方法								
DKRL （Xie 等， 2016）	16.0	12.0	18.1	22.9	23.1	32.0	54.6	
开普勒（Wang 等人， 2021c）	21.0	17.3	22.4	27.7	40.2	22.2	51.4	73.0
BLP-ComplEx （Daza 等人， 2021）	-	-	-	-	48.9	26.2	66.4	49.3 28.9 87.7
BLP-SimplE （Daza 等人， 2021）	-	-	-	-	63.9			86.6
SimKGCIB	35.3	30.1	37.4	44.8	60.3	39.5	77.8	92.3
SimKGCIB+PB	35.4	30.2	37.3	44.8	35.6	31.0	37.3	60.2 39.4 77.7 92.4
SimKGCIB+SN	43.9	35.8	31.3	37.6	44.1			71.3 60.7 78.7 91.3
SimKGCIB+PB+SN								71.4 60.9 78.5 91.7

表 2:Wikidata5M 数据集的主要结果。“IB”、“PB”和“SN”分别表示批次内负样本、批次前样本
负数和自负数。基于嵌入的方法本质上无法执行归纳
KGC。根据Wang 等人（2021c)的评估协议,归纳设置仅对 7,475 个实体进行排名
在测试集中,而传导设置对 460 万个实体进行排序,因此报告的归纳设置指标
设置要高得多。结果在配对学生 t 检验下具有统计学意义,p 值为 0.05。

方法	WN18RR				FB15k-237			
	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10
基于嵌入的方法								
TransE （Bordes 等人， 2013） † 24.3 4.3 44.1 DistMult （Yang 等人， 2015） † 44.4 41.2 47.0 RotatE （Sun 等人， 2019b） † 47.6 42.8 49.2 TuckER （Balazevic 等人， 2019） † 47 .0 44.3 48.2 基于文本的方法				53.2	27.9	19.8	37.6	44.1
				50.4	28.1	19.9	30.1	44.6
				57.1	33.8	24.1	37.5	53.3
				52.6	35.8	26.6	39.4	54.4
KG-BERT （Yao 等人， 2019）	21.6	4.1	30.2	33.1 20.3	52.4	-	-	- 42.0
MTL-KGC （Kim 等人， 2020 年）	38.3	40.1	24.3	49.1	59.7	26.7	17.2	29.8 45.8
StAR （Wang 等人， 2021a）					70.9	29.6	20.5	32.2 48.2
SimKGCIB	67.1	58.5	73.1	66.6 57.8	81.7	33.3	24.6	36.2 51.0
SimKGCIB+PB	72.3	66.7	58.8	72.1 66.6	81.7	33.4	24.6	36.5 51.1
SimKGCIB+SN	58.7	71.7			80.5	33.4	24.7	36.3 50.9
SimKGCIB+PB+SN					80.0	33.6	24.9	36.2 51.1

表 3:WN18RR 和 FB15k-237 数据集的主要结果。 † :数字来自Wang 等人（2021a）。

我们报告了各种组合的结果
底片。仅使用批内底片， SimKGCIB 的性能已经相当强大,这要归功于
我们使用较大的批量大小（1024）。添加自负值往往会改善 H@1,但会损害 H@10。
我们假设自我否定使得模型
较少依赖简单的文本匹配。因此,它们对强调召回率的指标有负面影响,例如
为 H@10。结合所有三种类型的负片
通常会有最好的结果,但并非总是如此。

与其他数据集相比，
FB15k-237 数据集更加密集（平均每个实体的度约为 37 ）,包含的实体较少（约为 15k）。为了表现良好,模型需要
学习可推广的推理规则,而不仅仅是
建模文本相关性。基于嵌入的
方法可能对此有优势
场景。我们的方法可以集成
与基于嵌入的类似,正如Wang 等人所做的那样。（2021a）。由于这不是本文的重点，
论文,我们将其留作未来的工作。此外， Cao 等人（2021)指出,FB15k-237 中的许多环节
数据集无法根据现有数据进行预测
信息。这两个原因有助于解释
SimKGC 的性能不令人满意。

添加自我否定对于
Wikidata5M 数据集的归纳设置,其中
MRR 从60.3%上升至71.3%。对于归纳
KGC,基于文本的模型更依赖于文本
匹配比传导设置。自负

可以防止模型简单地预测
给定的头部实体。
就推理时间而言，最昂贵的部分是 BERT 的前向传递。对于
Wikidata5M-Trans 数据集，SimKGC 需要 40
分钟计算 460万个嵌入
具有 2 个 GPU，而交叉编码器模型如
KG-BERT (Yao et al., 2019)预计需要3000小时。我们并不是第一个
能够进行快速推理的工作，例如
ConvE （Dettmers 等人， 2018)和 Star （Wang et al., 2021a)也具有类似优势。这里
我们只是想再次强调
设计时的推理效率和可扩展性
新模型。

5 分析

我们进行了一系列分析，以进一步了解
对我们提出的模型和 KGC 的见解
任务。

5.1 什么使得 SimKGC 如此出色?

与现有的基于文本的方法相比，
SimKGC 进行了两项重大更改：使用更多
负面因素，以及从基于边际的
排名损失到 InfoNCE 损失。指导未来
致力于知识图谱的补全，这至关重要
了解哪些因素对
SimKGC 的卓越性能。

损失	负 MRR 数	H@1 H@3 H@10	
信息NCE 255		64.4 53.8 71.7 48.8 31.9	82.8
InfoNCE 边		60.2 39.5 28.5 44.4	80.3
际边际-τ	5 255		61.2
	5	38.0 27.5 42.8	58.7
	255	57.8 48.5 63.7 74.9	

表 4：损失函数及样本数量分析
WN18RR 数据集上的负面结果。

在表4 中，我们使用 SimKGCIB ，批量大小
256作为基准。通过将负数从255减少到5，MRR 从64.4下降到48.8。
将损失函数从 InfoNCE 更改为
下列保证金损失导致 MRR 下降至 39.5：

$$\frac{1}{|\mathbb{N}|}\sum_{i=1}^{|\mathbb{N}|} \text{最大值}(0, \lambda + \phi(h, r, t'_i) - \phi(h, r, t)) \quad (6)$$

与公式5 一致， $\phi(h, r, t'_i)$是余弦
候选三元组的相似度得分，且λ = 0.8。
总而言之，InfoNCE 损失和较大的
底片的数量是重要因素，而
损失函数似乎有更大的影响。对于
InfoNCE 损失，困难负样本自然会贡献更大的梯度，并且添加更多的负样本
可以带来更稳健的表征。Wang和
Liu (2021)也得出了类似的结论：
硬度感知特性对于成功至关重要
对比损失。
我们还提出了一种变体“margin-τ ”损失，其方法是
将公式6中的权重从$\exp(s(t'_i)/\tau)/\sum_{j=1}^{|\mathbb{N}|}\exp(s(t'_j)/\tau)$
到 $\frac{1}{|\mathbb{N}|}$，其中$s(t'_i) = \text{最大值}(0, \lambda + \phi(h, r, t'_i) - \phi(h, r, t))$且τ = 0.05。类似于
InfoNCE 损失，“margin-τ ”损失使得模型
更加关注负面因素，从而导致
如表4所示，性能更好。
类似于“自我对抗负采样”
由Sun et al. (2019b)提出。大多数超参数都是基于
InfoNCE 损失进行调整的。我们期望
边际τ损失，以获得更好的结果
更多的超参数优化。

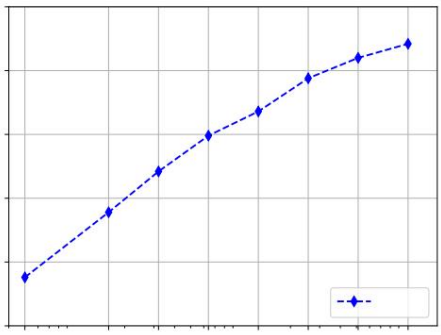

图 2：使用 SimKGCIB 的 WN18RR 数据集上的 MRR 与负样本数量的关系。我们使用批量大小
1024 的所有实验，并改变数量
使用 softmax logits 上的二进制掩码对负数进行处理。

在图2 中，我们定量地说明了
随着更多负面因素的增加，MRR 也会发生变化。
很明显的趋势是，表现从48.8稳步提高到67.1。然而，增加更多
底片需要更多的 GPU 内存，并且可能
造成优化困难(You et al.， 2020；
Chen 等人， 2020 年）。我们没有尝试
批次大小大于 1024。

5.2 重新排序的消融

我们提出的重新排序策略是一种简单的方法
将拓扑信息纳入知识图谱中。对于连通性模式
表现出空间局部性，重新排序可能会有所帮助。

三重证据……位于纽约州SimKGC 大理石城	（Rest Plaus 历史街区,位于纽约）阿尔斯特县马布尔敦的国家历史街区……
三重证据威廉·道格拉斯·格思里（William Douglas Guthrie,1967 年 1 月 17 日出生于密苏里州圣路易斯)是一名职业拳击手。...SimKGC 威廉·道格拉斯·格思里	（蒂莫西·P·格林,出生地,圣路易斯）
三重证据...机构用来处理SimKGC http 服务器	（TLS 终止代理、网络软件实例）传入 TLS 连接的代理服务器...
三次（1997 年 IBF 世界锦标赛,随后是1999 年 IBF 世界锦标赛）证据SimKGC 2000 IBF 世界青少年锦标赛	第十届国际羽联世界锦标赛（羽毛球)在苏格兰格拉斯哥举行，1997年5月24日至6月1日之间。

表 5:Wikidata5M-Trans 数据集测试集上的 SimKGC 预测结果示例。
预测以粗体显示。由于篇幅原因,我们只在“证据”行中显示一小段相关文本。

	MRR	H@1	H@3	H@10
重新排序	35.8	31.3	37.6	44.1
不重新排名	35.5	31.0	37.3	43.9

表 6:Wikidata5M-Trans 数据集上的重新排序的消融。

数据集	1-1	1-n	n-1	nn
Wikidata5M-Trans	30.4	8.3	71.1	10.6
维基数据 5M-工业	83.5	71.1	80.0	54.7

表 8:不同关系类型的 MRR
带有SimKGCIB+PB+SN的 Wikidata5M 数据集。

在表6 中,我们看到
Wikidata5M-Trans 数据集上的所有指标。注意
这种重新排序策略不适用于归纳KGC,因为测试集中的
实体从未出现
在训练数据中。探索更有效的方法
例如图神经网络（Wu et al.， 2019）
而不是简单的重新排名将是未来的发展方向。

5.3 细粒度分析

1-1	1-n
配偶	孩子
的首都	有部分
湖泊流入量	杰出的作品
政府首脑	副作用
n-1	**nn**
实例	演员
出生地	成员
给定的名称	受...影响
工作地点	提名

表 7:不同关系类别的示例
在 Wikidata5M-Trans 数据集上。

我们将所有关系分为四类
基于头尾参数的基数
遵循Bordes 等人（2013 年)的规则:一对一（1-1）、
一对多（1-n）、多对一（n-1），
以及多对多（nn）。示例如下

表7. 如表 8 所示,预测
“n” 侧通常更难,因为有
许多看似合理的答案会使模型混乱。另一个主要原因是

知识图谱的完整性。一些预测
根据人工评估,三元组可能是正确的，
尤其是对于主实体预测中的 1-n 关系,例如“实例”、“出生地”等。

在表5 中,对于第一个例子“Marbletown”，
“阿尔斯特县”和“纽约”都是正确答案。第二个例子
说明了
出生地关系案例:很多人
出生地相同,有些三胞胎可能
知识图谱中不存在。这有助于解释“1-n”关系在

Wikidata5M-Trans 数据集。在第三个例子中，
SimKGC 预测的结果与此密切相关,但并不准确
实体“http 服务器”。

5.4 人工评估

上述分析表明， MRR 等自动评估指标往往会低估

模型的性能。为了获得更准确的
为了评估绩效,我们进行人力
评估并将结果列于表9。根据

根据人类注释者的说法,H@1 是正确的。如果
考虑到这一点,我们提出的 H@1
模型会高得多。如何准确地

	正确	错误	未知
（左、右、左）	24%	54%	22%
（？ ,r,t）	74%	14%	12%
平均	49%	34%	17%

表 9:Wikidata5M-Trans 数据集上的人工评估结果。(h, r, ?) 和 (?, r, t) 表示尾部实体
和头部实体预测。我们随机
根据 H@1 抽样 100 个错误预测
测试集。“未知”类别表示注释者
无法确定预测是否正确
或者根据文本信息判断是错误的。

衡量KGC系统的性能也是
一个有趣的未来研究方向。

5.5 实体可视化

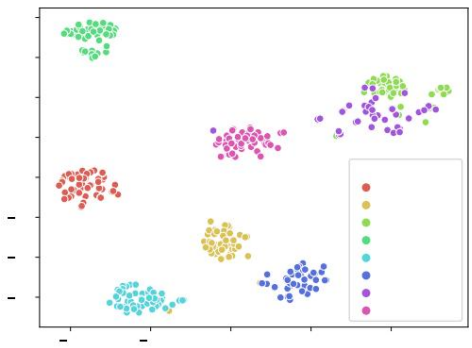

图 3:使用 t-SNE 对 Wikidata5M-Trans 数据集中的实体嵌入进行二维可视化
（Maaten 和 Hinton， 2008 年）。

为了定性地检验我们提出的模型，
我们对8 个最大实体的嵌入进行可视化
每个类别随机选择50 个实体
类别。实体嵌入是通过以下方式计算的
BERTt在第 3.2 节中。在图3 中，不同类别很好地分开，表明
学习到的嵌入的质量。一个有趣的
现象是“社区”和“村庄”两个类别有一些重叠。这是

这是合理的，因为这两个概念并不互相排斥。

6 结论

本文提出了一种简单的方法 SimKGC 来
改进基于文本的知识图谱完成。
我们认为关键问题是如何执行
高效的对比学习。利用对比学习领域的最新进展，

SimKGC 采用双编码器架构，结合了三种类型的负样本。
WN18RR、FB15k-237 和 Wikidata5M 数据集
表明 SimKGC 的表现远远优于最先进的方法。

对于未来的工作，一个方向是改进
SimKGC 的可解释性。在 RotatE (Sun et al., 2019b)和
TransE (Bordes et al.,
2013），三元组可以建模为复空间中的旋转或关系平移，而 SimKGC
无法进行如此易于理解的解释。另一个方向是探索有效的

处理假阴性的方法（Huynh 等人，
2020)是由于知识图谱的不完整性而导致的。

7 更广泛的影响

未来的工作可以使用 SimKGC 作为坚实的基线，以继续改进基于文本的知识图谱
完井系统。我们的实验结果和
分析还揭示了一些有前途的研究方向。例如，如何将全球

以更原则的方式理解图结构？有没有
有没有比 InfoNCE 损失表现更好的其他损失函数？
对于知识密集型任务，如
作为知识库问答（KBQA），
信息检索、基于知识的响应生成等，

探索改进的知识图谱完成系统带来的新机遇。

致谢

我们要感谢匿名评论者和
区域主席的宝贵意见，以及 ACL
滚动审查组织者的努力。

参考

Ivana Balazevic、Carl Allen 和 Timothy Hospedales。
2019. TuckER：知识的张量分解
图形完成。 2019 年自然语言实证方法会议论文集

处理和第 9 届国际自然语言处理联合会议 (EMNLP-IJCNLP)，
第 5185-5194 页，香港，中国。计算语言学协会。

库尔特·博拉克、科林·埃文斯、普拉文·帕里托什、蒂姆
Sturge 和 Jamie Taylor。2008 年。Freebase：一个用于构建
人类知识的协作创建的图形数据库。2008 年 ACM 会议论文集

5我们利用“实例”关系来确定实体
类别。

SIGMOD国际管理会议
数据,第 1247-1250 页。

安托万·博德斯、尼古拉斯·乌苏尼尔、阿尔贝托·加西亚-杜兰、杰森·韦斯顿和奥克萨娜·雅赫年科。
2013.翻译嵌入以建模多关系数据。神经信息进展

处理系统 26:第 27 届年度会议
神经信息处理系统 2013。2013 年 12 月 5 日至 8 日举行的会议论文集,
美国内华达州太浩湖,第 2787-88 页
2795.

汤姆·B·布朗、本杰明·曼、尼克·莱德、梅兰妮
苏比亚、贾里德·卡普兰、普拉芙拉·达里瓦尔、阿尔文德
尼拉坎坦、普拉纳夫·希亚姆、吉里什·萨斯特里、阿曼达
Askell、Sandhini Agarwal、Ariel Herbert-Voss、
格雷琴·克鲁格、汤姆·海尼汉、雷文·查尔德、
Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、
克莱门斯·温特、克里斯托弗·赫塞、马克·陈、
埃里克·西格勒、马特乌斯·利特文、斯科特·格雷、本杰明
Chess、Jack Clark、Christopher Berner、Sam Mc-
Candlish、Alec Radford、Ilya Sutskever 和 Dario
Amodei。2020.语言模型是小样本学习者。神经信息处理进展

Systems 33:2020 年神经信息处理系统年度会议,NeurIPS 2020,
2020 年 12 月 6 日至 12 日,虚拟。

曹益鑫、吉翔、吕鑫、李娟子、温永刚、
和张汉旺。2021年。缺失环节是可预测的吗?知识的推理基准

图形完成。在计算数学协会第 59 届年会论文集

语言学与第 11 届国际自然语言处理联合会议(第 1 卷:

长篇论文 (Long Papers),第 6855–6865 页,在线。计算语言学协会。

陈婷、Simon Kornblith、Mohammad Norouzi、
和 Geoffrey E. Hinton。2020 年。一个简单的框架
用于视觉表征的对比学习。在
第 37 届国际会议论文集
机器学习,ICML 2020,2020 年 7 月 13 日至 18 日,
虚拟事件,机器学习研究文集中第 119 卷,第 1597-1607 页。PMLR。

丹尼尔·达萨、迈克尔·科切斯和保罗·格罗斯。 2021 年。
通过链接从文本中归纳实体表示
预测。在网络会议论文集
2021 年,第 798-808 页。

蒂姆·戴特默斯、帕斯夸莱·米勒维尼、庞图斯·斯特内托普、
和 Sebastian Riedel。2018 年。卷积 2d
知识图谱嵌入。在诉讼程序中
第三十二届 AAAI 人工智能大会
智能(AAAI-18)、第30届人工智能创新应用(IAAI-18)和

第八届AAAI教育进步研讨会
新奥尔良人工智能(EAAI-18)
美国路易斯安那州,2018 年 2 月 2 日至 7 日,第 1811 页
1818.AAAI 出版社。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和
Kristina Toutanova。2019 年。BERT:
用于语言理解的深度双向转换器。 2019 年会议论文集

协会北美分会
计算语言学:人类语言
技术,第 1 卷(长篇和短篇论文),
第 4171–4186 页,明尼苏达州明尼阿波利斯。计算语言学协会。

艾米丽·迪南、斯蒂芬·罗勒、库尔特·舒斯特、安吉拉
范、迈克尔·奥利和杰森·韦斯顿。2019 年。巫师
维基百科:知识驱动的对话
代理。在路易斯安那州新奥尔良举行的第七届国际学习表征会议 ICLR
2019 上,
美国,2019 年 5 月 6 日至 9 日。OpenReview.net。

高天宇、姚兴成、陈丹琪。 2021 年。
Simcse:句子嵌入的简单对比学习。 ArXiv 预印本,abs/2104.08821。

Jean-Bastien Grill、Florian Strub、Florent Altché、
科伦坦·塔莱克、皮埃尔·H·里奇蒙德、埃琳娜
布哈茨卡娅、卡尔·多尔施、贝尔纳多·阿维拉·皮雷斯、
郭兆汉、Mohammad Gheshlaghi Azar、Bilal
Piot、Koray Kavukcuoglu、Rémi Munos 和 Michal
Valko。2020.发挥自己的潜力 - 一个新的
自我监督学习的方法。进展
神经信息处理系统 33:神经信息处理年度会议

Systems 2020、NeurIPS 2020,2020 年 12 月 6 日至 12 日,
虚拟的。

Beliz Gunel、杜京飞、Alexis Conneau 和 Veselin
Stoyanov。2021 年。监督对比学习
预训练语言模型微调。在第九届国际学习表征会议上,

ICLR 2021,虚拟活动,奥地利,2021 年 5 月 3 日至 7 日。
OpenReview.net。

何凯明、范浩琪、吴雨馨、谢赛宁、
罗斯·B·吉尔希克。2020 年。无监督视觉表征学习的动量对比。 2020
年
IEEE/CVF 计算机视觉与模式识别会议,CVPR 2020,美国华盛顿州西
雅图,
2020 年 6 月 13 日至 19 日,第 9726–9735 页。IEEE。

黄金、赵鑫、窦宏建、继荣
Wen 和 Edward Y. Chang。2018 年。通过知识增强改进序列推荐

记忆网络。在第 41 届国际 ACM 会议上
SIGIR 研究与开发会议
信息检索,SIGIR 2018,密歇根州安娜堡,
美国,2018 年 7 月 8 日至 12 日,第 505–514 页。ACM。

Tri Huynh、Simon Kornblith、Matthew R. Wal-ter、Michael
Maire 和 Maryam Khademi。 2020.
促进对比自监督学习
假阴性消除。 ArXiv 预印本,
绝对值/2011.11765。

蒋正宝、Frank F. Xu、荒木淳和 Graham
Neubig。2020.我们如何知道什么语言
模特知道吗?协会会刊
计算语言学,8:423-438。

Jeff Johnson、M. Douze 和 H. Jégou。2021 年。使用 GPU 进行十亿级相似性搜索。IEEE 大数据汇刊,7:535–547。

弗拉基米尔·卡普欣、巴拉斯·奥古兹、Sewon Min、帕特里克 Lewis、Ledell Wu、Sergey Edunov、Danqi Chen 和 Wen-tau Yih。2020.密集通道检索开放领域问答。在诉讼程序中 2020 年自然语言处理经验方法会议 (EMNLP),第 67-69 页

6781,在线。计算语言学协会。

Bosung Kim、Taesuk Hong、Youngjoong Ko 和 Jungyun Seo。2020.使用预训练语言进行知识图谱补全的多任务学习模型。在第 28 届国际会议论文集 计算语言学会议,页数 1737–1743 年,西班牙巴塞罗那(在线)。国际计算语言学委员会。

Phuc H Le-Khac、Graham Healy 和 Alan F Smeaton。2020. 对比表征学习:框架与回顾。IEEE Access。

Jinhyuk Lee、Mujeen Sung、Jaewoo Kang 和 Danqi 陈。2021.学习密集表示按比例的短语。在第 59 届年度计算语言学协会会议和第11届国际联合会议

自然语言处理 (第 1 卷:长论文),第 6634–6647 页,在线。计算语言学。

刘松、范浩琪、钱胜胜、亦如陈,丁,王中元。 2021 年。成功案例:具有动量的分层变压器用于视频文本检索的对比。 ArXiv 预印本,绝对值/2103.15049。

刘伟杰、周鹏、赵哲、王志若、鞠琪、邓浩堂、王平。2020 年。K -BERT:利用知识图谱实现语言表示。在第三十四届 AAAI 大会上

人工智能会议,AAAI 2020,第三十二届人工智能创新应用会议,IAAI 2020,第十届 AAAI 人工智能教育进步研讨会,EAAI 2020,美国纽约州纽约,2 月

7-12,2020 年,第 2901-2908 页。AAAI 出版社。

LVD Maaten 和 Geoffrey E. Hinton。2008 年。使用 t-sne 可视化数据。机器学习杂志研究,9:2579-2605。

Chaitanya Malaviya、Chandra Bhagavatula、安托万 Bosselut 和 Yejin Choi。2020 年。Commonsense 具有结构和语义上下文的知识库完成。在 AAAI 中。

George A. Miller。1992 年。WordNet :词汇数据库英语。演讲与自然语言:新奥尔良哈里曼大学研讨会论文集

约克,1992 年 2 月 23 日至 26 日。

倪建模、Noah Constant、马季、Keith B Hall、丹尼尔 Cer,杨银飞,等。 2021.句子-t5:可扩展来自预先训练的文本到文本模型的句子编码器。 ArXiv 预印本,abs/2108.08877。

马克西米利安·尼克尔、沃尔克·特雷斯普和汉斯·彼得 Kriegel。2011年。集体多关系数据的学习。诉讼中 第 28 届国际机器视觉大会学习,ICML 2011,美国华盛顿州贝尔维尤,2011 年 6 月 28 日至 7 月 2 日,第 809–816 页。Omnipress。

法比奥·彼得罗尼、蒂姆·洛克塔舍尔、塞巴斯蒂安·里德尔、帕特里克·刘易斯 (Patrick Lewis)、安东·巴赫金 (Anton Bakhtin)、吴宇翔 (Yuxiang Wu) 和亚历山大·米勒。2019 年。语言模型作为知识库? 2019 年自然语言实证方法会议论文集

处理和第 9 届国际自然语言处理联合会议 (EMNLP-IJCNLP),第 2463-2473 页,香港,中国。计算语言学协会。

曲英奇、丁雨辰、刘静、刘凯、瑞阳任、赵鑫、董大向、吴华、和王海峰。2021. RocketQA:一种针对密集段落检索的优化训练方法用于开放域问答。在 2021 年北美计算数学协会会议论文集上

语言学:人类语言技术,页 5835–5847,在线。计算科学协会语言学。

亚历克·拉德福德 (Alec Radford)、金钟旭 (Jong Wook Kim)、克里斯·哈拉西 (Chris Hallacy)、阿迪亚 (Aditya) Ramesh、Gabriel Goh、Sandhini Agarwal、Girish 萨斯特里、阿曼达·阿斯凯尔、帕梅拉·米什金、杰克·克拉克、Gretchen Krueger 和 Ilya Sutskever。2021 年。从自然语言监督中学习可转移的视觉模型。在第 38 届国际机器学习会议 ICML 论文集上

2021 年,2021 年 7 月 18 日至 24 日,虚拟活动,第 139 卷机器学习研究论文集,页数 8748–8763。PMLR。

亚历克·雷德福、卡蒂克·纳拉辛汉、蒂姆·萨利曼斯和 Ilya Sutskever。2018 年。通过生成式预训练提高语言理解能力。

科林·拉斐尔、诺姆·沙泽尔、亚当·罗伯茨、凯瑟琳·李、莎兰·纳朗、迈克尔·马特纳、Yanqi Zhou、W. Li 和 Peter J. Liu。2019 年。探索迁移学习的极限统一的文本到文本转换器。 ArXiv 预印本,绝对值/1910.10683。

Nils Reimers 和 Iryna Gurevych。2019 年。Sentence -BERT:使用暹罗 BERT网络的句子嵌入。 2019 年会议论文集

自然语言处理中的经验方法以及第九届国际自然语言处理联合会议(EMNLP-IJCNLP),第 3982–3992,香港,中国。计算语言学。

泰勒·辛 (Taylor Shin)、亚萨曼·拉泽吉 (Yasaman Razeghi)、罗伯特·洛根四世 (Robert L. Logan IV)、Eric Wallace 和 Sameer Singh。2020 年。自动提示:

从语言模型中获取知识
自动生成的提示。 2020 年实证方法会议论文集

自然语言处理 (EMNLP) 页
4222–4235,在线。计算数学协会
语言学。

Robyn Speer、Joshua Chin 和 Catherine Havasi。2017 年。
Conceptnet 5.5:一个开放的多语言通用知识图谱。在第三十一届会
议的议事录中
AAAI 人工智能会议,2017 年 2 月 4 日至 9 日,美国加利福尼亚州旧
金山,
第 4444–4451 页。AAAI 出版社。

Fabian M. Suchanek、Gjergji Kasneci 和 Gerhard
Weikum。2007.Yago :语义知识的核心。
第 16 届国际会议论文集
在万维网上,WWW 2007,班夫,艾伯塔省,
加拿大,2007 年 5 月 8-12 日,第 697-706 页。ACM。

Haitian Sun、Tania Bedrax-Weiss 和 William Cohen。
2019a. PullNet:开放领域问答
对知识库和文本进行迭代检索。
2019 年实证研究会议论文集
自然语言处理方法与
第九届国际自然语言处理联合会议（EMNLP-IJCNLP）,第
2380-2381 页
2390,香港,中国。计算语言学协会。

孙志清、邓志宏、聂建云、Jian
Tang. 2019b.旋转 :知识图谱嵌入
通过复空间中的关系旋转。在第七届国际学习表征会议上,

ICLR 2019,美国路易斯安那州新奥尔良,2019 年 5 月 6-9 日。
OpenReview.net。

Komal Teru、Etienne Denis 和 Will Hamilton。2020 年。
通过子图推理进行归纳关系预测。
第 37 届国际会议论文集
机器学习,ICML 2020,2020 年 7 月 13 日至 18 日,
虚拟事件,机器学习研究文集中第 119 卷,第 9448-9457 页。PMLR。

Kristina Toutanova、Danqi Chen、Patrick Pantel、Hoi-fung
Poon、Pallavi Choudhury 和 Michael Gamon。
2015.表示文本以进行文本联合嵌入
和知识库。在 2015 年会议记录中
自然语言处理经验方法会议,第 1499-1509 页,葡萄牙里斯本。计算
语言学协会。

泰奥·特鲁永、约翰内斯·韦尔布尔、塞巴斯蒂安·里德尔、埃里克
Gaussier 和 Guillaume Bouchard。2016 年。用于简单链接预测的
复杂嵌入。在
第 33 届国际会议论文集
机器学习,ICML 2016,纽约,
美国纽约,2016 年 6 月 19-24 日,JMLR 第 48 卷
研讨会和会议论文集,第 2071-2072 页
2080.JMLR.org。

Denny Vrandeci ˇ c 和 Markus Krötzsch。2014 年。Wiki-data :免费
协作知识库。ACM 通讯,57(10):78–85。

王博、沉涛、龙国栋、周天一、
Ying Wang 和 Yi Chang。2021a。结构增强文本表示学习以实现高
效
知识图谱补全。在
2021 年网络会议,第 1737-1748 页。

王峰和刘华平。2021 年。了解
对比损失的行为。2021 年 IEEE/CVF 计算机视觉和模式识别会议

（CVPR）,第 2495-2504 页。

王亮、赵伟、刘敬明。 2021b。
将跨语言句子表征与
双重动量对比。在 EMNLP 中。

王晓志、高天宇、朱兆成、正艳
张,刘志远,李娟子,唐健。
2021c. 开普勒 :知识的统一模型
嵌入和预训练的语言表示。
计算数学协会会刊
语言学,9:176-194。

王震、张建文、冯建林、郑
Chen. 2014.通过在超平面上平移实现知识图谱嵌入。在第二十八
届 AAAI 人工智能大会论文集上,

2014 年 7 月 27 日至 31 日,加拿大魁北克省魁北克市,
第 1112–1119 页。AAAI 出版社。

吴宗翰、潘诗睿、陈凤文、国栋
Long, C. Zhang 和 Philip S. Yu。2019 年。图神经网络综合调查。
IEEE
神经网络和学习系统学报,32:4-24。

谢若冰、刘志远、佳佳、栾焕波等
孙茂松. 2016.表征学习
具有实体描述的知识图谱。在 2016 年 2 月 12 日至 17 日于凤凰城
举行的第三十届 AAAI 人工智能会议论文集上,

美国亚利桑那州,第 2659-2665 页。 AAAI出版社。

李雄、陈彦雄、叶莉、邓国锋、
Jialin Liu、Paul N. Bennett、Junaid Ahmed 和
阿诺德·奥弗维克 (Arnold Overwijk)。2021 年。用于密集文本检索
的近似最近邻负对比学习。第九届国际学习会议

代表,ICLR 2021,虚拟活动,奥地利,
2021 年 5 月 3 日至 7 日。OpenReview.net。

杨碧山、叶文涛、何晓东、剑峰
Gao 和 Li Deng。2015 年。嵌入实体和
知识学习与推理的关系
基地。第三届国际学习会议
代表,ICLR 2015,美国加利福尼亚州圣地亚哥,
2015 年 5 月 7 日至 9 日,会议论文集。

杨寅飞、古斯塔沃·埃尔南德斯·阿夫雷戈、史蒂夫·袁、
Mandy 郭、沉沁兰、Daniel Cer、云轩
Sung, Brian Strope 和 Ray Kurzweil。2019 年。使用带有附加边
距 soft- max 的双向双编码器改进多语言句子嵌入。在第二十八届
国际人工智能联合会议 IJ-CAI 2019 论文集（中国澳门,2019 年 8
月 10 日至 16 日)中,第

5370–5378。ijcai.org。

姚亮、毛成胜、罗远。2019.Kg - bert:用于知识图谱补全的 Bert。论文集

预印本,abs/1909.03193。

杨游、李静、Sashank J. Reddi、Jonathan Hseu、
Sanjiv Kumar、Srinadh Bhojanapalli、宋晓丹、
詹姆斯·戴梅尔、库尔特·科策和谢卓瑞。
2020.深度学习的大批量优化:76 分钟内训练 BERT。在第八届国际
学习表征会议上,

ICLR 2020,埃塞俄比亚亚的斯亚贝巴,4 月 26-30 日
2020 年。OpenReview.net。

查瀚文、陈志宇、严西峰。 2021. bert 的归纳关系预测。 ArXiv 预印本,

绝对值/2103.07102。

张正彦、韩旭、刘志远、蒋欣、
孙茂松、刘群。2019. ERNIE:通过信息实体增强语言表示。在第 57 届
计算语言学协会年会论文集,第 1441-1451 页,意大利佛罗伦萨。协
会

计算语言学。

钟泽轩、丹·弗里德曼、陈丹琪。 2021 年。
事实探究是[MASK]:学习与学习
回想一下。在 2021 年会议论文集
北美分会
计算语言学:人类语言技术,第 5017-5033 页,在线。

计算语言学。

超参数的详细信息

GPU 的超参数数量	价值
	4
初始温度 τ 梯度剪辑 预热	0.05
步骤 批量大小	10
最大 token 数量	400
	1024
	50
重新排序的权重 α 0.05	
辍学 0.1	
权重衰减	10−4
InfoNCE 边缘池化	0.02
	意思是

表 10:我们提出的共享超参数
SimKGC 模型。

在表10 中,我们展示了
在所有数据集中共享。对于学习
率,我们使用5 × 10−5 ，10−5，和3 × 10−5
WN18RR、FB15k-237 和 Wikidata5M 数据集,
对于重新排序,我们对 WN18RR 使用 5 跳邻居,对其他使用 2 跳
邻居
数据集。每个时期需要大约 3分钟
WN18RR， FB15k-237约 12分钟,约 12
小时（两种设置）。我们的实现基于开源项目transformers −1
6
。

对于逆关系r ，我们添加一个前缀词
与 r 的描述“相反” 。例如,如果r
= “实例”,则 r $^{-1}$ = “逆实例”。

WN18RR 和 FB15k-237 中的一些实体
数据集的文本描述非常简短。我们
将它们与训练集中邻居的实体名称连接起来。为了避免训练期间的
标签泄漏,我们动态地排除正确的

输入文本中的实体。

B 更多分析结果

批量大小 MRR H@1 H@3 H@10	
	33.8 28.7 35.8 43.1
256512	34.6 29.4 36.7 43.7
1024	35.3 30.1 37.4 44.8

表 11:批量大小对使用 SimKGCIB 的 Wikidata5M-Trans 数据集的影
响。

批量大小 MRR H@1 H@3 H@10	
256	32.4 23.3 35.4 50.9
512	32.7 23.7 35.6 51.0
1024	33.3 24.6 36.2 51.0

表 12:批次大小对 FB15k-237 的影响
使用 SimKGCIB 的数据集。

裕度γ MRR H@1 H@3 H@10		
	33.4 24.8 36.0 33.6 24.9	50.9
0	36.2 33.6 25.0 36.2	51.1
0.02 0.05		50.9

表 13:FB15k-237 数据集上 In-foNCE 损失的附加边际 γ 的消融。

在表11和表 12 中,我们展示了批次
尺寸会影响 Wikidata5M- Trans 和 FB15k-237 数据集上的模型性能。

在公式5 中,我们使用 InfoNCE 损失的变体
具有附加边际γ。在我们的实验中，
此类变体的表现始终优于
标准 InfoNCE 损失,尽管改进
相当边际,如表 13 所示。

在表14 中,我们展示了更多示例
SimKGC 对 Wikidata5M-Trans 的预测

6https://github.com/huggingface/
变压器

三倍	（俘虏国家（电影）、电影实例）
证据	《俘虏国度》是一部 2019 年美国犯罪科幻惊悚片，由鲁伯特·瓦耶特执导 并由怀亚特 (Wyatt) 和埃里卡·比尼 (Erica Beeney) 共同撰写。...
SimKGC 3-D	电影
三倍	（莱昂内尔·贝拉斯科，职业，作曲家）
证据	莱昂内尔·贝拉斯科（Lionel Belasco，1881 年 - 1967 年 6 月 24 日）是一位杰出的钢琴家、作曲家和乐队指挥， 以他的卡里普索音乐录音而闻名。
SimKGC 乐队领队	
三重（Johan Nordhagen，国籍挪威）	
证据	瓦卡斯·艾哈迈德（Waqas Ahmed，1991 年 6 月 9 日出生）是一名挪威板球运动员。...
西姆·KGC·瓦卡斯·艾哈迈德	
三人（卡洛斯·佩尼亚·罗慕洛，担任菲律宾常驻专员）	
证据	弗朗西斯·伯顿·哈里森是一位出生于美国的菲律宾政治家，曾在美国任职 众议院议员，并被任命为菲律宾总督……
SimKGC 弗朗西斯·伯顿·哈里森	

表 14：Wikidata5M-Trans 测试集上的 SimKGC 预测结果更多示例。

数据集，以帮助更好地理解我们的模型的行为。测试
数据集上的完整模型预测是
可在我们的公共代码库中使用。