[知识图谱中的结构感知负采样 2](#_Toc46720)

[Abstract摘要 2](#_Toc46721)

[1 Introduction简介 2](#_Toc46722)

[2 Related Work相关工作 3](#_Toc46723)

[3 Structure Aware Negative Sampling结构感知负采样 4](#_Toc46724)

[3.1 Variants of SANS SANS算法变体 5](#_Toc46725)

[4 Experiments实验 5](#_Toc46726)

[4.1 Results结果 6](#_Toc46727)

[5 Conclusion and Future Directions结论与未来方向 8](#_Toc46728)

[Acknowledgments致谢 8](#_Toc46729)

[A Experimental Settings实验设置 8](#_Toc46730)

[A.1 Datasets数据集 8](#_Toc46731)

[A.2 Evaluation Protocols评估标准 9](#_Toc46732)

[B Implementation Details实现细节 9](#_Toc46733)

[B.1 Hyperparameters超参数 9](#_Toc46734)

[B.2 Preprocessing预处理 10](#_Toc46735)

[B.3 Infrastructure Settings基础设置 10](#_Toc46736)

[C Experimental Results实验结果 10](#_Toc46737)

[C.1 Qualitative Assessment of Negative Samples负样本定性评估 10](#_Toc46738)

[C.2 SOTA Algorithms 11](#_Toc46739)

[C.3 SANS Algorithms 11](#_Toc46740)

[D Computational Complexity复杂度计算 12](#_Toc46741)

[文献原文 13](#_Toc46742)

知识图谱中的结构感知负采样

## Abstract

Learning low-dimensional representations for entities and relations in knowledge graphs using contrastive estimation represents a scalable and effective method for inferring connectivity patterns. A crucial aspect of contrastive learning approaches is the choice of corruption distribution that generates hard negative samples, which force the embedding model to learn discriminative representations and find critical characteristics of observed data. While earlier methods either employ too simple corruption distributions, i.e. uniform, yielding easy uninformative negatives or sophisticated adversarial distributions with challenging optimization schemes, they do not explicitly incorporate known graph structure resulting in suboptimal negatives. In this paper, we propose Structure Aware Negative Sampling (SANS), an inexpensive negative sampling strategy that utilizes the rich graph structure by selecting negative samples from a node’s k-hop neighborhood. Empirically, we demonstrate that SANS finds semantically meaningful negatives and is competitive with SOTA approaches while requires no additional parameters nor difficult adversarial optimization.

采用对比评估方法来学习知识图谱中实体和关系的低维表示，是一种可扩展的、有效的推理实体之间联系的途径。对比学习方法的一个关键点在于选择产生 hard-negative（难以判别的负样本）的负采样方案，这迫使嵌入模型去学习高价值的样本并发现数据的关键特征。

早期的方法要么采用了过于简单的随机替换分布，即均匀负采样，它产生容易区分的、无有效语义信息的负样本；要么采用具有挑战性的优化方案，也即复杂的对抗式采样方案。二者都没有明确地包含知识图谱中已知的图结构信息，从而导致生成的负样本都不够优秀。

在本文中，我们提出了结构感知负采样(SANS)，这是一种耗费很低的负采样策略，它通过从节点的 k-hop 邻域中选择负样本，从而来利用丰富的图结构信息。从我们的经验可以得出，我们证明了 SANS 可以发现在语义上有意义的负样本。并且，与 SOTA 方法对比，它不需要额外的参数，也不需要做困难的对抗优化。

## 1 Introduction

知识图谱是信息的存储库，存储实体三元组（h，r，t）组成的信息。其中头实体 h 和尾实体 t 通过特定的关系 r 连接。事实上，知识图谱在许多领域都已经有了广泛的应用，如问答系统（Yao 和 Van Durme，2014；Hao 等人。，2017；摩尔多瓦和俄罗斯，2001）和机器阅读（Weissenborn 等人，2018；Yang 和 Mitchell，2017 年）等，在自然语言处理领域中拥有丰富的历史（Berant 等人，2013；Y u 和 Dredze，2014；Col-lobert 和 Weston，2008 年；Peters 等人，2019 年）。

通常已经存在的大型 KGs，比如 FreeBase (Bollacker 等人，2008 年)和 WordNet (Miller,

1995 年)都是不完整的。因此，通过链接预测来补全 KG 构成了一个基础研究课题，这改善了重要的自然语言处理任务的实践(Sun et al.， 2019;安杰利和曼宁，2013 年)。

近年来，使用图嵌入技术的方法激增，这些方法将 KGs 编码到低维向量空间，从而简化数据操作(Zhang et al.， 2019)，同时也是处理数据稀疏性和图谱不完整性的一个有吸引力的框架(Wang et al.， 2018)。为了学习这些嵌入方法，对比学习方法已经成为事实上的黄金标准。事实上，对比学习方法比那些需要在一个大的候选集上计算精确的 softmax 的方法有显著的计算优势（例如在给定头实体和关系的所有可能的尾部实体集上）。另一个重要的考虑因素是建模需求，因为将某些假设表示为基于边际或非规范化概率模型中的一些得分或能量值是最佳的选择(Smith 和 Eisner, 2005)。例如，以观察到的实体-关系-实体三元组，将实体关系建模为向量空间中的平移或旋转，自然会导致基于距离的得分被最小化 (Bordes et al.，2013)。

利用对比评估方法训练 KG 嵌入模型也包括模型的优化，即根据观察到的正三元组提高其权值，同时在负三元组上降低权值。因此，负采样分布方案的如何选择，在普遍采用简单随机抽样的模型应用现状中起着至关重要的作用。噪声对比估计(NCE) (Gutmann 和 Hyv arinen, 2010) -产生的负样本很容易就能够分类，并且以梯度信号的形式提供的信息也很少。如果采用更复杂的负采样策略（例如，对抗式生成负采样）将能够补救这一点。然而，对抗性的负抽样方法在计算上是昂贵的，而更容易处理的方法，比如基于缓存的方法(Zhang et al.， 2019) 则并没有契合 KG 的结构设置，因为它不能将已知的图结构作为采样过程的一部分。这就提出了一个重要的问题，即我们是否可以在从 KGs 丰富的图结构中获益的同时，获得一个计算上廉价的负采样策略。

**Present Work**：在这项工作中，我们引入了结构感知负采样(SANS)，这是一种利用 KG 的图结构来寻找 hard- negative 的算法。具体来说，SANS 使用限制在头部或尾部实体的 khop 邻域内的实体子集构建负样本。我们假设，彼此相邻但没有直接关系的实体有更高的机会相互关联，因此是负样本替换集的良好候选对象。我们还实验了一种基于随机游动的动态采样方案来近似节点的局部邻域。从经验上看，我们发现使用 SANS 的负抽样相比随机均匀抽样和复杂的生成对抗网络(Goodfellow 等人，2014)(GAN)有明显提升，它只需要很小的计算成本，并且没有额外添加参数，相比 SOTA 方法更具有竞争力。

## 2 Related Work

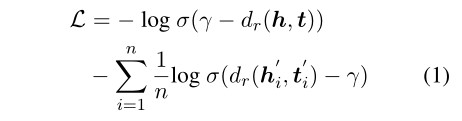
**Negative Sampling.** 负采样是一种可用于对数线性模型缩放的方法。从本质上讲，负抽样解决了计算归一化常数的计算困难，它将任务改为区分观察到的真实正样本和破坏正样本而产生的虚构的负样本。这种一般的方法是对 NCE 的简化，它基于 Importance Sampling (IS) 中使用的配分函数的蒙特卡洛近似(Bengio et al.，2003)。

**Non-Fixed Negative Sampling.** 正如(Mikolov et al.， 2013)提出的，可以使用均匀采样方案产生负三元组。然而，这种统一且固定的采样方案在训练过程中导致了容易分类的负样本，没有提供任何有意义的信息(Sun et al.， 2019;张等，2019)。因此，随着训练的进行，大部分采样的负三元组都将得到很小的分数和几乎为零的梯度，经过少量的几次迭代后就会对训练图嵌入模型产生阻碍。

为了解决产生 easy-negative 的问题，Sun et al.(2019)提出了自我对抗的负面抽样：在嵌入模型下，根据每个采样的负样本的概率对其进行加权。另外，(Wang et al.， 2018)和(Cai and Wang, 2018)的作者尝试通过利用 GANs 生成高质量的负样本，GANs 虽然有效，但训练费用昂贵，需要黑箱梯度估计技术（black-box gradient estimation techniques）。与基于 GAN 的方法相比，另一种使用较少参数且更容易训练的优雅方法是 NSCaching (Zhang et al.， 2019)，它需要使用高质量负三元组缓存——即，高分者。

## 3 Structure Aware Negative Sampling

给定一个观察到的正三元组(h, r, t)，一个负样本可以通过破坏头部或尾部实体形成一个新的三元组——即(h0, r, t0) -其中 h0, t0∈E，其中 E 是 KG 中所有实体的集合。另外，我们假设图嵌入模型使用如下形式的损失函数进行训练:



式中 dr(h, t)为 r 关系下头尾实体兼容性的评分，γ 为固定边界，σ 为 sigmoid 函数，n 为负样本数。

在本文中，我们试图在生成负三元组时显式地使用围绕特定节点的局部图结构。我们的方法是基于之前学习词嵌入的工作(Mikolov 等人，2013 年)的观察得出的，该工作在历史上出现了负抽样这一思路，但在知识图谱的设置中缺乏可以轻松获取的丰富的图结构。

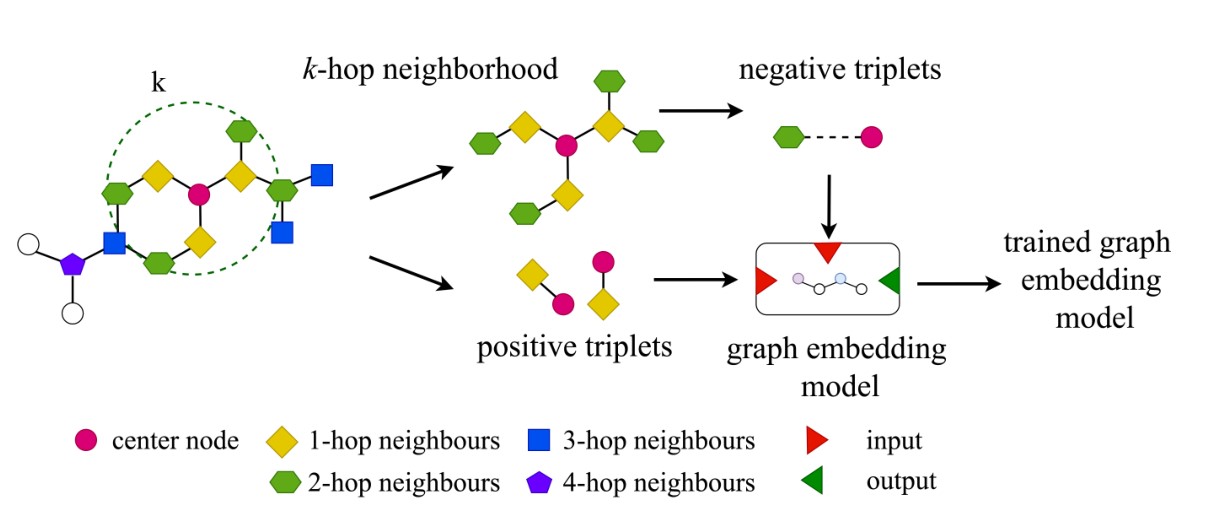


Figure 1

我们假设，用图结构信息来丰富负采样算法的过程可以产生更困难的负样本，这对学习有效的嵌入模式至关重要。图 1 演示了我们的方法，它要求在第一步为每个节点构建 K -hop 邻域(K)：



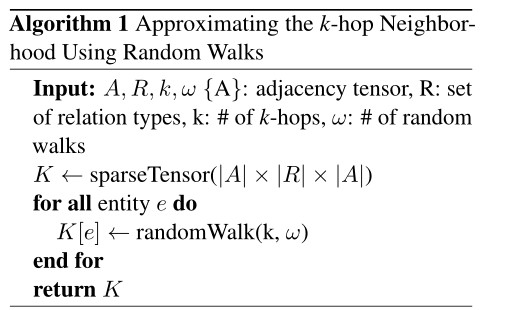
其中 k > 0，其中 k 为整数，表示邻域半径，A 为 KG 的邻接矩阵，S+为元素级（elementwise）函数，如果有路径存在，则设为 1，否则设为 0。

为了构造负三元组，我们现在可以简单地从 K 的非零单元中采样，K 表示即每个节点的图谱中所有实体范围内的邻域子集，且 K1E×E。直观地, SANS 利用了实体邻域的局域性，其中负样本被定义为没有直接链接到关系 r 下，但可以通过一条最长长度为 k 的路径访问的实体。我们认为，这种局部产生的负样本更难区分，使得嵌入模型的训练评估分数更高。构造K的一个重要技术细节是存在多种关系类型，这需要一个额外的维数来表示图的连通性，例如：邻接 K -hop 张量。

### 3.1 Variants of SANS 算法变体

虽然 SANS 需要一次性的预处理步骤构建公式 2 定义的 K，但这对于大型且密集的知识图谱可能仍然是昂贵的。为了解决这一效率低下的问题,我们将介绍算法 1 描述的 RW-

SANS。它使用 ω 随机漫步(Perozzi et al ., 2014)的长度 K 邻接张量来近似 k-hop 邻域。



由于 SANS 构建了一个局部邻域，从中提取负样本，它也可以与其他负样本方法相结合。在这项工作中，我们扩展了(Sun et al.， 2019)中的自我对抗方法，并通过将负三元组候选集限制为 k-hop 邻域将其与 SANS 结合起来。在接下来的章节中，我们将这种技术称为自对抗**(Self-Adv)SANS**，而前一种方法被称为 **Uniform SANS**。

## 4 Experiments

我们研究了 SANS 生成的负样本在嵌入训练模型（如 Trans、DistMult 和 RotatE 模型）下，在知识图谱补全任务中的应用。我们在标准基准上评估我们提出的方法，包括 FB15K237 (Bollacker et al.， 2008)， WN18 和 WN18RR (Miller, 1995)。从我们的实验中，我们试图回答以下问题:

(Q1) Hard Negatives:我们能单纯地使用图结构对进行负采样吗?

(Q2) 我们能否将图结构与其他 SOTA 负采样方法相结合?

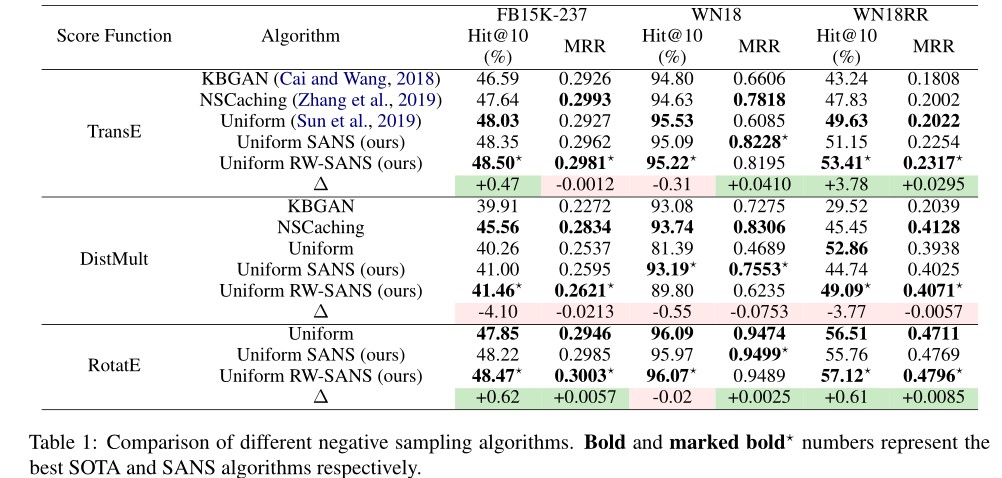
(Q3) 我们能有效地用 ω 随机游动来近似邻域张量**K**吗?

在我们的实验中，我们采用三个具有代表性的基准采样方法，即均匀负采样(Bordes et al.，2013)、KBGAN (Cai and Wang, 2018)和 NSCaching (Zhang et al.， 2019)。我们还用自对抗负采样中与当前的 SOTA 方法进行比较(Sun 等人，2019)，并测试局部图结构是否也可以在这种设置中使用。

### 4.1 Results

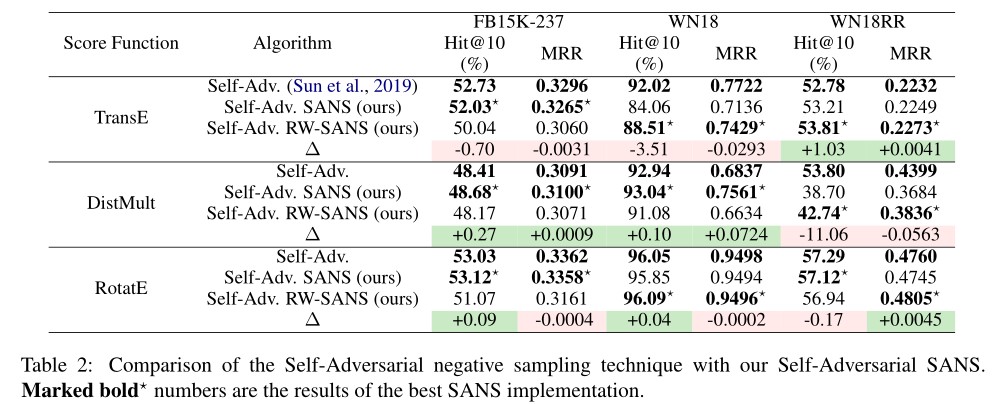
我们现在讨论实验的核心问题。

Q1:表 1 总结了我们的主要定量结果，同时强调了 SANS 和 RW-SANS 的结果数据。我们还计算了 SANS 的最佳变体与 Δ 行标注的表现最好的基准方法之间的差异。总的来说，我们发现在所有三个数据集上，SANS 负采样几乎总是能产生相比均匀负采样和 KBGAN 负采样更 hard 的负样本。

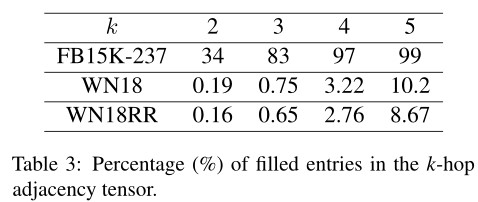
 此外，SANS 在结合 TransE 模型的情况下，相比 NSCaching 算法实现了更有竞争力的性能效率，并且在不需要额外参数的情况下，与 DistMult 模型结合使用，SANS 也能保持第二优秀算法位置。我们观察到，TransE、DisMult 和 RotateE 模型下的 Δ 行 MRR 平均值分别为 0.0231、-0.0341 和 0.0056，这证实了我们的方法相对于 SOTA 的有效性，同时又保持了计算效率。

我们也定性地研究了 SANS 与通过统一抽样产生的负样本的语义上的“难度”。例如，以 WN18RR 数据集的中心节点“arachnoid（蛛网膜）”为例，通过 SANS 在 2-hop 邻近区域内采样的负样本是“蛛形纲动物”、“生物学”、“节虫”、“狼蛛”和“花园蜘蛛”，而通过统一采样的负样本是“用餐者”、“拒绝者”、“景观者”、“崛起者”和“哺养者”。

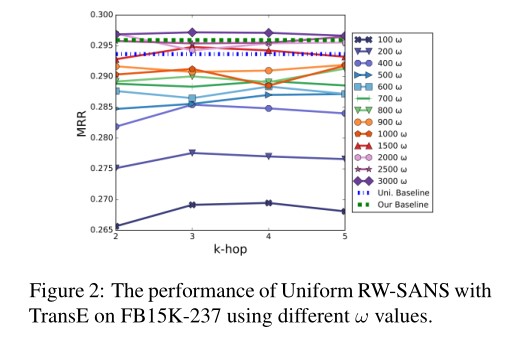
显然，通过 SANS 发现的负采样在语义上很难区分，因此，这些结果也证实了将图结构整合到负抽样中以生成 hard- Negative 的重要性。C.1 中可以找到对 SANS 产生的负样本的更详细的定性分析——包括不同邻域大小的影响。



Q2 : 我们现在将我们的方法 SANS 与自我对抗的负抽样相结合(Sun 等人，2019)。我们的结果显示在基于 Self-Adv SANS 和 Self-Adv RW-SANS 的表 2 中，两者都重新对负三元组赋权 (Sun et al.，2019)。我们观察到两种方法的性能相当，但关键的是，这主要是通过 WN18 和 WN18RR 等数据集中 0.2%至 9%的实体来实现的，如表 3 所示。考虑到部分填充邻接张量提高了计算的可行性，减少了对内存的要求，并允许进行稀疏张量运算，这无疑进一步强调了在选择负样本时结合图结构的吸引力。



Q3 : 我们现在分析使用随机漫步近似局部邻域的影响。图 2 描述了在不同半径和 MRR 范围内随机行走次数(ω)的影响。我们记录了两个基线方法，一个是均匀采样的性能，另一个是Uniform SANS与 TransE 结合时所获得的最佳性能，在这里 k-hop 张量是显式计算的。



有趣的是，我们发现 k-hop 张量不仅可以用 3000 次随机漫步很好地近似，而且 RWSANS 也优于这两个基线方法的性能。我们注意到某些节点由于与中心节点共享较多的路径而具有较高的抽样概率来解释这一结果，从而引出了隐式加权负抽样方案。

## 5 Conclusion and Future Directions

在这项工作中，我们引入了新的负抽样策略 SANS，它直接利用 k-hop 邻域的信息来选择负样本。我们的工作揭示了在为 KGs 设计负采样时，整合图结构的必要性和重要性，而

SANS 可以被视为一种廉价但强大的基线方法，不需要额外的参数或困难的优化。

从经验上看，我们发现基于 SANS 的负样本与 SOTA 方法具有相当的性能表现，甚至优于之前复杂的基于 GAN 的方法。

## Acknowledgments

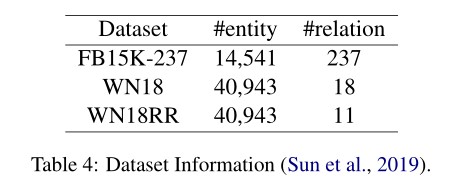
作者要感谢匿名 EMNLP 审稿人提供的有帮助和建设性的反馈。本研究得到加拿大 CIFAR 人工智能主席和 NSERC 发现基金 RGPIN-2019-0512 的支持。Avishek Joey Bose 也通过 IV ADO 博士奖学金得到了慷慨的支持。

## A Experimental Settings

本节提供了用于复现我们的结果的数据集和评估标准的概述。

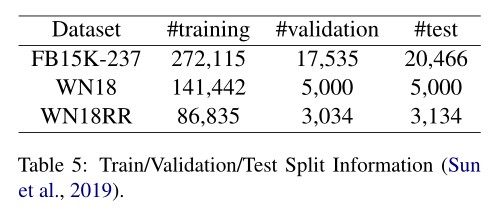
### A.1 Datasets

我们使用数据集 FB15K-237、WN18 和 WN18RR 对我们提出的方法进行实验。FB15K237 是 FB15K 的子集，FB15K 来源于 FreeBase 知识库(KB) (Bollacker et al.， 2008)，这是一个包含有许多不同关系类型的关于世界的一般事实的大型数据库。另一方面，WN18RR 是 WN18 的一个子集，而 WN18 是由 WordNet 知识库(Miller, 1995)衍生而来的，WordNet 知识库是一个捕捉词汇关系的大型英语词汇数据库，例如。词语之间的上下级关系。WN18 和 FB15K 首先在(Bordes et al.， 2013)中被引入，并在大多数与 kg 相关的研究中被使用。相比之下，WN18 和 WN18RR 包含的关系类型比 FB15K-237 少。表 4 提供了与每个数据集对应的实体数量和关系类型的摘要。



### A.2 Evaluation Protocols

为了评估我们的负采样方法，我们使用标准的评价指标，包括平均倒数秩(MRR)和 Hits at N (H@N)。训练/验证/测试分离信息如表 5 所示。

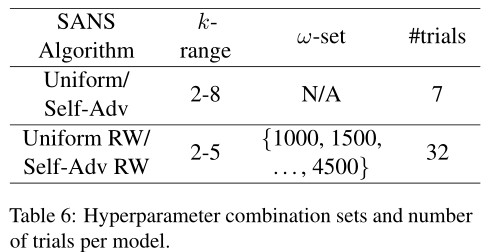


## B Implementation Details

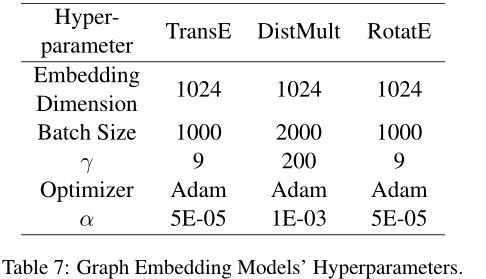
这部分补充介绍了我们的 RW-SANS 算法的实现细节——即 Uniform RW-SANS 和 SelfAdv RW-SANS，它使用随机游动来近似 k-hop 邻接张量。本文将进一步详细介绍其他实验设置。

### B.1 Hyperparameters

k 和 ω(当 k-hop 邻域被算法 1 近似时)是我们负采样算法中的超参数。为了在不同数据集的验证集上找到最优的超参数值，在负采样步骤中分别使用 k 值 2 ~ 8 和 ω 值 1000 ~ 5000 进行测试。换句话说，在我们的实证研究中，验证集的最佳性能是通过手动调整超参数发现的。关于实验试验的更多信息可以在表 6 中找到，其中也列出了在每个数据集上训练每个图嵌入模型的试验总数。



此外，表 7 列出了用于训练不同图嵌入模型以使其在验证集上达到最佳性能的超参数。



### B.2 Preprocessing

KG内节点的k-hop邻域的建立可以看作是实现SANS所必需的预处理步骤。在本文中，我们提出了两种实现此目的的技术:

* 通过公式 2 显式计算 k-hop 邻域，同时考虑不同的关系类型。
* 使用随机游动近似 k-hop 邻域，详见算法 1。

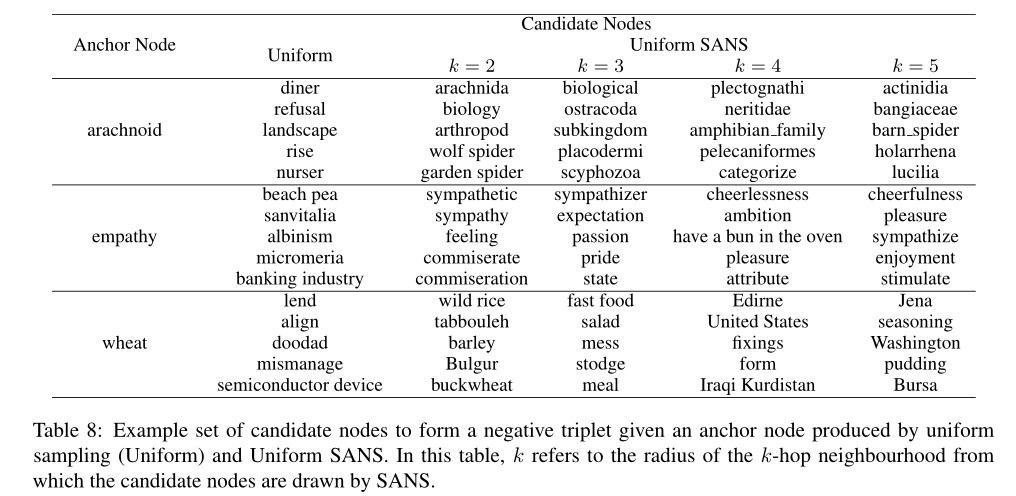
### B.3 Infrastructure Settings

我们的实验是在一台服务器上进行的，该服务器有一个 NVIDIA V100 GPU, 10 个 CPU核，46GB RAM。

## C Experimental Results

### C.1 Qualitative Assessment of Negative Samples

在本节中，我们使用 WN18RR 数据集评估 Uniform SANS 和统一抽样产生的负样本的语义意义。

 就像表 8 中例子所展现的，与统一抽样相比，Uniform SANS 会生成在语义上更难区分的负样本。我们还注意到，随着增加 k-hop 邻域的规模，负样本的语义意义将下降。这一观察结果确实是预期的，因为随着邻域大小的增加(即 k→∞)，Uniform SANS 将变得类似于均匀抽样。

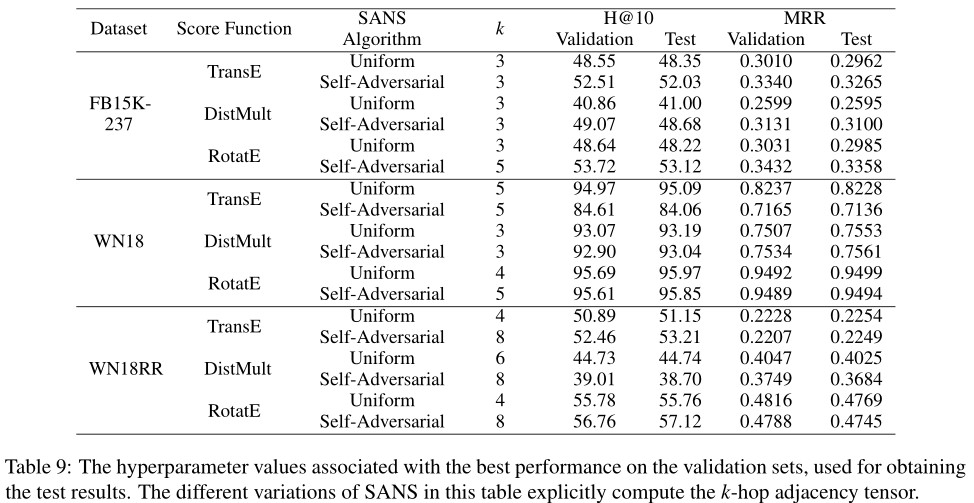
### C.2 SOTA Algorithms

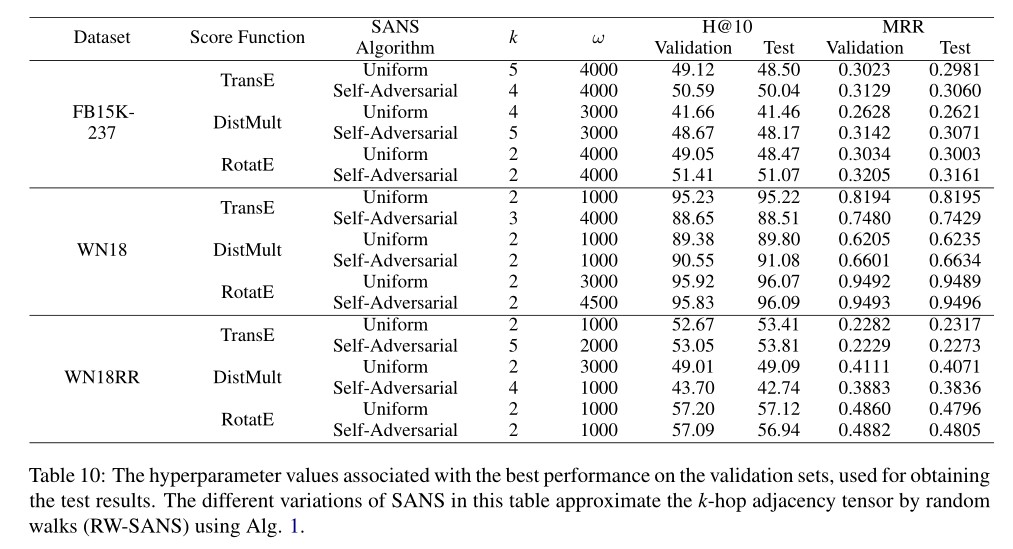
表 1 和表 2 中的 Uniform 算法和 Self-Adv 算法的结果分别是通过使用他们在不同数据集验证集上报告的最佳性能超参数，重新运行(Sun 等人，2019)提供的代码获得的。此外，表 1 中的 KBGAN 和 NSCaching 结果是直接从(Zhang et al.， 2019)获得的粗略结果。

### C.3 SANS Algorithms

表 9 和表 10 展示了与我们的负采样技术融合的图嵌入模型在验证和测试集上关于评价

指标的性能。此外，它们列出了 Uniform/Self-Adv 和 Uniform/Self Adv. RW-SANS 在验证集上取得最佳性能相对应的超参数值。



 基于我们的结果，我们可以得出假设，在近似 k-hop 邻域时使用随机游动方法，会隐式地导致移除到中心节点的游动次数最少的节点——即离群值的节点，这些节点与中心节点的联系（可达路径）并不是很密切。

## D Computational Complexity

表 11 代表了包括 SANS 在内的不同负抽样方法的时间和空间复杂性。

