SimplE Embedding for Link Prediction in Knowledge Graphs报告

201250203 陈张熠

(1) 论文摘要 abstract 和 introduction 翻译

Abstract部分：

知识图表包含关于世界的知识，并提供了这些知识的结构化表示。当前的知识图表只包含世界上真实情况的一小部分。鉴于实体之间的现有链接，链接预测方法旨在预测知识图的新链接。事实证明，张量分解方法对此类链接预测问题很有希望。1927年提出的规范多元（CP）分解是最早的张量分解方法之一。CP在链路预测方面通常表现不佳，因为它为每个实体学习了两个独立的嵌入向量，而它们实际上是绑定的。我们提出了CP（我们称之为SimplE）的简单增强，以允许依赖性地学习每个实体的两个嵌入。SimplE的复杂性随着嵌入的大小而线性增长。通过SimplE学习的嵌入是可以解释的，某些类型的背景知识可以
通过重量绑定被纳入这些嵌入中。我们证明SimplE是完全表达的，并根据其嵌入的大小导出一个界限，以获得完全表达性。我们实证表明，尽管SimplE简单，但它的表现优于几种最先进的张量分解技术。SimplE的代码可在GitHub上访问https://github.com/Mehran-k/SimplE。

Introduction部分：

在过去的两个阶段中，几种（也许是大概率）包含世界事实的 知识图谱（KGs）已经被建成。这些知识图谱已经在几个领域有应用，包含搜索、问答、自然语言处理、推荐系统等。由于可以断言关于我们世界的大量事实以及访问和存储所有这些事实的难度，KG是不完整的。但是，可以根据现有链接预测 KG 中的新链接。在统计关系学习（SRL）的保护伞下，研究了链接预测和其他几个旨在与实体和关系进行推理的相关问题[12、31、7]。关于为知识图谱进行链接预测的问题也被称作 知识图谱完成（？）。一个知识图谱可以用一个三元组（头，关系，尾）所代表。知识图谱完成的问题可以被看作在现有的三元组上预测新的三元组。

张量分解方法被证明是在知识图谱完成这一问题上一种有效的 统计关系学习（SRL）方法[29,4,39,26]。这些方法考虑了每个实体和每个关系的嵌入。为了预测三元组是否成立，他们使用一个函数，该函数将头部和尾部实体的嵌入和关系作为输入，并输出一个数字，指示预测概率。关于这些方法的细节和讨论可以在最近的几篇调查中找到[27,43]。

最早的张量分解方法就是规范多元（CP）分解[15]。这种方法为每个关系学习一个嵌入向量，为每个实体学习两个嵌入向量，一个在实体是头部时使用，一个在实体是尾部时使用。一个实体的头嵌入的学习是独立于（无关于）它的尾嵌入。这种独立性导致CP方法在KG完成问题上表现差[40]。在这篇论文中，我们在CP方法之上解决了实体中两嵌入向量的独立问题，发展了一个张量分解方法。因为我们模型的简单性，我们称它为 SimplE（Simple Embedding）。

我们证明SimplE：1–可以被认为是一个双线性模型，2–是完全表现力的，3–能够通过参数共享（又名权重绑定）将背景知识编码到其嵌入中，尽管（或可能是因为它）简单，但4在经验上表现非常好。我们还讨论了其他现有方法的几个缺点。我们证明，许多现有的翻译方法（例如，[4、17、41、26]）没有充分表达，我们确定了对它们所能代表什么的严格限制。我们还表明，ComplEx [39, 40]中使用的函数是一种最先进的链路预测方法，涉及冗余计算。

(2) 问题描述

解决了实体中两嵌入向量的独立问题，在CP方法之上完成一个新的张量分解方法

(3) 输入、输出、模型算法描述(附框架图0

对于trainer部分，输入：数据集 输出：模型

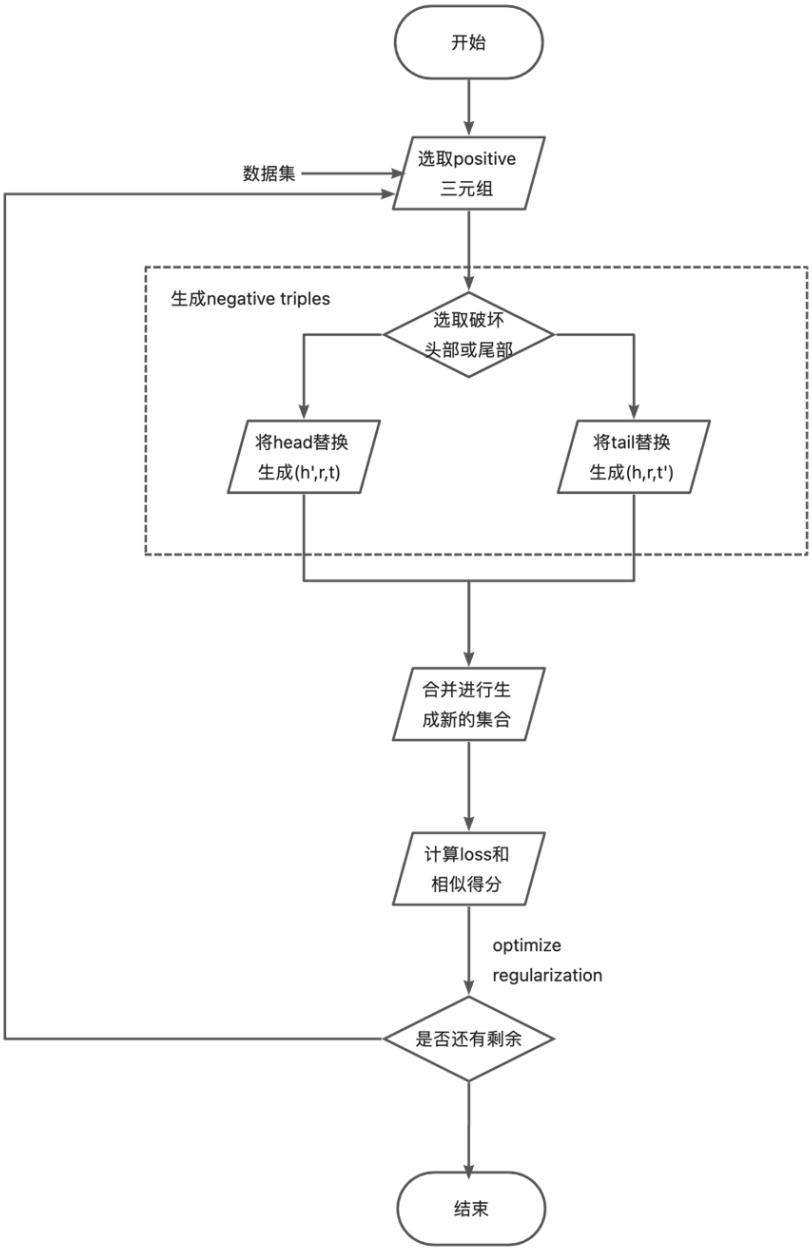
对于tester部分，输入：数据集，模型 输出：MRR评价指标，以及hit率

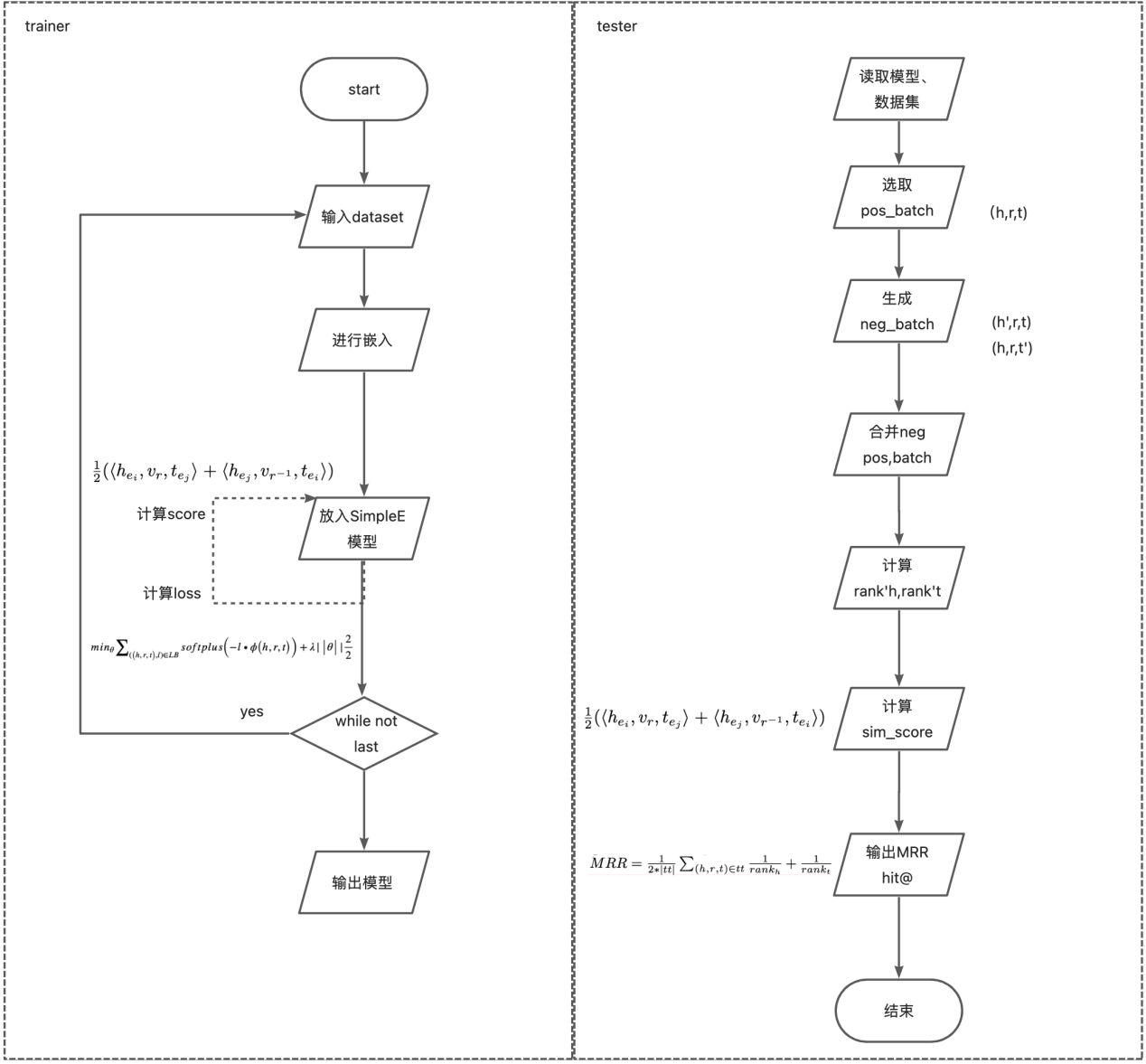
模型算法描述：在每次学习迭代中，我们迭代从数据集中提取一批三元组，然后对于批次中的每一个三元组（h，r，t），我们随机决定损坏头部或尾部。如果选择头部，我们将三元组中的h替换为从E − {h}中随机选择的实体h′，并生成损坏的三元组（h′，r，t）。如果选择尾部，我们将三元组中的t替换为从E − {t}随机选择的实体t′，并生成负三元组（h，r，t ′），一共生成n个负三元组。

一旦我们有一个标记的批次，按照我们优化了批次的公式,以此为指标进行模型的训练。

$$min_\theta \sum_{((h,r,t),l)\in LB} softplus\left(-l \bullet \phi(h,r,t)\right) + \lambda \| \theta \| \frac{2}{2}$$

其中θ表示模型的参数（嵌入中的参数），l表示三元组的标签，ϕ(h , r , t) 表示三元组（h , r , t ）的相似性分数，λ 是参数，softplus(x) = log(1 + exp(x)) 。





(4) 评价指标及其计算公式

计算平均倒数排名（MRR）这些排名作为排名逆的平均值：

$$MRR = \frac{1}{2*|tt|}\sum_{(h,r,t)\in tt} \frac{1}{rank_h} + \frac{1}{rank_t}$$

其中tt表示测试三元组。

MRR是一个比平均排名更可靠的衡量标准，因为单个糟糕的排名在很大程度上会影响平均排名。

除了MRR，还有命中概率，作为评价指标。模型的hit@k计算为排名小于或等于k的测试三元组的百分比。

(5) 对比方法及引用出处

CP方法对比：

这种方法为每个关系学习一个嵌入向量，为每个实体学习两个嵌入向量，一个在实体是头部时使用，一个在实体是尾部时使用。一个实体的头嵌入的学习是独立于（无关于）它的尾嵌入。这种独立性导致CP方法在KG完成问题上表现差

而SimplE同时为头尾部实体进行学习。

$$\frac{1}{2}(\langle h_{e_i}, v_r, t_{e_j}\rangle + \langle h_{e_j}, v_{r^{-1}}, t_{e_i}\rangle)$$

TransE——Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In NIPS, pages 2787–2795, 2013.

ComplEx——Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In ICML, pages 2071–2080, 2016.

TransR——Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In AAAI, pages 2181–2187, 2015.

DistMult——Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. ICLR, 2015.

NTN——Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In NIPS, 2013.

STransE——Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu, and Mark Johnson. Stranse: a novel embedding model of entities and relationships in knowledge bases. In NAACL-HLT, 2016.

ER-MLP——Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In ACM SIGKDD, pages 601–610. ACM, 2014.

(6) 结果 （在数据集WN18下 训练100次）

	Hit@1	Hit@3	Hit@10	MRR
SimplE	0.9197	0.9265	0.9254	0.9198
CP	0.071	0.072	0.075	0.074
ComplexE	0.9131	0.9152	0.9198	0.9189