**《大数据分析》第二次实验报告**

201250070郁博文

所选论文：《Relation Prediction as an Auxiliary Training Objective for Improving Multi-Relational Graph Representations》

1. 论文摘要和介绍翻译

摘要

在多关系图上学习良好的表示对于知识库补全(KBC)至关重要。本文提出了一个新的多关系图表示学习的自监督训练目标，通过简单地将关系预测纳入常用的1vsAll目标。新的训练目标不仅包含用于预测给定三元组的主客体的术语，还包含用于预测关系类型的术语。我们分析了这一新的目标如何影响KBC中的多关系学习:在各种数据集和模型上的实验表明，关系预测可以显著提高实体排名，这是KBC最广泛使用的评估任务，在FB15k-237上产生了6.1%的MRR增长和9.9%的Hits@1增长，在Aristo-v4上产生了3.1%的MRR增长和3.4%的Hits@1增长。此外，我们观察到所提出的目标在高度多关系数据集(即具有大量谓词的数据集)上特别有效，并在使用更大的嵌入尺寸时生成更好的表示。代码将在https://github.com/facebookresearch/ssl-relation-prediction上提供。

介绍

知识库补全(Knowledge Base Completion, KBC)，又称链路预测，旨在补全缺失条目，在构建大规模知识图谱中起着至关重要的作用[Nickel等人，2016,Ji等人，2020,Li等人，2020]。在过去的几年里，大部分关于KBC的研究都集中在知识图谱嵌入模型上，这些模型学习知识图谱中所有实体和关系的表示，并使用它们对是否存在边进行评分[Nickel et al.， 2016]。文献中提出了许多模型和架构创新，包括但不限于基于翻译的模型[Bordes等人，2013]，潜在因子分解模型[Nickel等人，2011,Trouillon等人，2016,Balazevic等人，2019]，以及基于神经网络的模型[detmers等人，2018,Schlichtkrull等人，2018,Xu等人，2020]。

其他更近期的研究一直在分析这些KBC模型的评估程序方面做出补充努力。例如，Sun等人[2020]呼吁将评估协议标准化;Kadlec等人[2017]、Ruffinelli等人[2020]和Jain等人[2020]强调了训练策略的重要性，并表明仔细的超参数调整可以产生更准确的结果

比采用更精细的模型架构;Lacroix等人[2018]认为，当适当选择训练目标时，一个简单的模型可以产生最先进的结果。

从这些发现中获得灵感，本文探索了关系预测:一个简单的辅助训练目标，可以显著提高几个KBC模型的多关系图表示学习。除了训练模型来预测知识图谱中三元组的主体和客体实体，我们还训练它们来预测关系类型，从而形成一个自监督的训练目标。直观地说，这种方法类似于使用类似于掩码语言模型的训练目标[Devlin et al.， 2019]而不是KBC常用的自回归训练目标。在我们的实验中，我们发现新的辅助训练目标显著提高了下游链路预测的精度。

对各种模型和数据集的实证评估支持了我们新的训练目标的有效性:在嵌入大小介于2K和4K之间的ComplEx-N3 [Trouillon et al.， 2016]和CP-N3 [Lacroix et al.，2018]上观察到最大的改进，在Hits@1上提供高达9.9%的提升，在FB15k-237上提供6.1%的提升，而计算开销可以忽略不计。

在具有不同数量谓词的数据集上进行了进一步的实验，发现当数据集高度多关系时，即包含大量谓词时，关系预测的帮助更大。此外，我们的定性分析表明，一些多对多[Bordes et al.， 2013]谓词的预测得到了改进，关系表示更加多样化。

1. 问题描述

这篇文章探索了，在与KBC有关的多关系图表示学习模型中，增加一个辅助训练目标，即关系预测任务的效果。这个训练目标类似于掩码语言模型的训练目标，而非是KBC模型的传统自回归训练目标。除了训练模型来预测知识图谱中三元组的主体和客体实体，作者们还训练它们来预测关系类型，从而形成一个自监督的训练目标。对于各个KBC相关模型和数据集的实证评估证明了该方法的有效性。

1. 输入、输出、模型算法描述

输入：

给定一个查找(s; p; ?) 或 (?; p; o)，s代表主语，p代表谓语，o代表宾语。

输出：

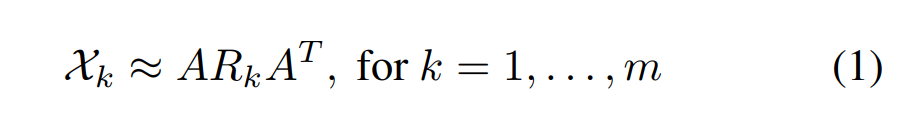
对于给定的查找，模型根据自身的评分函数（scoring function），为预测出的查找结果进行评分，然后将评分后的实体集进行排序，返回一个实体名单。

模型算法描述：

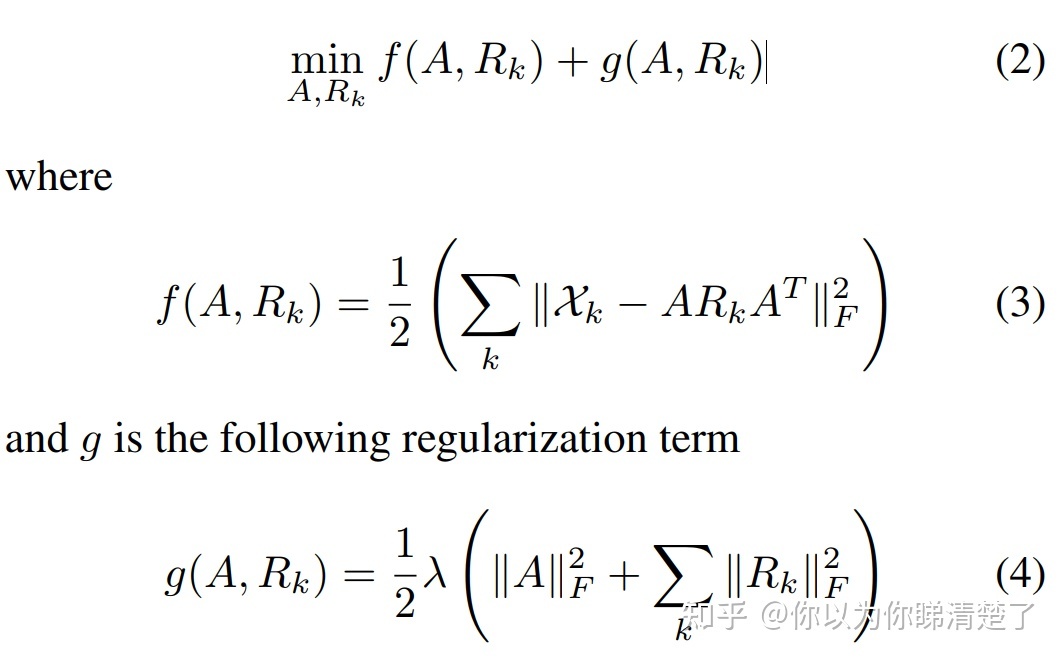
由于这是一篇实证论文，作者总共使用了四个模型来测试他们的新辅助训练目标，这里我选取了RESCAL来描述算法。

知识图谱中的表示一般为语义网络的RDF形式三元组，如（头实体，关系，尾实体），本质上这是一种二元关系数据，为了建模这种二元关系数据为张量，论文作者提出了一种三维张量因式分解方法。假设有一个三维张量 χ ，其二维切平面为串联的实体向量各自表示，而第三维则为实体之间的关系谓词，如果两个实体之间存在着关系，则三维张量对应的点为1，否则为0，这很容易联想到数据结构中的邻接矩阵的概念，论文提出的三维张量可以看作是 k （假设存在 k 个关系谓词）个邻接矩阵作为二维切平面拼接起来。

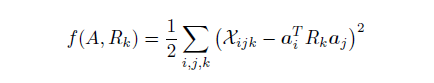
关系学习主张，对于同一个域的实体集，获取一个实体的信息不仅可以通过直接使用该节点的附加信息或者考虑特定学习任务这些手段，还可以根据与该实体节点相连接的其他实体节点和关系边之间潜在的拓扑关系来获取。比如想知道美国总统的党派，只需要知道总统的下级副总统的党派就能推断出总统的党派，因为他们通常是同一党派。作者认为这种探索由相关实体节点提供的信息而不考虑手头上特定学习任务的机制叫做集体学习（Collective Learning），为了将集体学习应用到多元关系数据上，作者提出了RESCAL这个模型。作者提出了如下的打分函数，A是包含了实体集的潜在表示信息的 n×r 矩阵， Rk 是一个 r×r 的不对称矩阵，它建模了在第k个关系谓词上实体之间的潜在信息的相互影响信息。



可以通过求解最小化正则化问题来计算因子矩阵 A 和 Rk ，如下所示，其中g为正则项：



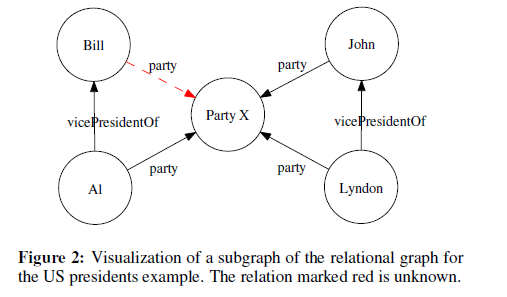
对于集体学习，关于公式（1）的一个重要方面以及它与其他张量分解方法(如CP甚至BCTF)的区别在于：域的实体有一个独特的潜在分量表示，不管它们是作为一个关系三元组中的头实体还是尾实体出现，因为它们都由矩阵A表示。通过观察(3)的基本公式，这种建模的效果变得更加明显：



ai 和 aj 表示 A 的第 i 行和第 j 行，因此是第 i 和第 j 个实体的潜在分量表示。将 aj 和 Rk 结合起来看，显然潜在分量表示 ai 依赖于 aj 和三元组（第 i 个实体，第 k 个关系谓词，第 j 个实体）。

并且，由于实体有自己唯一的潜在成分表示， aj 也保存了那些与第 j 个实体节点相关的作为头实体或者尾实体的实体节点的信息。因此，所有直接关系和间接关系对 ai 的计算都有决定性的影响。就如实体由 A 矩阵表示，每个关系谓词由矩阵 Rk 表示一样，它建模了潜在成分信息是如何在各自的关系中交互的，并且矩阵 Rk 的不对称性也考虑了一个潜在成分信息是作为头实体出现还是尾实体出现。

可以用一个简单的例子来说明这个机理，如图：



对于上图，AI和Lyndon的潜在成分信息表示是类似的，他们的潜在信息都会反映他们是属于党X的，因此，Bill和John也应该有着相同的潜在成分信息表示。所以， αBillTRpartyαPartyX 和 αJohnTRpartyαPartyX 应该有着类似的值，这样这种残缺的边关系就能被正确预测出来。

综上所述，典型的关系学习任务可以通过RESCAL来解决，具体如下：

对于第 k 个谓词，要预测两个实体 ei , ej 之间是否存在关系边连接，查看相应的二维切平面 Xk 的降秩重构 Xk∧=ARkAT 就足够了。同时，链接预测可以通过将 Xijk∧>θ 与某个给定的阈值 θ 进行比较，或者根据链接存在的可能性对实体进行排位。对于集体分类问题，其可以转换为链路预测的子任务，因为对于实体类可以通过引入类关系并将类作为“实体”来使用张量建模。因此，分类问题也可以通过重构类关系的适当切片来解决。最后，对于实体之间的相似度问题，这是根据它们在多个关系中的相似度来计算的。

1. 评价指标及其计算公式

MRR(Mean Reciprocal Rank)平均倒数排名：对于一个给定的查找(s; p; ?) 或 (?; p; o)，模型会根据其自身的评价函数给定一个所有候选实体的名单。如果正确的查找实体在这张名单中排名为r，那么它的倒数排名就是1/r。而平均倒数排名就是将数据集中所有查找实体的倒数排名取平均数。

Hits@K（本文中K取1，3，10）：对于上述排名清单，如果遍历前K个实体就能查找到预期实体，则视为命中。换而言之，该指标在直觉上就是用户需要阅读多少个实体才能找到想要的实体。

1. 对比方法及引用出处

1vsAll: 给定一个主语s和一个谓语p，预测正确的宾语o的任务被映射成一个多维的多标签分类任务，一个标签对应一个特定的宾语o，多个标签可以被分配给给定的<s, p>对，这被称为K vs All. 如果每个<s, p>对只能被分配一个标签，那么便称为1 vs All.

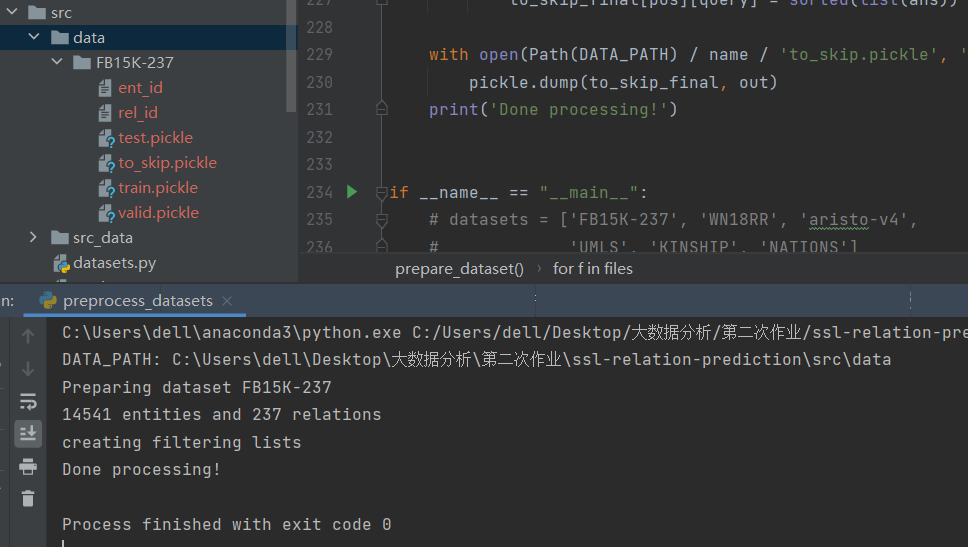
出处：D. Ruffinelli, Samuel Broscheit, and Rainer Gemulla. You can teach an old dog new tricks! on training knowledge graph embeddings. In ICLR, 2020.

这篇文章只是在上述方法的基础上增加了一个辅助训练目标，并没有提出新的技术。因此所谓的对比只是在1vsall方法的基础上有没有使用这篇文章提出的辅助训练目标。

1. 结果
2. 数据集处理

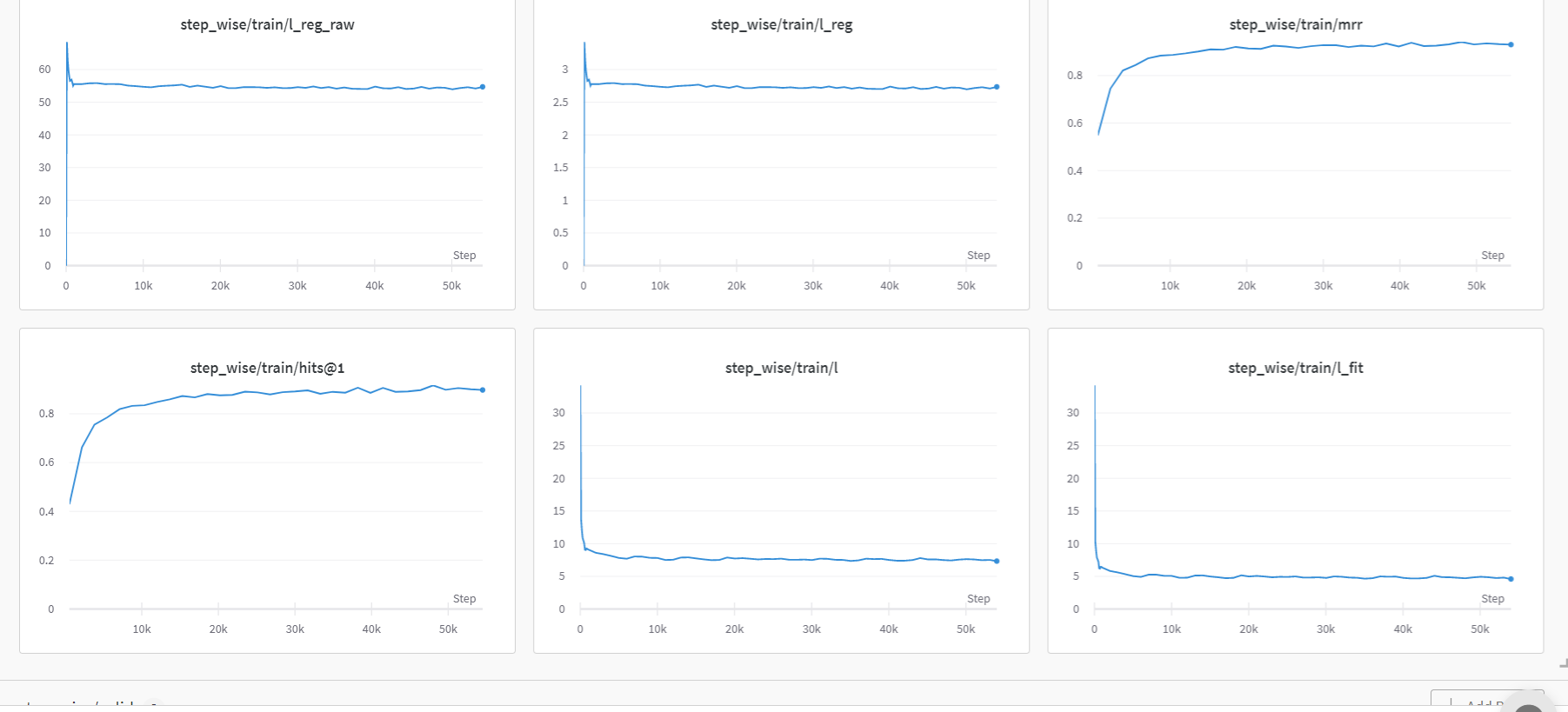
首先，从项目的readme中提供的链接中下载数据集FB15K-237，然后运行preprocess\_data.py文件处理数据，得到训练集、验证集、测试集的pickle文件。

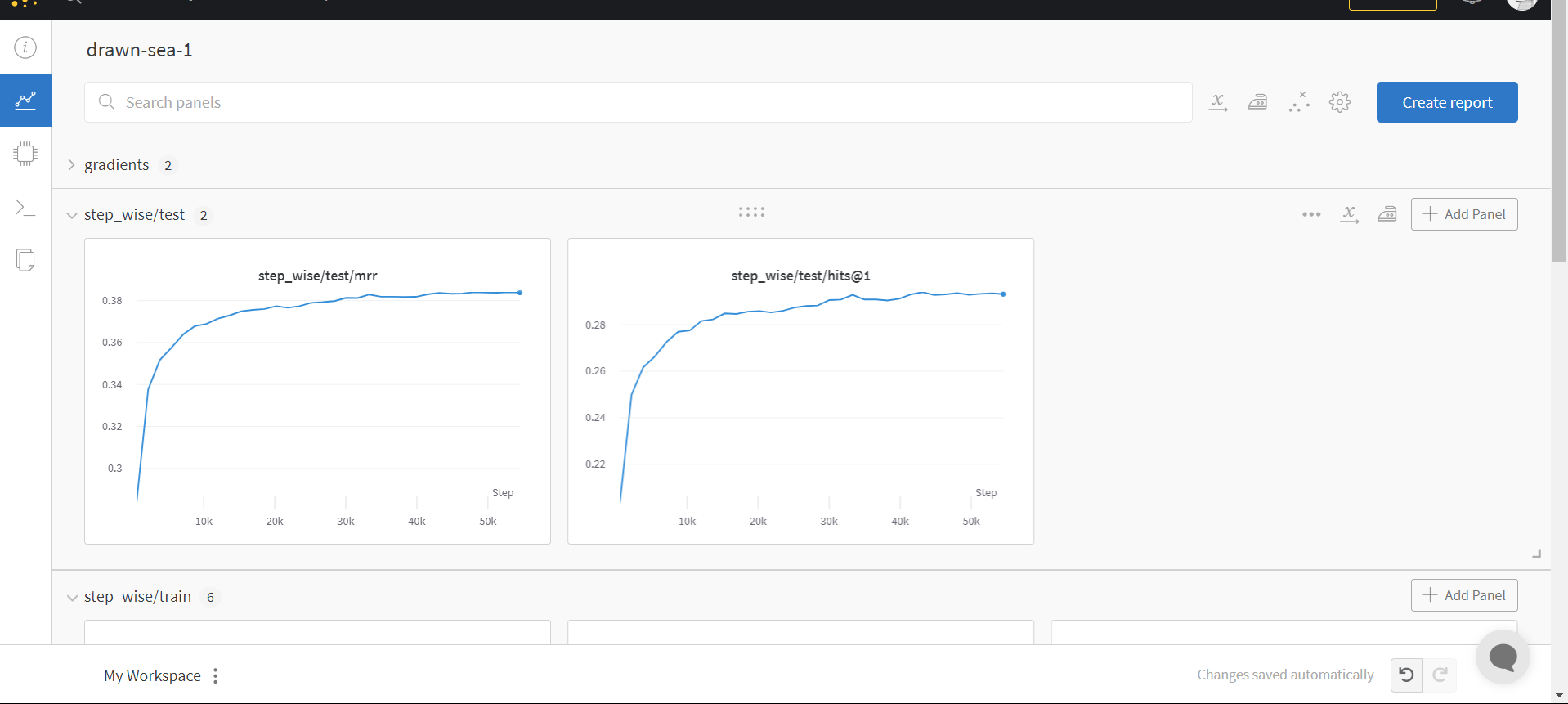
运行结果如下图：



1. 训练模型

以下是在FB15K-237上，使用辅助训练目标的运行总结：







以下是在FB15K-237上，不使用辅助训练目标的运行总结：

