《智能信息处理》课程考试

基于知识嵌入的链接预测研究进展

刘冬帅

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2020 年 12 月 8 日

基于知识嵌入的链接预测研究进展

刘冬帅

（大连海事大学 信息科学技术学院，辽宁省大连市 中国 116026）

摘 要： 目前知识图谱（Knowledge Graphs）已经在工业和学术有了许多应用场景，然后，KGs 中常存在不完全性，即 KGs 无法包含全部的已知知识。此时知识库补全技术在应对此种情形时就显得尤为重要, 任何现有的知识图谱都需要通过补全来不断完善知识本身, 甚至可以推理出新的知识.链接预测（Link Prediction）是一项预测实体之间缺失事实的任务，旨在解决知识图谱的不完全性。目前已有的 LP 技术中，基于知识图谱嵌入的 LP 技术在一些基准测试中取得了非常好的表现。本文从 KG 嵌入的角度出发，将解决 LP 问题的方法分为以下三类主要模型：1)张量分解模型;2)几何模型;3)深度学习模型。并对以上链接预测技术研究历程、发展现状和最新进展进行了回顾与探讨, 最后提出了未来该技术需要应对的挑战和相关方向的发展前景.

关键词： 知识图谱；知识图谱补全；链路预测；深度学习；

Review of Knowledge Graph Embedding for Link Prediction

Liu Dong Shuai

(School of Information Science and Technology, Dalian maritime university, Liaoning Dalian 116026, China)

Abstract: At present, there are many application scenarios of knowledge graphs in industry and academia. However, there is often incompleteness in kgs, that is, kgs cannot contain all the known knowledge. At this time, the completion technology of knowledge base is particularly important in dealing with this situation. Any existing knowledge map needs to constantly improve the knowledge itself through completion, and even can infer new knowledge. Link prediction is a task to predict the missing facts between entities, aiming to solve the incompleteness of knowledge map. Among the existing LP technologies, the LP technology based on knowledge map embedding has achieved very good performance in some benchmark tests. In this paper, from the perspective of kg embedding, the methods to solve LP problems are divided into the following three types: 1) tensor decomposition model; 2) geometric model; 3) deep learning model. The research history, development status and the latest progress of link prediction technology are reviewed and discussed. Finally, the challenges and development prospects of this technology in the future are proposed.

Key words: knowledge graph; knowledge graph completion; link prediction; deep learning;

1 概述

知识图谱（KGs）是对现实世界信息的结构化表示。在 KG 中，节点表示实体，比如人和地点;标签表示一种关系，可以把实体联系起来;边是用某种关系连接两个实体的特殊事实。由于 KGs 能够以机器可读的方式对结构化的复杂数据进行建模，因此 KGs 如今广泛应用于各个领域，从问答系统到信息检索，以及基于内容的推荐系统，它们对任何语义 web 项目都至关重要。比较典型的例子有 FreeBase，WikiData ，DBPedia， Yago，谷歌 KG ，和 Facebook Graph Search。这些大规模的 KGs 可以包含数百万个实体和数十亿个事实。

我们将一个 KG 定义为一个具有标签的有向图

KG = (E,R,G):

E:一组表示实体的节点集合;

R:一组表示关系的标签集合;

G⊆E×R×E:一组表示连接实体对事实的一组边的集合。每一个事实都是三元组⟨ h r t⟩ ，h 是头部,r 是关系,和 t 是尾部。

虽然知识图谱能提供高质量的结构化数据，但是大部分开放知识图谱，例如 Freebase，DBpedia 都由人工或者半自动的方式构建，这些图谱通常比较稀疏，大量实体之间隐含的关系没有被充分地挖掘出来。在 Freebase 中，有 71%的人没有确切的出生日期，75%的人没有国籍信息。由于知识图谱具有高质量的结构化数据，是很多人工智能应用的基石，因此，近期很多工作都在研究如何利用机器学习算法更好地表示知识图谱，并以此为基础进行知识图谱补全，从而扩大知识图谱的规模。

知识图谱补全旨在从外部来源(如 Web 语料库)提取新的事实，或者从已存在于 KG 中的事实中推断缺失的事实来实现对现有 KG 的扩充。其中，最新的方法称为链接预测(LP)，是本文分析的重点。

LP 受益于机器学习和深度学习技术的兴起,最近已经成为非常活跃的研究领域。链接预测是利用一个 KG 中已有的事实来推断缺失的事实的任务。这相当于预测出正确的实体完成⟨ h r ?⟩ (尾实体预测)或⟨ ? r t⟩ (头实体预测)。目前绝大多数 LP 模型都使用原始的 KG 元素进行学习称为知识图嵌入的低维表示，然后利用它们来推断新的事实。受 RESCAL 和 TransE 等一些开创性作品的启发，在短短几年时间里，研究人员开发了几十种基于非常不同架构的新模型，大致可以分为以下三类：1)张量分解模型;2)几何模型;3)深度学习模型。

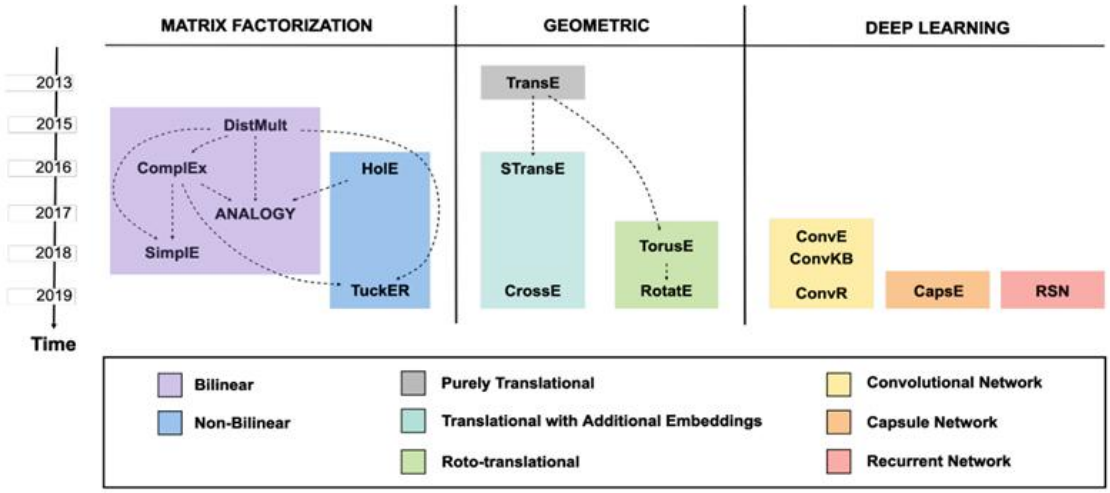

图 1 知识图谱的体系架构

2 链接预测中的深度学习模型

深度学习模型使用深度神经网络来完成 LP 任务。其中，模型中的参数比如权值和偏差等通过从数据中学习获得。深度神经网络通常将每一个单独层都有一组独立的参数，并用非线性的激活函数处理。随着技术的发展，增加了许多类型的层，对输入数据进行不同的操作。例如，稠密的层将把输入数据 X 与权重 w 结合起来，并添加一个偏差 B:W×X+B。

在 LP 任务中，知识图谱的嵌入表示通常是与各层的权值和偏差一起学习的;这些共享的参数使这些模型更有表现力，但可能更复杂，更难训练。根据目前已使用的神经结构，我们确定了三种类型:卷积神经网络，胶囊神经网络和循环神经网络。

2.1 卷积神经网络

这些卷积神经网络模型使用一个或多个卷积层：这些层中的每一层都应用低维滤波器 ω 对输入数据进行卷积（例如在训练事实中嵌入 KG 元素）。结果是一个特征映射，然后通常传递到额外的密集层以计算事实得分。

2.1.1 ConvE

ConvE 将头实体和关系向量重组成矩阵作为卷积层的输入，与不同形状的卷积核进行卷积生成多个特征图， 最终映射成向量和尾实体点积， 得分用于判定三元组正确性。其中输入实体和关系之间的交互作用由卷积层和完全连接层建模。该模型的主要特点是分数由二维形状嵌入的卷积来定义。

ConvE 评分函数如下：

$$\psi_r(\mathbf{e}_s,\mathbf{e}_o) = \quad f(\mathrm{vec}(f([\overline{\mathbf{e}_s};\overline{\mathbf{r}_r}]*\omega))\mathbf{W})\mathbf{e}_o, \qquad (1)$$

ConvE 损失函数如下：

$$\mathcal{L}(p,t) = -\frac{1}{N}\sum_i (t_i\cdot\log(p_i)+(1-t_i)\cdot\log(1-p_i)), \quad (2)$$

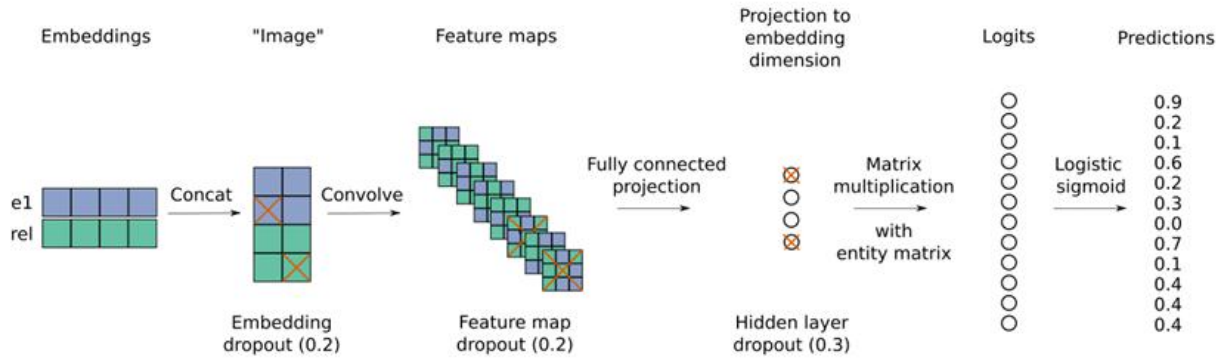

图 2 ConvE 架构图

2.1.2 ConvKB

ConvKB 模型将三元组矩阵$[v_h, v_r, v_t]$作为输入，和不同滤波器进行卷积操作，通过打分函数得到每个三元组的得分，作为判断三元组正确的依据。上述两个模型均利用卷积神经网络提取实体和关系的嵌入表示，捕获实体间的复杂关系，适用于大规模知识图谱补全。ConvKB 使用卷积层对知识图谱中三元组信息进行编码，但输入层和输出层神经元过于简单，不能深层地挖掘实体和关系的嵌入表示。结构如下图所示：

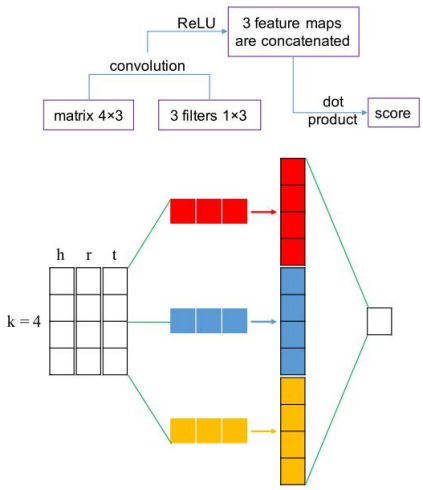

图 3 ConvKB 架构图

得分函数如下：

$$f(h,r,t) = \text{concat}\left(g\left([\boldsymbol{v}_h, \boldsymbol{v}_r, \boldsymbol{v}_t] * \boldsymbol{\Omega}\right)\right) \cdot \mathbf{w} \tag{3}$$

损失函数如下:

$$\mathcal{L} = \sum_{(h,r,t)\in\{\mathcal{G}\cup\mathcal{G}'\}} \log\left(1+\exp\left(l_{(h,r,t)} \cdot f(h,r,t)\right)\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

$$\text{in which, } l_{(h,r,t)} = \begin{cases} 1 \text{ for } (h,r,t) \in \mathcal{G} \\ -1 \text{ for } (h,r,t) \in \mathcal{G}' \end{cases} \tag{4}$$

2.2 胶囊神经网络

胶囊神经网络（CapsNet）模型引入胶囊概念，将胶囊定义为一组用向量表示一种特定类型实体的实例化参数的神经元，它可以表示图像中特定实体(或关系)的各种特征。该模型利用胶囊捕获图像中的实体，通过路由操作指定从上一层胶囊到下一层胶囊的连接，并提出利用胶囊网络对三元组进行补全操作。

由于胶囊使用向量表示实体或关系，取代了以往单个神经元数值的表示，从实体各个属性(维度)进行表示，表示能力更强。结构如下：

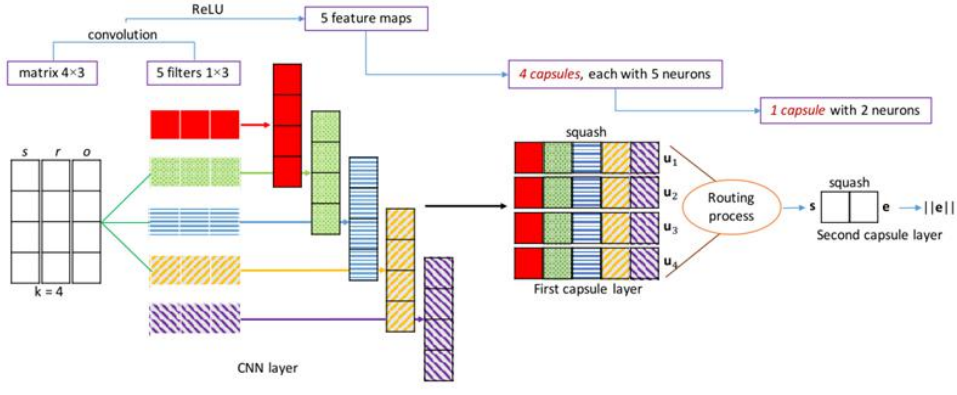

图 4 CapsNet 结构图

得分函数如下：

$$f(s,r,o) = \|\text{capsnet}(g([\boldsymbol{v}_s, \boldsymbol{v}_r, \boldsymbol{v}_o] * \boldsymbol{\Omega}))\| \quad (5)$$

损失函数如下：

$$\mathcal{L} = \sum_{(s,r,o)\in\{\mathcal{G}\cup\mathcal{G}'\}} \log\left(1+\exp\left(-t_{(s,r,o)}\cdot f(s,r,o)\right)\right)$$
$$\text{in which, } t_{(s,r,o)} = \begin{cases} 1 \text{ for } (s,r,o)\in\mathcal{G} \\ -1 \text{ for } (s,r,o)\in\mathcal{G}' \end{cases} \quad (6)$$

2.3 循环神经网络（RNNs）

这类模型使用一个或多个递归层来分析从训练集中提取的整个路径（事实序列），而不是仅仅单独处理单个事实。RNNs 的目的使用来处理序列数据。在传统的神经网络模型中，是从输入层到隐含层再到输出层。层与层之间是全连接的，每层之间的节点是无连接的。但是这种普通的神经网络对于很多问题却无能无力。

RNNs 之所以称为循环神经网络，即一个序列当前的输出与前面的输出也有关。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。

3. 实验结果

本文所有的实验，以及对每一个模型的训练和评估，都是在服务器环境下进行的，使用的是 88 CPUs Intel Core(TM) i7-3820 at 3.60GH, 516GB RAM 和 4 GPUs NVIDIA Tesla P100-SXM2, 操作系统是 Ubuntu17.10。测试结果如下：其中文中的 5 个模型，分别在 5 个数据集 FB15K，WN18，FB15K-237，WN18RR，YAGO3-10 上进行测试。

		FB15k				WN18				FB15k-237				WN18RR				YAGO3-10			
		H@1	H@10	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10	MR	MRR	H@1	H@10	MR	MRR
Deep Learning Models	ConvE	59.46	84.94	51	0.688	93.89	95.68	413	0.945	21.90	47.62	281	0.305	38.99	50.75	4944	0.427	39.93	65.75	2429	0.488
	ConvKB	11.44	40.83	324	0.211	52.89	94.89	**202**	0.709	13.98	41.46	309	0.230	5.63	52.50	3429	0.249	32.16	60.47	1683	0.420
	ConvR	70.57	88.55	70	0.773	94.56	95.85	471	0.950	25.56	52.63	251	0.346	43.73	52.68	5646	0.467	44.62	67.33	2582	0.527
	CapsE	1.93	21.78	610	0.087	84.55	95.08	233	0.890	7.34	35.60	405	0.160	33.69	55.98	**720**	0.415	0.00	0.00	60676	0.000
	RSN	72.34	87.01	51	0.777	91.23	95.10	346	0.928	19.84	44.44	248	0.280	34.59	48.34	4210	0.395	42.65	66.43	1339	0.511

图 5 实验结果

4. 总结与展望

本文对基于 KG 嵌入的 LP 模型进行了广泛的比较分析。研究了 5 个代表不同技术和架构的 L-P 模型，并在文献中最流行的 5 个数据集上分析了它们的效率和有效性。

本文引入了一组表征训练数据的结构特性，并展示了强有力的实验证据，证明它们对预测性能具有重要影响。在这样做的时候，我们调查了使模型能够令人满意地运行的情况，同时确定了研究仍有改进空间的领域。

本文深入讨论了目前的评价做法，证实它们可以依赖不同的低水平条件，产生优秀的表现结果，不过在某些情况下，会产生误导性的结果。由此分析了使模型对这些策略最敏感的组成部分，为未来的研究提供了有用的观察结果。

参 考 文 献

[1] D. Q. Nguyen, T. D. Nguyen, D. Q. Nguyen, and D. Q. Phung. A novel embedding model for knowledge base completion based on convolutional neural network. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT,New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), pages 327 - 333, 2018.

[2] X. Jiang, Q. Wang, and B. Wang. Adaptive convolution for multi-relational learning. In Proceedings of the 2019 Conference o-f the North AmericanChapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019,Volume 1 (Long and Short Papers), pages 978 - 987, 2019. URL https://www.aclweb.org/anthology/N19-1103/.

[3] T. Detmers, P. Minervini, P. Stenetorp, and S. Riedel. Convolutional 2d knowledge graph embeddings. In AAAI Conference on Artifcial Intelligence,2018.

[4] D. Q. Nguyen, T. Vu, T. D. Nguyen, D. Q. Nguyen, and D. Q. Phung.

A capsule network-based embedding model for knowledge graph completion and search personalization. In NAACL-HLT (1),pages 2180 - 2189. Association for Computational Linguistics, 2019

[5] J. J. Hopfeld. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences, 79(8):2554－2558, 1982.

[6] L. Guo, Z. Sun, and W. Hu. Learning to exploit long-term relational dependencies in knowledge g-raphs. In ICML, volume 97 of Proceedings of Machine Learning Research, pages 2505－2514. PMLR, 2019.