基于 MDistMult 模型的新冠科研抗病毒药物图谱的链接预测

王维川，谢志文，赵焜松，刘进*

武汉大学计算机学院，湖北武汉，430072
{tandaocmm,xiezhiwen,kszhao,jinliu}@whu.edu.cn

摘要 链接预测任务，需要通过知识图谱中给定的三元组来预测未知的实体间的关系。针对新冠科研抗病毒药物图谱的链接预测任务，本文基于 DistMult 模型进行改进设计了 MDistMult 模型，来预测新冠科研抗病毒药物图谱中实体间的未知关系。改进后的 MDistMult 模型同其他已有的知识图谱嵌入模型比较，在新冠科研抗病毒药物图谱的链接预测任务验证集上的 MRR 指标上在取得了 0.244 的最好成绩，在测试集上 Top10 结果的 MRR 达到 0.2384, 排名第一。除此之外，我们提出的 MDistMult 模型在其他链接预测相关指标上也取得了最好的结果，同时在对实体嵌入维度的探究实验中，我们证明了我们提出的 MDistMult 模型具有良好的扩展性。

关键词: 知识图谱，链接预测，DistMult，MDistMult

1 引言

在互联网软硬件相关技术飞速发展的今天，知识图谱作为承载底层海量知识并支持上层智能应用的重要载体，在智能时代中扮演了极其重要的角色。同时在 2020 年，新冠病毒疫情爆发，给社会和人类带来不小的灾难。在 OpenKG 总体组织和协调下，部分相关企业院校使用自动化的技术，以新型冠状病毒为核心构建了包括新冠百科、健康、防控等多个高质量的知识图谱。本文针对 CCKS 的第三个任务：链接预测任务设计对应的 MDistMult 模型，推断实体之间存在的关系，以及实体通过特定关系可以链接到哪些实体。

知识图谱中的链接预测任务是知识图谱嵌入的应用之一。知识图谱嵌入模型中，有人们所熟知的翻译系列知识表示模型，把关系当做头尾实体之间的平移，例如 TransE [1],TransH [2],TransR [3] 等，通过设计评分函数对三元组进行训练，最后得到三元组中头实体、关系和尾实体的向量表示；还有

* 通讯作者

双线性模型，如 RESCAL [4]、DistMult [5]、ComplEx [6] 等，这些模型将关系用矩阵进行表示，在训练时，实体和关系的信息可以进行深层次交互，非常具有表现力；随着神经网络的兴起，也有神经网络知识图谱嵌入模型，例如 ConvE [7] 等，ConvE 利用了卷积神经网络的特性来对知识图谱嵌入进行建模；此外，还有旋转模型，即把关系当做头实体和尾实体之间的旋转，如 RotatE [8]，训练后的旋转模型可以表示出多种实体间的关系模式，如对称/反对称、翻转、组合关系。在进行链接预测任务时，可以将实体嵌入表示带入设计的对应模型评价函数进行计算得到对应分数，然后对分数排序，根据任务指标设定计算对应指标的值。

虽然上述的模型在链接预测通用数据集上取得了一定的效果，但是考虑到新冠科研抗病毒药物知识图谱中，实体和关系的数量和分布与通用数据集差距较大，取得不到预想的结果，需要设计新的链接预测模型或者改进现有的知识图谱嵌入模型。

DistMult 模型的主要优点在于计算效率高，参数数量少，但是 DistMult 模型只能解决知识图谱中的对称关系，无法建模非对称的关系模式。本文对 DistMult 模型进行了改进，提出了 MDistMult 模型，将多个 DistMult 模型联合起来训练，并使用共享的尾实体嵌入解决 DistMult 无法处理非对称关系的缺陷。另外, 与 DistMult 利用负采样的训练方式不同，我们使用更高效的基于 softmax 的最大对数似然损失函数，以达到更好的预测效果。

我们参与了 CCKS2020 新冠科研抗病毒药物图谱的链接预测任务，基于 DistMult 模型便于计算的优点改进设计了 MDistMult 模型用于链接预测任务，在 CCKS 提供的新冠科研抗病毒药物图谱测试集上，我们 Top10 结果的 MRR（平均倒数排名）分数取得了 0.2384 的结果。接下来，本文分别对方法、任务定义、实验与实验分析及结论进行介绍。

2 任务定义

对于给定的新冠科研抗病毒药物图谱，其中存在 n 个已知的三元组，任务的目标是给定三元组的头实体和关系，或者给定三元组的关系和尾实体，即（h,r,?）或（?,r,t）的形式，来预测 “?” 所代表的实体。在 CCKS2020 新冠科研抗病毒药物图谱的链接预测中，考虑多种关系预测：病毒-药物关系预测、蛋白-蛋白交互预测，病毒-病毒蛋白交互预测等，这些共同构成全局的关系预测。

3 方法

本文基于 DistMult 模型进行改进，提出了 MDistMult 模型，通过多个 DistMult 的联合训练提高模型的表达能力。本节首先介绍了 DistMult 模型，然后对我们提出的 MDistMult 模型结构以及训练方法进行了详细介绍。

3.1 DistMult 模型

DistMult 模型是一种双线性模型，计算实体和关系在向量空间中潜在语义的可信度。它包含以下三个部分：实体表示，关系表示和参数学习。

实体表示： 定义实体 e_1 和实体 e_2 的输入向量为 $\mathbf{x}_{e_1}$，$\mathbf{x}_{e_2}$，W 为第一层投影矩阵，则需要学习的实体表示 $\mathbf{y}_{e_1}$ 和 $\mathbf{y}_{e_2}$ 可以表示为：

$$\mathbf{y}_{e_1} = f(W\mathbf{x}_{e_1}), \mathbf{y}_{e_2} = f(W\mathbf{x}_{e_2}) \tag{1}$$

其中 f 可以是一个线性或非线性函数，W 是一个参数矩阵，可以随机初始化或者使用预先训练的向量初始化。

关系表示： 获得关系表示需要定义模型对应的评分函数，这里的关系表示为一个对角矩阵 $diag(r)$，令 DistMult 模型的评分函数为 $\mathbf{f}(h,r,t)$, 则对应评分函数表示为：

$$\mathbf{f}(h,r,t) = \mathbf{h}^T \cdot diag(\mathbf{r}) \cdot \mathbf{t} \tag{2}$$

其中 $\mathbf{h}$ 和 $\mathbf{t}$ 为头实体和尾实体的向量表示。

参数学习： 给定一个三元组正样本集合 T, 我们可以通过替换三元组其中的任一个实体构建一个三元组负样本集合 $T^{'}$:

$$\begin{aligned} T' =& \{(e_1', r, e_2) | e_1' \in E, (e_1', r, e_2) \notin T\} \cup \\ & \{(e_1, r, e_2') | e_2' \in E, (e_1, r, e_2') \notin T\} \end{aligned} \tag{3}$$

我们把三元组 (e_1, r, e_2) 的评分函数定义为 $f(e_1, r, e_2)$. 则训练目标是最小化下述损失函数：

$$L(\Omega) = \sum_{(e_1,r,e_2)\in T} \sum_{\left(e_1^{'},r,e_2^{'}\right)\in T'} \max\left\{f(e_1^{'}, r, e_2^{'}) - f(e_1, r, e_2) + 1.0\right\} \tag{4}$$

3.2 MDistMult 模型

DistMult 模型对 RESCAL 模型进行了简化，将关系矩阵表示为对角矩阵，由对角矩阵的性质可知，$\mathbf{h}^T diag(r)\mathbf{t} = \mathbf{t}^T diag(r)\mathbf{h}$。可以看到，DistMult 设计的参数学习目标是让正确实体的分数远远大于错误实体的分数，并且头尾实体实际上默认存在对称关系，这种参数学习目标会导致一个严重问题，即默认知识图谱中三元组存在一个对称三元组，即如果存在三元组 (h, r, t)，那么也存在三元组 (t, r, h)。然而实际实验中发现，新冠科研抗病毒药物图谱中关系为非对称关系，故需要进行改进来来减弱 DistMult 中这种对称关系的影响。

基于上述分析我们提出了 MDistMult 模型，首先，我们将多个 DistMult 模型放在一起进行联合训练，并通过共享尾实体嵌入，可以解决 DistMult 只能处理对称关系的问题，如图 1所示。同时为了更高效的训练模型，我们使用 Softmax 函数计算每个候选实体的概率，并使用最大对数似然目标函数对模型进行优化。另外，我们对每个三元组 (h, r, t) 引入了反转的三元组 $(t, r\_reverse, h)$, 这样 $(?, r, t)$ 等价于预测 $(t, r\_reverse, ?)$。

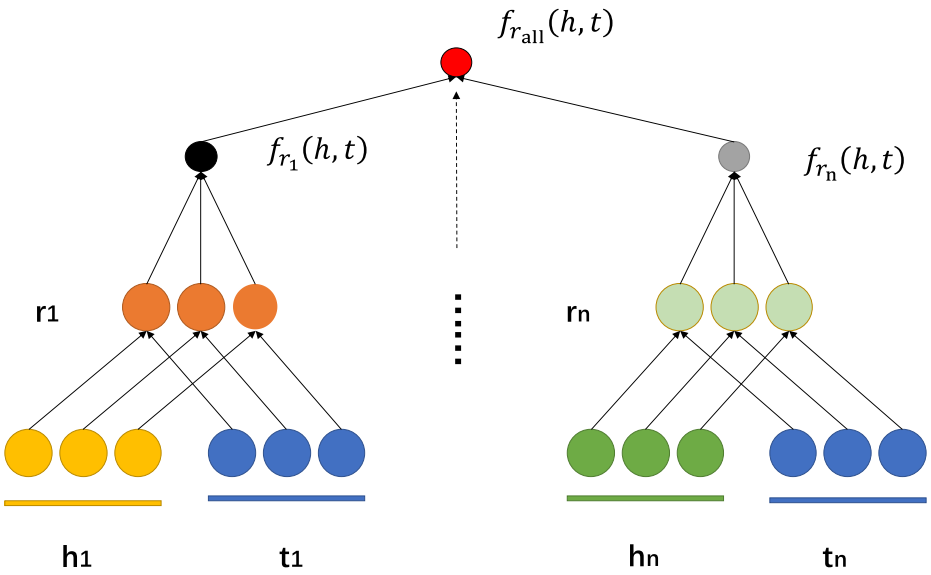

图 1. MDistMult 模型结构图.

改进后的 MDistMult 模型其中每个 DistMult 模型的实体表示和关系表示同上述的 DistMult 模型对应表示一样。只是需要注意的是多个 DistMult 模型的尾实体嵌入共享且训练的关系对角矩阵各不相同，同时参数训练目标进行了改变。

其中，每个 DistMult 模型的参数训练目标改为最大对数似然作为损失函数：

$$P_i(t|h,r) = \frac{exp(f_i(h,r,t))}{\sum_t exp(f_i(h,r,t))} \tag{5}$$

$$loss_i = -\log P_i(t|h,r) \tag{6}$$

对于整体 MDistMult 而言，将每个 DistMult 的评价函数相加可以得到一个整体的评价函数：

$$f_{all}(h,r,t) = \sum_i^N f_i(h,r,t) \tag{7}$$

其中 N 表示 DistMult 模型的个数。其对应的最大对数似然损失函数为：

$$loss_{all} = -\log P_{all}(t|h,r) \tag{8}$$

我们最终将整体的损失函数和每个 DistMult 计算的损失函数相加得到最终的 MDistMult 的损失函数：

$$loss = loss_{all} + \sum_i loss_i \tag{9}$$

4 实验与结果分析

4.1 数据集

数据集为 CCKS2020 新冠科研抗病毒药物图谱的链接预测数据集[1]，新冠科研抗病毒药物图谱包括药物、病毒、病毒蛋白和药物蛋白四类实体，以及药效、产物、结合和相互作用四种关系。整个数据集共包括 7844 个实体。我们将 CCKS2020 官方提供的 40000 个三元组按照 9:1 的比例进一步划分为训练集和验证集，测试集使用 CCKS2020 官方提供的第一次测试集的 4256 个三元组，如表 1所示。

[1] https://www.biendata.xyz/competition/ccks_2020_7_3/

表 1. 新冠科研抗病毒药物图谱实验数据

	训练集	验证集	测试集
实体数量		7844	
三元组数量	36000	4000	4562

表 2. 参数设置

参数名称	参数取值
每个 DistMult 实体嵌入维度	2000
优化函数	Adam, 学习率 0.0005
Dropout	0.5
L2 正则化	$1e^{-5}$
Batch_size	256
DistMult 模型个数	2,3,4

4.2 实验设置

我们依照 CCKS2020 任务三制定的 MRR 指标来进行模型的调参，最终确定的实验设置参数信息如下：实体嵌入维度设置为 2000 维，优化函数我们选择了 Adam，同时学习率是 $5e^{-4}$，我们使用了 Dropout 并将其参数设置为 0.5，同时我们为了加快模型的收敛速度，使用了 L2 正则化，参数为 $1e^{-5}$，我们把数据的 Batch size 设定为 256，DistMult 模型的个数分别取 $N = 2, 3, 4$，表 2展示了我们的各项参数。

4.3 实验结果

除了 CCKS2020 任务三要求的 MRR 指标外，我们还增加了链接预测任务常用的评测指标，包括 MR(正确的实体评分函数的平均排名)，H@1（正确的实体排名在前 1 的比例）、H@3（正确的实体排名在前 3 的比例）和 H@10（正确的实体排名在前 10 的比例）. 具体的实验结果如表 3所示. 表 3中的粗体数字是本实验中各测量指标的最佳值。可以清楚地看到，无论取多大的 N 值，MDistMult 的结果都优于其他知识图谱嵌入模型。除了在 CCKS2020 任务三规定的 MRR 指标中我们提出的 MDistMult 模型取得最好结果 **0.244** 外，我们的模型在其他链接预测相关指标中也取得了最好的成绩。

表 3. 模型结果比对

	MRR	MR	H@1	H@3	H@10
TransE	0.142	487.59	0.040	0.188	0.334
TransR	0.104	892	0.033	0.120	0.245
TransH	0.105	773.39	0.027	0.126	0.263
TransD	0.103	813.33	0.026	0.124	0.257
DisMult	0.090	1274.04	0.038	0.098	0.197
complEx	0.073	1658.56	0.027	0.066	0.172
SimplE	0.084	747.28	0.015	0.093	0.238
ConvE	0.180	758.17	0.102	0.201	0.340
QuatE	0.105	620.31	0.023	0.118	0.280
Tucker	0.096	627.00	0.045	0.098	0.193
RotatE	0.200	521.50	0.113	0.233	0.369
MDistMult(N=2)	0.243	459.17	0.150	0.275	0.429
MDistMult(N=3)	0.244	**455.35**	**0.152**	0.277	**0.432**
MDistMult(N=4)	**0.244**	458.88	0.150	**0.278**	0.430

4.4 实体嵌入维度对实验结果的影响

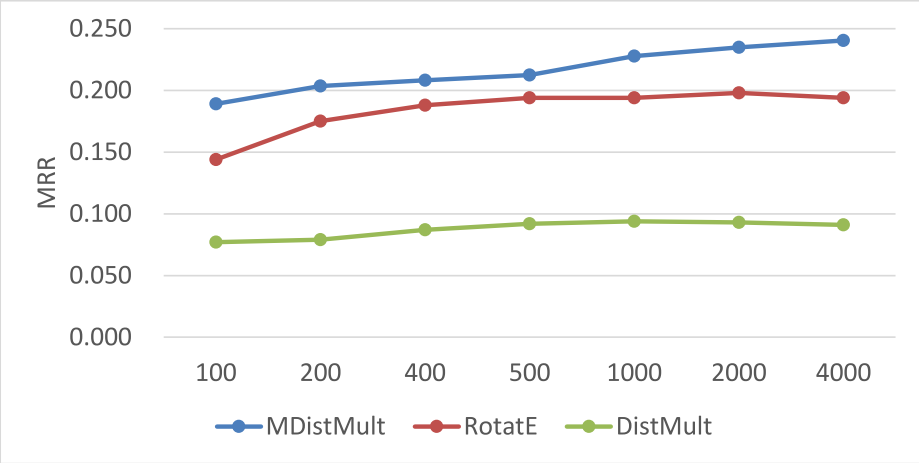

图 2. MDistMult、RotatE、DistMult 三种模型 MRR 值随实体嵌入维度变化表.

除了测试基础的链接预测任务之外，我们还探究了在链接预测过程中实体嵌入维度对模型性能的影响，我们选择 CCKS2020 任务三中给定的实验

指标 MRR 作为对照指标，同时选取了 DistMult 模型，和除 MDistMult 模型之外表现最好的 RotatE 模型与我们的 MDistMult 模型进行比较，具体效果如图 2所示。从图中可以看出，DistMult 模型和 RotatE 模型在实体嵌入分别达到 1000 维度和 2000 维度之前，模型效果随着实体嵌入维度的增加而变好，但是超出 1000 维度和 2000 维度后实体的效果逐渐变差，而我们提出的 MDistMult 模型维度会随着实体嵌入维度的增加而变好，超出 500 维度后增加效果变缓，这说明我们提出的 MDistMult 模型具有较好的扩展性，可以通过维度提升来进一步提高最终的实验结果。

5 结论

本文基于 CCKS2020 提供的新冠科研抗病毒药物图谱，在链接预测任务上基于 DistMult 模型改进设计了一种新的知识图谱嵌入模型 MDistMult 模型，从结果部分我们可以看到，我们提出的模型不仅仅在 MRR 上取得了最好的成绩 0.244，在其他链接预测指标上也取得了最好的效果。同时在实体嵌入维度的实验中，也可以看到不同于其他知识图谱嵌入模型，我们提出的 MDistMult 模型的最终效果会随着实体嵌入维度的增加而提高，这进一步说明了我们模型具有良好的扩展性。考虑到新冠科研抗病毒药物图谱中许多实体属性并没有用到，我们后续会探索如何增加实体属性信息到实体嵌入中，并进一步提升链接预测的效果。

参考文献

1. Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. 2013, 12 2013.
2. Jianlin Feng. Knowledge graph embedding by translating on hyperplanes. 06 2014.
3. Y. Lin, Zhiyuan Liu, M. Sun, Y. Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. *Proceedings of AAAI*, pages 2181–2187, 01 2015.
4. Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. pages 809–816, 01 2011.
5. Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and li Deng. Embedding entities and relations for learning and inference in knowledge bases. 12 2014.

6. T Trouillon, J Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. 10 2016.
7. T Dettmers, Pasquale Minervini, P Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. 02 2018.
8. Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space, 02 2019.