天津大学

本科生毕业论文

题目：领域知识图谱补全系统的设计与实现

学　　院　智能与计算学部

专　　业　计算机科学与技术专业

年　　级　2018 级

姓　　名　张鑫

学　　号　3018216156

指导教师　饶国政

独创性声明

本人声明：所呈交的毕业设计（论文），是本人在指导教师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本毕业设计（论文）中不包含任何他人已经发表或撰写过的研究成果。对本毕业设计（论文）所涉及的研究工作做出贡献的其他个人和集体，均已在论文中作了明确的说明。本毕业设计（论文）原创性声明的法律责任由本人承担。

论文作者签名：

年　　月　　日

本人声明：本毕业设计（论文）是本人指导学生完成的研究成果，已经审阅过论文的全部内容。

论文指导教师签名：

年　　月　　日

摘　要

近年来，由于知识图谱对于知识问答等下游任务的重要性，越来越多的学者开始从事知识图谱领域的研究。但是，目前的知识库并不可能完全包含所有的实体与关系，知识图谱中存在大量缺失的事实，这对下游任务的发展产生了很大的影响，因此，知识图谱补全也逐渐成为人们关注的焦点。

知识图谱补全的目的是根据知识图谱来学习实体与关系的分布式表示，从而对未知三元组进行评分来判断该三元组是否为事实。本文选择了当前表现较好的两个模型 MRotatE 与 CompGCN，并对 MRotatE 的负样本采样策略与 CompGCN 的编码器进行了改进。MRotatE 模型是将基于低维向量空间的分布式表示模型与复杂向量空间结合在一起的联合模型，本文提出了一种全新的负样本采样策略，并采用该策略对 MRotatE 模型进行改进，改进后的模型称为 MRotatE+，该负样本采样策略会记录分类错误的负样本并为其赋予权重，在计算 loss 值时权重会影响该负样本的 loss 最终对模型训练结果产生影响。CompGCN 则是为关系赋予特征向量并将其结合进迭代公式的图神经网络模型，对于该模型，本文提出了一种全新的实体与关系结合编码器，并将改进后的模型称为 CompGCN+。随后，本文对 CompGCN、 CompGCN+、MRotatE 以及 MRotatE+的三元组预测结果进行测试，并基于 MRotatE+模型设计并实现了一个知识图谱补全可视化系统。

经过测试，本文提出的负采样策略优化方法与实体与关系结合编码器确实对模型的性能有所提升。

关键词：知识图谱补全，可视化，图神经网络，负样本采样

ABSTRACT

In recent years, more and more scholars have begun to engage in the field of knowledge graph due to the importance of knowledge graph for downstream tasks. For example, Question answering over knowledge graph relies heavily on knowledge graph. However, the current knowledge base cannot completely include all entities and relationships. There are a lot of lack of facts in the knowledge graph, which has a great impact on the development of downstream tasks. Therefore, the knowledge graph completion has gradually become people's attention.

The purpose of the knowledge graph completion is to learn the distributed representation of entities and relationships based on the knowledge graph. Subsequently, the model uses the distributed representation to score the unknown triples. In this article, I selected two state-of-the-art models, MRotatE and CompGCN, and improved the negative sample strategy of MRotatE and the encoder of CompGCN. The MRotatE model is a combination of low-dimensional vector space and complex vector space. This paper proposed a new negative sampling strategy and applied to the MRotatE model. The improved model is called MRotatE+ in this article. In MRotatE+, the weight of the negative sample will affect the loss of the model. CompGCN is a graph neural network model that embeds feature vectors of relation and combines them into iterative formulas. For this model, this paper proposes a new encoder for combining entities and relationships, and calls the improved model CompGCN+.Subsequently, I compare CompGCN, CompGCN+, MRotatE and MRotatE+ to several state-of-the-art model in link prediction, and design a knowledge graph visualization system to base on the MRotatE+.

After testing, the negative sampling strategy optimization method and the new encoder proposed in this paper does improve the performance of the model.

KEY WORDS: Knowledge graph completion,Visualization, GNN, Negative sample

目　录

第一章　绪论

1.1　研究背景以及意义

知识图谱是一种可以表示实体之间关系的大规模的图结构数据库，内部由大量的三元组组成。目前已经存在一些包含大量事实的知识图谱数据库，例如由结构化数据构成的 Freebase[1]，将字典中的单词整合为单词网络的 WordNet[2]，整合了维基百科与 WordNet 的 YAGO[3]，以及从维基百科中抽取结构化信息并不断更新的 DBpedia[4]。由于其灵活的表示方式，知识图谱常用于其他下游任务，例如自然语言处理[5]，知识问答[6]等。尽管知识图谱有着广泛的应用，目前的知识库仍然缺少大量的事实，这对下游任务有着极大的影响，因此，知识图谱补全技术十分重要。

知识图谱补全领域内部又被划分为了很多的子领域，根据需要补全的知识图谱中是否可能增加未知结点，知识图谱补全又被分为静态知识图谱补全与动态知识图谱补全。

静态知识图谱补全是指被预测三元组的实体与关系已经包含在了训练集中，在预测时并不会引入新的结点与关系，例如 TransE[7]模型。显而易见，这种知识图谱补全方法有着一个极大的缺陷：当知识图谱发生变化时，需要重新对整个知识图谱进行训练，但现实生活中的知识图谱很难不引入新的结点与关系，这就导致了静态知识图谱补全在现实生活中的应用性能较差。

为了解决静态知识图谱补全的缺陷，有学者提出了动态知识图谱补全方法。由于动态知识图谱补全中新增的实体与关系往往是模型从未接触过的数据，因此动态知识图谱补全也可以看成零数据的迁移学习，原有的知识图谱中的实体与关系可以看作源域，新增加的实体与关系可以看作目标域。在迁移学习中，源域与目标域之间需要具有相关性，否则迁移学习的效果会较差，因此动态知识图谱补全往往不只关注知识图谱的结构关系，还将一些额外信息纳入考虑。例如新增加的实体与原本的知识图谱具有三元组关系或者新实体包含丰富文本信息（实体名称，实体描述，类型等），使用这些信息来对新实体与关系构建特征向量，就可以提升模型对新实体与关系的表现能力，从而无需对整个知识图谱重新训练即可获得较好的知识图谱补全效果。

随着动态知识图谱的发展，时序知识图谱也被提了出来，这种方法在传统知识图谱的三元组中加入了第四个元素 t，用来表示该三元组出现的时间，即三元

组(s，r，o)转变为四元组(s，r，o，t)（为了与第四元时间作区分，这里的三元组使用(s，r，o)进行表示，下文中的所有三元组均使用（h，r，t）表示），在增加了时间维度后，知识图谱中的三元组便具有了时效性，这时的知识图谱更接近于真实世界中的知识图谱。目前的时序知识图谱其时间基本单位为一年，因为如果时间单位过小的话，同一时间出现的三元组数量会较少，模型很难学到三元组之间的关联信息。此时，如果将时间相同的三元组提取出来，就可以构成某一时刻的知识图谱，因此，时序知识图谱可以看做一个知识图谱的序列。由于增加了时间维度，四元组预测任务还必须考虑时间维度的信息，如果限制训练集与测试集之间的时间维度没有交集，那么就可以赋予模型预测未来的能力，如果不进行限制，那么就是普通的四元组预测。

1.2 主要研究目的

为了进一步提高知识图谱补全模型的效果，本文中对两个当前效果较好的不同方法类别的模型 MRotatE[8]与 CompGCN[9]进行了改进并对改进结果进行分析。MRotatE 是一个将低维向量空间的分布式表示模型与复数向量空间模型整合起来的联合模型，该模型在训练的过程中既能学习知识图谱中的关系模式，又能学习一对多，多对一与多对多关系。CompGCN 是一个基于图神经网络的模型，与传统的神经网络模型不同，该模型不仅为实体构建特征向量，还为实体与实体之间的关系构建特征向量，通过将关系结合进图神经网络的迭代公式，该模型缓解了图神经网络中常见的过拟合问题。本文提出的改进策略将在下文详细描述。

对于 MRotatE 模型，本文提出了一种改进的负样本采样策略，改进后的新模型在本文中称为 MRotatE+。该策略在随机负样本采样后会记录评分高于正样本的负样本，并根据差异程度来确定该负样本的权重，在下一轮训练中，该负样本有更高的概率被选择并且在计算 loss 值时具有更高的权重，这样就能够改变由于负样本采样策略的随机性导致的模型训练的不稳定性，此外，更高的 loss 惩罚能使模型对错误分类的负样本有更好的学习效果。如果一个负样本再次被选中，在训练结束后会更新该负样本的权重，如果再次训练的评分低于正样本的评分，该负样本的权重会逐渐降低。但是，仅仅记录分类错误的负样本，会使记录负样本权重的参数逐渐增加，训练程序所占的内存逐渐增加，此外，参数数量增加意味着模型训练时间的延长，当训练轮数过多时，训练时间急剧增加。因此，本文还为每一个负样本记录设置了一个生命周期，当负样本记录超出生命周期后，就将其从权重参数中移除。

对于 CompGCN 模型，受到 MRotatE 模型的启发，本文将两种简单编码器进行结合作为新模型 CompGCN+的编码器。通过组合简单编码器，CompGCN+的训练时间远小于 CompGCN，但训练效果并没有降低，反而略有提高。

本文的主要贡献如下：

(1) 提出了一种更优的负样本采样策略，并将该策略应用在 MRotatE 模型上，创建出 MRotatE+模型。

(2) 提出了由两种简单编码器组合的联合编码器，并将该编码器应用在 CompGCN 模型上，创建出 CompGCN+模型。

(3) 设计并实现了一个知识图谱补全可视化系统。

1.3 文章结构描述

本文在序章中分别介绍了该研究的研究背景与意义，以及研究目的，随后在第二章相关理论研究中，本文对知识图谱，知识图谱补全的几种类型以及其经典模型分别进行了详细的描述，此外，该章还介绍了本文使用的知识图谱可视化系统的主要技术：Django 与 d3.js。在第三章基于负采样改进的 MRotatE 模型中首先详细介绍了 MRotatE 模型的结构以及改进策略，接着该章对实验数据集，评价指标，基准线与参数选择进行介绍，然后对实验结果进行分析，最后对改进后的模型进行进一步分析。在第四章基于编码器改进的 CompGCN 模型中首先详细介绍了 CompGCN 模型的结构以及改进策略，接着该章对实验数据集，评价指标，基准线与参数选择进行介绍，然后对实验结果进行分析，最后对改进后的模型进行进一步分析。第五章知识图谱补全系统的设计与实现则是对应用上述模型的知识图谱补全系统的功能与各模块的实现的详细讲解。

第二章　相关理论研究

知识图谱最早是由 google 在 2012 年提出的，其目的是提升搜索引擎的性能，使其更容易获取搜索的相关内容。随着时间的推移，知识图谱的规模也越来愈大，靠人工构建知识图谱逐渐成为一件费时费力的工作，研究者便提出了知识图谱补全这一技术。

2.1　知识图谱补全

知识图谱中的三元组通常表示为（头实体，关系，尾实体），由于三元组具有三部分，知识图谱补全任务也被分为了三个方面：头实体预测，关系预测与尾实体预测。仅给出头实体（尾实体）与关系预测尾实体（头实体）的任务称为头实体（尾实体）预测，仅给出头实体与尾实体预测关系的任务称为关系预测。由于现实生活中实体的数量远远大于关系的数量，因此头实体（尾实体）预测更为重要。

知识图谱补全主要包含两部分：知识图谱嵌入与三元组预测，这两部分都有着大量的研究，但知识图谱嵌入的相关研究更加丰富。知识图谱嵌入的主要目的是将三元组中的实体与关系投影到低维向量空间或者复数向量空间，从而学习其分布式表示，这些分布式表示则提供给后续的三元组预测部分。三元组预测的内容是选择合适的评分函数，使得正确三元组的评分尽可能的高于错误三元组，从而能够在大量的候选三元组中选择出正确三元组，以达到知识图谱补全的目的。

在知识图谱嵌入中，基于低维向量空间的模型主要分为基于距离的模型，例如 TransE，TransH[10]等，以及基于神经网络的模型，例如本文中改进的 CompGCN 模型。而基于复数向量空间的模型主要包括 ComplEx[11]，RotatE[12]等。但是，基于低维向量空间的模型很难学习到不同关系的关系模式（如 a 是 b 的父亲与 b 是 a 的儿子），而基于复数向量空间的模型虽然能够很好的表现关系模式，但很难表示一对多，多对一与多对多关系。

2.2　基于低维向量空间的分布式表示模型

基于低维向量空间的分布式表示模型是最早提出的关于知识图谱嵌入的模型，由于提出时间较早，这一领域已经包含了大量的知识图谱嵌入模型。

2.2.1 基于距离的模型

在基于距离的知识图谱嵌入模型中，TransE 是最经典的模型，其核心思想是对于一个三元组(h,r,t)∈S，应当存在关系 h+r=t，但是对于不同的关系，实体的表示是相同的，因此当出现自反关系，即(t,r,h)∈S 时就会出现 r=0，h=t 的结果，此外当处理一对多，多对一与多对多关系时，相同的 h 与 r 无法指向不同的 t，因此该模型的效果有限。随后 TransH 作为 TransE 模型的改进被提了出来，该模型的核心思想是为每一个关系定义一个超平面，当进行 h+r=t 的运算时，h 与 t 是在关系 r 超平面上的投影，这样就可以保证同一个实体在不同关系中的表示可以不同，而不同实体在同一关系中的表示可以相同，从而解决了 TransE 模型中的问题。

TransR[13]在 TransE 与 TransH 的基础之上，提出了一种新的投影方式，在 TransH 中，作者为每个关系设定了一个超平面，这样实体与关系就被限制在了同一语义空间，实体与关系的分布式表示长度必须相同，而 TransR 的作者为每个关系设定了一个语义空间，实体通过一个投影矩阵投影到关系的语义空间，这样两者的维度就可以不同，从而丰富了语义空间。然而，采用投影矩阵使得 TransR 的训练参数数量急剧增加，其训练的时间复杂度与空间复杂度也随之增加，为此，TransD[14]应运而生，该模型为实体与关系均设定了一个投影向量 P_h，P_r ，P_t，在进行投影时由实体与关系的投影向量相乘来生成投影矩阵，即 $M = P_h * P_r^T$，这样就大大减少了参数数量，此外，通过对实体也设定参数而非单纯使用关系的参数，使得不同实体对于同一个关系可以投影到不同的语义空间，从而进一步提高模型效果。

除了上述模型外，还有通过对向量低维表示的每一维度赋予权重的 TransA[15]，采用额外语义信息的联合对齐模型 L[16]等基于低维向量空间的分布式表示模型。

2.2.2 基于图神经网络的模型

随着深度学习的发展，使用神经网络的训练模型也逐渐显现出了他们的优点，其主要架构是图神经网络。

图神经网络是一种基于深度学习来处理图信息的模型，最早是由 Gori[17]等人在 2005 年提出的，这种模型一般将整张知识图谱作为输入，通过结点与结点之间传递并学习特征信息，从而为知识图谱中的每一个结点生成其特征向量，但是当时的计算水平很难满足这种复杂的神经网络，因此投入该领域的研究人员不多。

随着卷积神经网络的发展以及其在计算机视觉方面所取得的巨大成就，图卷积神经网络逐渐发展了起来，图卷积神经网络主要是通过结点以及其相邻结点的

特征向量来更新结点的特征向量，在多次迭代的过程中改变所有结点的特征向量，这样就可以将已知结点的信息传递到未知结点上，提升未知结点在后续任务中的表现。

虽然图卷积神经网络在处理有关图的问题时具有优异的表现，但是其仍然有一定的局限性，首先，由于图卷积神经网络训练极度依赖邻接结点，因此这种模型只能应用于静态图，对于动态变换的图完全没有办法处理，其次，图卷积神经网络的参数共享策略使其对于不同的邻接结点很难分配权重，因此其对有向图的处理能力有限。为了突破图卷积神经网络的限制，图注意力神经网络逐渐兴起，该模型在图卷积神经网络的基础之上，根据结点与邻接结点的相似度为两者之间的关系赋予权重，在后续的结点迭代更新的过程中，将权重加入其中。

本文中改进的 CompGCN 模型便是基于图神经网络，与普通的图神经网络不同的是，该模型不仅对实体进行编码，还对关系进行编码，在执行图神经网络更新策略时，结点的编码受其相邻结点与对应关系的共同影响，这样就可以摆脱图神经网络训练过程中常见的过拟合问题。

DisenKGAT[18]则是一个图注意力神经网络，区分于普通的图注意力神经网络，该模型不再把每一个相邻结点作为注意力机制的最小单元，而是将相邻结点依据其特征向量相似度分为几个集合，然后对每个集合分配不同的权重。

2.3 基于复数向量空间的模型

由于基于低维向量空间的分布式表示模型大部分使用的是点积，而点积是可交换的，因此当某个关系成立时，其对称关系必然成立，所以基于低维向量空间的分布式表示模型无法很好的学习对称与反对称关系模式，但是，目前的知识图谱数据库中，非对称关系的数量远大于对称关系，从而导致基于低维向量空间的分布式表示模型训练效果不尽人意。

为了解决基于低维向量空间的分布式表示模型难以学习对称与反对称关系模式的问题，由实部与虚部组成的复数向量空间模型被提了出来。因为虚部的存在，复数向量空间模型相较于低维向量空间的分布式表示模型拥有更丰富的特征向量信息，因此能够更好的表示实体与关系。ComplEx 模型是最早提出复数向量空间的模型，在该模型中，特征向量的每一维均为复数，基于低维向量空间的分布式表示模型的点积在该模型中就变为了埃尔米特乘积，由于实部是对称的，虚部是不对称的，埃尔米特乘积就是不对称的，所以该模型能够很好的学习对称与非对称关系模式。但是，仅使用埃尔米特乘积使 ComplEx 对其他关系模式（如组合）的表示效果较差。

RotatE 的思想是将关系作为复数向量空间中头实体与尾实体之间的旋转，通过相位的变化，从而能够表示各种关系模式。在模型 RotatE 中，当且仅当关系 r 中每一个元素的角度均为 0 或者 π 时，该关系为对称关系，当且仅当关系 r_1 与关系 r_2 中每一个元素共轭时，两者为相反关系，当且仅当关系 r_1 与关系 r_2 的每一个元素角度之和等于关系 r_3 时，r_3 为 r_1 与 r_2 的组合。尽管 RotatE 模型能够表现各种关系模式，其仍然具有一定的缺陷。RotatE 模型仅对头实体与尾实体的相位进行了限制，并没有对两者的距离进行限制，这就导致了 RotatE 模型对一对一关系表现较好，而对一对多，多对一与多对多关系表现较差。

2.4 其他知识图谱补全模型与方法

基于低维向量空间的分布式表示模型难以表示复杂的关系模式，而基于复数向量空间的模型难以表示一对多，多对一与多对多关系，为了综合考虑上述两方面，本文改进的 MRotatE 模型被提了出来，该模型借鉴了 TransH 模型与 RotatE 模型，将模型分为了两部分，包含实体部分与关系部分，从而能够同时学习知识图谱中的关系模式与多重关系。但是，该模型的负样本采样策略仍然采用的 RotatE 模型的采样策略，即随机挑选负样本进行自对抗训练，通过对负样本计算 loss 值，与正样本 loss 值相加得到最终的 loss 值，这种随机采样策略就会使模型的训练变得不稳定，因此，本文提出了一种改进的负样本采样策略，对该模型进行改进。

在知识图谱嵌入模型充分发展的同时，也有学者对其他领域进行了研究，例如放弃负样本采样策略而采用全样本采样[19],使用更严格的关系约束[20]等。

2.5 Django

Django 是一个开源的 web 开发框架，作为一个轻量化的开发框架，Django 拥有完善的核心组件，包括 auth，中间件，ORM 系统等。其最早的开发团队是 2003 年至 2005 年的报纸网站维护团队，在经历了大量的重复劳动后，该团队开始将一些常见的代码组装成框架，后来逐渐演变为了 Django 框架。由于 Django 为开源框架，其内容不断由大众丰富，一直发展至今。

2.6 d3.js

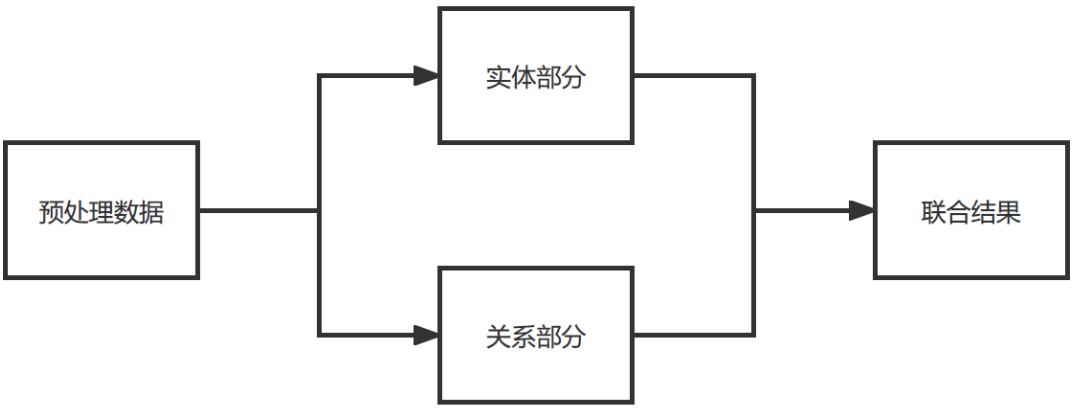
d3.js 是一个使用动态图形进行可视化的 javascript 库，可以在 html 页面上生动的展示知识图谱可视化的效果。相较于其他的可视化工具（如 Echarts）将数据传给相应函数后就可直接生成图表，d3.js 需要首先在 html 界面中规划布局，然后再根据获得的数据在之前规划好的布局上添加图形，因此其复杂度相对较高，但是因为 d3.js 可以自行规划，其编写可视化界面十分灵活。

第三章　基于负采样改进的 MRotatE 模型

3.1　模型架构

MRotatE 模型可以同时学习关系模式与多重关系，其建立在 TransH 与 RotatE 模型的基础上，所以 MRotatE 模型由两部分构成，分别为实体部分与关系部分，整体架构如图 3-1，下面将分别详细介绍。

图 3 – 1 ： MRotatE 模型整体架构。



3.1.1　关系部分

关系部分主要是基于 RotatE 模型，在这一部分，关系被视为头实体到尾实体之间的旋转，其具体过程如下：

首先，头实体 h 与尾实体 t 被投影到复数向量空间，得到其向量表示 $\mathbf{h_c}$，$\mathbf{t_c}$，而关系 r 的向量表示为 $\mathbf{r_c}$，对于一个正确的三元组（h，r，t），应当具有关系：

$$\mathbf{t_c} = \mathbf{h_c} * \mathbf{r_c} \qquad (3-1)$$

其中$|\mathbf{r_{ci}}|=1$，*表示哈达玛积，$\mathbf{r_{ci}}$ 表示 $\mathbf{r_c}$ 中的每一个元素，$|\mathbf{r_{ci}}|=1$ 的目的是使关系仅影响头实体与尾实体之间的相位差。此时，当且仅当关系 r 中每一个元素的角度均为 0 或者 π 时，该关系为对称关系，当且仅当关系 r_1 与关系 r_2 中每一个元素共轭时，两者为相反关系，当且仅当关系 r_1 与关系 r_2 的每一个元素角度之和等于关系 r_3 时，r_3 为 r_1 与 r_2 的组合。

由上述公式就能得到关系部分的评分函数：

$$f(\mathbf{h_c}, \mathbf{t_c}) = \| \mathbf{h_c} * \mathbf{r_c} - \mathbf{t_c} \| \qquad (3-2)$$

关系部分能够很好的表示各种关系模式，但其并没有考虑一对多，多对一与多对多关系的表示。

3.1.2 实体部分

实体部分的主要目的是通过限制头实体与尾实体之间的距离来构建三元组特征，该部分主要是基于 TransH 模型，在这一部分只能中头实体 h 与尾实体 t 被投影到低维向量空间，得到其向量表示 $\mathbf{h_e}$，$\mathbf{t_e}$，而关系 r 的向量表示为 $\mathbf{r_e}$，对于一个正确的三元组（h，r，t），应当具有关系：

$$\| \mathbf{h_e} - \mathbf{t_e} \| = \| \mathbf{r_e} \| \qquad (3-3)$$

此时，头实体与尾实体之间的距离便具有了限制，模型不仅能表示一对一关系，还能表示一对多，多对一与多对多关系。

基于上述公式，我们能够得到实体部分的评分函数：

$$f(\mathbf{h_e}, \mathbf{t_e}) = \| \mathbf{h_e} - \mathbf{t_e} \| - \| \mathbf{r_e} \| \qquad (3-4)$$

3.1.3 整体模型架构

实体部分仅能表示多重关系而关系部分仅能表示关系模式，因此我们将两部分合二为一，得到的评分函数为：

$$\begin{aligned} f(\mathbf{h}, \mathbf{t}) &= f(\mathbf{h_c}, \mathbf{t_c}) + f(\mathbf{h_e}, \mathbf{t_e}) \\ &= \| \mathbf{h_c} * \mathbf{r_c} - \mathbf{t_c} \| + (\| \mathbf{h_e} - \mathbf{t_e} \| - \| \mathbf{r_e} \|) \end{aligned} \qquad (3-5)$$

其中*表示哈达玛积。

3.2 优化负样本采样策略

在 MRotatE 模型中，作者使用了与 RotatE 相同的负样本采样策略，即随机采样后进行自对抗训练，其损失函数如下所示：

$$L = -\log\sigma(\gamma - f(\mathbf{h}, \mathbf{t})) - \sum_{i=1}^{n} p\left(h_i', r, t_i'\right) \log \sigma\left(f\left(\mathbf{h_i'}, \mathbf{t_i'}\right) - \gamma\right) \qquad (3-6)$$

其中 γ 是一个固定参数，在提供的范围内进行选择，σ 表示 sigmoid 函数，(h_i', r, t_i') 表示第 i 个负样本，而$p(h_i', r, t_i')$表示该负样本在当前批次中的权重，其公式如下：

$$p\left(h_i', r, t_i'\right) = \frac{\exp \alpha * f\left(\mathbf{h_j'}, \mathbf{t_j'}\right)}{\sum_{i=1}^{n} \exp \alpha * f\left(\mathbf{h_i'}, \mathbf{t_i'}\right)} \qquad (3-7)$$

其中 α 表示以h_i'作为头向量，以t_i'作为尾向量的正样本三元组的数量。但是，随机采样策略并不是完美的，由于负样本采样策略的随机性，在模型训练过程中很可能无法提供稳定的模型性能，因此，本文提出了一种改进的负样本采样策略，

该策略首先进行随机负样本采样，在经过一轮训练后，对得分$f(\mathbf{h}_i', \mathbf{t}_i')$大于正样本得分 $f(\mathbf{h}, \mathbf{t})$的负样本进行记录，并为其生成权重 $W(h_i', r, t_i')$，其权重生成公式为

$$W(h_i', r, t_i') = \sigma(f(\mathbf{h}_i', \mathbf{t}_i') - f(\mathbf{h}, \mathbf{t})) \quad (3-8)$$

其中 σ 表示 sigmoid 函数，对于尚未记录的负样本，依据 sigmoid 函数的性质，将其权重设为 0.5。随后，在下一轮训练中，提升该负样本的随机概率，即在生成负样本序列后，每个有记录的负样本有$(W(h_i', r, t_i') - 0.5)$（当$(W(h_i', r, t_i') - 0.5) < 0$时，不进行替换）的概率替换掉负样本序列中的随机一个负样本。

在计算负样本的 loss 值时，为其赋予相应权重，改进后的损失函数为

$$L = -\log\sigma(\gamma - f(\mathbf{h}, \mathbf{t})) - \sum_{i=1}^{n} W(h_i', r, t_i')\, p(h_i', r, t_i') \log\sigma(f(\mathbf{h}_i', \mathbf{t}_i') - \gamma) \quad (3-9)$$

在经过一轮训练后，如果有记录的负样本再次被选中，则需要更新该负样本的权重，负样本权重的更新公式为：

$$W_{k+1}(h_i', r, t_i') = (W_k(h_i', r, t_i') + W_k'(h_i', r, t_i'))/2 \quad (3-10)$$

其中 $W_{k+1}(h_i', r, t_i')$表示第 k+1 轮中负样本(h_i', r, t_i')的记录权重，$W_k(h_i', r, t_i')$ 表示第 k 轮中负样本(h_i', r, t_i')的记录权重，$W_k'(h_i', r, t_i')$表示第 k 轮中负样本(h_i', r, t_i')参与训练得到的权重，由公式可知，当负样本在下一轮训练中得分小于正样本后，其权重会逐渐降低。

但是，仅仅记录分类错误的负样本，会使记录负样本权重的参数逐渐增加，训练程序所占的内存逐渐增加，此外，参数数量增加意味着模型训练时间的延长，当训练轮数过多时，单轮训练时间急剧增加。因此，本文还为每一个负样本记录设置了一个生命周期，当负样本记录超出生命周期后，就将其从权重参数中移除。这样就可以保证负样本的权重参数数量保持在一定范围之内。

通过上述方法，能够增加模型的训练稳定性，此外，通过增加错误负样本的 loss 惩罚，能够使模型得到更好的性能。

3.3 基于 MRotatE 与 MRotatE+模型的相关实验

3.3.1 数据集

为了测试模型的效果，本文在两个数据集上对 MRotatE 与 MRotatE+进行了测试，这两个数据集分别是 FB15k-237 与 WN18RR，我接下来将详细介绍这两个数据集，两个数据集的详细数据如表 3-1。

FB15K-237 是 FB15K 的一个子数据集，而 FB15K 是 Freebase 的一个子数据集，该数据集去除了 FB15K 中所有的相反关系，并将验证集与测试集中与训练集直接相连三元组全部删除，因此，其相对于 FB15K 具有更好的测试效果。

WN18RR 是 WN18 的一个子数据集，而 WN18 是 WordNet 的一个子数据集，相对于 FB15K-237 的改动，WN18RR 的改动相对较少，该数据集仅删除了 WN18 中的所有相反关系。

3.3.2 评价指标

在本文中，本文选择三种常用的评价指标，分别是平均排名(MR)，平均倒数排名(MRR)，以及正样本排名前 N 的概率。其中平均排名是指所有正样本评分排名的平均值，平均倒数排名是指所有正样本评分排名倒数的平均值，正样本排名前 N 的概率的概率指的是正样本排名前 N 的次数除以正样本总数，在本文中，N 值的选择为 1，3，10。

表 3－1 ： FB15K-237 与 WN18RR 数据集统计数据

Dataset	entity	relation	train	valid	test
FB15K-237	14541	237	272115	17535	20466
WN18RR	40943	11	86835	3034	3134

3.3.3 基准线

本文选择了一些表现较好的模型与本文改进的模型进行对比，包含 TransE，TransH 这两个基于低维向量空间的分布式表示模型，ComplEx，RotatE 这两个基于复数向量空间的模型，以及 DistMust[21]， ConvE[22]这两个从实体与关系的嵌入方式入手的知识图谱补全模型。

3.3.4 参数选择

对于 MRotatE 与 MRotatE+本文选择 Adam[23]作为模型优化器，其中实体与

关系的特征向量维度从[125，250，500，1000]中进行选择，批量大小从[256，512，1024]中进行选择，固定参数 γ 从[6，9，12，18]中进行选择，此外，在关系部分，本文对关系的相位进行了限制，使其在 0 至 2π 的范围内。

3.3.5 链接预测

表 3 – 2 ： FB15K-237 数据集上测试结果

其中表现最好的数据用粗体表示

Model	MR	MRR	Hit@1	Hit@3	Hit@10
TransE	357	.294	-	-	.465
TransH	**173**	.331	.232	.317	.529
DistMust	254	.241	.155	.263	.419
ComplEx	339	.247	.158	.275	.428
ConvE	244	.325	.237	.356	.501
RotatE	177	**.338**	**.241**	**.375**	.501
MRotatE	197	.332	.235	.364	.519
MRotatE+	195	.334	**.241**	.373	**.535**

表 3 – 3 ： WN18RR 数据集上测试结果

其中表现最好的数据用粗体表示

Model	MR	MRR	Hit@1	Hit@3	Hit@10
TransE	3384	.226	-	-	.501
TransH	3748	.212	.008	.386	.496
DistMust	5110	.430	.390	.440	.490
ComplEx	5261	.440	.410	.460	.510
ConvE	4187	.430	.400	.440	.520
RotatE	**3340**	.467	.428	.492	**.571**
MRotatE	4873	.477	.440	.492	.558
MRotatE+	4631	**.479**	**.443**	**.496**	.567

本文将 MRotatE， MRotatE+与基准线模型进行链接预测，其测试结果如表 3 – 2 与表 3 - 3（所有基准线模型的结果均来自其原文展示结果）。

从表中可以看出 MRotatE+在两个数据集上的表现均优于 MRotatE，这说明

本文提出的负样本采样策略是有效的，此外。本文测试的 MRotatE 在 FB15K-237 上的测试结果低于原文所给出的测试结果而在 WN18RR 数据上的部分参数高于原文所给出的测试结果，这就表示了随机选取的负样本采样策略对模型稳定性具有一定的影响。由测试数据还能看出，MRotatE 与 MRotatE+在 WN18RR 上超出了所有的基准模型，而只有 MRotatE+的部分指标在 FB15K-237 上超出了基准模型，其中 FB15K-237 中的关系数量远远超出 WN18RR，这说明 MRotatE 与 MRotatE+对关系种类较少的数据集分类效果较好而对关系种类较多的数据集分类效果较差。

3.4 进一步分析

3.4.1 内部关系分析

由于 MRotatE 由实体部分与关系部分组成，为了进一步揭示实体部分与关系部分的作用，本文对实体部分与关系部分分别添加权重，其增加权重后的评分函数为：

$$f(\mathbf{h}, \mathbf{t}) = a * f(\mathbf{h_e}, \mathbf{t_e}) + (1 - a) * f(\mathbf{h_c}, \mathbf{t_c})$$

$$= a * (\| \mathbf{h_e} - \mathbf{t_e} \| - \| \mathbf{r_e} \|) + (1 - a) * \| \mathbf{h_c} * \mathbf{r_c} - \mathbf{t_c} \| \quad (3-20)$$

其中 a 从 0.1 增长至 0.9，在两个数据集上的测试结果如图 3-2 与图 3-3。

图 3 - 2

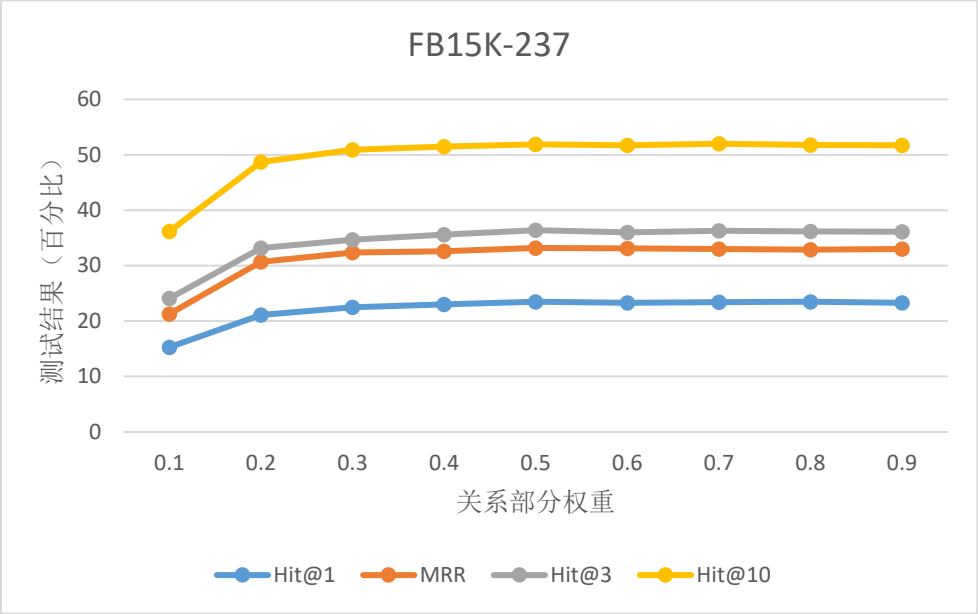
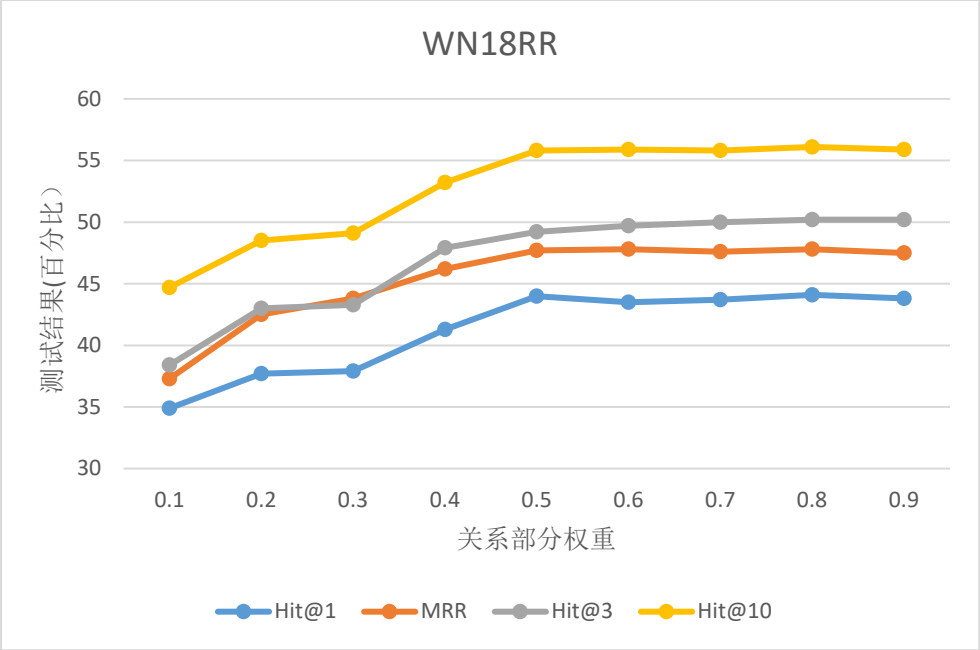


图 3 - 3



从图中可以看出，随着关系部分的权重增加，模型的表现逐渐提升，其中在 0.1 至 0.2 的过程提升最大，对于 FB15K-237 数据集，当 a=0.5 时取得最优表现效果，而在 WN18RR 数据集上，a=0.8 时大部分测试结果优于 a=0.5 时的测试结果，这说明该模型中实体部分与关系部分的重要性会随着数据集的不同而不同。因此，为了在不同数据集上取得最优表现效果，应分别设定其实体部分与关系部分的权重。

3.4.2 优化负采样策略参数分析

由于本文中提出的负采样优化策略需要为负样本权重设置存活周期，因此存活周期的选择会对模型训练所需的时间以及模型结果产生影响。本文在 FB15K-237 数据集上进行测试，其中存活周期 e 在[10，50，100]中选择，测试硬件环境为 RTX2080，测试结果如下：

表 3－4 ： FB15K-237 数据集上测试结果

Model	MR	MRR	Hit@1	Hit@3	Hit@10	Time
MRotatE	197	.332	.235	.364	.519	5
MRotatE+(e=10)	195	.334	.241	.373	.535	12
MRotatE+(e=50)	194	.334	.243	.376	.537	25
MRotatE+(e=100)	192	.335	.244	.380	.541	61

从表中可以看出，随着存活周期 e 的增加，训练的结果的部分指标有所提升，但是训练时间也急剧增加，为了综合考虑训练的时间成本与训练效果，本文选择的

存活周期 e 为 10。

3.4.3 模型稳定性分析

本文中提出的优化负样本采样策略由于能够获取负样本的历史信息，因此相对于原模型的随机负采样，本文提出的优化负采样模型训练更加稳定，为了研究改进后模型与改进前模型稳定性的差异，本文分别使用最佳参数的 MRotatE 与 MRotatE+在 FB15K-237 数据集上训练三次，训练结果如表 3 – 5 与表 3 – 6。

由于两个模型在 MR 与 MRR 上训练结果变化较小，本文选择使用两个模型的 Hit@10 的方差进行比较，为了表示方便，这里对 Hit@10 乘以 1000 化为整数，由训练结果可以得出，MRotatE 的 Hit@10 方差为 32.335，而 MRotatE+的 Hit@10 方差为 7，由此可以看出本文提出的负样本采样策略对模型稳定性有所提升。

表 3 – 5 ： MRotatE 模型三次训练结果

MR	MRR	Hit@1	Hit@3	Hit@10
197	.332	.235	.364	.519
200	.331	.232	.360	.511
198	.332	.234	.363	.522

表 3 – 6 ： MRotatE+模型三次训练结果

MR	MRR	Hit@1	Hit@3	Hit@10
195	.334	.241	.373	.535
196	.334	.240	.375	.534
195	.334	.239	.372	.530

第四章　基于编码器改进的 CompGCN 模型

4.1　模型架构

与大部分的图神经网络模型不同，CompGCN 不仅学习实体的低维向量表示，还学习关系的低维向量表示，其模型原理如图 4-1 与图 4-2，下面将详细介绍该模型。

图 4－1 ： 展示 CompGCN 模型原理所用的知识图谱示例

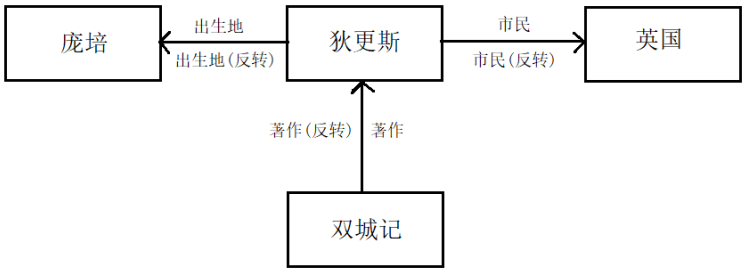
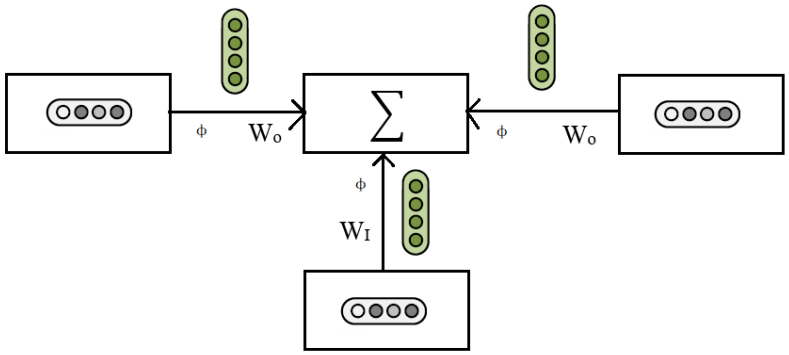


图 4－2 ： CompGCN 模型原理，两种不同颜色的向量分别代表实体与关系的低维分布式表示，ϕ 表示实体与关系之间的编码器，W_O 与 W_I 分别表示原关系与相反关系的参数矩阵，更新的向量结果为所有邻接向量编码后求和。



4.1.1　数据集预处理与初始化

由于 CompGCN 是基于传统图神经网络进行改进的，该模型就具有图神经网络的缺点：只能处理无向图。因此，为了对知识图谱更好的学习，该模型需要对数据集进行预处理。本文选择的 FB15K-237 与 WN18RR 数据集均对其母数据集

进行了处理，将其中的相反关系删除，因此，对于 FB15K-237 与 WN18RR 数据集只需要为每个关系添加了其相反关系，并对每个实体添加了一个指向自身的关系，就可以构建出一张无向图，其公式表示为：

$$E = E \cup \{(h, t, r^{-1}) \mid (h, t, r) \in E\} \cup \{(h, h, T) \mid h \in V\} \quad (4-1)$$

$$R = R \cup R^{-1} \cup T \quad (4-2)$$

其中 E 表示初始三元组集合，h，r，t 分别表示头实体，关系与尾实体，r^{-1} 表示与关系 r 相反的关系，T 表示指向自身的关系，V 表示实体集，R 表示关系集。

由于该模型需要学习关系的低维向量表示，为了避免关系初始化时需要的随机参数过多，该模型选择了一定数量的基向量，然后为每个基向量随机化权重，从而解决了关系过多时初始化所需时间过长的问题，其公式表示为：

$$\mathbf{r}^{k=0} = \sum_{i=1}^{n} \alpha_i * v_i \quad (4-3)$$

其中α_i表示每个关系对于第 i 个基向量的随机化权重，v_i表示第 i 个基向量。

4.1.2 模型算法

由于 CompGCN 不仅学习实体的低维向量表示，还学习关系的低维向量表示，该模型能够缓解图神经网络训练时普遍产生的过拟合现象。为了将关系的低维向量表示应用到图神经网络的训练过程中，该模型使用了如下公式：

$$\mathbf{t}^{k=1} = \varphi(\mathbf{h}^{k=0}, \mathbf{r}^{k=0}) \quad (4-4)$$

其中 φ 是一个 $R^d * R^d = R^d$ 的编码器，$\mathbf{h}^{k=0}$，$\mathbf{r}^{k=0}$ 与 $\mathbf{t}^{k=0}$ 分别表示头实体，关系与尾实体的低维向量表示。

普通的图神经网络模型的更新策略如下所示：

$$\mathbf{t}^{k=1} = f\left(\sum_{(h,r)\in N(v)} W_r * \boldsymbol{h}^{k=0}\right) \quad (4-5)$$

其中 N(V)是所有指向尾实体 t 的头实体与关系组成的集合，这种更新策略很容易受到过拟合的影响，因此，CompGCN 将关系的低维向量表示加入了更新策略，改变后的更新策略如下：

$$\mathbf{t}^{k=1} = f\left(\sum_{(h,r)\in N(v)} W_r * \varphi(\boldsymbol{h}^{k=0}, \boldsymbol{r}^{k=0})\right) \quad (4-6)$$

其中 $\mathbf{h}^{k=0}$ 与 $\mathbf{r}^{k=0}$ 分别表示头实体与关系的初始化低维向量表示，$\mathbf{t}^{k=0}$ 表示更新后尾实体的低维向量表示，此外，W_r 是一个与关系模式有关的参数矩阵，该参数矩阵会随着模型迭代进行学习，其选择依据如下：

当 $r \in R$ 时，$W_r = W_o$.

当 $r \in R_{inv}$ 时，$W_r = W_I$.

当 $r \in T$ 时，$W_r = W_S$.

在对尾实体更新后，还需要对关系进行更新，其更新公式为：

$$\mathbf{r}^{k=1} = W_{rel} * \mathbf{r}^{k=0} \quad (4-7)$$

其中 W_{rel} 是一个将关系投影到实体所在向量空间的投影矩阵，该矩阵会在迭代的过程中进行学习。

对于多次迭代运算，其头实体与关系的更新公式为：

$$\mathbf{t}^{k=n+1} = f\left(\sum_{(h,r)\in N(v)} W_r^n * \varphi\left(\boldsymbol{h}^{k=n}, \boldsymbol{r}^{k=n} \right) \right) \quad (4-8)$$

$$\mathbf{r}^{k=n+1} = W_{rel}^{n} * \mathbf{r}^{k=n} \quad (4-9)$$

其中$\mathbf{h}^{k=n}$表示经过 n 次迭代后的头实体低维向量表示，$\mathbf{r}^{k=n}$表示经过 n 次迭代后的关系低维向量表示，W_r^n表示经过 n 次迭代后的参数矩阵，W_{rel}^{n} 表示经过 n 次迭代后的关系投影矩阵。

4.2 改进编码器

由公式(4 – 4)可知，CompGCN 将关系结合进模型迭代公式需要编码器 φ，而编码器 φ 的选择则关乎着模型的性能。CompGCN 模型分别对三种编码器进行了测试，这三种编码器分别是：

减法编码器：$\mathbf{t}^{k=1} = \varphi(\mathbf{h}^{k=0}, \mathbf{r}^{k=0}) = \mathbf{h}^{k=0} - \mathbf{r}^{k=0}$

乘法编码器：$\mathbf{t}^{k=1} = \varphi(\mathbf{h}^{k=0}, \mathbf{r}^{k=0}) = \mathbf{h}^{k=0} * \mathbf{r}^{k=0}$

循环相关编码器：$\mathbf{t}^{k=1} = \varphi(\mathbf{h}^{k=0}, \mathbf{r}^{k=0}) = \mathbf{h}^{k=0} \circ \mathbf{r}^{k=0}$

受到 MRotatE 模型的启发，实体与关系之间不仅需要考虑角度关系，还需要考虑距离关系，因此，本文将减法编码器与乘法编码器进行结合，得到了一个联合编码器：

$$\mathbf{t}^{k=1} = \varphi(\mathbf{h}^{k=0}, \mathbf{r}^{k=0}) = 0.5*(\mathbf{h}^{k=0} - \mathbf{r}^{k=0}) + 0.5*(\mathbf{h}^{k=0} * \mathbf{r}^{k=0})$$

然后将该联合编码器应用到 CompGCN 模型中，得到 CompGCN+模型

4.3 基于 CompGCN 模型的相关实验

4.3.1 数据集

为了测试模型的效果，本文在两个数据集上对 CompGCN 进行了测试，这两个数据集分别是 FB15k-237 与 WN18RR，关于两个数据集的详细介绍已经包含在 3.3.1 中，本节不再赘述。

4.3.2 评价指标

在本文中，本文选择三种常用的评价指标，分别是平均排名(MR)，平均倒数排名(MRR)，以及正样本排名前 N 的概率。其中平均排名是指所有正样本评分排名的平均值，平均倒数排名是指所有正样本评分排名倒数的平均值，正样本排

名前 N 的概率的概率指的是正样本排名前 N 的次数除以正样本总数，在本文中，N 值的选择为 1，3，10。

4.3.3 基准线

本文选择了一些表现较好的模型与本文改进的模型进行对比，包含 TransE，TransH 这两个基于低维向量空间的分布式表示模型，ComplEx，RotatE 这两个基于复数向量空间的模型，以及 DistMust， ConvE 这两个从实体与关系的嵌入方式入手的知识图谱补全模型，此外，为了更好的体现 CompGCN 模型的效果，在本章中还加入了 R-GCNp[24]与 KBGAN[25]这两个图神经网络模型。

表 4 – 1 ： FB15K-237 数据集上测试结果

其中表现最好的数据用粗体表示

Model	MR	MRR	Hit@1	Hit@3	Hit@10
TransE	357	.294	-	-	.465
TransH	**173**	.331	.232	.317	.529
DistMust	254	.241	.155	.263	.419
ComplEx	339	.247	.158	.275	.428
ConvE	244	.325	.237	.356	.501
RotatE	177	.338	.241	.375	.501
R-GCN	-	.248	.151	-	.417
KBGAN	-	.278	-	-	.458
CompGCN	202	.353	.263	.387	.530
CompGCN+	197	**.355**	**.265**	**.392**	**.542**

4.3.4 参数选择

对于 CompGCN 与 CompGCN+，本文选择 Adam 作为模型优化器，其中实体与关系的特征向量维度为从[200，400，600，800]中选择，批量大小为从[128，256，512]中选择。

4.3.5 链接预测

本文将 CompGCN，CompGCN+与基准线模型进行链接预测，其测试结果如表 4 – 1 与表 4 - 2（所有基准线模型的结果均来自其原文展示结果）。

表 4 – 2 ： WN18RR 数据集上测试结果

其中表现最好的数据用粗体表示

Model	MR	MRR	Hit@1	Hit@3	Hit@10
TransE	3384	.226	-	-	.501
TransH	3748	.212	.008	.386	.496
DistMust	5110	.430	.390	.440	.490
ComplEx	5261	.440	.410	.460	.510
ConvE	4187	.430	.400	.440	.520
RotatE	**3340**	.467	.428	**.492**	**.571**
R-GCN	-	-	-	-	-
KBGAN	-	.214	-	-	.472
CompGCN	3577	.478	.441	.490	.538
CompGCN+	3452	**.480**	**.443**	**.492**	.549

由测试数据可以看出，改进后的 CompGCN+模型在所有指标上均超出了 CompGCN 模型，这说明本文提出的编码器是有效的。此外，CompGCN+在 FB15K-237 上超出了所有基准模型而在 WN18RR 上有部分指标超出了基准模型，其中 FB15K-237 中的关系数量远远超出 WN18RR，这说明 CompGCN 对关系种类较多的数据集分类效果较好而对关系种类较少的数据集分类效果较差。

4.4 进一步分析

表 4 – 3 ：四种编码器在 WN18RR 数据集上测试结果

其中表现最好的数据用粗体表示

编码器	MR	MRR	Hit@1	Hit@3	Hit@10	Time
$\mathbf{h}^{k=0} - \mathbf{r}^{k=0}$	207	.350	.261	.384	.526	6
$\mathbf{h}^{k=0} * \mathbf{r}^{k=0}$	220	.352	.261	.386	.527	6
$\mathbf{h}^{k=0} \circ \mathbf{r}^{k=0}$	202	.353	.263	.387	.530	19
联合编码器	**197**	**.355**	**.265**	**.392**	**.542**	8

对于 CompGCN 模型，由于其将关系的低维向量表示结合进模型的迭代公式中，因此不同的编码器会产生不同效果。为进一步体现联合编码器的效果，本文在 FB15K-237 数据集上分别测试了四种种编码器，其实验结果如表 4-3 所示。

由图中数据可以看出，本文中提出的联合编码器相较于循环相关编码器不仅训练时间更短，表现效果也更优。

第五章　知识图谱补全系统的设计与实现

5.1　功能描述

本文提出的知识图谱补全系统的功能应该包括如下内容：

(1) 拥有文件上传界面，能够以 txt 格式的文件上传待补全知识图谱的三元组。

(2) 具有数据处理能力，能够将上传的 txt 格式的三元组构建出模型能够处理的数据集。

(3) 能够将数据集传递给知识图谱补全模型，并对其进行补全，得到未知三元组的评分。

(4) 将补全后的知识图谱进行可视化展示，展示的内容中补全的关系应当与初始知识图谱中的三元组具有明显区别。

5.2　系统功能模块具体设计与实现

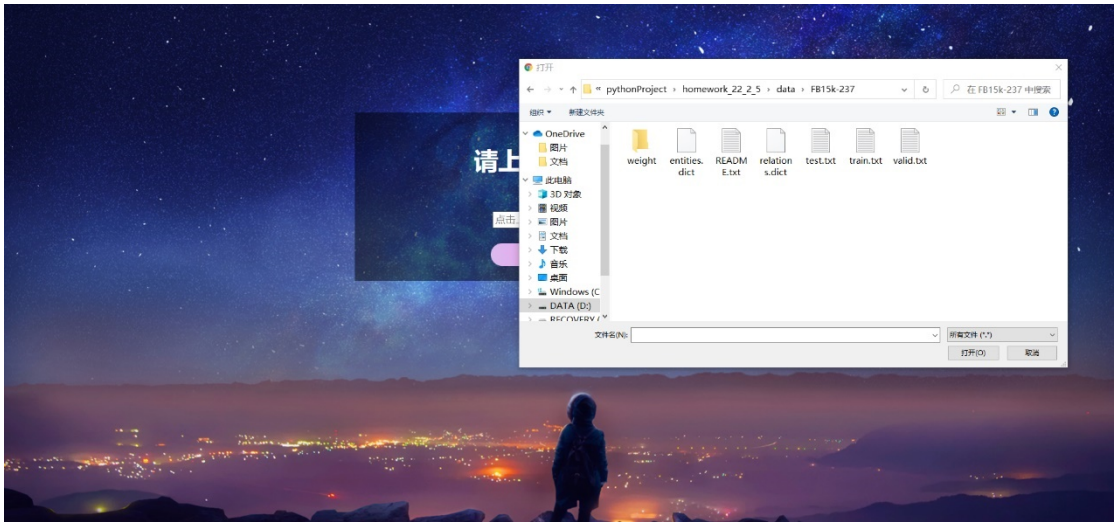
5.2.1　数据上传模块

该模块的目的是设计出一个优秀的人机交互界面，能够以 txt 格式的文件上传待补全知识图谱的三元组。该模块实现所需要的主要技术为 Django 框架，通过部署 Django 使其监听本地窗口，随后访问本地路由，发送请求，Django 框架接收到请求后在 urls.py 文件中找到指定路由进行访问，访问到对应控制器后，控制器返回所指定的前端 html 界面，随后在界面的输入框中选择要上传的文件，点击上传按钮后，通过 form 表单上传 Django 中的后端处理程序，将上传的文件存储到服务器端，其效果图如下图所示：

图 5 – 1 ：文件上传界面

点击上传文件后可以进行文件选择：

图 5 – 2 ： 文件选择界面



选择想要上传的文件后，会显示上传的文件名：

图 5 – 3 ： 选择文件后的界面

点击上传按钮即可上传文件。

5.2.2 数据处理模块

该模块的目的是通过后端处理程序能够从上传的 txt 文件中读取出三元组，并根据模型训练所需的要求生成指定的数据格式。该模块实现的主要技术为 python 对于 txt 文件的读取。对于本文使用的 MRotatE+模型，数据处理模块首先要把上传的数据整理为 MRotatE+模型能够读取的数据格式，再将数据传给知识图谱补全模块。

在进行三元组训练的过程中，MRotatE+模型首先将实体与关系进行编号，然后将初始三元组转变为对应的编号三元组，这样在训练的过程中就能节约大量的内存空间。因此，首先将上传的 txt 文件读取到内存中，接着分别统计其中的实体与关系的数量，然后构建实体与关系的映射字典，最后再将上传的三元组转变为对应的编号三元组进行训练。

5.2.3 知识图谱补全模块

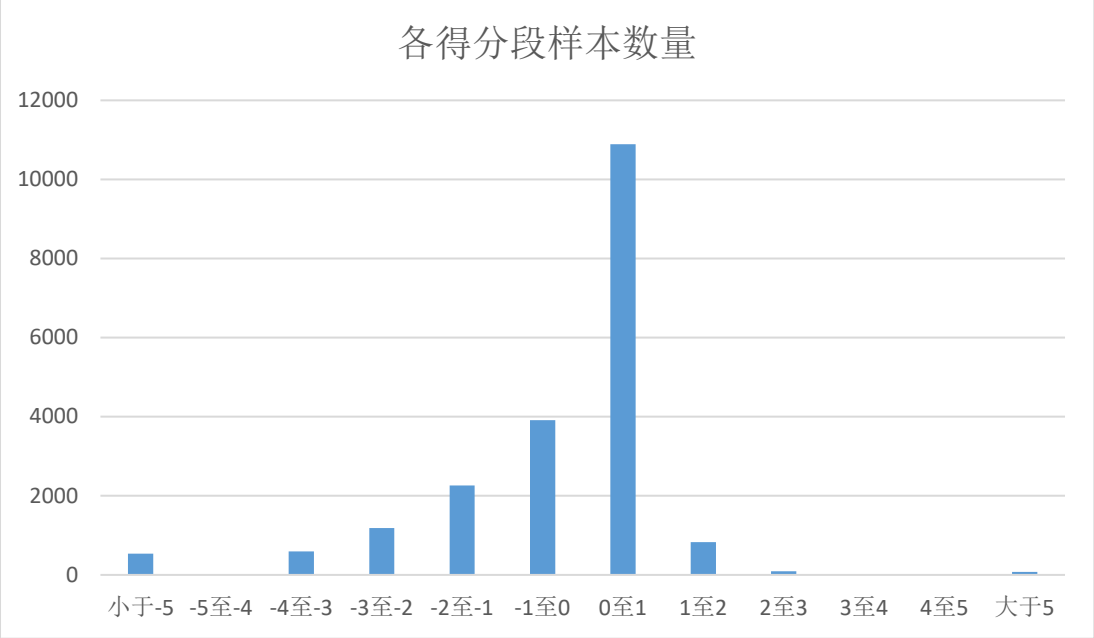
该模块的主要目的是将经过数据处理模块处理后的训练集，输入到知识图谱补全模型中，对其进行训练。该模块实现的主要技术为 pytorch，将数据输入到知识图谱补全模型后，会得到上传数据集中不包含的三元组的得分，由于本系统对学习的知识图谱并没有先验知识，因此需要计算所有不包含的三元组的得分。

在本文中提出了两个改进模型 MRotatE+与 CompGCN+，但 CompGCN+仅对无向图有较好的处理性能，对于知识图谱补全任务来说，其表现不如 MRotatE+模型，因此该模块选择 MRotatE+作为知识图谱补全模型。

随后，我们选择评分介于-1 与 1 之间的未知三元组作为补全的三元组，选择

该评分的原因是本文使用 MRotatE+模型对 FB15K-237 中测试集的所有正样本得分进行统计，其结果如下图所示：

图 5－4：各得分段正样本数量



其中 72%正样本的评分均介于-1 与 1 之间，因此选择评分介于-1 与 1 之间的未知三元组作为补全的三元组。

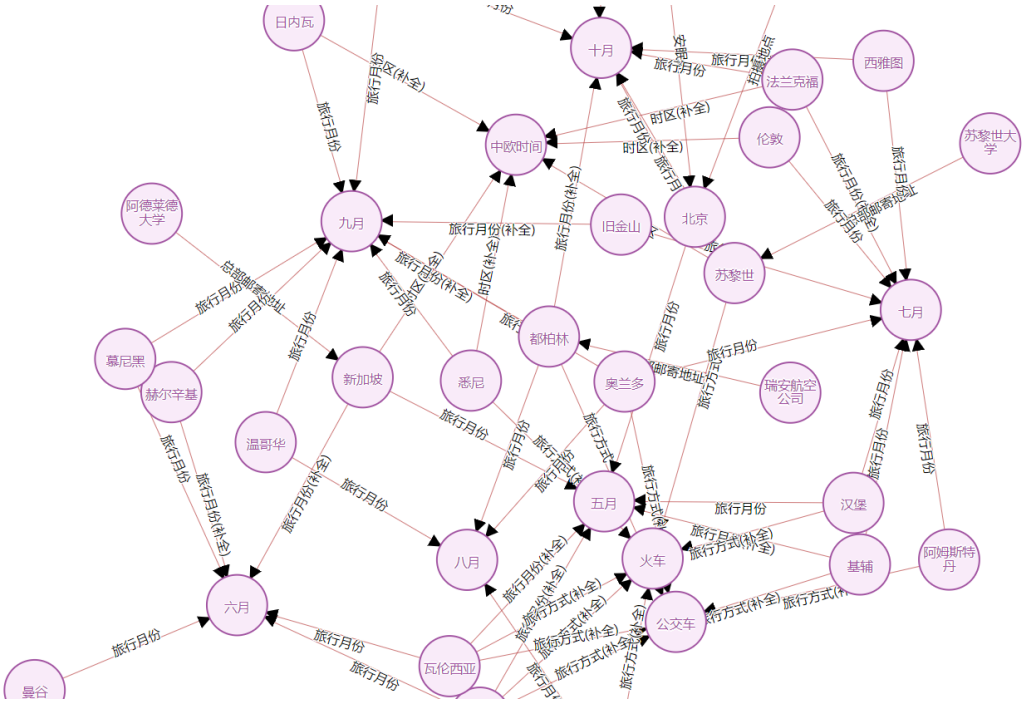
5.2.4 知识图谱可视化模块

该模块的主要目的是对补全后的知识图谱进行可视化，并使补全生成的三元组与原始三元组产生明显的区别，从而更直观的看到知识图谱补全效果。该模块实现的主要技术为 d3.js。在知识图谱补全模型输出结果后，选择评分大于-1 的未知三元组与原始三元组共同组成前端可视化所需的数据格式，即包含头实体，关系与尾实体的 json 格式。

由于本文进行可视化的数据集 FB15K-237 中仅包含实体代码而不包含具体的实体名称，为了可视化更加直观，本文使用谷歌提供的实体代码与 wiki 网址对应文件来获取具体实体名称，由于谷歌提供的对应文件中并不包含所有的实体，所以获取到的具体实体名称数量为 11479，占 FB15K-237 数据集中实体数量的 79%。对于对可获取的实体，本文根据其 wiki 网址编写了网络爬虫来生成对应的中文实体名称。此外，对于补全生成的未知三元组，其关系在可视化时会标明补全字样，从而实现与原始三元组的区别。生成 json 数据后，通过 django 的 render 函数将 json 数据传递给前端界面，前端调用 d3.js 库分别设定实体与关系的样式

并根据 json 数据构建相应的知识图谱，然后进行渲染后得到可视化界面。可视化界面的效果如下图：

图 5 – 5 ： 知识图谱补全可视化界面



第六章　总结与展望

本文基于 MRotatE 模型与 CompGCN 模型提出了两个改进模型 MRotatE+与 CompGCN+，其中 MRotatE 模型是将实体与关系分别嵌入到低维向量空间与复数向量空间中，从而能够同时学习知识图谱中的关系模式与多重关系，而 CompGCN 模型则是将关系也进行编码，并将其合并入图神经网络的迭代公式，从而缓解图神经网络中普遍存在的过拟合现象。对于 MRotatE 模型，本文还分别对其实体部分与关系部分添加权重，以测试两部分的重要性，而对于 CompGCN 模型，本文分别测试了不同的编码器对模型效果的影响。

对于 MRotatE 模型，本文提出的一种改进的负样本采样策略，该策略在随机负样本采样后会记录评分高于正样本的负样本，并根据差异程度来确定该负样本的权重，在下一轮训练中，该负样本有更高的概率被选择并且在计算 loss 值时具有更高的权重，这样就能够改变由于负样本采样策略的随机性导致的模型训练的不稳定性，此外，更高的 loss 惩罚能使模型对错误分类的负样本有更好的学习效果。为了解决该负样本采样策略在训练轮数过多后导致负样本权重记录过多，从而极大增加训练程序的内存占用与训练时间的问题，本文还为每个负样本权重记录增加了存活周期，当负样本记录超出存活周期后，就将该负样本记录从负样本权重参数中删除。通过实验结果可以看出，随着存活周期的增加，模型的表现效果逐渐提升，但其训练时间也急剧增加，因此选择合适的存活周期才能兼顾模型性能与训练时间。

对于 CompGCN 模型，受到 MRotatE 模型的启发，本文提出了将两个简单编码器合并为一个编码器来替换原模型中的复杂编码器。通过组合简单编码器，改进后的模型训练时间远小于原模型，但训练效果不但没有降低，反而有所提升。

最后，本文设计并实现了一个知识图谱补全可视化系统，该系统的主要框架是 Django，其前端可视化界面的主要依赖技术为 d3.js，而内部的知识图谱补全模型为本文改进的 MRotatE+模型。通过该知识图谱补全可视化系统，用户能够直观的感受到知识图谱补全的效果。

由于本文提出的负样本采样策略需要手动设定存活周期，存活周期的选择仍是一个巨大的问题，这一问题或许可以通过根据每个负样本的初始权重来为其分配存活周期，分配的方式可以是一些简单的数学工具，也可以是一个简单的前向神经网络，因此，该负样本采样策略仍需后续的工作进行改进。

参考文献

[1] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. 2008: 1247-1250.

[2] Miller G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.

[3] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th international conference on World Wide Web. 2007: 697-706.

[4] Lehmann J, Isele R, Jakob M, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia[J]. Semantic web, 2015, 6(2): 167-195.

[5] 华一雄. 基于自然语言处理的摩擦学知识图谱构建及软件系统设计[D].上海交通大学,2020.

[6] Zheng W, Cheng H, Yu J X, et al. Interactive natural language question answering over knowledge graphs[J]. Information Sciences, 2019.

[7] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[J]. Advances in neural information processing systems, 2013, 26.

[8] Huang X, Tang J, Tan Z, et al. Knowledge graph embedding by relational and entity rotation[J]. Knowledge-Based Systems, 2021, 229: 107310.

[9] Vashishth S, Sanyal S, Nitin V, et al. Composition-based multi-relational graph convolutional networks[J]. arXiv preprint arXiv:1911.03082, 2019.

[10] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2014, 28(1).

[11] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[C]//International conference on machine learning. PMLR, 2016: 2071-2080.

[12] Sun Z, Deng Z H, Nie J Y, et al. Rotate: Knowledge graph embedding by relational rotation in complex space[J]. arXiv preprint arXiv:1902.10197, 2019.

[13] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Twenty-ninth AAAI conference on artificial intelligence. 2015.

[14] Ji G, He S, Xu L, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers). 2015: 687-696.

[15] Xiao H, Huang M, Hao Y, et al. TransA: An adaptive approach for knowledge graph embedding[J]. arXiv preprint arXiv:1509.05490, 2015.

[16] Wang Z, Zhang J, Feng J, et al. Knowledge graph and text jointly embedding[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1591-1601.

[17] Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains[C]//Proceedings. 2005 IEEE international joint conference on neural networks. 2005, 2(2005): 729-734.

[18] Wu J, Shi W, Cao X, et al. DisenKGAT: Knowledge Graph Embedding with Disentangled Graph Attention Network[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 2140-2149.

[19] Li Z, Ji J, Fu Z, et al. Efficient Non-Sampling Knowledge Graph Embedding[C]//Proceedings of the Web Conference 2021. 2021: 1727-1736.

[20] Li M, Sun Z, Zhang S, et al. Enhancing knowledge graph embedding with relational constraints[J]. Neurocomputing, 2021, 429: 77-88.

[21] Yang B, Yih W, He X, et al. Embedding entities and relations for learning and inference in knowledge bases[J]. arXiv preprint arXiv:1412.6575, 2014.

[22] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2d knowledge graph embeddings[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).

[23] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[24] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[C]//European semantic web conference. Springer, Cham, 2018: 593-607.

[25] Cai L, Wang W Y. Kbgan: Adversarial learning for knowledge graph embeddings[J]. arXiv preprint arXiv:1711.04071, 2017.

致　谢

首先我要感谢我的论文指导老师饶国政老师，饶国政老师为我指引了论文的研究方向并在我撰写论文的过程中提供了悉心的指导，提出了许多改善的意见。

其次，我还要感谢一同由饶国政老师指导的 2018 级天津大学智能与计算学部的学生，在与大家互帮互助，共同交流的过程中，我度过了一段难忘的时光。

最后，谢谢论文评阅老师的辛苦工作以及我的家人朋友们对我的支持与鼓励。