



学校代码 10459  
学号或申请号 201912172014292  
密 级 公 开

# 郑州大学

## 硕士学位论文

图卷积网络的知识表示学习研究

作者姓名: 方海川

导师姓名: 叶阳东 田侦

学科门类: 工 学

专业名称: 软件工程

培养院系: 计算机与人工智能学院

完成时间: 2022年4月

A thesis submitted to  
Zhengzhou University  
for the degree of Master

**Research on Knowledge Representation Learning  
based on Graph Convolutional Networks**

By Haichuan Fang

Supervisor: Yangdong Ye and Zhen Tian

Software engineering

School of Computer and Artificial Intelligence

April 2022

## 学位论文原创性声明

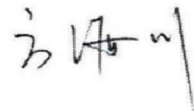
本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究  
所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集  
体已经发表或撰写过的科研成果。对本文的研究作出重要贡献的个人和集体，均  
已在文中以明确方式标明。本声明的法律责任由本人承担。

学位论文作者： 

日期：2022年5月28日

## 学位论文使用授权声明

本人在导师指导下完成的论文及相关的职务作品，知识产权归属郑州大学。  
根据郑州大学有关保留、使用学位论文的规定，同意学校保留或向国家有关部门  
或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权郑州大学  
可以将本学位论文的全部或部分编入有关数据库进行检索，可以采用影印、缩印  
或者其他复制手段保存论文和汇编本学位论文。本人离校后发表、使用学位论文  
或与该学位论文直接相关的学术论文或成果时，第一署名单位仍然为郑州大学。  
保密论文在解密后应遵守此规定。

学位论文作者： 

日期：2022年5月28日

## 摘要

知识图谱是知识库中大量三元组链接而成的数据库，广泛应用于搜索引擎、问答系统、推荐系统等知识驱动任务。知识图谱的不断扩充使得基于知识图谱的知识推理难以有效进行，因此，知识表示学习应运而生。知识表示学习在向量空间中学习实体和关系的低维稠密向量表示，从而对知识图谱进行建模以保存其本质的结构信息和语义信息。知识表示学习在很大程度上缓解了数据稀疏和传统方法面临的计算效率低下问题，进而提升了知识驱动任务的性能。因此，研究知识表示学习具有重要意义和应用价值。

基于三元组的知识表示学习模型在建模知识图谱时总是独立地处理每个三元组，忽略了知识图谱的结构信息。知识图谱的结构信息充分的反映了实体和关系的邻域信息，进一步刻画了实体与关系之间和实体与实体之间的语义交互。因此，本文从知识图谱的结构信息入手，结合实体和关系所处的邻域，分别探索了融合部分邻域信息和融合全部邻域信息知识表示学习方法。研究内容如下：

(1) 针对融合部分邻域信息，提出基于聚合的图卷积知识表示学习方法。首先，使用基于聚合的图卷积网络编码器聚合邻居实体及相连关系的信息；其次，将聚合的表示用于更新中心实体的表示，并且使关系表示进行自我更新。该方法极大缓解了大部分基于三元组的模型在知识表示学习过程中无法有效融入实体所在邻域的信息这一问题。

(2) 针对融合全部邻域信息，提出基于双注意力的图卷积知识表示学习方法。首先，在图卷积网络编码器中设计两个注意力机制同时评估邻域的重要性；其次，根据不同重要性融合邻域信息；最后，根据融合的邻域信息更新实体和关系的向量表示。该方法同时考虑了邻域对实体和关系的表示学习造成的影响，极大促进了实体与关系之间和实体与实体之间的语义交互。

本文对基因本体数据进行建模，结合基因功能相似度分析任务，评估了基于聚合的图卷积知识表示学习方法的性能，同时在标准的链接预测数据集上测试了基于双注意力的图卷积知识表示学习方法的性能。实验结果表明，在知识表示学习中利用图卷积网络融入知识图谱的结构信息能充分地对知识图谱进行建模。

**关键词：**知识图谱；知识表示学习；图卷积网络；结构信息；邻域信息

## Abstract

Knowledge graphs are databases linked by many triples in knowledge bases and are widely used in knowledge-driven tasks such as search engines, question answering, and recommendation systems. The continuous expansion of knowledge graphs makes knowledge reasoning based on knowledge graphs difficult to carry out effectively. Therefore, knowledge representation learning emerges as the times require. Knowledge representation learning learns low-dimensional and dense vector representations of entities and relations in vector spaces, thereby modeling knowledge graphs to preserve their essential structural and semantic information. Knowledge representation learning largely alleviates the problems of data sparsity and computational inefficiency faced by traditional methods, thereby improving the performance of knowledge-driven tasks. Therefore, studying knowledge representation learning is of great significance and application value.

The triple-based knowledge representation learning models always process each triple independently when modeling knowledge graphs, ignoring the structural information of knowledge graphs. The structural information of knowledge graphs fully reflects the neighborhood information of entities and relations, and further describes the semantic interaction between entities and relations and between entities. Therefore, this thesis starts with the structural information of the knowledge graph, combined with the neighborhoods where entities and relations are located, and explores knowledge representation learning methods that fuse partial neighborhood information and fuse complete neighborhood information. The main contributions of this thesis are summarized as follows.

(1) Aiming at fusing partial neighborhood information, an aggregation-based graph convolutional knowledge representation learning method is proposed. First, an aggregation-based graph convolutional network encoder is used to aggregate the information of neighbor entities and connected relations. Second, the aggregated representation is used to update the representation of the central entity, and relation

representations perform self-update. This method dramatically alleviates the problem that most triple-based models cannot effectively incorporate the neighborhood information of entities during the knowledge representation learning process.

(2) Aiming at fusing complete neighborhood information, a dual-attention-based graph convolutional knowledge representation learning method is proposed. First, two attention mechanisms are designed in the graph convolutional network encoder to simultaneously evaluate the importance of the neighborhood. Second, the neighborhood information is fused according to different importance. Finally, the vector representations of entities and relations are updated according to the fused neighborhood information. This method simultaneously considers the influence of neighborhood on the representation learning of entities and relations, which significantly promotes the semantic interaction between entities and relations and between entities.

This thesis evaluates the performance of the aggregation-based graph convolutional knowledge representation learning method on the gene functional similarity analysis task by modeling gene ontology, while testing the performance of the dual-attention-based graph convolutional knowledge representation learning method on standard link prediction datasets. The experimental results demonstrate that incorporating structural information of knowledge graphs using graph convolutional networks in knowledge representation learning can adequately model the knowledge graphs to learn effective entity and relation representations.

**Key Words:** Knowledge graph; Knowledge representation learning; Graph convolutional networks; Structural information; Neighborhood information

## 目录

摘要 .....	IV
Abstract.....	V
目录 .....	VII
符号说明 .....	X
英文缩略语 .....	XI
图目录 .....	XII
表目录 .....	XIII
1 绪论 .....	1
1.1 研究背景及意义 .....	1
1.2 本文研究内容 .....	3
1.3 组织结构 .....	4
2 相关工作 .....	5
2.1 知识表示学习 .....	5
2.1.1 基于三元组的模型 .....	5
2.1.2 融合外部信息的模型 .....	12
2.2 图卷积网络 .....	13
2.2.1 谱域图卷积 .....	13
2.2.2 空间域图卷积 .....	15
2.3 本章小结 .....	17

3	基于聚合的图卷积知识表示学习方法 .....	17
3.1	问题引入 .....	18
3.2	基于基因本体的相关基因功能相似度分析方法 .....	20
3.3	基于聚合的图卷积知识表示学习 .....	21
3.3.1	符号体系 .....	22
3.3.2	模型架构与学习框架 .....	22
3.4	术语间的语义相似度计算 .....	24
3.5	基因间的功能相似度计算 .....	25
3.6	实验设计与结果分析 .....	25
3.6.1	实验数据 .....	25
3.6.2	实验设置 .....	27
3.6.3	对比方法 .....	28
3.6.4	评价指标 .....	28
3.6.5	酵母和人类蛋白质相互作用实验 .....	30
3.6.6	酵母基因表达数据实验 .....	31
3.6.7	CESSM 数据集实验 .....	32
3.6.8	酵母基因生物通路实验 .....	33
3.6.9	消融实验 .....	36
3.7	本章小结 .....	38
4	基于双注意力的图卷积知识表示学习方法 .....	39
4.1	问题引入 .....	39
4.2	基于双注意力的图卷积知识表示学习 .....	41
4.2.1	符号体系 .....	42
4.2.2	模型架构与学习框架 .....	42
4.3	实验设计及结果分析 .....	47
4.3.1	实验数据 .....	47
4.3.2	实验设置 .....	47



目录

---

4.3.3	对比方法 .....	48
4.3.4	评价指标 .....	49
4.3.5	整体结果 .....	50
4.3.6	建模不同类型关系的 Hit@10 结果.....	52
4.3.7	收敛性分析 .....	54
4.3.8	参数敏感性分析 .....	54
4.3.9	消融实验 .....	56
4.3.10	案例分析 .....	57
4.4	本章小结 .....	57
5	总结与展望 .....	59
5.1	本文工作总结 .....	59
5.2	研究展望 .....	60
	参考文献 .....	61
	个人简历、在学期间发表的学术论文与研究成果 .....	67
	致谢 .....	68

## 符号说明

$\mathcal{G}$	知识图谱
$\mathcal{E}, \mathcal{R}, \mathcal{T}$	实体, 关系及三元组集合
$(h, r, t)$	头实体、关系及尾实体组成的三元组
$(\mathbf{h}, \mathbf{r}, \mathbf{t})$	头实体、关系及尾实体的向量表示
$\mathbf{M}_r$	关系的矩阵表示
$f(h, r, t)$	得分函数
$\sigma(\cdot)$	非线性激活函数
$\text{vec}(\cdot)$	向量化
$\omega$	卷积核
$\mathcal{L}$	损失函数
$\mathbb{R}$	实数空间
$\mathbb{R}^d$	$d$ 维实数向量空间
$\mathbb{R}^{m \times n}$	$m \times n$ 维的实数矩阵空间
$\mathbb{C}^d$	$d$ 维复数向量空间
$\langle \mathbf{a}, \mathbf{b} \rangle$	向量 $\mathbf{a}$ , $\mathbf{b}$ 的内积
$\mathbf{a} \circ \mathbf{b}$	向量 $\mathbf{a}$ , $\mathbf{b}$ 的哈达玛积
$*$	卷积运算
$\bar{\mathbf{t}}$	向量 $\mathbf{t}$ 的共轭或重塑表示

## 英文缩略语

KB	Knowledge Base
KG	Knowledge Graph
RDF	Resource Description Framework
KRL	Knowledge Representation Learning
KGE	Knowledge Graph Embedding
CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
GNN	Graph Neural Network
GCN	Graph Convolutional Network
GAT	Graph Attention Network
GO	Gene Ontology
CC	Cellular Component
BP	Biological Process
MF	Molecular Function
IC	Information Content
BMA	Best Match Average
PPI	Protein-Protein Interaction
GOGCN	Graph Convolutional Network on Gene Ontology for functional similarity analysis of genes
D-AEN	Learning knowledge graph embedding with a dual-attention embedding network

## 图目录

图 1.1	以‘Steve Jobs’为中心的简要知识图谱.....	2
图 3.1	以‘Mark Zuckerberg’为中心的简要知识图谱.....	19
图 3.2	GOGCN 算法结构图.....	22
图 3.3	单一生物通路中基于基因功能的基因分类结果.....	35
图 3.4	不同聚合操作对模型性能的影响结果.....	37
图 3.5	关系的不同处理方式对模型性能的影响结果.....	37
图 3.6	编码器和解码器对模型性能的影响.....	38
图 4.1	以‘Oliver Stone’为中心的简要知识图谱.....	40
图 4.2	D-AEN 模型中的单层编码器模型示意图.....	42
图 4.3	在 FB15K-237 和 WN18RR 数据集上的收敛性分析.....	54
图 4.4	在 WN18RR 和 Kinship 数据集上探究初始嵌入维度对模型性能的影响结果.....	55
图 4.5	在 Kinship 数据集上探究注意力头数量和负采样数量对模型性能的影响.....	55

## 表目录

表 3.1	GOGCN 模型的最佳参数配置 .....	27
表 3.2	酵母蛋白质互作数据实验的 AUC 值 .....	31
表 3.3	人类蛋白质互作数据实验的 AUC 值 .....	31
表 3.4	酵母基因表达数据与基因功能相似度间的皮尔逊相关系数 .....	32
表 3.5	蛋白质间的 Pfam 相似度与基因功能相似度间的皮尔逊相关系数 .....	33
表 3.6	生物通路‘L-tyrosine degradation III’中的基因及其详细信息 .....	34
表 3.7	实验所用生物通路集合中 KEGG 通路及其详细信息 .....	35
表 3.8	基于 KEGG 通路的集判别能力实验结果 .....	36
表 4.1	数据集信息 .....	47
表 4.2	D-AEN 模型在各数据集上的最佳参数配置 .....	48
表 4.3	在 FB15k-237 数据集上的实验结果 .....	50
表 4.4	在 WN18RR 数据集上的实验结果 .....	51
表 4.5	在 Kinship 数据集上的实验结果 .....	52
表 4.6	在 FB15K-237 数据集上对不同类型的关系进行建模的实验结果 .....	53
表 4.7	在 WN18RR 数据集上对不同类型的关系进行建模的实验结果 .....	53
表 4.8	消融实验结果 .....	56
表 4.9	在 FB15k-237 数据集上进行案例分析的实验结果 .....	57

# 1 绪论

## 1.1 研究背景及意义

知识库 (Knowledge Base, KB) 以结构化的形式存储人类知识, 通常由机器自动标注和专家标注等方法构建而来, 其包括基本事实、规则和其它有关信息, 常用于知识求解任务如智能问答、关系抽取和知识推理。随着大数据时代的飞速发展, 互联网产生的数据信息呈指数级增长。为了有效的存储与利用其中包含的知识数据, 国内外互联网公司构建了大量面向领域的知识库如 Wikidata<sup>[1]</sup>, Freebase<sup>[2]</sup>, DBpedia<sup>[3]</sup>, Google Knowledge Graph, YAGO<sup>[4]</sup>, WordNet<sup>[5]</sup>, 百度知心等。知识库已经成为大量知识驱动任务的基础, 在个性化推荐<sup>[6]</sup>、搜索引擎<sup>[7]</sup>、问答系统<sup>[8]</sup>等人工智能领域中广泛应用。

根据 Guha 在 1997 年提出的资源描述框架<sup>[9]</sup> (Resource Description Framework, RDF) 及相关标准, 现实世界中的事实以三元组的形式存储在知识库中。三元组由头实体、关系和尾实体组成, 表示头实体通过关系指向尾实体, 如三元组 (C 语言, 属于, 编程语言), 描述了 C 语言属于编程语言这一事实, 其中“C 语言”和“编程语言”分别对应头实体和尾实体, “属于”对应头、尾实体之间的关系。知识库中的某一个实体可能存在于多个三元组中, 同时, 不同的实体对之间可能存在相同的关系。如图 1.1 所示, 实体 ‘Steve Jobs’ 和 ‘Reed Jobs’ 存在于不止一个三元组事实中, 关系 ‘Nationality’ 也同时连接了不同的实体对。知识库中的众多三元组相互链接形成一种复杂的图结构数据——知识图谱 (Knowledge Graph, KG), 其节点和边分别对应知识库中的实体和关系。

互联网以及现实世界中数据的不断增长, 使得现有知识图谱中的事实不断增加。庞大的数据量使得如何有效地利用知识并对知识进行储存与表示成为如今亟待解决的问题。以 Wikidata 为例, 截至目前, Wikidata 中已经存在了超过 9600 万个三元组。传统的知识图谱表示方法将实体和关系表示为唯一的符号, 在具体任务中通常使用独热表示 (One-hot Representation), 即将实体和关系表示为一个长向量, 将相连实体和关系对应的维度设为非零而其他维度置为零, 但是基于符号的表示难以满足当今数据的产生速度, 面临以下两个难题: (1) 计算效率低: 现存的知识图谱包含大规模数量的数据, 且在不断地更新, 基于符号的表

示难以适应这种情况，同时，三元组包含了大量语义信息，其内部的头、尾实体和关系两两之间都存在语义交互，用符号对它们进行表示会忽略这种语义交互。由此导致后续的知识查询与推理任务变得困难，对相关算法要求较高，并且此类算法复杂度较高，计算效率较低，可扩展性较差。（2）数据稀疏：长尾分布是大部分大规模知识图谱具有的特征，即大部分关系往往只存在于小部分实体对之间，或者说大部分实体往往只与少部分关系相连。针对这类实体和关系进行查询与推理，基于符号的表示方法达到的准确率往往不尽人意。

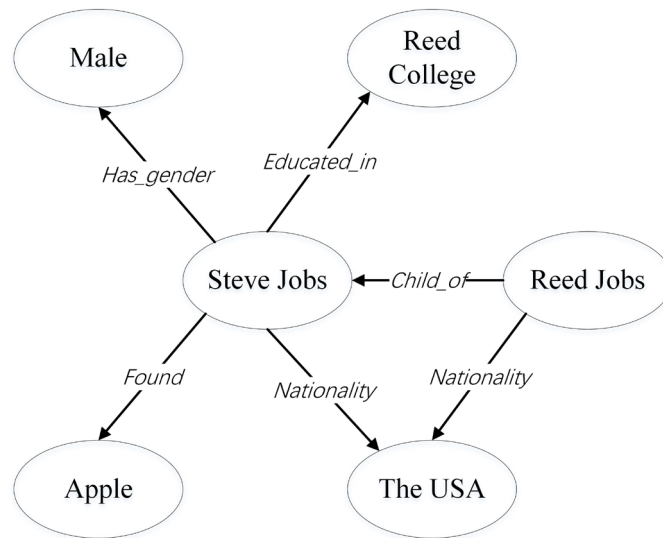


图 1.1 以 ‘Steve Jobs’ 为中心的简要知识图谱

为解决以上两个难题，研究者们受 word2vec<sup>[10]</sup>方法的启发，提出了建模大规模知识图谱的知识表示学习（Knowledge Representation Learning, KRL），也叫知识图谱嵌入（Knowledge Graph Embedding, KGE）。知识表示学习得益于分布式表示（Distribution Representation）的思想，在低维、稠密的向量空间中建模实体和关系，旨在保存知识图谱包含的语义信息。知识表示学习有效缓解了符号表示方法面临的难题。第一，知识表示学习利用低维、稠密的实值向量建模实体和关系，有效缓解了符号表示的维度灾难，且向量封装了实体和关系对应的语义信息，显著提高了计算效率以及知识查询与推理的能力。第二，高频实体与关系和低频实体与关系都被表示为相同维度的向量，利用向量的相似性可以表达三元组中实体与关系的语义联系，极大地缓解了数据稀疏性问题。

作为知识图谱构建与知识驱动的智能任务之间的桥梁，知识表示学习能够将知识图谱中包含的多源异质信息进行融合，保留知识图谱的本质结构信息，用

于搜索引擎、问答系统和推荐系统等知识驱动任务，有效缓解了数据稀疏性问题，计算效率较高，并且能反向应用于知识图谱补全<sup>[11]</sup>、实体分类<sup>[12]</sup>、关系抽取<sup>[13]</sup>等知识图谱构建任务，因此，研究知识表示学习近年来成为了知识图谱研究中的一项重要课题。

## 1.2 本文研究内容

知识表示学习相关方法在低维、稠密的向量空间中建模实体和关系，旨在保存知识图谱的信息用于相关知识驱动任务。近年来，知识表示学习已经吸引了大量海内外学者的目光，由此产生了大量的研究成果。目前，大部分知识表示学习方法只关注三元组本身的信息，忽略了三元组之间的语义关联，将每个三元组独立地进行表示学习。知识图谱中的实体和关系大多存在于多个三元组中，由此构成了实体和关系所处的众多邻域，独立地对每个三元组进行表示学习会导致语义丢失，从而忽略知识图谱本质的结构信息。因此，结合知识图谱的结构信息，本文首先针对融合实体的部分邻域信息提出了基于聚合的图卷积知识表示学习方法，其次针对融合实体和关系的全部邻域信息提出了基于双注意力的图卷积知识表示学习方法，其中实体和关系所在的整个三元组被认为是实体和关系的全部邻域，实体的邻居实体及相连关系被认为是实体的部分邻域。

(1) **基于聚合的图卷积知识表示学习方法。**知识图谱由三元组相互链接而成，实体通过不同的关系与大量的实体相连，表达了实体间的不同语义关系，进一步说明一个实体的语义不是只受某一个三元组的影响，而是受到包含该实体的所有三元组的影响。从知识图谱的结构来看，一个实体的语义会受到其所有邻居实体以及相连关系的影响，但是大多数知识表示学习模型独立地对每个三元组进行表示学习，忽略了实体与邻居实体及关系的语义交互。为此，本文结合实体所处的邻域，聚合中心实体所处邻域中包含的邻居实体及相连关系的信息以更新中心实体的表示，结合关系表示的自动更新，作为后续基于三元组进行独立表示学习的基础。该方法在实体表示学习过程中融合了邻居节点及相连关系的信息，能有效捕获知识图谱的部分结构信息与语义信息。

(2) **基于双注意力的图卷积知识表示学习方法。**知识图谱中实体的语义会受到邻域的影响，但是实体所处的不同邻域对中心实体造成的影响并不是完全相同，因此在进行实体的表示学习时，本文利用注意力机制对实体所处的整个邻



域的重要性进行建模，动态分配不同权重给中心实体的邻域结构。同时，知识图谱中的关系表示头尾实体对之间的语义关联，因此也包含了独特的语义，使关系进行自我更新会丢失相连实体对对其造成的影响。为解决这个问题，本文融合关系所处的整个邻域信息对关系进行表示学习。由于一个关系可能存在于不同的邻域，为了建模不同邻域对关系的影响，本文设计基于关系的注意力机制，分配不同的权重给关系的邻域结构，对关系进行有效地表示学习。该方法利用实体及关系所处的全部邻域结构信息同时对实体及关系进行表示学习，结合两个注意力机制，促进了实体及关系之间的语义交互，最大化保留了知识图谱的结构及语义信息。

### 1.3 组织结构

本文分为五个章节，组织结构如下：

第一章：绪论。本章从整体上介绍本文的研究背景、研究意义及相关概念的定义，并对研究内容及组织结构进行了进一步的阐述。

第二章：相关工作。本章首先介绍知识表示学习相关工作，包括基于三元组的模型和融合外部信息的模型，其次介绍图卷积相关工作，包括谱域图卷积和空间域图卷积的演变和发展。

第三章：基于聚合的图卷积知识表示学习方法。本章详细介绍基于聚合的图卷积知识表示学习方法。首先引入问题，交代本章的动机及主要创新，其次简要介绍基于基因本体的相关基因功能相似度分析方法，随后详细描述模型细节，包括符号体系和模型整体架构，接着介绍根据模型输出的实体表示进行基因功能相似度计算的方法，最后设计实验并针对实验结果进行分析。

第四章：基于双注意力的图卷积知识表示学习方法。本章详细介绍了基于双注意力的图卷积知识表示学习方法。首先引入问题，交代本章的动机及主要创新，其次详细描述模型细节，包括符号体系和模型整体架构，最后设计实验并针对实验结果进行分析。

第五章：总结及展望。本章总结了本文的主要工作，并结合不足对未来的研究方向进行展望。

## 2 相关工作

### 2.1 知识表示学习

近十年来,知识表示学习吸引了大量研究者的关注,每年的人工智能、自然语言处理及表示学习的顶会和期刊(比如 AAAI, ACL, ICLR, IJCAI, TNNLS, KBS, ESWA)都发表了很多关于知识表示学习的研究成果。知识表示学习模型主要分为基于三元组的模型和融合外部信息的模型,接下来详细介绍这两类模型的代表性成果。

#### 2.1.1 基于三元组的模型

基于三元组的模型独立地处理每个三元组,专注于三元组内部的语义交互。具体来讲,对于三元组  $(h, r, t)$ ,这类方法的设计分为三个部分:第一,初始化头、尾实体的向量表示  $\mathbf{h}, \mathbf{t} \in \mathbb{R}^d$  和关系的向量表示  $\mathbf{r} \in \mathbb{R}^d$  或矩阵表示  $\mathbf{M}_r \in \mathbb{R}^{d \times d}$ 。第二,基于实体和关系的表示设计得分函数  $f(h, r, t): \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow \mathbb{R}$ ,根据得分评价每个三元组的真实程度,得分越高,说明给定三元组越真实。第三,设计损失函数或目标函数,根据三元组的得分训练模型,从而学习实体和关系的最终表示。基于三元组的模型主要分为翻译模型及其扩展模型、张量分解模型及深度学习模型。

**翻译模型**受 word2vec 方法的启发,根据词向量在嵌入空间中的平移不变性,在向量空间中将关系当作头实体至尾实体的“翻译”。例如,在三元组  $(Beijing, capital\_of, China)$  中,‘*capital\_of*’是头实体‘*Beijing*’和尾实体‘*China*’之间的关系,翻译模型假定加法运算‘*Beijing + capital\_of = China*’成立,将‘*capital\_of*’当作‘*Beijing*’至‘*China*’的翻译操作。根据这种假定,Brodes 等人<sup>[1]</sup>在 2013 年提出经典的 TransE 模型,给定三元组  $(h, r, t)$ ,其得分函数定义如下:

$$f(h, r, t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_{L_1/L_2} \quad (2.1)$$

其中,  $L_1$  和  $L_2$  表示 L1 范数和 L2 范数。TransE 的本质思想是利用向量之间的距离来衡量三元组的得分,其目的是使得真实的样本得分尽可能趋于 0,于是定义了如下损失函数使模型能尽可能区分正负样本。

$$\mathcal{L} = \sum_{(h,r,t) \in \mathcal{T}} \sum_{(h',r,t') \in \mathcal{T}'_{(h,r,t)}} \max(\gamma + f(h,r,t) - f(h',r,t'), 0) \quad (2.2)$$

其中,  $\gamma > 0$  代表正负样本之间人为指定的间隔参数,  $\mathcal{T}$  和  $\mathcal{T}'$  分别代表正样本及负采样生成的负样本。负样本  $\mathcal{T}'_{(h,r,t)}$  通过替换真实三元组的头或尾实体得到, 采样方式如下:

$$\mathcal{T}'_{(h,r,t)} = \{(h',r,t) | h' \in \mathcal{E}\} \cup \{(h,r,t') | t' \in \mathcal{E}\} \quad (2.3)$$

虽然 TransE 在建模知识图谱时简单高效, 且参数量较少, 但是它仅在面对一对一关系时具有较好性能, 在处理其他类型的关系时表现乏力。比如, 对于多对多关系 ‘act’ 来说, 其相关的三元组 (*Leonardo DiCaprio, act, Titanic*) 和 (*Kate Winslet, act, Titanic*), 且有 ‘*Leonardo DiCaprio + act = Titanic*’ 和 ‘*Kate Winslet + act = Titanic*’ 成立, TransE 就会为 ‘*Leonardo DiCaprio*’ 和 ‘*Kate Winslet*’ 学习到几乎相同的向量, 但他们显然是不同的实体, 具有不同的语义。

为了建模复杂关系, TransH<sup>[14]</sup>, TransR<sup>[15]</sup>, TransD<sup>[16]</sup>, RotatE<sup>[17]</sup>, HAKE<sup>[18]</sup> 等 TransE 的扩展模型相继被提出。其中 TransH 引入基于关系的超平面变换, 利用超平面的法向量将头、尾实体投影到超平面上。对于给定三元组  $(h,r,t)$ , 头、尾实体在超平面上的投影向量表示为:

$$\mathbf{h}_\perp = \mathbf{h} - \mathbf{w}_r^T \mathbf{h} \mathbf{w}_r, \quad \mathbf{t}_\perp = \mathbf{t} - \mathbf{w}_r^T \mathbf{t} \mathbf{w}_r \quad (2.4)$$

其中,  $\mathbf{h}_\perp$  和  $\mathbf{t}_\perp$  分别表示头、尾实体在关系超平面上的投影向量,  $\mathbf{w}_r$  表示关系  $r$  对应超平面的法向量, 由此可看出实体在不同的关系超平面上会得到不同的投影向量。根据 TransE 的思想, 在关系超平面上进行基于关系的翻译操作, 得到 TransH 的得分函数为:

$$f(h,r,t) = \|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_{L_1/L_2} \quad (2.5)$$

即使引入了关系超平面, TransH 仍在相同的语义空间建模实体和关系。TransR 则引入实体空间和关系空间分别对实体和关系进行建模。为了在相同的空间基于关系实现头、尾实体之间的翻译操作, TransR 将实体空间中的头、尾实体向量通过关系投影矩阵  $\mathbf{M}_r$  映射到关系空间。对于给定三元组  $(h,r,t)$ , 头、尾实体在关系空间的投影向量为:

$$\mathbf{h}_\perp = \mathbf{M}_r \mathbf{h}, \quad \mathbf{t}_\perp = \mathbf{M}_r \mathbf{t} \quad (2.6)$$

同时定义如公式(2.5)所示的得分函数。由于 TransR 引入不同的嵌入空间对实体和关系进行建模，因此关系和实体的嵌入维度可以不同，在一定程度上提升了模型的表达性能。

由于 TransR 在不同向量空间建模实体和关系，引入了投影矩阵，使得模型的复杂度较高，TransD 通过以向量分解投影矩阵的方式来简化 TransR。简单来说，对于给定三元组  $(h, r, t)$ ，TransD 除了为实体和关系学习向量表示以外，同时，为它们学习了一个对应的权重向量  $w_h$ ， $w_r$  和  $w_t$ ，为每一个三元组中的头、尾实体构造了对应的投影矩阵，表示如下：

$$M_r^1 = w_r w_h^T + I, \quad M_r^2 = w_r w_t^T + I \quad (2.7)$$

其中， $I$  表示单位阵。头、尾实体对应的投影向量则更新为：

$$h_{\perp} = M_r^1 h, \quad t_{\perp} = M_r^2 t \quad (2.8)$$

TransD 通过以上变换形式，大大的减少了 TransR 的参数量与计算量，极大地提升了 TransR 的性能。

为了处理知识图谱中的对称、非对称、逆反和组合等关系，Sun 等人提出 RotatE 模型，在复数空间建模实体和关系，将关系当作头实体至尾实体的“旋转”，由此扩展了 TransE。具体来说，给定三元组  $(h, r, t)$ ，RotatE 假设 ‘ $t = h \circ r$ ’ 成立，定义如下得分函数：

$$f(h, r, t) = \|h \circ r - t\| \quad (2.9)$$

不同于 TransE，RotatE 定义如下损失函数优化模型训练。

$$\mathcal{L} = -\log \sigma(\gamma - f(h, r, t)) - \sum_{i=1}^n \frac{1}{k} \log \sigma(f(h'_i, r, t'_i) - \gamma) \quad (2.10)$$

其中， $(h'_i, r, t'_i)$  表示第  $i$  个负样本。除此之外，RotatE 注意到对负样本进行均匀采样会导致效率低下，因为某些采到的负样本明显是错误三元组，这类负样本并不能为模型训练提供有效帮助。因此，RotatE 设计了一种自对抗负采样方式，采样的分布如下：

$$p(h'_j, r, t'_j | \{(h_i, r_i, t_i)\}) = \frac{\exp \alpha f(h'_j, r, t'_j)}{\sum_i \exp \alpha f(h'_i, r, t'_i)} \quad (2.11)$$

其中， $\alpha$  表示采样温度。经过自对抗负采样之后，RotatE 同步以自对抗方式训练

模型，模型的损失函数更新为：

$$\mathcal{L} = -\log \sigma(\gamma - f(h, r, t)) - \sum_{i=1}^n p(h'_i, r, t'_i) \log \sigma(f(h'_i, r, t'_i) - \gamma) \quad (2.12)$$

知识图谱中除了关系存在多种类型以外，实体之间还存在语义层次，例如，实体“树”包含实体“乔木”和“棕榈”，显然，“树”拥有比“乔木”和“棕榈”更高的语义层次。为了建模知识图谱中实体的语义层次，Zhang 等人提出了在极坐标系下建模实体的模型 HAKE。对于给定三元组  $(h, r, t)$ ，HAKE 为头、尾实体和关系分别学习两个不同的向量  $\mathbf{h}_m$ ,  $\mathbf{h}_p$ ,  $\mathbf{t}_m$ ,  $\mathbf{t}_p$ ,  $\mathbf{r}_m$  和  $\mathbf{r}_p$ ，其中  $\mathbf{h}_m$  和  $\mathbf{t}_m$  对应极坐标系下头、尾实体向量的模长，代表实体的语义层次， $\mathbf{h}_p$  和  $\mathbf{t}_p$  对应极坐标系下头、尾实体向量的圆心角，用于区分同一层次的不同实体， $\mathbf{r}_m$  和  $\mathbf{r}_p$  表示头、尾实体间对应模长和圆心角的变换。HAKE 假设 ‘ $\mathbf{t}_m = \mathbf{h}_m \circ \mathbf{r}_m$ ’ 和 ‘ $\mathbf{t}_p = (\mathbf{h}_p + \mathbf{r}_p) \bmod 2\pi$ ’ 成立。HAKE 定义如下得分函数：

$$f(h, r, t) = f_m(h, r, t) + f_p(h, r, t) \quad (2.13)$$

其中， $f_m(h, r, t)$  和  $f_p(h, r, t)$  定义为：

$$\begin{aligned} f_m(h, r, t) &= \|\mathbf{h}_m \circ \mathbf{r}_m - \mathbf{t}_m\|_2 \\ f_p(h, r, t) &= \left\| \sin\left((\mathbf{h}_p + \mathbf{r}_p - \mathbf{t}_p) / 2\right) \right\|_1 \end{aligned} \quad (2.14)$$

其中，模长部分遵循 RotatE 的思想，将关系当作头实体至尾实体的“旋转”，并且采用与 RotatE 相同的训练方式与负采样方式。HAKE 能同时区分不同语义层次和相同语义层次的实体，以此完成对整个知识图谱的语义层次进行的建模。

**张量分解模型**使头、尾实体通过关系矩阵实现双线性变换来分解知识图谱代表的三阶张量，其中关系被分解为一个矩阵。张量分解模型实质是根据关系衡量头、尾实体的相似程度，以此评价每个三元组的真实程度。RESCAL<sup>[12]</sup>将实体映射为实值向量，将关系映射为头、尾实体向量之间的关联矩阵，使头、尾实体向量通过关系关联矩阵进行相似度匹配。给定三元组  $(h, r, t)$ ，RESCAL 定义如下得分函数：

$$f(h, r, t) = \mathbf{h}^T \mathbf{M}_r \mathbf{t} \quad (2.15)$$

RESCAL 有效捕捉到了实体间的语义交互，取得了良好的性能，但是模型中包含过多参数，使得训练容易出现过拟合，且无法对对称关系进行有效建模。

为了减少 RESCAL 模型的参数量及计算量，同时避免模型过拟合，DistMult<sup>[19]</sup>将关系矩阵约束为对角矩阵，定义如下得分函数：

$$f(h, r, t) = \mathbf{h}^T \text{diag}(\mathbf{r}) \mathbf{t} \quad (2.16)$$

其中， $\text{diag}(\mathbf{r})$ 表示对关系向量的对角化。DistMult 引入对角矩阵改进 RESCAL 不能处理对称关系的问题，导致 DistMult 只适用于建模对称关系，无法建模常规知识图谱。

Complex<sup>[20]</sup>通过在复数嵌入空间建模实体和关系改进了 DistMult，定义了如下基于复数空间的相似度得分函数：

$$f(h, r, t) = \text{Re}(\mathbf{h}^T \text{diag}(\mathbf{r}) \bar{\mathbf{t}}) \quad (2.17)$$

其中， $\text{Re}(\cdot)$ 表示复数向量的实部， $\bar{\mathbf{t}}$ 表示 $\mathbf{t}$ 的共轭。Complex 有建模对称/非对称、逆反等复杂关系的能力，进而在建模知识图谱时表现较好。

HolE<sup>[21]</sup>引入循环相关操作将头、尾实体的向量表示进行组合，把关系建模为实值向量表示而非矩阵表示，减少了模型的参数量且极大地促进了实体间的语义交互。给定三元组 $(h, r, t)$ ，HolE 定义如下得分函数：

$$f(h, r, t) = \sigma(\mathbf{r}^T (\mathbf{h} \star \mathbf{t})) \quad (2.18)$$

其中， $\star$ 表示循环相关操作，定义为：

$$[\mathbf{h} \star \mathbf{t}]_k = \sum_{i=0}^{d-1} \mathbf{h}_{[i]} \mathbf{t}_{[(k+i) \bmod d]} \quad (2.19)$$

其中， $\mathbf{h}_{[i]}$ 代表向量 $\mathbf{h}$ 的第 $i$ 个位置的值。由于循环相关操作具有非对称性，使得 HolE 在建模非对称关系时表现出色。

ANALOGY<sup>[22]</sup>引入类比的思想建模实体间的关系，比如“北京之于中国就如伦敦之于英国”，在 RESCAL 的基础上对关系的矩阵表示进行如下正则性和可交换性约束：

$$\begin{aligned} \mathbf{M}_r \mathbf{M}_r^T &= \mathbf{M}_r^T \mathbf{M}_r, \quad \forall r \in \mathcal{R} \\ \mathbf{M}_r \mathbf{M}_{r'}^T &= \mathbf{M}_{r'}^T \mathbf{M}_r, \quad \forall r, r' \in \mathcal{R} \end{aligned} \quad (2.20)$$

ANALOGY 可以通过指定特定关系矩阵泛化为 RESCAL, DistMult, Complex 和 HolE，具有很强的包容性，且模型复杂程度不高。

Tucker<sup>[23]</sup>通过 Tucker 分解，分别建模实体表示和关系表示，结合共享核矩

阵存储部分知识，实现实体与关系间的知识共享，捕获实体间的语义交互。

另外，鉴于神经网络在人工智能领域的蓬勃发展，Socher 等人<sup>[24]</sup>利用双线性张量分解层代替传统神经网络的线性层而提出了神经张量网络模型 NTN，使头、尾实体能通过关系矩阵在向量空间中进行多维度的语义交互。给定三元组  $(h, r, t)$ ，NTN 定义如下评分函数：

$$f(h, r, t) = \mathbf{r}^T \sigma(\mathbf{h}^T \mathbf{M}_r \mathbf{t} + \mathbf{M}_r^1 \mathbf{h} + \mathbf{M}_r^2 \mathbf{t} + \mathbf{b}_r) \quad (2.21)$$

其中， $\mathbf{M}_r$  表示关系对应的三阶张量， $\mathbf{M}_r^1$  和  $\mathbf{M}_r^2$  分别表示头、尾实体基于关系的投影矩阵， $\mathbf{b}_r$  表示偏置向量。NTN 为每一个关系定义了四个相关的参数，头、尾实体通过关系张量进行双线性变换，可以在不同维度上进行交互，且头、尾实体分别通过对应的关系矩阵进行线性变换，充分挖掘了三元组内部的语义信息，因此在建模知识图谱中具有出色表现。但是，由于 NTN 模型包含较多的可学习参数，使得模型复杂度较高，计算效率较低。

**深度学习模型**利用深度神经网络建模知识图谱以学习实体和关系的非线性特征，充分实现三元组内部的语义交互。用于知识表示学习的深度神经网络有卷积神经网络 (Convolutional Neural Network, CNN)、胶囊网络 (Capsule Network) 和生成对抗网络 (Generative Adversarial Network, GAN) 等。

Dettmers 等人<sup>[25]</sup>最早提出使用 CNN 建模三元组的模型 ConvE。ConvE 利用二维卷积神经网络提取头实体和关系重塑后表示的深层特征，通过线性层将量化的特征表示进行特征变换，与尾实体向量进行内积操作来评价三元组的真实程度。给定三元组  $(h, r, t)$ ，ConvE 定义如下评分函数：

$$f(h, r, t) = \sigma \left( \text{vec} \left( \sigma([\bar{\mathbf{h}}; \bar{\mathbf{r}}] * \omega) \right) \mathbf{W} \right) \mathbf{t} \quad (2.22)$$

其中， $[\cdot; \cdot]$  表示矩阵的拼接操作， $\mathbf{W}$  表示线性变换的可学习权重矩阵， $\bar{\mathbf{h}}$  和  $\bar{\mathbf{r}}$  表示  $\mathbf{h}$  和  $\mathbf{r}$  的重塑。ConvE 采用二分类交叉熵作为损失函数来优化模型。虽然 ConvE 利用神经网络建模了三元组，但是局部来看，ConvE 实现了头实体与关系之间的充分交互，而尾实体则作为最后的独立模块，没有充分参与三元组内部的语义交互，同时，卷积核也是一个可学习的参数，导致模型负载较大。

为了充分实现三元组内部的语义交互，ConvKB<sup>[26]</sup>将三元组中的头、尾实体和关系的向量表示拼接起来，利用二维卷积神经网络提取拼接向量的特征，对三元组整体进行建模。给定三元组  $(h, r, t)$ ，ConvKB 定义得分函数如下：

$$f(h, r, t) = \text{vec}(\sigma([\mathbf{h}; \mathbf{r}; \mathbf{t}] * \omega)) \mathbf{W} \quad (2.23)$$

同时, ConvKB 采用带 L2 正则化约束参数的损失函数优化实体和关系的表示学习。

$$\mathcal{L} = \sum_{(h,r,t) \in \mathcal{T} \cup \mathcal{T}'} \log(1 + \exp(l_{(h,r,t)} \cdot f(h, r, t))) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 \quad (2.24)$$

其中,  $l_{(h,r,t)}$  表示三元组  $(h, r, t)$  的标签, 正样本属于集合  $\mathcal{T}$ , 标签为 1, 负样本属于集合  $\mathcal{T}'$ , 标签为 0。ConvKB 将三元组作为一个整体进行建模, 展现了强大的表达能力和泛化能力。

ConvR<sup>[27]</sup>在 ConvE 的基础上将三元组中的关系定义为卷积核, 自适应提取关系的表示用于生成头实体的非线性特征, 减少了模型的参数, 极大地促进了头实体和关系的语义交互。

InteractE<sup>[28]</sup>将普通二维卷积神经网络替换为循环卷积神经网络。相较二维卷积神经网络, 循环神经网络能提取更多头实体和关系的非线性特征用于匹配尾实体, 在 ConvE 的基础上扩展了头实体和关系之间的语义交互。

Nguyen 等人<sup>[29]</sup>提出用胶囊网络来建模三元组的 CapsE 模型。CapsE 首先利用二维卷积神经网络对三元组中头、尾实体和关系的拼接表示进行特征提取生成特征表示, 其次利用胶囊网络中的路由模块对每个特征表示进行压缩, 最后用压缩后向量的模长来评价三元组的真实程度。给定三元组  $(h, r, t)$ , CapsE 定义得分函数如下:

$$f(h, r, t) = \|\text{capsnet}(\sigma([\mathbf{h}; \mathbf{r}; \mathbf{t}] * \omega))\| \quad (2.25)$$

其中,  $\text{capsnet}$  表示胶囊网络。CapsE 采用和 ConvKB 一样的训练方式和不带 L2 正则化约束参数的损失函数来优化模型参数。

由于大多模型通过已有负采样方式采集到的负样本质量较低, Cai 等人<sup>[30]</sup>提出 KBGAN, 将生成对抗网络 GAN 用于知识表示学习。KBGAN 采用两个知识表示学习模型分别作为生成器和判别器, 生成器用于生成负样本, 判别器则基于生成的负样本和已有的正样本对模型进行训练, 并且反馈到生成器。KBGAN 的目标是使判别器能将知识图谱中的真实三元组和生成器生成的负样本完美区分开。虽然 KBGAN 建模知识图谱的性能不是非常出色, 但是其提出的负采样方式非常新颖, 可以应用于其他模型。



### 2.1.2 融合外部信息的模型

近年来,大量相关工作通过融合外部相关信息以扩充知识表示学习模型,比如融合实体类别信息,文本描述信息,图片信息及图结构信息。基于三元组的模型仅针对知识图谱中的三元组进行建模,忽略了这些外部相关信息。

融合实体类别信息的模型能在知识表示学习中捕获实体的类别信息,除了实现三元组内部的语义交互以外,还能最大化的区分不同实体。**SSE**<sup>[31]</sup>模型引入实体类别信息建模知识图谱,假设相同类别的实体在嵌入空间中距离更近。为了拟合这种平滑假设,**SSE**采用局部线性嵌入和拉普拉斯特征映射约束向量空间的几何结构。较于基于三元组的模型,**SSE**具有更好的表示能力且针对某些特定的下游任务有更好的执行能力。真实世界中知识图谱所包含实体的类别信息具有层次结构,比如熊猫属于哺乳动物类别同时也属于动物这一类别,而动物包含哺乳动物,说明实体可以属于多种类别且所属类别可能具有层次结构。即使**SSE**引入了实体的类别信息,但无法对这种层次结构进行建模。**TKRL**<sup>[32]</sup>融合实体的类别层次信息,基于关系构建针对不同实体类别的映射矩阵,同时结合翻译模型的思想,在不同类别下学习实体的不同表示。

知识图谱中的实体仅仅包含少量信息,融合文本描述信息的模型引入丰富的实体描述信息,为实体学习语义更加丰富的表示。**Xie**等人<sup>[33]</sup>提出**DKRL**模型,结合翻译模型,为实体学习基于三元组和基于文本描述的表示,其中基于文本描述的表示由连续词袋模型结合神经网络编码得到。另外,**TEKE**<sup>[34]</sup>结合知识图谱与特定语料库,根据实体在语料库中的信息提取实体的上下文信息并将其融入翻译模型。

融合图片信息的模型在知识表示中融入实体的图片信息。**IKRL**<sup>[35]</sup>融合实体的图像信息,学习跨模态的知识表示。具体来说,**IKRL**首先引入基于注意力机制的图片自编码器自动选择高质量的实体图片,接着学习实体基于图片的表示,最后遵循翻译模型的思想训练模型以充分建模知识图谱。

融合图结构信息的模型充分考虑知识图谱本质的结构信息,在知识表示学习中引入实体和关系所在的邻域信息。较于基于三元组的模型,融合图结构信息的模型充分考虑了知识图谱的结构,使实体和关系进行更加充分的语义交互。**GAKE**<sup>[36]</sup>类比文本中词的上下文结构为知识图谱中的实体定义了三种上下文结构,分别对应与实体相关的关系、邻居和路径。关系上下文指与实体连接的关系集合,邻居上下文指实体的邻居实体集合,路径上下文指特定步长下随机游走所

得路径的集合。GAKE 采用注意力机制动态地融合这三种上下文结构信息，学习实体和关系的向量表示，极大提升了知识表示的性能。R-GCN<sup>[37]</sup>首次将图卷积网络（Graph Convolutional Network, GCN）用于建模知识图谱，采用 GCN 编码器-解码器模式学习实体和关系的向量表示，其中，GCN 编码器融合实体的一阶邻居实体信息来更新中心的表示，通过叠加多层编码器聚合高阶邻居的信息，解码器则采用任意基于三元组的知识表示学习模型如 TransE, DistMult, ConvE 等。Shang 等人<sup>[38]</sup>注意到 R-GCN 在编码器中以相同权重融合邻居实体的信息来更新中心实体，由此提出了端到端的结构感知模型 SACN。SACN 包含一个加权图卷积网络编码器 WGCN 用于自适应地融合邻居实体的信息，解码器 Conv-TransE 同时遵循 ConvE 和 TransE 的思想，在满足“翻译”思想的前提下扩展了实体和关系的交互。KBGAT<sup>[39]</sup>则在编码器中引入注意力机制自适应地融合实体的邻域结构信息来更新中心实体的表示。RGHAT<sup>[40]</sup>在 KBGAT 的基础上考虑知识图谱中实体的结构信息学习实体基于结构的向量表示。

## 2.2 图卷积网络

CNN 在建模图像、文本、语音和视频等欧式数据上具有巨大优势并成功应用于图像分类、目标检测和机器翻译等下游任务。CNN 的强大能力在于其能学习局部固定结构，通过局部卷积核有规律地平移进行卷积操作，生成数据的多尺度分层特征。局部卷积核具有平移不变性，能独立于数据所在的空间位置而识别出相同的特征，比如，在一张画布上识别出一张猫的图像，不管猫的图像处于这张画布的哪个位置，CNN 总能为该图像生成相同的特征。CNN 可以在欧式数据和灰度网络等数据上高效且简便地定义卷积操作，但是针对真实世界中的大量非欧数据如社交网络和知识图谱，卷积核不能在这些数据上有规律地移动并提取特征，因此将卷积操作有效地泛化到这些数据上是一个巨大的挑战。针对这个问题，大量学者开始了基于图的卷积研究，根据研究基准的不同，图卷积分为谱域图卷积和空间域图卷积。

### 2.2.1 谱域图卷积

谱域图卷积通过图傅里叶变换和卷积理论定义卷积。给定一个图  $G = (V, E, W, X)$ ，其中  $V$  表示节点集合，节点数量  $n = |V|$ ， $E$  表示边的集合，

$W \in \mathbb{R}^{n \times n}$  表示带权邻接矩阵, 如果是不带权图,  $W$  则只包含 0, 1。  $X \in \mathbb{R}^{n \times d}$  表示节点特征矩阵, 每个节点具有  $d$  维特征,  $X$  的每一列表示当前维度所有节点的信号, 信号用  $x$  表示。

针对图  $G$ , 谱域图卷积首先对其进行图拉普拉斯变换, 表示如下:

$$L = D - W \quad (2.26)$$

其中,  $D$  是一个对角矩阵, 对角线元素  $D_{ii} = \sum_j W_{ij}$ 。结合单位阵  $I$  对图拉普拉斯矩阵进行正则化, 得到如下可对角化的正则化图拉普拉斯矩阵。

$$L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \quad (2.27)$$

其次, 谱域图卷积定义图的傅里叶基和傅里叶变换。根据矩阵的特征分解方法, 求得正则化图拉普拉斯矩阵  $L$  的特征值集和对应的特征向量集分别为  $\{\lambda_l\}_{l=1}^n$  和  $\{u_l\}_{l=1}^n$ , 则  $L$  可被对角化为:

$$L = U \Lambda U^T \quad (2.28)$$

其中,  $U$  表示傅里叶基, 记为  $U = [u_1, u_2, \dots, u_n]$ ,  $\Lambda = \text{diag}([\lambda_1, \lambda_2, \dots, \lambda_n])$ 。对于信号  $x \in \mathbb{R}^n$  来说, 图的傅里叶变换及其逆变换记为:

$$\begin{aligned} \hat{x} &= U^T x \\ x &= U \hat{x} \end{aligned} \quad (2.29)$$

最后, 谱域图卷积定义两个信号的卷积等于这两个信号对应傅里叶变换的哈达玛积的傅里叶逆变换。根据这一卷积理论, 给定信号  $x$  和  $y$ , 谱域图卷积定义为:

$$x *_G y = U \left( (U^T y) \circ (U^T x) \right) \quad (2.30)$$

其中,  $x *_G y$  称作  $x$  相对  $y$  的卷积,  $U^T y$  称作谱域图卷积的卷积核。记  $U^T y = [\theta_0, \theta_1, \dots, \theta_{n-1}]^T$ ,  $g_\theta = \text{diag}([\theta_0, \theta_1, \dots, \theta_{n-1}])$ , 则公式(2.30)可更新为:

$$x *_G y = U g_\theta U^T x \quad (2.31)$$

通过堆叠多层谱域图卷积, Bruna 等人<sup>[41]</sup>提出如下节点表示更新方式:

$$x_{k+1,j} = \sigma \left( \sum_{i=1}^{f_k} U F_{k,i,j} U^T x_{k,i} \right), \quad j = 1, \dots, f_{k+1} \quad (2.32)$$

其中,  $k$  表示图卷积的层数,  $f_k$  表示第  $k$  层的卷积核数量,  $F_{k,i,j}$  表示第  $k$  层的卷积核,  $x_{k,i}$  表示第  $k$  层的信号。即使谱域图卷积定义了图上的卷积操作, 但是

计算复杂，需要对拉普拉斯矩阵进行特征分解，并且其卷积核不能在图的顶点域中局部化。

Defferrard 等人<sup>[42]</sup>提出 ChebyNet 模型将卷积核在图的顶点域中局部化，并且极大降低了计算复杂度，其定义信号  $x$  相对  $y$  的卷积为：

$$x *_G y = U g_\beta(\Lambda) U^T x \quad (2.33)$$

ChebyNet 的核心是通过限制卷积核为特征值矩阵  $\Lambda$  的多项式函数将卷积核  $g_\theta$  替换为  $g_\beta(\Lambda)$ ， $g_\beta(\Lambda)$  定义为：

$$g_\beta(\Lambda) = \sum_{k=0}^{k-1} \beta_k \Lambda^k \quad (2.34)$$

其中， $k$  表示卷积半径，在图中可以理解为邻居节点相对于中心节点的跳数。

图小波神经网络 GWNN<sup>[43]</sup>通过替换图的傅里叶变换为图小波变换构造小波基，极大减少了模型的参数量且实现了卷积核在顶点域中的局部化。

### 2.2.2 空间域图卷积

空间域图卷积基于图的顶点域实现卷积操作，通过聚合邻居节点的特征来更新中心节点的特征，经过多层图卷积堆叠，达到聚合高阶邻居节点特征的目的。

最早，Niepert 等人<sup>[44]</sup>类比 CNN 的卷积思想在空间域中实现图的卷积操作，选择中心节点的邻居节点，对邻居节点进行排序构建感受野，通过卷积操作对每个节点的感受野进行卷积。其中每个节点的感受野使用相同卷积核进行卷积，实现参数共享。

GraphSAGE<sup>[45]</sup>则在为中心节点选择邻居节点的前提下，通过聚合函数聚合邻居节点的特征与节点原始特征进行融合提出了图神经网络（Graph Neural Network, GNN）的通用框架。对图  $G$  中的节点  $v$ ，GraphSAGE 定义如下聚合邻居特征的方式：

$$h_{\mathcal{N}(v)}^k = \text{AGGREGATE}^k \left( \{h_u^{k-1} \mid u \in \mathcal{N}(v)\} \right) \quad (2.35)$$

其中， $\mathcal{N}(v)$  表示  $v$  的邻居节点。通过连接原始特征和聚合的邻居节点特征，节点  $v$  的特征更新为：

$$h^k = \sigma \left( W^k \cdot \text{CONCAT} \left( h^{k-1}, h_{\mathcal{N}(v)}^k \right) \right) \quad (2.36)$$

较于 GraphSAGE，Kipf 等人<sup>[46]</sup>通过正则化的拉普拉斯矩阵聚合邻居节点的

信息，经过堆叠多层 GCN 且共享特征变换的参数，实现了图上的卷积操作，正式提出了图卷积网络 GCN 的概念。GCN 定义如下层传播公式：

$$\mathbf{H}^{k+1} = \sigma \left( \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}^k \mathbf{W}^k \right). \quad (2.37)$$

其中， $k$  表示 GCN 的层数， $\mathbf{H}^0 = \mathbf{X}$ ， $\mathbf{H}^k$  表示第  $k$  层的节点特征矩阵， $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$  表示添加了自回环的邻接矩阵， $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}$ ， $\mathbf{W}^k$  表示第  $k$  层对应的权重参数。GCN 通过定义层间传播公式，使节点特征和包含的消息在图中流动，从而实现了卷积操作，且 GCN 参数较少，易于训练，在很多下游任务中展示了不错的表现。

图注意力网络<sup>[47]</sup>（Graph Attention Network, GAT）在 GCN 中引入注意力机制自适应地聚合邻居节点的信息，使中心节点能聚合更重要的邻居节点的信息。GAT 定义如下方式聚合邻居信息：

$$\mathbf{h}_i^{k+1} = \sigma \left( \sum_{j \in \mathcal{N}(i)} \alpha_{ij}^k \mathbf{W}^k \mathbf{h}_j^k \right) \quad (2.38)$$

其中， $\alpha_{ij}^k$  表示第  $k$  层节点  $j$  对于节点  $i$  的注意力，定义为：

$$\alpha_{ij}^k = \frac{\exp \left( \text{LeakyReLU} \left( \mathbf{a}^T [\mathbf{W}^k \mathbf{h}_i^k \parallel \mathbf{W}^k \mathbf{h}_j^k] \right) \right)}{\sum_{m \in \mathcal{N}(i)} \exp \left( \text{LeakyReLU} \left( \mathbf{a}^T [\mathbf{W}^k \mathbf{h}_i^k \parallel \mathbf{W}^k \mathbf{h}_m^k] \right) \right)} \quad (2.39)$$

其中， $\mathbf{a}$  表示注意力向量。

由于 GCN 在处理图数据上只针对节点进行建模，Schlichtkrull 等人<sup>[37]</sup>提出 R-GCN 将 GCN 推广到了多关系图领域，定义了基于关系的可学习权重参数作为卷积核，实现了在多关系图上的卷积操作。针对知识表示学习，R-GCN 提出了 GCN 编码器-解码器模型，编码器基于关系聚合邻居节点的信息对中心节点的表示进行第一次更新，解码器模型将编码器的输出作为输入对节点和关系的表示进行二次学习。随后，SACN<sup>[38]</sup>，CompGCN<sup>[48]</sup>和 KEGCN<sup>[49]</sup>等基于 GCN 的知识表示学习模型相继出现。KBGAT<sup>[39]</sup>，RGHAT<sup>[40]</sup>和 EIGAT<sup>[50]</sup>等模型将注意力机制引入基于 GCN 的知识表示学习中，引起了大量学者的研究兴趣，推动了 GCN 在知识表示学习中的发展。

### 2.3 本章小结

本章对知识表示学习模型的代表性成果进行了仔细的介绍。知识表示学习模型分为基于三元组的模型和融合外部信息的模型，其中基于三元组的模型独立地对三元组进行建模，融合外部信息的模型除了对知识图谱已包含信息进行建模外，还额外融入了实体或关系相关的外部信息，提升了原有模型的性能。另外，本章通过分析谱域图卷积和空间域图卷积，详细回顾了图卷积的理论知识，并且介绍了空间域图卷积在知识表示学习中的引入和发展。

## 3 基于聚合的图卷积知识表示学习方法

### 3.1 问题引入

知识图谱由真实世界的事实三元组组成，具有丰富的语义和结构信息，广泛应用于信息搜索、问答系统和推荐系统等人工智能领域。如何有效地表示知识图谱中包含的知识从而用于相关知识驱动任务一直以来都是一个热点问题，因此针对知识表示学习方法的研究在近些年引起了大量学者的兴趣。

当前主流的基于三元组的方法针对单个三元组进行建模，只考虑三元组内部的语义交互。翻译模型将三元组中的关系视为头实体至尾实体的“翻译”。张量分解模型通过张量分解操作学习实体和关系的表示，将知识图谱构造成三阶张量。深度学习模型利用深度神经网络如卷积神经网络学习实体和关系的表示来捕获三元组的内部语义交互。由于大量三元组相互链接，形成了一个庞大且复杂的图结构，其中一个节点所在的一个三元组被认为是该节点的邻域，基于三元组的方法忽略了这种图结构信息，从而忽略了节点所在的邻域对其造成的影响。

从图的结构来看，一个实体通过相同或不同的关系与邻居实体相连，由此具有不同的邻域，而不同的邻域则反映了该实体包含的不同语义信息。因此，中心实体的邻居实体从另一个角度反映了其蕴含的语义信息，同时，具有相同邻居实体的实体在语义上具有一定联系。举例来讲，如图 3.1 所示，‘Randi Zuckerberg’的国籍为‘The USA’，且‘Randi Zuckerberg’是‘Mark Zuckerberg’的姐姐，则可以推断‘Mark Zuckerberg’的国籍很大概率上也为‘The USA’，这说明‘Randi Zuckerberg’可以从侧面反映出‘Mark Zuckerberg’的部分语义信息。同时，‘Aaron Sittig’和‘Kevin colleran’都在‘Facebook’工作，则可以推断‘Aaron Sittig’和‘Kevin colleran’是同事关系，并且很大概率上熟识，这从侧面反映了具有相同邻居实体的实体在语义上具有一定联系。

真实世界中除知识图谱外，还具有大量的图结构数据，例如基因本体、社交网络和论文索引结构。为了对图数据进行表示学习用于后续的分类或链接预测任务，图神经网络以融合邻居节点信息来更新中心节点的表示这一思路，为知识图谱的表示学习带来了新的改变。在知识图谱中，结合图神经网络的思想，研究者提出解码器-编码器这一新模式用于知识表示学习，其中编码器由一定数

量的图神经网络层组成，用于第一次更新知识图谱中所有节点的表示，解码器则采用传统类型的知识表示学习模型，即翻译模型、张量分解模型或深度学习模型，对单个三元组进行建模，在此过程中实现实体表示的二次学习。

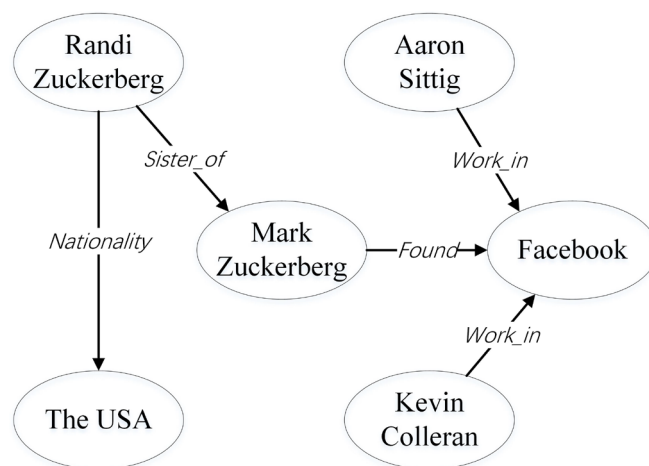


图 3.1 以 ‘Mark Zuckerberg’ 为中心的简要知识图谱

然而，当前大部分基于图神经网络的知识表示学习模型存在以下两个问题：

(1) 在编码器模型中学习实体表示时，只考虑邻居实体的信息，忽略了实体间相连的关系信息，而事实上关系也蕴含其特定的语义信息。仍以图 3.1 为例，通过 ‘Mark Zuckerberg’ 创建了 ‘Facebook’ 和 ‘Aaron Sittig’ 在 ‘Facebook’ 工作可知 ‘Mark Zuckerberg’ 是 ‘Aaron Sittig’ 的老板，由此可以发现，通过实体间相连的关系可以大概推断两个独立实体间的关系，这对补全知识图谱有很大的帮助。(2) 在编码器模型中只学习实体的表示，却忽略关系的表示学习。在后续解码器模型中，实体表示由编码器输出得到，关系表示却需要重新初始化，这在一定程度上造成了知识图谱的语义丢失。

结合以上分析，本章利用图卷积网络，在编码器模型中聚合邻居节点及相连关系的信息用于更新中心实体的表示，同时，除了学习实体的表示外，还针对关系表示进行学习。具体来讲，学习实体表示时，采用循环相关、加和乘等操作聚合实体及关系的表示从而得到一个新的表示，结合中心实体包含的邻域结构，用相连关系及实体经过聚合后的表示来更新中心实体的表示。学习关系表示时，为每一层图卷积网络定义对应的层传播参数，用于实现关系的自我更新。本章在基因本体<sup>[51]</sup> (Gene Ontology, GO) 上进行了实验设计和分析，提出了一种基于聚合的图卷积网络知识表示学习方法并在基因本体上进行基因功能相似度分析的



方法 (Graph Convolutional Network on Gene Ontology for functional similarity analysis of genes, GOGCN), 整体结构如图 3.2 所示。本章工作的主要创新总结如下:

(1) 提出一种基于聚合的图卷积网络知识表示学习方法, 利用邻居实体及相连关系聚合后的表示更新中心实体的表示, 同时对关系进行表示学习。与经典模型相比, 在知识表示学习模型中创新性地融合知识图谱的本质结构信息能学习到表达能力更强的实体和关系表示。

(2) 提出一种全新的基因功能相似度分析方法, 利用图卷积知识表示学习模型学习实体和关系的表示, 在语义上分析基因之间的功能相似度。

## 3.2 基于基因本体的相关基因功能相似度分析方法

基因功能相似度广泛应用于计算生物领域, 比如基因功能分析及预测、基因聚类 and 蛋白质相互作用预测。当前大部分基因功能相似度分析方法主要基于基因本体。在基因本体知识图谱中存在三种类别的实体, 分别为: 细胞成分 (Cellular Component, CC), 分子功能 (Molecular Function, MF) 和生物过程 (Biological Process, BP)。基因本体中的实体又称为“术语”, 用于注释基因或者蛋白质, 分别对基因在以上三个方面的功能进行说明。术语间存在七种不同的关系, 表示术语间的语义关联。用于说明注释基因的术语的数据库称作基因本体注释<sup>[52]</sup> (Gene Ontology Annotation, GOA), 基因本体注释详细标注了注释每个基因的三类术语。例如, 基因 ‘IGKV3-7’ 由术语 ‘GO:0002250’, ‘GO:0005886’ 和 ‘GO:0019814’ 等术语注释, 每个术语属于 ‘BP’, ‘CC’ 和 ‘MF’ 中的一类。事实上, 具有相似注释术语的基因在功能上也具有相似性。因此, 传统的基因功能相似度分析方法主要利用注释基因的术语在基因本体中的层次、祖先和后代信息, 结合简单的数学计算方法计算基因功能相似度, 主要分为两类: 成对比较法和成组比较法。

成对比较法分为两步, 第一步计算术语间的语义相似度, 第二步通过整合术语间的语义相似度计算最终的基因功能相似度。Resnik 等人<sup>[53]</sup>提出了利用术语的最低公共祖先具有的语义信息量 (Information Content, IC) 表示术语间的语义相似度, 接着用该方法计算两个基因的注释术语集合中术语两两之间的语义相似度, 最后求平均相似度作为最终的基因功能相似度。由于 Resnik 方法在计算术语间的语义相似度时只考虑最低公共祖先, 因此 Jiang 等人<sup>[54]</sup>和 Lin 等人<sup>[55]</sup>将

术语本身在基因本体中的特异性考虑进去，进而改进了 Resnik 方法。Pesaranghader 等人<sup>[56]</sup>则根据术语在特定注释语料库中的定义计算术语间的语义相似度。

成组比较法认为如果一个术语注释了一个基因，那么该术语的祖先术语也注释该基因，其计算过程主要分为两步，首先将两个基因对应的注释术语集合中术语的祖先术语添加到对应的注释集合中，而后通过计算这两个注释术语集合的相似度作为最终的基因功能相似度。Gentleman 等人<sup>[57]</sup>提出 SimUI 方法，将两个术语集合交集与并集中术语数量的比值作为基因功能相似度。但 SimUI 没有考虑术语的特异性，Pesquita 等人<sup>[58]</sup>则在 SimUI 的基础上将两个术语集合交集与并集中术语的数量比转化为了术语的语义信息量之和的比值。Sánchez 等人<sup>[59]</sup>结合术语在基因本体中的所处位置，考虑节点的特异性，改进了术语语义信息量的计算方式。SORA<sup>[60]</sup>，WIS<sup>[61]</sup>等方法在 Sánchez 等人的基础上，考虑术语的祖先术语和子孙术语关系，进一步深化了术语的特异性。STE<sup>[62]</sup>则同时考虑术语及其祖先在基因本体中的位置信息，同时结合子孙术语的数量，对术语的语义信息量进行更充分的定义，并且，STE 综合考虑术语在基因本体中的继承语义信息，对边的权重进行了全新的定义，以此改进了基因功能相似度计算方式。Benabderrahmane 等人<sup>[63]</sup>和 Zhang 等人<sup>[64]</sup>将基因的注释术语集合利用独热编码形成了注释向量，计算向量间的相似度作为最终得基因功能相似度。

以上方法只能捕获基因本体少量的结构信息，对术语特异性的衡量不够充分，而本章使用基于聚合图卷积网络的知识表示学习模型对基因本体进行建模，学习到术语和关系特定的表示，该表示充分融合了基因本体的结构信息，使术语在基因本体中的特异性得到了充分的体现。

### 3.3 基于聚合的图卷积知识表示学习

如图 3.2 (a) 所示，GOGCN 初始化术语和关系的表示，根据基因本体中的三元组及其结构信息，设计基于聚合的图卷积知识表示学习模型，对术语和关系的表示进行更新。其中，编码器模型由图卷积网络实现，用于聚合中心实体的邻居实体及相连关系的信息并进行融合，解码器模型以编码器模型的输出作为输入，利用基于三元组的知识表示模型以链接预测作为任务对实体和关系的表示进行二次学习。

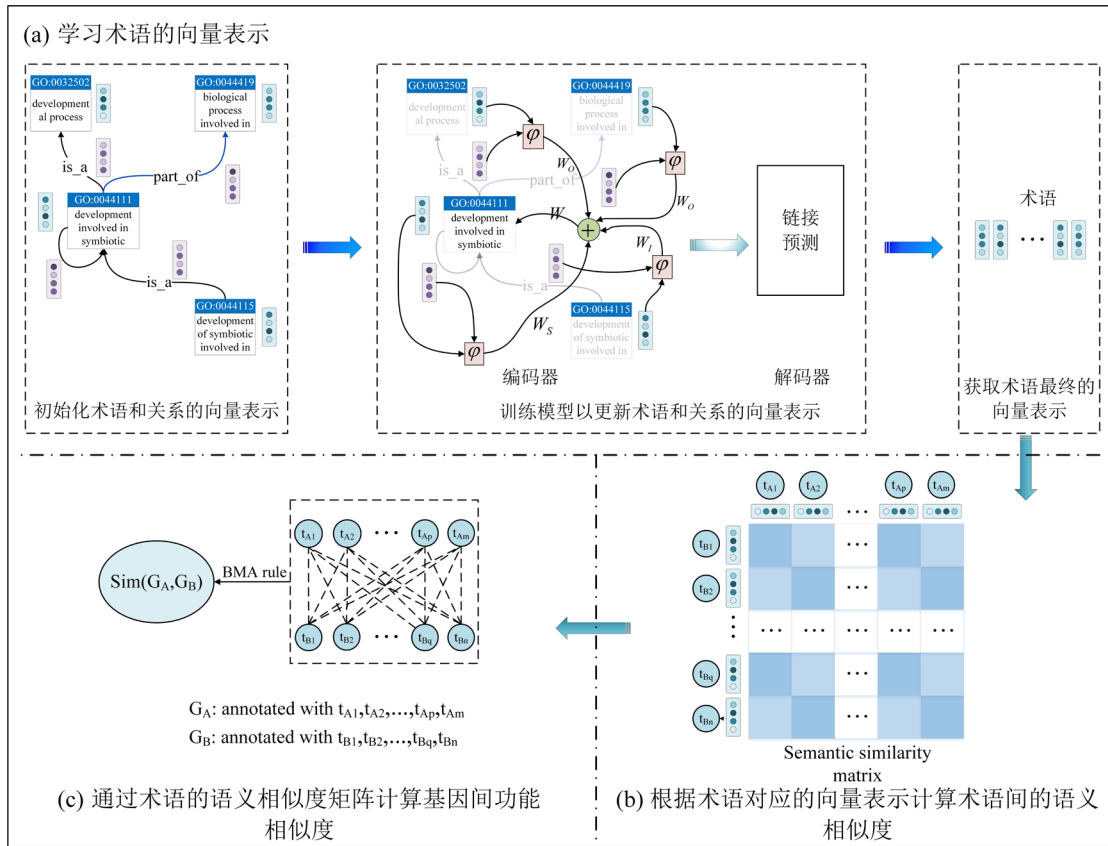


图 3.2 GOGCN 算法结构图

### 3.3.1 符号体系

本节对本章所用的符号体系进行介绍。用  $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T}, \mathcal{X}, \mathcal{Z})$  表示基因本体， $\mathcal{E}$  表示实体集合，也就是术语集合， $\mathcal{R}$  表示关系集合， $\mathcal{T}$  表示基因本体中三元组的集合，也叫边的集合， $\mathcal{X}$  表示术语的初始特征， $\mathcal{Z}$  表示关系的初始特征。同时，本章引入了关系的逆关系，因此对于基因本体中给定三元组  $(h, r, t)$ ，引入逆关系得到三元组的逆三元组  $(t, r^{-1}, h)$ ，扩展了关系集合，即  $\mathcal{R}' = \mathcal{R} \cup \mathcal{R}_{inv}$ ，其中  $\mathcal{R}_{inv}$  表示逆关系集合，有  $\mathcal{R}_{inv} = \{r^{-1} | r \in \mathcal{R}\}$ 。

### 3.3.2 模型架构与学习框架

**编码器模型**用于捕获基因本体的结构信息，首先聚合邻居实体及相连关系的表示得到聚合表示，然后采用图卷积核对聚合表示进行卷积，最后进行信息融合，通过层传播权重参数实现图卷积网络层与层之间的信息传播，定义如下节点表示更新方式：

$$\mathbf{h}_i^{k+1} = \sigma \left( \mathbf{W}^k \left( \sum_{(j,r) \in \mathcal{N}(i)} \mathbf{W}_{t(r)}^k \text{Aggr}(\mathbf{h}_j^k, \mathbf{h}_r^k) + \mathbf{W}_s \mathbf{h}_i^k \right) \right) \quad (3.1)$$

其中， $\mathcal{N}(i)$ 表示实体*i*的一阶出边邻居实体及相连关系组合的集合。在学习中心实体的表示时，对于*i*的入边，即从邻居实体指向中心实体的关系，将其转化为逆关系，则*i*的入边被转化为出边并且不改变三元组表达的语义。此时针对一个中心实体进行表示学习时，因其只存在出边邻居，则相连关系由出边和入边两种类型转化为原始关系（Original relations）和逆关系（Inverse relations）两种类型。*r*和*j*分别表示与*i*相连的关系及邻居实体。 $\mathbf{h}_j^k$ 和 $\mathbf{h}_r^k$ 分别表示实体*j*和关系*r*经过*k*层图卷积网络后更新的向量表示，实体*j*和关系*r*的初始化表示 $\mathbf{h}_j^0$ 和 $\mathbf{h}_r^0$ 分别对应 $\mathcal{X}$ 和 $\mathcal{Z}$ 中第*j*个实体和第*r*个关系的表示。 $\mathbf{W}^k$ 表示第*k*层图卷积网络到第*k*+1层图卷积网络的层间传播参数。 $\mathbf{W}_s$ 表示用于节点自更新的卷积核。*Aggr*表示聚合操作，用于聚合邻居实体及相连关系的信息得到聚合表示，本章采用HolE<sup>[21]</sup>中的循环相关操作聚合，定义如下：

$$\begin{aligned} \text{Aggr}(\mathbf{h}_j, \mathbf{h}_r) &= \mathbf{h}_j \star \mathbf{h}_r \\ [\mathbf{h}_j \star \mathbf{h}_r]_{[k]} &= \sum_{m=0}^{d-1} \mathbf{h}_{j:[m]} \mathbf{h}_{r:[(k+m) \bmod d]} \end{aligned} \quad (3.2)$$

其中， $\star$ 代表循环相关操作， $\mathbf{h}_{[m]}$ 表示向量 $\mathbf{h}$ 在第*m*个位置上的值。 $\mathbf{W}_{t(r)}^k$ 表示基于特定关系类型（原始关系或逆关系）的图卷积核，用于融合邻居实体及相连关系的聚合表示，针对不同类型的关系， $\mathbf{W}_{t(r)}^k$ 有不同的定义，具体如下：

$$\mathbf{W}_{t(r)}^k = \begin{cases} \mathbf{W}_O, & r \in \mathcal{R} \\ \mathbf{W}_I, & r \in \mathcal{R}_{inv} \end{cases} \quad (3.3)$$

同时，在编码器模型中也对关系表示进行学习，使其进行自我更新从而在图卷积网络层间实现信息传播，关系表示更新方式定义如下：

$$\mathbf{h}_r^{k+1} = \mathbf{W}_{rel}^k \mathbf{h}_r^k \quad (3.4)$$

其中， $\mathbf{W}_{rel}^k$ 表示更新关系表示时的图卷积网络层间传播参数。

**解码器模型**以编码器模型输出的实体和关系表示作为输入，独立地对每个三元组进行建模，以链接预测作为任务对节点和关系的表示进行二次更新。本章

选择深度学习模型中基于卷积神经网络的 ConvE 作为解码器模型，定义如下得分函数：

$$f(h, r, t) = \sigma \left( \text{vec} \left( \sigma([\bar{h}; \bar{r}] * \omega) \right) \mathbf{W} \right) t \quad (3.5)$$

其中， $[\cdot; \cdot]$ 表示矩阵的拼接操作， $\mathbf{W}$ 表示线性变换的可学习权重矩阵， $\bar{h}$ 和 $\bar{r}$ 表示 $h$ 和 $r$ 的重塑， $\omega$ 表示卷积核， $\text{vec}(\cdot)$ 表示向量化操作。

在模型训练过程中，本章采用一步训练策略，即将编码器模型和解码器模型集成起来，在解码器独立地处理每个三元组之前，用编码器模型更新所有实体和关系的表示。本章采用二分类交叉熵损失作为模型损失函数优化模型参数。损失函数定义如下：

$$\mathcal{L} = -\frac{1}{N} \sum_i (t_i \cdot \log(p_i) + (1-t_i) \cdot \log(1-p_i)) \quad (3.6)$$

其中， $N$ 表示基因本体中的真实三元组和对应的负例三元组的数量之和， $t_i$ 表示第 $i$ 个三元组的标签，真实三元组标签为1，负例三元组为0， $p_i$ 表示第 $i$ 个三元组的得分。模型训练的目的在于使真实三元组的得分尽可能高而负例三元组的得分尽可能低。本章采用1-N采样策略对负例三元组进行采样，即对于每一个真实三元组，用基因本体中剩下的所有实体替代其头和尾实体，同时排除替代后依然存在于训练集中的三元组，保证采到的负例三元组为错误三元组。

经过以上训练过程，模型使实体间进行了充分的语义交互，同时学到的实体和关系的表示能有效地保留知识图谱的结构信息，从而使实体的特异性有效地封装在了对应的向量表示中。

### 3.4 术语间的语义相似度计算

传统方法利用简单的数学方法，同时结合术语在基因本体中的结构来计算术语间的语义相似度，但是传统方法没有深入挖掘基因本体的结构信息。本章通过基于聚合的图卷积知识表示学习模型建模基因本体，学习基因本体中术语的向量表示，能够有效地保存基因本体的结构信息。如图3.2(b)所示，术语间的语义相似度通过学习到的术语向量间的相似性来度量，本章计算术语对应的向量表示之间的余弦相似度作为术语间的语义相似度，给定两个术语 $t_1$ 和 $t_2$ 及对应的向量表示 $\mathbf{t}_1$ 和 $\mathbf{t}_2$ ，两者语义相似度定义为：

$$S_T(t_1, t_2) = \frac{t_1 \cdot t_2}{|t_1| \cdot |t_2|} \quad (3.7)$$

### 3.5 基因间的功能相似度计算

基于得到的术语间的语义相似度，如图 3.2 (c) 所示，本章利用最佳匹配平均 (Best Match Average, BMA) 规则对其进行整合来计算基因间的功能相似度。特别地，给定两个基因  $G_A$  和  $G_B$  及对应的注释术语集合  $T_A = \{t_{A1}, \dots, t_{Am}\}$  和  $T_B = \{t_{B1}, \dots, t_{Bn}\}$ ，BMA 规则首先整合其中每个集合内术语与另一个集合内所有术语的相似度，通常称为术语与基因间的相似度，以术语  $t_{A1}$  与基因  $G_B$  为例，它们之间的相似度定义为：

$$S_{TG}(t_{A1}, G_B) = \max_{1 \leq i \leq n} (S_T(t_{A1}, t_{Bi})) \quad (3.8)$$

随后，基因  $G_A$  和  $G_B$  间的功能相似度定义为：

$$S_G(G_A, G_B) = \frac{\sum_{i=1}^m S_{TG}(t_{Ai}, G_B) + \sum_{j=1}^n S_{TG}(t_{Bj}, G_A)}{m + n} \quad (3.9)$$

### 3.6 实验设计与结果分析

本节在基因本体数据集上进行基因功能相似度计算实验来评估模型的性能。首先介绍实验所用的数据集，包括基因本体数据和基因本体注释数据、酵母和人类蛋白质相互作用数据集、酵母基因表达数据、CESSM 基于基因本体的语义相似度量度的协同评价 (Collaborative Evaluation of GO-based Semantic Similarity Measures, CESSM) 数据集和酵母的生物通路数据。其次介绍了实验设置、对比方法和评价指标。最后根据实验结果对模型性能进行分析。

#### 3.6.1 实验数据

##### (1) 基因本体数据和基因本体注释数据

基因本体数据下载自基因本体数据库 (2020 年 9 月版)，其中包含 28,922 个 BP 术语，4,193 个 CC 术语，11,157 个 MF 术语和 7 个关系。本章基于基因本体进行基因间功能相似度的计算，和已有文献保持相同的设置，只考虑 ‘is a’ 和

‘part of’ 两种关系而忽略其他关系，因此获得 81,952 个三元组。

同时，酵母基因及人类基因的基因本体注释数据也下载自基因本体数据库（2020 年 10 月版），用于除 CESSM 实验外的其他实验，对于 CESSM 实验，本文从 UniPort 数据库中下载了包含所有物种的基因本体注释数据。基因本体注释数据标识了注释每个基因的术语名称及类别，同时，基因本体注释数据包含电子注释和非电子注释，非电子注释主要通过生物实验观察得来，而电子数据则是根据当前已有注释推断而来，本文将电子注释记为 IEA+，非电子注释记为 IEA-。

#### （2）酵母和人类蛋白质相互作用数据集

蛋白质相互作用（Protein-Protein Interaction, PPI）数据集由众多蛋白质对组成，而每一个蛋白质对应一个基因，由此可以得到对应的基因对。每个基因对的标签为 0 或 1, 0 表示基因对之间不存在相互作用，该对基因被称为负例基因对，1 代表基因对之间存在相互作用，该对基因被称为正例基因对。本文从 Zhang 等人<sup>[64]</sup>的数据集中收集酵母蛋白质相互作用数据，从 Tian 等人<sup>[61]</sup>的数据集中收集人类蛋白质相互作用数据，在去除已经被当前基因本体注释数据库删除的基因以外，本文使用的酵母和人类蛋白质相互作用数据集分别包含 1,113 和 1,251 个正例基因对，提取出以上正例基因对后，随时生成与正例数量相同的负例。

#### （3）酵母基因表达数据

在生物学中，参与相同生物过程或共同实现某种功能的基因之间通常具有较高的表达值，通过评估基因之间表达值和功能相似度的相关性是检验基因功能相似度计算方法是否有效的一个重要方法。基因表达数据由众多基因对及对应的表达值组成，本文从 Jain 等人<sup>[65]</sup>的数据集中下载基因表达数据，由 6,000 个基因对及对应表达值组成，其中分别包含 2,000 个对应 BP, CC, MF 注释的基因对。

#### （4）CESSM 数据集

CESSM 是一个评估基因功能相似度与序列和蛋白质家族（Protein Family, Pfam）相似度之间相关性的在线工具，由于其在 2008 年就已停止更新，比起一直更新的基因本体注释数据，其中的数据相对陈旧，因此，本文从 Tian 等人<sup>[61]</sup>的数据集中下载了 CESSM 数据集中的基因对并进行更新得到不同物种的 10,774 个蛋白质对，随后从在线数据库 UniPort 上获取了数据集中所有基因的 Pfam 注释信息，并在 BP, CC, MF 三个类别上计算 Pfam 相似度。

### (5) 酵母的生物通路数据

酵母的生物通路包含多个反应阶段，每个反应阶段由多个基因共同实现一个具体的生物功能，处于相同反应阶段的基因在功能上应该比处于不同反应阶段的基因间在功能上更具有相似性。为此，本文从 SGD 数据库中收集了 1 条单一的酵母生物通路，用于比较同一生物通路中所有基因之间的功能相似度。同时，酵母包含多个生物通路，处于相同生物通路的基因之间在功能上应该比处于不同生物通路中的基因之间更具有相似性，因此，本文从 KEGG (Kyoto Encyclopedia of Genes and Genomes) 数据库中收集了 5 条不同的生物通路，每条通路包含 11 至 14 个不同的基因，用于比较所有通路中包含的基因间的功能相似度。

### 3.6.2 实验设置

本节主要对 GOGCN 中基于聚合的图卷积知识表示学习模型在基因本体上的最佳超参数进行详细的说明。GOGCN 用 Xavier<sup>[66]</sup>为术语和关系初始化向量表示，术语和关系的初始化嵌入维度为 100，编码器采用单层图卷积网络融合实体和关系的表示，且经过图卷积网络后，术语和关系的嵌入维度设置为 200，学习率从 {0.0001, 0.0005, 0.001} 中选择，经过图卷积网络后术语和关系表示的 dropout<sup>[67]</sup>比率从 {0.1, 0.2, 0.3} 中选择，批量大小从 {64, 128, 256} 中选择，训练迭代次数为 50，用 Adam<sup>[68]</sup>优化器对模型进行优化。经过对超参数进行网格搜索，得到如表 3.1 所示的最佳参数配置。

表 3.1 GOGCN 模型的最佳参数配置

超参数	取值
Initial embedding size	100
GCN embedding size	200
Learning rate	0.001
GCN layers	1
GCN dropout	0.1
Epochs	50
Batch size	128
Optimizer	Adam



### 3.6.3 对比方法

本章选择当前主流的基因功能相似度计算方法与 GOGCN 进行对比，简要说明如下。

**Resnik<sup>[53]</sup>**: Resnik 是最经典的成对比较法，其基于术语在语料库中出现的频率计算术语的语义信息量。

**Wang<sup>[69]</sup>**: Wang 是一个创新性的成对比较法，其根据术语在基因本体中的结构信息来衡量术语间的语义相似度。

**VSM<sup>[64]</sup>**: VSM 是最经典的向量方法，其将基因的注释术语集合通过独热表示转化为一个向量，而后根据向量间的相似度衡量基因间的功能相似度。

**simUI<sup>[57]</sup>**: simUI 是简单的成组比较法，其通过计算基因注释术语集合的交集与并集来衡量基因间的功能相似度。

**simGIC<sup>[58]</sup>**: simGIC 是 simUI 的一个改进方法，其通过基因注释术语集合的语义重叠率来衡量基因间的功能相似度。

**SORA<sup>[60]</sup>**: SORA 充分利用了基因本体的结构信息来计算术语的语义信息量，进而有效地衡量基因间的功能相似度。

**STE<sup>[62]</sup>**: STE 是本章在前期工作中基于传统方法的改进，是一个成组比较法，充分考虑了术语和关系的特异性。

### 3.6.4 评价指标

本章在四种数据集上进行了四个实验来验证 GOGCN 的性能，分别为酵母和人类蛋白质相互作用实验，酵母基因表达数据实验，CESSM 数据集实验以及酵母基因生物通路实验。本节分别对以上四个实验的评价指标进行详细说明。

针对**酵母和人类蛋白质相互作用实验**，用受试者工作特征曲线（Receiver Operating Characteristic curve, ROC curve）的曲线下面积（Area Under Curve, AUC）值作为评价指标。对于蛋白质相互作用数据集来说，通过基因功能相似度计算方法可以计算出每个蛋白质对的功能相似度，而后从 0 到 1 设置阈值，根据功能相似度和阈值的大小比较，结合蛋白质对的标签，就可以得到每个阈值下的真阳性、假阳性、真阴性和假阴性样本的数量，则可以算得真阳性率（True Positive Rate, TPR）、假阳性率（False Positive Rate, FPR）、真阴性率（True Negative Rate, TNR）和假阴性率（False Negative Rate, FNR）。根据每个阈值下的 FPR 和 TPR 值，可以画出对应的 ROC 曲线，进一步可以求出对应的 AUC 值大小。AUC 值

的大小反映了一个基因功能相似度分析方法区分当前蛋白质相互作用数据的能力，AUC 值越大，说明该方法能更准确地判断蛋白质对的真实标签。

针对**酵母基因表达数据实验**，计算出所有基因对之间的功能相似度并将其整合成一个对应的向量，结合基因对间的表达值构造而成的向量，计算这两个向量之间的皮尔逊相关系数。皮尔逊相关系数越大，说明该方法计算出来的基因功能相似度和基因表达值越相关。

针对**CESSM 数据集实验**，与酵母基因表达数据实验类似，计算出所有基因对之间的功能相似度并将其整合成一个对应的向量，结合基因对间的 Pfam 相似度构造而成的向量，计算这两个向量之间的皮尔逊相关系数。皮尔逊相关系数越大，说明该方法与 Pfam 相似度相关性越大。

针对**酵母基因生物通路实验**，对于单一通路实验，计算出通路中所有基因间的功能相似度，构造成相似度矩阵的形式，如果处于同一反应阶段的基因间功能相似度大于处于不同反应阶段的基因间功能相似度，则说明该方法能根据基因实现的功能将它们区分开来。对于多个通路实验，以集判别能力（Set Discriminating Power）作为评价指标，记为 DP。具体来讲，给定包含  $n$  条 KEGG 生物通路的集合  $P = \{P_1, \dots, P_n\}$  及基因功能相似度计算方法  $S_G$ 。以  $P$  中的其中一条通路  $P_k$  为例， $\{g_{k1}, \dots, g_{kp}\}$  表示该通路中包含的基因集合。为了计算  $P_k$  的 DP 值，首先定义如下集内平均相似度  $sim_{Intra}$ ：

$$sim_{Intra}(P_k) = \frac{\sum_{i=1}^{kp} \sum_{j=1}^{kp} S_G(g_{ki}, g_{kj})}{kp^2} \quad (3.10)$$

通过以上计算方式可以计算出每条生物通路中所有基因间的平均相似度。而后，以  $P_k$  和另一条通路  $P_l$  为例， $\{g_{l1}, \dots, g_{lq}\}$  表示  $P_l$  中包含的基因集合，定义两条通路间的集间平均相似度  $sim_{Inter}$ ：

$$sim_{Inter}(P_k, P_l) = \frac{\sum_{i=1}^{kp} \sum_{j=1}^{lq} S_G(g_{ki}, g_{lj})}{kp \times lq} \quad (3.11)$$

最终定义  $P_k$  的 DP 值：

$$DP(P_k) = \frac{(n-1) \times sim_{Intra}(P_k)}{\sum_{i=1, i \neq k}^n sim_{Inter}(P_k, P_i)} \quad (3.12)$$

一条生物通路的集判别能力反映了当前基因功能相似度算法区分该通路与其它生物通路的能力。若一条生物通路中基因间的功能相似度远远大于该通路中基因与其它生物通路中基因之间的相似度,说明该通路的 DP 值越大,进一步说明该基因功能相似度计算方法的有效性。

#### 3.6.5 酵母和人类蛋白质相互作用实验

本节在酵母和人类蛋白质相互作用数据集上进行实验,通过计算每个方法对应的 AUC 值作为实验结果,在酵母蛋白质相互作用数据集上的实验结果见表 3.2,在人类蛋白质相互作用数据集上的实验结果见表 3.3,其中 BP\_IEA+表示针对基因的注释术语只选取 BP 类别的电子注释术语,BP\_IEA-表示针对基因的注释术语只选取 BP 类别的非电子注释术语,其它类似。

从实验结果可以看出,在酵母蛋白质相互作用实验中,GOGCN 在 5 个子实验上均取得了最佳结果,其中在基因注释术语类别为 BP 和 MF 时表现出较大优势,比如在子任务 MF\_IEA-上,GOGCN 的 AUC 值为 0.8312,而第二则为 0.8000。另外,即使在子任务 CC\_IEA-上没有取得最好的结果,对应结果与 simGIC 取得的最好结果差距也不大。在人类蛋白质相互作用实验中,GOGCN 同样在 5 个子任务上均取得了最好的结果。虽然方法 Wang 和 simGIC 表现不错,但是仍与 GOGCN 有差距,比如,在子任务 CC\_IEA-上,GOGCN 分别比方法 Wang 和 simGIC 高 1.30%和 6.07%,在子任务 MF\_IEA+上,GOGCN 分别比方法 Wang 和 simGIC 高 1.89%和 5.88%。另外,STE 在整体上也取得了不错的结果,但是和本章提出的 GOGCN 仍有差距。

总的来说,GOGCN 整体上在酵母和人类蛋白质相互作用实验中取得了最好的结果,同时从实验结果中有以下发现:

(1) 不管是在酵母蛋白质相互作用实验中还是人类蛋白质相互作用实验中,实验结果总是在术语类别 BP,CC 和 MF 上依次降低,说明 BP 类别的注释术语在推测蛋白质相互作用任务上能做出更多的贡献。

(2) GOGCN 本质上属于成对比较法,和方法 Wang 相比,GOGCN 是通过设计基于聚合的图卷积知识表示学习模型来学习术语的向量表示,而后根据向量间的相似度计算术语间的语义相似度,实验结果表明 GOGCN 表现比方法 Wang 更好。同时,即使其他某些传统方法也表现出不错的性能,比如 simGIC 和 STE,GOGCN 仍取得较大优势。这些现象说明通过基于聚合的图卷积知识表示

### 3 基于聚合的图卷积知识表示学习方法

学习能学习到有效的术语表示, 充分挖掘术语的特异性, 并进一步提升对基因本体进行建模的性能。

(3) 虽然 GOGCN 取得了最好的结果, 但是从整体上来看, 成组比较法 (simUI, simGIC, SORA 和 VSM) 在某些情况下表现比成对比较法 (Resnik, Wang 和 GOGCN) 更好, 这和方法 SORA 得到的结论处于一致。

表 3.2 酵母蛋白质互作数据实验的 AUC 值

对比方法	BP_IEA+	BP_IEA-	CC_IEA+	CC_IEA-	MF_IEA+	MF_IEA-
Resnik	0.7926	0.7977	0.7852	0.7762	0.7506	0.7611
Wang	0.8718	0.8676	0.8416	0.8138	0.7699	0.8000
simUI	0.8515	0.8376	0.8002	0.7809	0.7600	0.7711
simGIC	0.8784	0.8680	0.8262	<b>0.8145</b>	0.7843	0.7940
SORA	0.8762	0.8653	0.8140	0.8031	0.7899	0.7985
STE	0.8401	0.8291	0.7996	0.7915	0.7564	0.7600
VSM	0.8545	0.8394	0.8010	0.7824	0.7615	0.7713
GOGCN	<b>0.8968</b>	<b>0.8837</b>	<b>0.8457</b>	0.8090	<b>0.8132</b>	<b>0.8312</b>

表 3.3 人类蛋白质互作数据实验的 AUC 值

对比方法	BP_IEA+	BP_IEA-	CC_IEA+	CC_IEA-	MF_IEA+	MF_IEA-
Resnik	0.8550	0.8604	0.7488	0.7536	0.7435	0.7615
Wang	0.9247	0.9195	0.8344	0.8311	0.7689	0.7545
simUI	0.8993	0.8922	0.7614	0.7605	0.6520	0.6551
simGIC	0.9227	0.9154	0.7955	0.7937	0.7399	0.7579
SORA	0.9147	0.9096	0.7676	0.7722	0.7014	0.7040
STE	0.9217	0.9140	0.7951	0.7951	0.7489	<b>0.7665</b>
VSM	0.9081	0.8986	0.7648	0.7641	0.6428	0.6519
GOGCN	<b>0.9323</b>	<b>0.9228</b>	<b>0.8378</b>	<b>0.8419</b>	<b>0.7834</b>	0.7653

#### 3.6.6 酵母基因表达数据实验

本节利用基因功能相似度计算方法计算出的所有基因对的功能相似度, 结合基因对对应的基因表达值, 使用皮尔逊相关系数计算方法评估它们之间的相

关性，结果如表 3.4 所示。

从结果来看，GOGCN 在四个子任务上取得了最好的结果，并且与基因表达值的相关性远远高于其他对比方法。同时，在子任务 BP\_IEA- 上，GOGCN 和取得第一的方法 simGIC 几乎并驾齐驱，在子任务 MF\_IEA+ 上，GOGCN 也取得了不错的结果。总的来说，GOGCN 与基因表达数据之间表现出了最高的相关性，这也在一定程度上证明了 GOGCN 的有效性。另外，大多数方法在注释术语类别为 CC 时与基因表达值具有较高相关性，说明 CC 类别的注释术语更能反应基因功能相似度与基因表达值间的关联。

表 3.4 酵母基因表达数据与基因功能相似度间的皮尔逊相关系数

对比方法	BP_IEA+	BP_IEA-	CC_IEA+	CC_IEA-	MF_IEA+	MF_IEA-
Resnik	0.3111	0.2690	0.2922	0.2976	0.3023	0.3127
Wang	0.3086	0.2939	0.4221	0.4322	0.3048	0.3084
simUI	0.3449	0.3542	0.3872	0.3465	0.3575	0.3069
simGIC	0.3892	<b>0.3917</b>	0.3583	0.3450	0.3529	0.3208
SORA	0.3128	0.3535	0.4017	0.3808	<b>0.3622</b>	0.3377
STE	0.3810	0.3897	0.3464	0.3252	0.3348	0.2801
VSM	0.2898	0.3017	0.4077	0.3611	0.3280	0.2861
GOGCN	<b>0.4051</b>	0.3864	<b>0.4329</b>	<b>0.4500</b>	0.3303	<b>0.3529</b>

### 3.6.7 CESSM 数据集实验

一个 Pfam 一般表示蛋白质的进化过程，那么具有相同 Pfam 注释的蛋白质在功能上更具有相似性。在 CESSM 数据集实验中，所有基因对之间的 Pfam 相似度采用杰卡德系数 (Jaccard index) 计算得到，即由两个蛋白质共享的 Pfam 数量与它们所属 Pfam 总数之比。而后，结合基因功能相似度计算方法计算得到的基因间的功能相似度，可以得到基因功能相似度与 Pfam 相似度之间的皮尔逊相关系数。实验结果如表 3.5 所示，其中 1,000, 2,000 和 5,000 表示将数据集中的基因对分成了对应的组数，组内多个蛋白质对的相似度取平均，然后构造成以对应组数为维度的向量。

整体来看，方法 simGIC, SORA, STE 和 GOGCN 的整体表现优于剩下的方法，尤其在子任务 BP (1000), CC (1000) 和 MF (1000) 上，实验结果均超

过了 0.8。同时，随着分组的数量增加，基因功能相似度与 Pfam 相似度的相关性逐渐降低，说明相关性与分组数量呈负相关。

进一步说，GOGCN 在 8 个指标上均取得了最好的结果，即使在子任务 MF (5000) 上，GOGCN 也比取得第一的 simGIC 低 0.001。实验结果说明比起大部分对比方法，GOGCN 方法更优越，进一步证明 GOGCN 对基因本体建模的有效性。

表 3.5 蛋白质间的 Pfam 相似度与基因功能相似度间的皮尔逊相关系数

对比方法	BP			CC			MF		
	1000	2000	5000	1000	2000	5000	1000	2000	5000
Resnik	0.695	0.576	0.414	0.725	0.631	0.474	0.727	0.623	0.467
Wang	0.836	0.733	0.564	0.861	0.774	0.610	0.859	0.782	0.626
simUI	0.851	0.753	0.596	0.877	0.807	0.669	0.881	0.808	0.665
simGIC	0.849	0.758	0.606	0.854	0.780	0.646	0.893	0.824	<b>0.683</b>
SORA	0.876	0.788	0.631	0.875	0.800	0.658	0.863	0.786	0.641
STE	0.871	0.789	0.599	0.872	0.810	0.648	0.893	0.785	0.663
VSM	0.860	0.762	0.599	0.884	0.812	0.672	0.846	0.765	0.618
GOGCN	<b>0.887</b>	<b>0.794</b>	<b>0.641</b>	<b>0.898</b>	<b>0.842</b>	<b>0.706</b>	<b>0.896</b>	<b>0.829</b>	0.682

### 3.6.8 酵母基因生物通路实验

本节设计两个酵母基因生物通路实验，其一是针对单一生物通路，检验基因功能相似度计算方法是否能区分同一生物通路中处于不同反应阶段的基因。其二是针对生物通路集合，检验基因功能相似度计算方法是否能区分处于不同生物通路中的基因。

**针对单一生物通路**，选择通路 ‘L-tyrosine degradation III’ 为例，其详细信息如表 3.6 所示，该通路包含 10 个处于三个反应阶段的基因。在本实验中，本章选择方法 Resnik, simUI 和 simGIC 基于 MF 类别的注释术语与 GOGCN 进行对比，计算这 10 个基因两两之间的基因功能相似度并构造成相似度矩阵，如图 3.3 所示。这四个方法均能将这 10 个基因通过基因间相似度分为准确的三类。

从 GOGCN 的角度来看，处于相同反应阶段的基因间的功能相似度远远高于处于不同反应阶段的基因间的功能相似度，并且处于相同反应阶段的基因间

的功能相似度最低为‘0.63’，处于不同反应阶段的基因间的功能相似度在‘0.2’至‘0.4’变化，这在很大程度上非常合理，因为处于相同阶段的基因间相似度应该比较大，其次即使两个基因处于不同的反应阶段，但是它们处于同一生物通路，它们之间的相似度也不应该很小，甚至为0。

对于方法 Resnik，即使它能准确的区分开 10 个基因，但是仍然存在两个问题，第一，以基因‘PDC1’和‘PDC5’为例，它们同时被相同的术语注释，但是 Resnik 计算出它们的相似度为‘0.79’而不为‘1’，第二，基因‘ADH1’与处于同一反应阶段的其它基因之间的相似度为 0.43，在一定程度上过低。

方法 simUI 整体上表现不错，但基因‘ADH1’与处于相同反应阶段的其他基因之间的相似度为‘0.54’，比起 GOGCN 计算出的‘0.63’有所不足。

方法 simGIC 虽然完美的区分了三类基因，但是对于处于不同反应阶段的基因之间，simGIC 计算出的相似度为‘0’，这在很大程度上是不合理的。

因此，总的来说，GOGCN 在单一生物通路中基因分类实验上取得了最好的表现。

表 3.6 生物通路‘L-tyrosine degradation III’中的基因及其详细信息

类别	反应阶段	基因名称
1	1.1.1.1	ADH1
		ADH2
		ADH3
		ADH4
		ADH5
2	2.6.1-	ARO8
		ARO9
3	4.1.1.80	PDC1
		PDC5
		PDC6

针对生物通路集合，选取 5 个 KEGG 生物通路进行实验，详细信息如表 3.7 所示，并选择 Resnik, Wang 和 VSM 作为对比方法与 GOGCN 一起计算每条通路的 DP 值，结果如表 3.8 所示。GOGCN 在 4 条生物通路上计算出的 DP 值均大于对比方法的结果。并且，尽管在通路‘sce00514’上 GOGCN 计算的 DP 值

### 3 基于聚合的图卷积知识表示学习方法

小于 Resnik 的结果，但是也优于其余两个对比方法。因此，从整体上可以看出 GOGCN 方法的优越性。

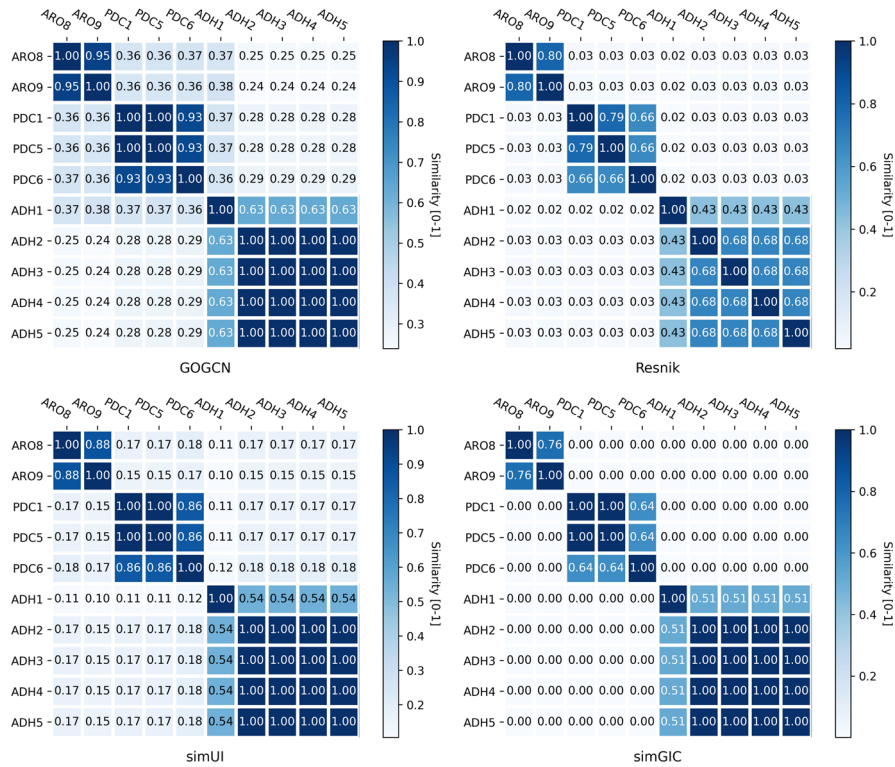


图 3.3 单一生物通路中基于基因功能的基因分类结果

表 3.7 实验所用生物通路集合中 KEGG 通路及其详细信息

通路 ID	通路名称	通路包含基因数量
sce00053	Ascorbate and aldarate metabolism	11
sce00290	Valine, leucine and isoleucine biosynthesis	12
sce00350	Tyrosine metabolism	13
sce00514	Other types of O-glycan biosynthesis	14
sce00790	Folate biosynthesis	11



表 3.8 基于 KEGG 通路的集判别能力实验结果

对比方法	sce00053	sce00290	sce00350	sce00514	sce00790
Resnik	0.9087	1.2750	0.9901	<b>2.3778</b>	1.1812
Wang	1.0003	1.2292	1.1840	1.6300	1.1705
VSM	1.0439	1.2285	1.1703	0.5872	1.1457
GOGCN	<b>1.0589</b>	<b>1.2886</b>	<b>1.2192</b>	1.7684	<b>1.1877</b>

### 3.6.9 消融实验

本节在人类蛋白质相互作用数据集上进行了消融实验探究 GOGCN 模型中组件的有效性。

(1) 首先, 本节探究在邻居实体及相连关系间采用不同的聚合操作对模型性能的影响, 引入了“加”操作和“乘”操作与本章采用的循环相关操作进行对比。“加”操作和“乘”操作分别记为 ‘add’ 和 ‘mult’, 定义如下:

$$\begin{aligned} Aggr(h_j, h_r)_{add} &= h_j + h_r \\ Aggr(h_j, h_r)_{mult} &= h_j * h_r \end{aligned} \quad (3.13)$$

实验结果如图 3.4 所示, 整体结果表明, 比起“加”和“乘”操作, 采用循环相关操作在 5 个子任务上均取得了最好的结果。虽然在子任务 ‘MF\_IEA-’ 采用“加”操作优于循环相关操作, 但是从整体看采用循环相关操作更有利于对邻居实体及相连关系的信息进行聚合, 进一步说明, 因为“加”和“乘”操作相对简单, 所以采用较复杂的聚合操作对 GOGCN 模型更有利。

(2) 其次, 本节研究对关系的不同处理方式对模型性能的影响。基于聚合的图卷积知识表示学习模型同时学习了术语和关系的向量表示, 但是在之后进行基因功能相似度计算的时候只使用了术语的向量表示而忽略了关系的表示。因此, 为了证明有必要对关系表示进行学习, 本节引入了 GOGCN 的两个变体, 其一, 在编码器中忽略关系的表示学习, 这种情况记为 ‘noRelation’, 其二, 由于只使用了关系 ‘is a’ 和 ‘part of’, 这里将它们视为一个关系引入 ‘oneRelation’ 变体与 GOGCN 对比探究它们之间的区别。实验结果如图 3.5 所示。

从结果来看, ‘oneRelation’ 较于 ‘noRelation’ 在 4 个子任务上均表现出较好的性能, 说明在模型里引入关系的表示学习能对术语的表示学习起到积极的作用。同时, GOGCN 在所有子任务上表现都优于 ‘noRelation’, 并且在 5 个子

任务上优于 ‘oneRelation’, 这说明不管是 ‘is a’ 还是 ‘part of’ 都具有它们自己特定的意义, 并且能对术语的表示学习起到关键的积极作用。

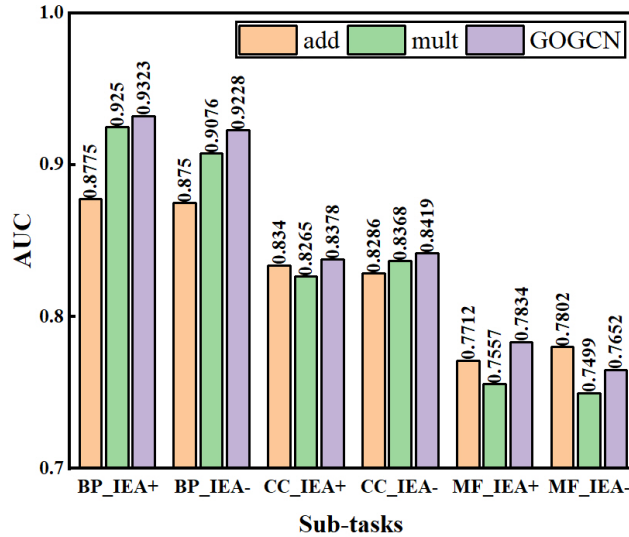


图 3.4 不同聚合操作对模型性能的影响结果

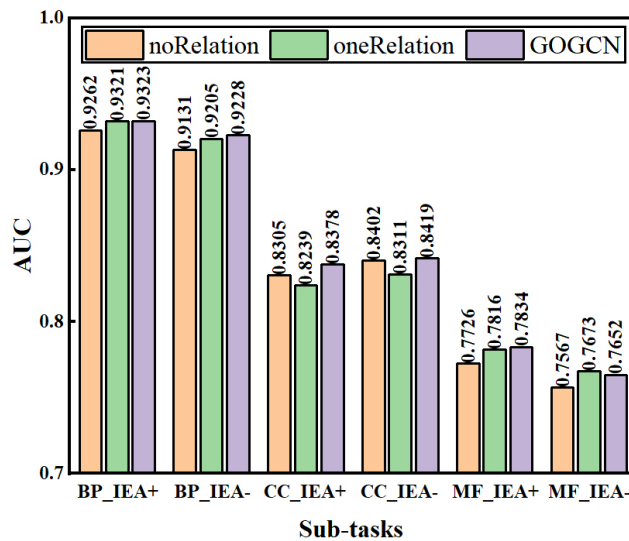


图 3.5 关系的不同处理方式对模型性能的影响结果

(3) 最后, 本节使用解码器模型 ConvE 建模基因本体并学习术语的表示用于基因功能相似度计算, 与 GOGCN 进行对比, 验证模型中编码器的有效性。

如图 3.6 所示, GOGCN 在 5 个子任务上都比 ConvE 表现出色, 尤其在注释术语类别为 BP 和 CC 时, GOGCN 远远优于 ConvE, 这说明了在编码器中设计

的基于聚合的图卷积网络模型能很好地对基因本体进行建模，通过编码器聚合的邻居实体及相连关系的信息非常有效，并在很大程度上将基因本体丰富的结构信息融入到术语和关系的向量表示中。

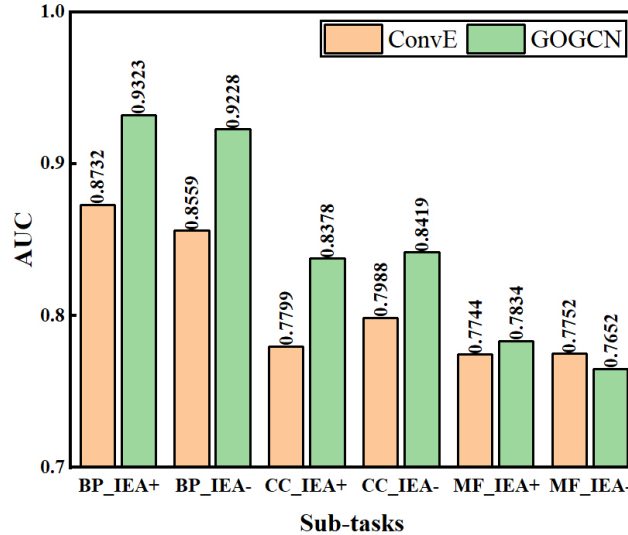


图 3.6 编码器和解码器对模型性能的影响

### 3.7 本章小结

知识图谱丰富的结构信息会对实体和关系的表示学习贡献大量有用的语义信息，主流的基于单个三元组的知识表示学习模型在建模知识图谱时总是独立地处理每个三元组，会忽略知识图谱中邻域结构对实体和关系表示学习的语义贡献，造成学到的实体和关系表示不能很好地蕴含知识图谱的语义信息。为了解决这个问题，本章提出基于聚合的图卷积网络知识表示学习方法，通过设计高效的图卷积网络模型，结合编码器-解码器结构，使实体表示蕴含更多的邻居实体及相连关系的信息，同时让关系表示进行自动更新。本章针对基因本体进行建模，学习术语的表示以评估基因间的功能相似度，经过丰富的实验证明了模型先进的性能。

## 4 基于双注意力的图卷积知识表示学习方法

### 4.1 问题引入

近年来,随着图神经网络在建模图数据上引起的广泛关注,大量研究者们将目光聚焦于利用图神经网络建模知识图谱。知识图谱本质的图结构信息成为了实现图神经网络的天然载体。本文在前一章探索了利用图卷积网络融合实体的部分邻域信息来学习有效的实体表示。本章关注如何利用每个实体和关系所处的整个邻域信息来提升建模知识图谱的性能,从而学习到融合更多知识图谱信息的实体和关系表示,以此促进知识图谱中的实体与实体之间和实体与关系之间的语义交互。

当前大多数利用图神经网络建模知识图谱的工作主要关注于实体的表示学习,以同等重要性融合邻居实体或邻域的信息来更新中心实体的表示,同时,忽略关系的表示学习或者使关系表示进行自我更新。基于图卷积网络的方法如 R-GCN<sup>[37]</sup>, VRGCN<sup>[70]</sup>, WGCN<sup>[38]</sup>和 CompGCN<sup>[48]</sup>,总以同等重要性聚合邻居实体的信息同时忽略关系的表示学习或使关系表示进行自我更新,该类方法不能很好地使学到的实体和关系表示蕴含知识图谱的信息。由于图注意力网络的崛起,基于图注意力网络的知识表示方法如 KBGAT<sup>[39]</sup>和 RGHAT<sup>[40]</sup>通过在模型中引入注意力机制以不同的重要性融合邻居实体或邻域的信息,该类方法在评估邻域或邻居实体的重要性时忽略了关系方向的影响,因为关系的方向在一定程度上具有语义信息,同时,它们仍然使关系表示进行自我更新,忽略关系所处邻域对其表示学习的影响,造成一定程度的语义丢失。下面以一个例子来说明实体和关系所在的邻域对它们的表示学习过程造成的影响

如图 4.1 (a) 所示,中心实体 ‘Oliver Stone’ 由不同的出边和入边与邻居实体联系在一起组成众多邻域,以学习中心实体 ‘Oliver Stone’ 的向量表示为例,可以观察到如下现象:(1) ‘Oliver Stone’ 存在众多邻居实体,但 ‘Director’ 可能为 ‘Oliver Stone’ 的表示学习贡献更多的信息,因为 ‘Oliver Stone’ 以电影导演闻名于世界。这个例子说明不同的邻域对中心实体的表示学习存在不同的贡献。(2) ‘New York’ 和 ‘USA’ 分别以出边关系 ‘Born\_in’ 和 ‘Nationality’ 与 ‘Oliver Stone’ 相连,可以推测出 ‘New York’ 很可能是 ‘USA’ 的一部分,同

时, ‘Snowden’ 以入边关系 ‘Directed\_by’ 与 ‘Oliver Stone’, 可以推测出 ‘Oliver Stone’ 很可能导演了不止一部电影。这个例子说明邻居实体以不同方向的关系 (入边或出边) 与中心实体相连很可能为中心实体的表示学习贡献不同的语义。因此, 本章提出基于关系方向的双向注意力机制动态评估邻域结构的重要性。与前一章类似, 本章引入逆关系将实体的入边转化为出边, 则一个实体与其邻居实体间的相连关系则由出边关系和入边关系转化为原始关系和逆关系。如图 4.1(b) 所示, 中心实体 ‘Oliver Stone’ 的入边关系 ‘Son\_of’ 和 ‘Directed\_by’ 被转化为相应的逆关系。具体来讲, 在实体表示学习中引入的双向注意力机制首先根据相连关系类型将邻居实体分成两个集合, 而后分别计算这两个集合内与中心实体组成的邻域的注意力值用于融合对应邻域的信息。

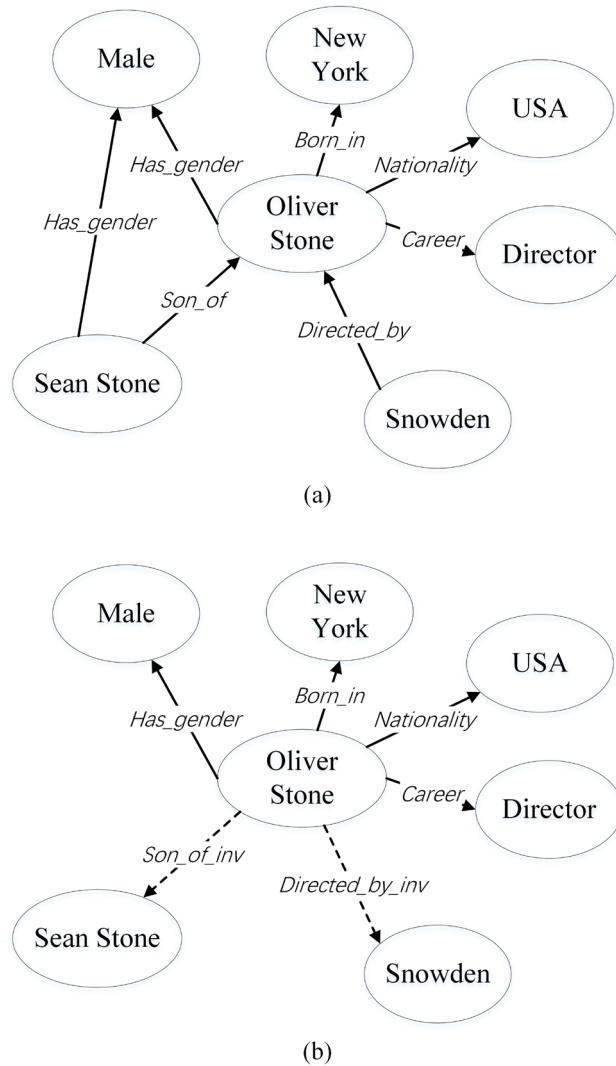


图 4.1 以 ‘Oliver Stone’ 为中心的简要知识图谱

对于关系的表示学习，仍以图 4.1 (a) 为例，可以观察到：(1) 以关系 ‘Has\_gender’ 和 ‘Born\_in’ 为例，它们在知识图谱中本身就具有特定的语义，因此为每个关系学习一个独特的表示能建模它们的不同，这是非常有意义的。(2) 关系 ‘Nationality’ 和 ‘Career’ 处于不同的邻域中，呈现出不同的语义，因此它们所在的特定邻域应该对它们的表示学习贡献特定的语义信息。同时，考虑邻域对关系表示学习的贡献，能反过来促进头实体与尾实体间的语义交互。(3) 实体 ‘Oliver Stone’ 和 ‘Sean Stone’ 均通过关系 ‘Has\_gender’ 与实体 ‘Male’ 相连，说明不止一个邻域包含关系 ‘Has\_gender’。该例子说明同一个关系所处的不同邻域应该对该关系的表示学习有不同的贡献。因此，本章提出关系注意力机制动态评估邻域结构的重要性用于关系的表示学习。

总的来说，本章通过融合知识图谱中的邻域信息同时学习实体和关系的向量表示，提出基于双注意力的图卷积知识表示学习模型 (Learning knowledge graph embedding with a dual-attention embedding network, D-AEN)。值得注意的是，本章提出的双向注意力机制和关系注意力机制通过一个 GCN 编码器同时评估邻域的重要性来分别学习实体和关系的表示，使实体与关系之间和实体与实体之间进行充分的语义交互，最后学到的实体和关系表示尽可能多地保存了知识图谱本质的结构信息。本章的主要创新总结如下：

(1) 提出了一个新颖的知识表示学习框架 D-AEN，其中实体和关系的表示以一种端到端的方式被同时学习且被用于促进彼此之间的优化。

(2) 提出一种基于关系方向的双向注意力机制以动态评估邻域的重要性用于学习实体的表示。

(3) 提出一种关系注意力机制以动态评估邻域的重要性用于学习关系的表示。

(4) 在标准链接预测数据集上的实验结果，相比对比模型具有显著的提升。

## 4.2 基于双注意力的图卷积知识表示学习

本章采用前一章使用的编码器-解码器结构设计 D-AEN 模型，本节主要对模型的整体框架进行详细地介绍，包括编码器模型，解码器模型和模型的训练与学习框架。同时，详细说明实验设计并针对实验结果进行分析。在这之前先介绍本章所用的符号体系。

### 4.2.1 符号体系

本节对本章所用符号体系进行介绍。用  $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$  表示知识图谱， $\mathcal{E}$  表示实体集合， $\mathcal{R}$  表示关系集合， $\mathcal{T}$  表示知识图谱中三元组的集合，也叫边的集合。 $t_{ij}^k = (e_i, r_k, e_j)$  表示知识图谱中第  $i$  和  $j$  个实体作为头和尾实体，第  $k$  个关系作为相连关系的三元组。与前一章一样，由于知识图谱是一个有向图，为了使实体和邻居实体间的关系方向一致，引入关系的逆关系同时构造了三元组的逆三元组，即  $\mathcal{R}' = \mathcal{R} \cup \mathcal{R}_{inv}$ ， $\mathcal{T}' = \mathcal{T} \cup \mathcal{T}_{inv}$ ，其中  $\mathcal{R}_{inv}$  和  $\mathcal{T}_{inv}$  分别表示逆关系集合与逆三元组集合，有  $\mathcal{R}_{inv} = \{r_k^{-1} | r_k \in \mathcal{R}\}$  和  $\mathcal{T}_{inv} = \{(e_j, r_k^{-1}, e_i) | (e_i, r_k, e_j) \in \mathcal{T}\}$ 。同时，本章将一个实体或关系所在的三元组称为该实体或关系所处的邻域。

### 4.2.2 模型架构与学习框架

编码器模型结构如图 4.2 所示，其中，本章除了考虑原始关系与原始三元组外，还将逆关系与逆三元组引入知识表示学习，并同时以适当的重要性将实体和关系所在的邻域信息融入对应的实体和关系表示中。为了简便，在图 4.2 和本节中只针对单层编码器模型进行详细说明。

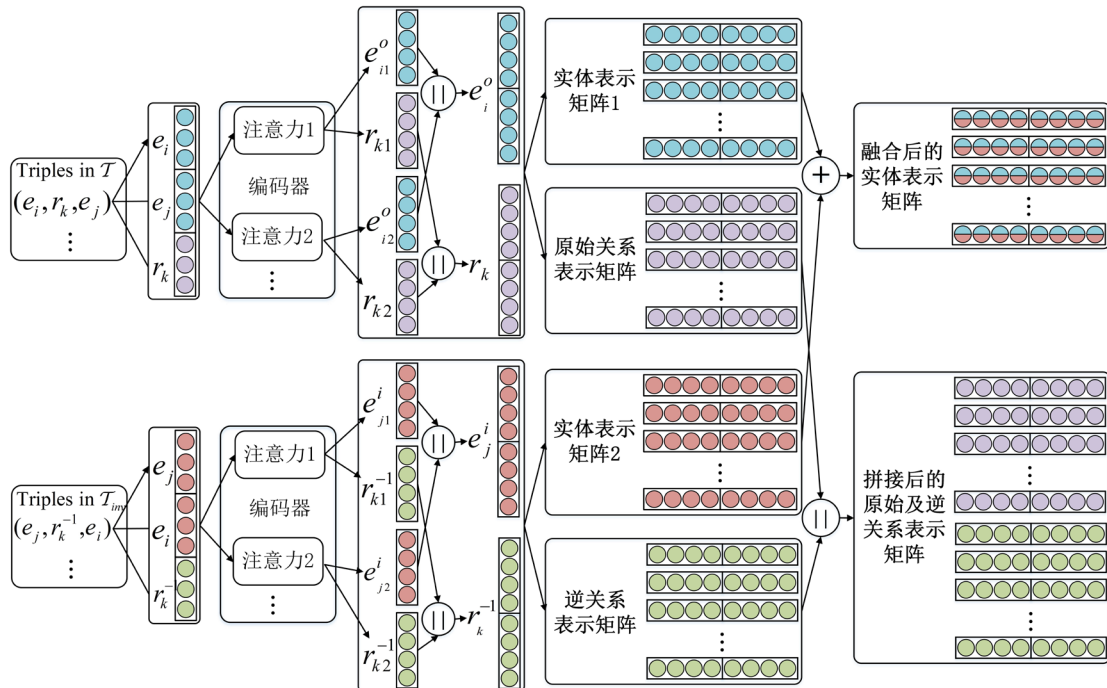


图 4.2 D-AEN 模型中的单层编码器模型示意图

具体来说, 和 KBGAT 一样, 给定  $\mathcal{T}'$  中的三元组  $t_{ij}^k = (e_i, r_k, e_j)$ , 三元组的表示定义为:

$$\mathbf{v}_{ijk} = \mathbf{W}_1 \cdot [\mathbf{h}_i \parallel \mathbf{h}_j \parallel \mathbf{g}_k] \quad (4.1)$$

其中,  $\mathbf{h}_i$  和  $\mathbf{h}_j$  分别表示实体  $e_i$  和  $e_j$  的初始向量表示,  $\mathbf{g}_k$  表示关系  $r_k$  的初始向量表示,  $[\parallel]$  表示向量的拼接操作,  $\mathbf{W}_1$  表示线性变换矩阵。

三元组  $t_{ij}^k$  的重要性表示为:

$$b_{ijk} = \text{LeakeyReLU}(\mathbf{a} \cdot \mathbf{v}_{ijk}) \quad (4.2)$$

其中,  $\mathbf{a}$  表示注意力向量, LeakeyReLU 表示非线性激活函数, 其负斜率取值为 0.2。

三元组  $t_{ij}^k$  的相对注意力值则通过在  $b_{ijk}$  上应用 Softmax 计算得到。为了使实体和关系进行充分的语义交互, 三元组的表示被同时引入实体和关系的表示学习。接下来对实体和关系的表示学习分别进行说明。

(1) 实体的表示学习: 在实体的表示学习中, 使用基于关系方向的双向注意力机制分别评估实体所在具有不同关系类型的邻域的重要性, 表示如下:

$$\begin{aligned} \alpha_{ijk} &= \text{softmax}(b_{ijk}) = \frac{\exp(b_{ijk})}{\sum_{(r, e_n) \in \mathcal{N}_{e(i)}} \exp(b_{inr})}, \quad (r_k, r_r \in \mathcal{R}) \\ \beta_{ijk} &= \text{softmax}(b_{ijk}) = \frac{\exp(b_{ijk})}{\sum_{(r, e_n) \in \mathcal{N}_{e(i)}} \exp(b_{inr})}, \quad (r_k, r_r \in \mathcal{R}_{inv}) \end{aligned} \quad (4.3)$$

其中,  $\mathcal{N}_{e(i)} = \{(r_k, e_j) \mid (e_i, r_k, e_j) \in \mathcal{T}'\}$  表示实体  $e_i$  的尾实体及相连关系组合的集合。

中心实体所处的某些邻域若具有代表性, 则在实体的表示学习中这些邻域对中心实体的表示学习影响更大。因此, 本章结合实体所在邻域的注意力值融合实体所在的邻域信息, 同时保存实体自身的信息定义如下的实体表示更新方式:

$$\mathbf{h}'_i = \mathbf{h}'_{i\_0} + \mathbf{h}'_{i\_1} + \mathbf{W}_e \cdot \mathbf{h}_i \quad (4.4)$$

其中,  $\mathbf{h}_i$  和  $\mathbf{h}'_i$  分别表示实体  $e_i$  的原始表示和更新后的表示,  $\mathbf{W}_e$  表示线性变换矩阵, 用于保存节点自身的信息,  $\mathbf{h}'_{i\_0}$  和  $\mathbf{h}'_{i\_1}$  分别表示融合实体  $e_i$  所在原始三元组与逆三元组的信息后的表示, 具体定义为:



$$\begin{aligned} \mathbf{h}'_{i_o} &= \sigma \left( \mathbf{W}_o \sum_{(r_k, e_j) \in \mathcal{N}_{e(i)}} \alpha_{ijk} \mathbf{v}_{ijk} \right), \quad (r_k \in \mathcal{R}) \\ \mathbf{h}'_{i_l} &= \sigma \left( \mathbf{W}_l \sum_{(r_k, e_j) \in \mathcal{N}_{e(i)}} \beta_{ijk} \mathbf{v}_{ijk} \right), \quad (r_k \in \mathcal{R}_{inv}) \end{aligned} \quad (4.5)$$

其中,  $\mathbf{W}_o$  和  $\mathbf{W}_l$  表示线性变换矩阵, 分别用于融合原始三元组的信息和逆三元组的信息,  $\sigma$  表示非线性激活函数。

为了尽可能多地融合邻域的信息且使学习过程更稳定, 本章运用和 GAT 一样的多头注意力机制到实体的表示学习过程中。具体来讲, 给定  $M$  个注意力头, 为每个节点学习  $M$  个独立的表示然后连接起来作为实体的最终表示, 定义为:

$$\mathbf{h}'_i = \parallel_{m=1}^M \mathbf{h}'_{im} \quad (4.6)$$

(2) 关系的表示学习: 与实体的表示学习过程类似, 对于原始关系与逆关系集合  $\mathcal{R}' = \mathcal{R} \cup \mathcal{R}_{inv}$  中的每一个关系  $r_k$ , 仍然以自适应的权重融合关系所在的邻域信息。使用关系注意力机制计算  $r_k$  所处的每个邻域的注意力值, 定义如下:

$$\gamma_{ijk} = \text{softmax}(b_{ijk}) = \frac{\exp(b_{ijk})}{\sum_{(e_m, e_n) \in \mathcal{N}_{r(k)}} \exp(b_{mnk})} \quad (4.7)$$

其中,  $\mathcal{N}_{r(i)} = \{(e_i, e_j) | (e_i, r_k, e_j) \in \mathcal{T}\}$  表示与关系  $r_k$  相连的头、尾实体对的集合。

结合关系所在邻域的注意力值融合关系所在的邻域信息, 同时保存关系自身的信息, 定义如下的关系表示更新方式:

$$\mathbf{g}'_k = \mathbf{g}'_{k_N} + \mathbf{W}_r \cdot \mathbf{g}_k \quad (4.8)$$

其中,  $\mathbf{g}_k$  和  $\mathbf{g}'_k$  分别表示关系  $r_k$  的原始表示和更新后的表示,  $\mathbf{W}_r$  表示线性变换矩阵, 用于保存关系自身的信息,  $\mathbf{g}'_{k_N}$  表示融合关系  $r_k$  所在邻域的信息后的表示, 定义为:

$$\mathbf{g}'_{k_N} = \sigma \left( \mathbf{W}_R \sum_{(e_i, e_j) \in \mathcal{N}_{r(i)}} \gamma_{ijk} \mathbf{v}_{ijk} \right) \quad (4.9)$$

其中,  $\mathbf{W}_R$  表示线性变换矩阵。

与学习实体表示类似, 仍然在关系的表示学习过程中运用多头注意力机制,

给定  $M$  个注意力头，关系  $r_k$  的最终表示定义为：

$$\mathbf{g}_k^i = \parallel_{m=1}^M \mathbf{g}_{km}^i \quad (4.10)$$

本章在解码器之前采用两层编码器进行实体和关系表示的更新，其中第一层采用多头注意力机制学习多个实体和关系的向量表示并连接起来，第二层采用单头注意力机制学习最终的实体和关系表示。结合前面详细描述的单层编码器结构，算法 4.1 详细描述了使用多层编码器实现实体和关系表示学习的过程。

**解码器模型**以编码器模型的输出作为输入。本章选择 ConvE<sup>[25]</sup>模型作为解码器模型，同时也尝试了其他模型如 DistMult<sup>[19]</sup>和 TransE<sup>[11]</sup>，但是发现 ConvE 效果最好。对知识图谱中的三元组  $t_{ij}^k = (e_i, r_k, e_j)$ ，ConvE 定义如下得分函数：

$$f(e_j, r_k, e_i) = \sigma \left( \text{vec} \left( \sigma([\bar{\mathbf{h}}_i; \bar{\mathbf{g}}_k] * \omega) \right) \mathbf{W} \right) \mathbf{h}_j \quad (4.11)$$

其中， $[\cdot; \cdot]$ 表示矩阵的拼接操作， $\mathbf{W}$ 表示线性变换的可学习权重矩阵， $\bar{\mathbf{h}}_i$ 和 $\bar{\mathbf{g}}_k$ 表示 $\mathbf{h}_i$ 和 $\mathbf{g}_k$ 的重塑， $\omega$ 表示卷积核， $\text{vec}(\cdot)$ 表示向量化操作。

**在模型训练过程中**，本章和前一章一样采用一步训练策略，将编码器与解码器模型联合起来一起训练，使用二分类交叉熵损失函数用于在训练过程中优化模型，定义如下：

$$\mathcal{L} = -\frac{1}{N} \sum_i (t_i \cdot \log(p_i) + (1-t_i) \cdot \log(1-p_i)) \quad (4.12)$$

其中， $N$ 表示基因本体中的真实三元组和对应的负例三元组的数量之和， $t_i$ 表示第 $i$ 个三元组的标签，真实三元组标签为1，负例三元组为0， $p_i$ 表示第 $i$ 个三元组的得分。模型训练的目的在于使真实三元组的得分尽可能高而负例三元组的得分尽可能低。我们为每一个三元组通过替换其头和尾实体采样一定数量的负例三元组，同时约束负采样得到的负例三元组不能出现在训练集、验证集和测试集中，负例三元组的数量由超参数控制，在模型中使用 Xavier<sup>[66]</sup>初始化实体和关系的向量表示，使用标签平滑<sup>[71]</sup>和 dropout<sup>[67]</sup>技术减少模型过拟合并提高泛化性能，使用批量归一化<sup>[72]</sup>技术用于稳定、监测和加快模型收敛速度，使用 Adam 优化器<sup>[68]</sup>用于优化损失函数。

**算法 4.1** 实体和关系在编码器中的表示学习过程

**输入:** 知识图谱  $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$ ; 编码器的层数  $L$ ; 注意力头的数量  $M_1, \dots, M_L$ ;

嵌入维度  $d^0, \dots, d^L$

**输出:** 更新后的实体表示, 原始关系表示和逆关系表示

- 1: 扩展关系集合  $\mathcal{R}' \leftarrow \mathcal{R} \cup \mathcal{R}_{inv}$ ,  $\mathcal{R}_{inv} = \{r_k^{-1} \mid r_k \in \mathcal{R}\}$
- 2: 扩展三元组集合  $\mathcal{T}' \leftarrow \mathcal{T} \cup \mathcal{T}_{inv}$ ,  $\mathcal{T}_{inv} = \{(e_j, r_k^{-1}, e_i) \mid (e_i, r_k, e_j) \in \mathcal{T}\}$
- 3: 初始化实体和关系的向量表示  $\mathbf{h}_i^0 \in \mathbb{R}^{d^0}, \forall e_i \in \mathcal{E}$ ;  $\mathbf{g}_k^0 \in \mathbb{R}^{d^0}, \forall r_k \in \mathcal{R}'$
- 4: **for**  $l \in L$  **do:**
- 5:     **for**  $m \in M$  **do:**
- 6:         **for**  $t_{ij}^k \in \mathcal{T}'$  **do:**
- 7:             以公式(4.1)学习三元组的表示;
- 8:             以公式(4.2)计算三元组的重要性;
- 9:         **for**  $e_i \in \mathcal{E}$  **do:**
- 10:             **for**  $(e_i, r_k, e_j) \in \mathcal{T}$  **do:**
- 11:                 以公式(4.3)计算原始三元组的注意力值  $\alpha_{ijk}^l$ ;
- 12:                 以公式(4.5)融合原始三元组的信息  $\mathbf{h}_{im\_0}^l$ ;
- 13:             **for**  $(e_i, r_k, e_j) \in \mathcal{T}_{inv}$  **do:**
- 14:                 以公式(4.3)计算原始三元组的注意力值  $\beta_{ijk}^l$ ;
- 15:                 以公式(4.5)融合原始三元组的信息  $\mathbf{h}_{im\_1}^l$ ;
- 16:                 以公式(4.4)更新实体的向量表示  $\mathbf{h}_{im}^l$ ;
- 17:         **for**  $r_k \in \mathcal{R}'$  **do:**
- 18:             **for**  $(e_i, r_k, e_j) \in \mathcal{T}'$  **do:**
- 19:                 以公式(4.7)计算所有三元组的注意力值  $\gamma_{ijk}^l$ ;
- 20:                 以公式(4.8)更新关系的向量表示  $\mathbf{g}_{km}^l$ ;
- 21:             以公式(4.6)和公式(4.10)连接所有注意力头学习到的实体和关系表示,  
得到  $\mathbf{h}_i^l \in \mathbb{R}^{d^l}, \forall e_i \in \mathcal{E}$ ;  $\mathbf{g}_k^l \in \mathbb{R}^{d^l}, \forall r_k \in \mathcal{R}'$ ;
- 22: 返回实体和关系的最终向量表示  $\mathbf{h}_i^L \in \mathbb{R}^{d^L}, \forall e_i \in \mathcal{E}$ ;  $\mathbf{g}_k^L \in \mathbb{R}^{d^L}, \forall r_k \in \mathcal{R}'$ 。

### 4.3 实验设计及结果分析

本章在标准的链接预测数据集 FB15k-237, WN18RR 和 Kinship 上进行实验,并详细的评估和解释了模型的性能,同时进行消融实验研究组件的作用以及参数敏感性实验研究超参数对模型性能的影响。在此之前先介绍数据集的详细信息、实验设置、对比方法以及评价指标。

#### 4.3.1 实验数据

本章使用 FB15k-237<sup>[25]</sup>, WN18RR<sup>[25]</sup>和 Kinship<sup>[73]</sup>三个标准的链接预测数据集进行实验。这三个数据集都包含了一定数量的实体和关系,它们的基本信息如表 4.1 所示。其中 FB15k-237 包含 14,541 个实体和 237 个关系,由 FB15k<sup>[11]</sup>数据集去除逆反关系构造而来,WN18RR 包含 40,943 个实体和 11 个关系,由 WN18<sup>[11]</sup>去除逆反关系构造而来。逆反关系在一定程度上可以看作是无向边,即不管是从头实体指向尾实体还是从尾实体只向头实体,它们所在的三元组都表示相同的语义。Kinship 包含 104 个实体和 25 个关系,描述了 Alyawarra 部落的亲属关系。

表 4.1 数据集信息

数据集	实体数量	关系数量	边的数量			
			训练集	验证集	测试集	总和
FB15k-237	14,541	237	272,115	17,535	20,466	310,116
WN18RR	40,943	11	86,835	3,034	3,134	93,003
Kinship	104	25	8,544	1,068	1,074	10,686

#### 4.3.2 实验设置

D-AEN 模型包含多个超参数,其中学习率从 {0.001,0.0005,0.0001} 中选择,批量大小从 {64,128,256} 中选择,标签平滑系数从 {0.1,0.2,0.3} 中选择,实体和关系的初始嵌入维度从 {100,200,300,400} 中选择,经过图卷积网络后,实体和关系的嵌入维度统一设置为 200,注意力头的数量从 {1,2,3,4} 中选择,解码器中的卷积核大小从 {3,5,7} 中选择,过滤器数量从 {100,200,300} 中选择,模型中所用的所有 dropout 参数从 0.0 至 0.7 中选择,负采样数量从 {20,30,40} 中选择。经过对超参数进行网格搜索,得到如表 4.2 所示的最佳参数配置。

#### 4 基于双注意力的图卷积知识表示学习方法

表 4.2 D-AEN 模型在各数据集上的最佳参数配置

数据集	FB15k-237	WN18RR	Kinship
Learning rate	0.0001	0.0001	0.0001
Epochs	500	800	500
Batch size	128	128	128
Label smooth	0.1	0.1	0.1
Initial embedding size	300	200	300
GCN embedding size	200	200	200
LeakeyReLU	0.2	0.2	0.2
GCN dropout	0.6	0.3	0.4
Attention heads	3	1	3
Embedding dropout	0.0	0.0	0.0
Hidden dropout	0.0	0.0	0.0
Feature dropout	0.0	0.1	0.0
Kernel size	5	5	5
Number of filters	300	300	300
Negative samples	40	40	40

#### 4.3.3 对比方法

本章选择了一些基于三元组的模型以及基于图神经网络的模型与 D-AEN 进行对比，下面对对比模型进行简要说明。

(1) TransE<sup>[11]</sup>: TransE 是最经典的知识表示学习模型，同时也是翻译模型的开篇之作。该方法将关系当作头实体至尾实体的“翻译”，定义了基于距离的得分函数。

(2) RotatE<sup>[17]</sup>: RotatE 是 TransE 的一个先进扩展，它将关系当作头实体至尾实体的“旋转”，从而可以建模不同关系模式。

(3) ModE<sup>[18]</sup>: ModE 是一个创新性的翻译模型，它在极坐标下建模知识图谱中的实体和关系。

(4) DistMult<sup>[19]</sup>: DistMult 是一个代表性的张量分解模型，通过定义双线性得分函数计算三元组的得分。

(5) ComplEx<sup>[20]</sup>: ComplEx 在 DistMult 的基础上引入复数嵌入空间使得模

型可以建模反对称关系。

(6) TuckER<sup>[23]</sup>: TuckER 是最近的一个张量分解模型, 通过 Tucker 分解建模三元组的二进制表示。

(7) ConvE<sup>[25]</sup>: ConvE 是第一个使用卷积神经网络去建模知识图谱的模型。

(8) ConvKB<sup>[26]</sup>: ConvKB 是一个利用卷积神经网络建模三元组的全局结构的模型, 同时保留了三元组内部的翻译性质。

(9) ConvR<sup>[27]</sup>: ConvR 是 ConvE 的一个先进扩展, 它将关系表示作为卷积神经网络的卷积核, 极大减少了模型参数。

(10) InteractE<sup>[28]</sup>: InteractE 也是 ConvE 的一个扩展, 它通过设计循环卷积操作捕获了实体和关系间的更多语义交互。

(11) R-GCN<sup>[37]</sup>: R-GCN 是第一个用图卷积网络建模知识图谱的模型, 提出了知识表示学习里的编码器-解码器结构。

(12) WGCN<sup>[38]</sup>: WGCN 在 R-GCN 的基础上提出加权图卷积网络, 定义线性变换矩阵融合邻居实体的信息。

(13) CompGCN<sup>[48]</sup>: CompGCN 组合邻居实体和关系的表示, 将组合后的表示融合起来更新中心实体的表示。

(14) KBGAT<sup>[39]</sup>: KBGAT 是一个强大的基于图网络的知识表示学习模型, 它在三元组的层次上运用注意力机制更新中心实体的表示, 同时通过将二阶邻居转化为一阶邻居扩展了知识图谱包含的语义信息。

#### 4.3.4 评价指标

与大多数对比模型一样, 本章使用三元组排名的评估模型的性能。具体来讲, 需要对测试集中的三元组进行头实体预测和尾实体预测。以预测一个三元组的尾实体为例, 首先用知识图谱的其他实体代替该三元组的尾实体构造负例集, 然后, 通过得分函数计算出原始三元组 and 对应负例三元组的得分, 根据它们的得分进行降序排列得到原始三元组的排名, 最后利用平均排名 (Mean Rank, MR) 指标、平均倒数排名指标 (Mean Reciprocal Rank, MRR) 和命中率指标 Hits@N (N=1,3,10), 根据所得三元组的排名计算这三个评价指标的具体值。我们希望得到较低的 MR 值, 较高的 MRR 值和 Hits@N 值。为了得到更为合理的结果, 本章像 TransE 一样在预测时使用 ‘filter’ 设置, 即负例三元组中存在于训练集、验证集和测试集中的三元组不参与排名。与预测尾实体一样, 预测头实体使用同

样的方法，最终实验结果取预测头实体和预测尾实体的平均值。

#### 4.3.5 整体结果

表 4.3、表 4.4 和表 4.5 分别展示了 D-AEN 和对比方法在 FB15k-237, WN18RR 和 Kinship 上的整体实验结果。其中，‘-’代表缺失值，每个评价指标上最好的结果用粗体标识，次好的结果用下划线标识。鉴于使用相同的实验数据集，大部分对比方法的实验结果从对应原始论文中收集而来，少数方法的实验结果从当前具有代表性的工作中收集而来。其中，在 FB15k-237 和 WN18RR 数据集上，TransE、DistMult 和 ComplEx 的实验结果来自 RotatE；在 Kinship 数据集上，TransE、DistMult、ComplEx、ConvKB 和 R-GCN 的实验结果来自 KBGAT；其他结果均来自对应的工作。由于有些对比方法没有在 Kinship 数据集上进行实验，因此在 Kinship 数据集上的实验结果中忽略了这些方法。

表 4.3 在 FB15k-237 数据集上的实验结果

数据集	FB15k-237				
评价指标	Hits@1	Hits@3	Hits@10	MR	MRR
TransE	-	-	0.465	357	0.294
RotatE	0.241	0.375	0.533	177	0.338
ModE	0.244	0.380	0.534	-	0.341
DistMult	0.155	0.263	0.419	254	0.241
ComplEx	0.158	0.275	0.428	339	0.247
TuckER	0.266	0.394	0.544	-	0.358
ConvE	0.239	0.350	0.491	246	0.316
ConvKB	-	-	0.517	257	0.396
ConvR	0.261	0.385	0.528	-	0.350
InteractE	0.263	-	0.535	<u>172</u>	0.354
R-GCN	0.151	0.264	0.417	-	0.249
WGCN	0.26	0.39	0.54	-	0.35
CompGCN	0.264	0.390	0.535	197	0.355
KBGAT	<b>0.460</b>	<b>0.540</b>	<b>0.626</b>	210	<b>0.518</b>
D-AEN	<u>0.337</u>	<u>0.471</u>	<u>0.611</u>	<b>164</b>	<u>0.429</u>

从实验结果中可以看出：

(1) 在 FB15k-237 数据集上，除了 KBGAT 以外，D-AEN 在所有指标上均取得了最好的结果。在 WN18RR 数据集上，D-AEN 在 3 个指标上取得最佳性能。在 Kinship 数据集上，D-AEN 在所有指标上均表现最佳。KBGAT 在 FB15k-237 数据集上取得最优表现的原因有两点，第一，KBGAT 将知识图谱中节点的二阶邻居转化为一阶邻居从而扩展了训练集，使得学到的实体表示包含了更多的邻域信息。第二，FB15k-237 数据集相对于 WN18RR 和 Kinship 显得更为复杂，包含更多的实体、关系和三元组，从而该数据集中的实体就包含更多的二阶邻居，因此，KBGAT 在 FB15k-237 数据集上表现出非常卓越的性能。尽管在 FB15k-237 数据集上 KBGAT 的性能优于 D-AEN，但是 D-AEN 在 WN18RR 和 Kinship 数据集上分别有 3 个指标和 5 个指标优于 KBGAT。因此总体的实验结果表明 D-AEN 是非常有效的。

表 4.4 在 WN18RR 数据集上的实验结果

数据集	WN18RR				
	Hits@1	Hits@3	Hits@10	MR	MRR
TransE	-	-	0.501	3384	0.226
RotatE	0.428	0.492	0.571	3340	0.476
ModE	0.427	0.486	0.564	-	0.472
DistMult	0.39	0.44	0.49	5110	0.43
ComplEx	0.41	0.46	0.51	5261	0.44
TuckER	<b>0.443</b>	0.482	0.526	-	0.470
ConvE	0.39	0.43	0.48	5277	0.46
ConvKB	-	-	0.525	2544	0.248
ConvR	0.433	0.489	0.537	-	0.475
InteractE	0.430	-	0.528	5202	0.463
R-GCN	-	-	-	-	-
WGCN	0.43	0.48	0.54	-	0.47
CompGCN	<b>0.443</b>	0.494	0.546	3533	<u>0.479</u>
KBGAT	0.361	0.483	<b>0.581</b>	<b>1940</b>	0.440
D-AEN	<b>0.443</b>	<b>0.500</b>	0.561	<u>2248</u>	<b>0.484</b>



(2) 与仅仅使用解码器模型的 ConvE 模型相比, D-AEN 在三个数据集上的所有指标上均表现出更优的性能,例如,在 FB15k-237 数据集的 MRR 指标上, D-AEN 相较于 ConvE 取得了 11.3% 的性能提升。这个现象强力地证明了编码器模型的先进性,同样说明由 D-AEN 融合的邻居信息具有很大的价值。同时, D-AEN 在大部分指标上远远优于同样使用 ConvE 作为解码器的 CompGCN 模型,这说明 D-AEN 的编码器模型具有很大优势,能捕获到更丰富的知识图谱结构信息,并进一步证明了 D-AEN 的有效性。

表 4.5 在 Kinship 数据集上的实验结果

数据集	Kinship				
评价指标	Hits@1	Hits@3	Hits@10	MR	MRR
TransE	0.009	0.643	0.841	6.8	0.309
DistMult	0.367	0.581	0.867	5.26	0.516
CompLex	0.733	0.899	0.971	2.48	0.823
ConvE	0.73	0.91	0.98	2	0.83
ConvKB	0.436	0.755	0.953	3.3	0.614
R-GCN	0.03	0.088	0.239	25.92	0.109
KBGAT	<u>0.859</u>	<u>0.941</u>	<u>0.980</u>	<u>1.94</u>	<u>0.904</u>
D-AEN	<b>0.968</b>	<b>0.984</b>	<b>0.990</b>	<b>1.52</b>	<b>0.977</b>

#### 4.3.6 建模不同类型关系的 Hit@10 结果

本节在 FB15k-237 和 WN18RR 数据集上对不同类型的关系进行建模,以 Hits@10 作为评价指标。直观上讲,知识图谱中的关系可以被分为四类:一对一(1-1)关系、多对一(N-1)关系、一对多(1-N)关系和多对多(N-N)关系。一对一关系指的是唯一的一个头实体和唯一的一个尾实体通过该关系相连,多对一关系指的是存在多个头实体和唯一的一个尾实体通过该关系相连,一对多关系指的是唯一的一个头实体和多个尾实体通过该关系相连,多对多关系指的是存在多个头实体和多个尾实体通过该关系相连。统计表明,FB15k-237 数据集包含的一对一关系、多对一关系、一对多关系和多对多关系的数量百分比为 7.2%, 34.2%, 11% 和 47.6%。WN18RR 数据集中包含的一对一关系、多对一关系、一对多关系和多对多关系的数量百分比为 18.2%, 27.3%, 36.3%, 18.2%。

表 4.6 和表 4.7 分别展示了 D-AEN 和几个对比方法在 FB15k-237 和 WN18RR 数据集上对不同关系类型进行建模的 Hits@10 结果, 其中 TransE, DistMult, ComplEx 和 ConvE 的结果收集自 Li 等人<sup>[74]</sup>的工作, Tail 指的是预测尾实体, Head 指的是预测头实体。

表 4.6 在 FB15K-237 数据集上对不同类型的关系进行建模的实验结果

模型	1-1		N-1		1-N		N-N	
	Tail	Head	Tail	Head	Tail	Head	Tail	Head
TransE	0.521	0.537	0.833	0.070	0.052	0.573	0.508	0.428
DistMult	0.182	0.193	0.793	0.031	0.039	0.514	0.485	0.403
ComplEx	0.411	0.411	0.818	0.050	0.050	0.551	0.533	0.456
ConvE	0.258	0.250	0.865	0.147	0.132	0.603	0.581	0.504
D-AEN	<b>0.563</b>	<b>0.589</b>	<b>0.872</b>	<b>0.564</b>	<b>0.211</b>	<b>0.625</b>	<b>0.640</b>	<b>0.598</b>

表 4.7 在 WN18RR 数据集上对不同类型的关系进行建模的实验结果

模型	1-1		N-1		1-N		N-N	
	Tail	Head	Tail	Head	Tail	Head	Tail	Head
TransE	<b>0.976</b>	<b>0.976</b>	0.190	0.022	0.061	0.276	0.941	0.942
DistMult	0.929	0.952	0.334	0.047	0.051	0.269	0.944	0.946
ComplEx	<b>0.976</b>	<b>0.976</b>	0.309	0.053	0.086	0.288	0.950	0.951
ConvE	<b>0.976</b>	<b>0.976</b>	0.303	0.107	0.190	0.451	0.948	0.948
D-AEN	<b>0.976</b>	<b>0.976</b>	<b>0.387</b>	<b>0.242</b>	<b>0.229</b>	<b>0.505</b>	<b>0.952</b>	<b>0.952</b>

从结果来看, 在 FB15k-237 数据集上, D-AEN 远远优于四个对比模型, 尤其在建模具有多个头实体的多对一关系时, 预测头实体的实验结果在 Hit@10 指标上取得了 0.564 的高分, 其余模型的最高得分才仅为 0.147。另外, 在建模具有多个尾实体的一对多关系时, 对比方法与 D-AEN 的性能均不理想, 但 D-AEN 相较对比模型仍然取得了不错的提升。这些结果说明 D-AEN 在建模具有多关系的知识图谱上有很大优势。在 WN18RR 数据集上, D-AEN 仍然在建模四种类型关系时取得了最好的结果。值得注意的是 TransE, ComplEx 和 ConvE 与 D-AEN 在建模一对一关系时同时取得了最好的结果 0.976, 并且在建模多对多关系时, D-AEN 虽然结果最好, 但相对于对比方法提升不大。造成这些现象的原因是因

为 WN18RR 数据集仅包含 2 个一对一关系和 2 个多对多关系，使得预测出正确的结果相对简单。

#### 4.3.7 收敛性分析

本节在 FB15k-237 和 WN18RR 数据集上研究 D-AEN 随着训练的迭代基于 MRR 指标的收敛情况。如图 4.3 所示，图中绿色线和蓝色线分别对应训练阶段在验证集上预测头实体和尾实体的 MRR 指标收敛情况，红色线对应它们的平均值。在 FB15k-237 数据集上，这三个值在刚开始的 50 轮迅速上升，随后大约在 350 轮以后达到一个稳定值。同样，在 WN18RR 数据集上这三个值在刚开始的 100 轮迅速上升，随后随着训练轮次的增加而缓慢提升。这些观察说明 D-AEN 收敛迅速，不容易过拟合，并且在实际运用中是可靠的。

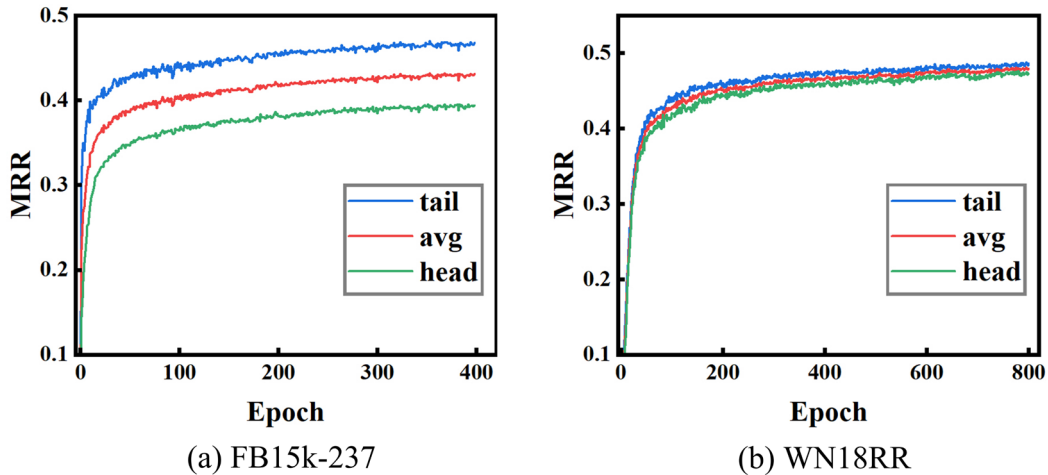


图 4.3 在 FB15K-237 和 WN18RR 数据集上的收敛性分析

#### 4.3.8 参数敏感性分析

本节主要研究模型的主要参数对模型性能的影响，包括实体和关系的初始嵌入维度大小、注意力头的数量和负采样的数量。下面将对这些参数对模型性能的影响进行详细说明。

(1) 如图 4.4 所示，本章在 WN18RR 和 Kinship 数据集上探究了不同维度的实体和关系初始嵌入对 MRR 指标的影响，其中初始嵌入维度为 {100,200,300,400}。从结果来看，模型在 WN18RR 和 Kinship 数据集上的实体和关系初始嵌入维度分别为 200 和 300 时取得最佳表现，其他较大或较小的取值都导致模型性能降低，出现这个现象的原因是初始化嵌入维度较小的实体和

关系表示不能保存知识图谱中足够的结构信息,另一方面,初始化嵌入维度较大的实体和关系表示可能会使学习到的实体和关系表示包含噪声,从而降低模型的泛化性能。

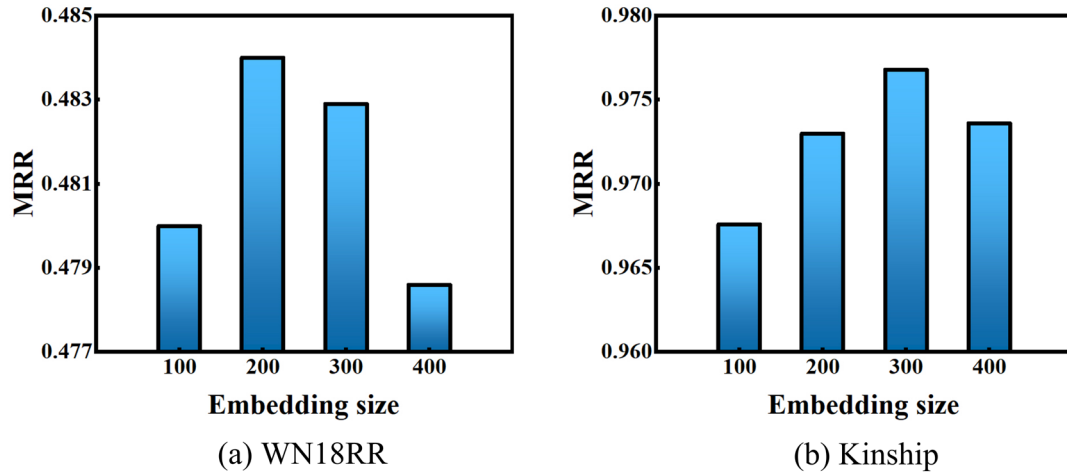


图 4.4 在 WN18RR 和 Kinship 数据集上探究初始嵌入维度对模型性能的影响结果

(2) 如图 4.5 (a) 所示,本章在 Kinship 数据集上探究不同的注意力头数量对 MRR 指标的影响,其中注意力头的数量取值为  $\{1,2,3,4\}$ 。从结果来看,模型在注意力头数量为 3 时取得最好结果,同时在注意力头分别取 1,2 和 3 时,模型表现逐步提升,这个现象说明适当的注意力头可以使模型融合更多有效的邻域信息用于更新实体和关系的表示,但是一直增加注意力头的数量可能会使模型融合的邻域信息包含很多无用的信息,从而影响模型性能。

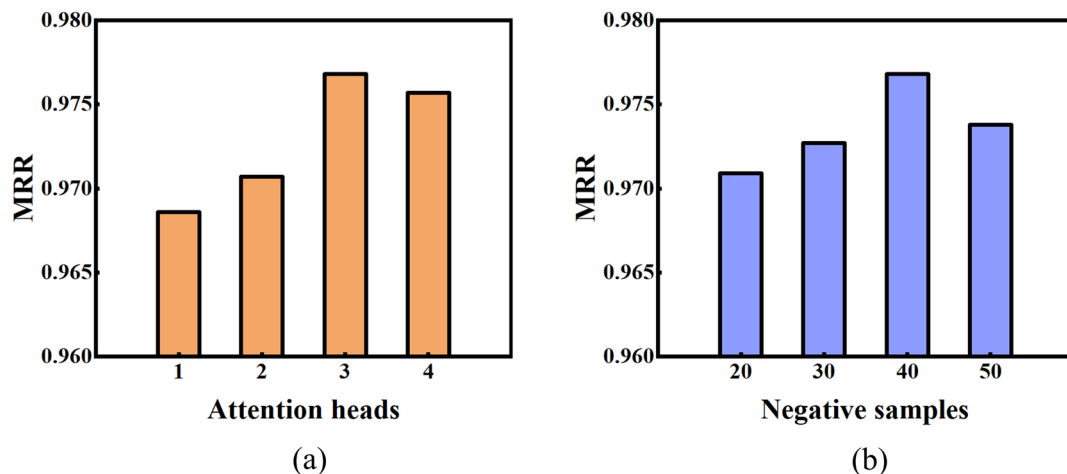


图 4.5 在 Kinship 数据集上探究注意力头数量和负采样数量对模型性能的影响

(3) 如图 4.5 (b) 所示, 本章在 Kinship 数据集上探究不同的负采样数量对 MRR 指标的影响, 其中负采样的数量取值为 {20,30,40,50}。从结果可以看出模型在负采样数量为 40 时表现出最好的性能, 而在这基础上增加或减少负采样数量都会降低模型的性能, 这很大程度上是因为较小的负采样数量不足以有效地训练模型, 而较大的负采样数量可能包含对模型训练没有帮助的负例三元组。

#### 4.3.9 消融实验

本节引入 D-AEN 的三个变体模型在 WN18RR 和 Kinship 数据集上进行消融实验用于验证模型组件的有效性, 评价指标包括 MRR 和 Hits@N (N=1,3,10)。D-AEN 的三个变体模型包括: (1) RemoveRA: 从 D-AEN 中移除了关系注意力机制, 使邻域信息以同等的重要性融入关系的新表示中。(2) RemoveR: 从 RemoveRA 中移除邻域对关系表示学习的影响, 使关系表示进行自我更新。(3) RemoveBR: 从 RemoveR 中移除双向注意力机制的双向信息, 将中心实体的出边邻居和入边邻居当成一个整体评估它们的重要性, 然后结合邻域信息融入中心实体的新表示中。表 4.8 展示了 D-AEN 及三个变体模型的实验结果。

表 4.8 消融实验结果

数据集	WN18RR				Kinship			
	Hits			MRR	Hits			MRR
	@1	@3	@10		@1	@3	@10	
RemoveBR	0.424	<u>0.497</u>	0.554	0.472	0.922	0.967	0.985	0.946
RemoveR	0.434	0.493	<u>0.555</u>	0.476	0.932	0.969	<u>0.988</u>	0.952
RemoveRA	<u>0.439</u>	<u>0.497</u>	0.553	<u>0.479</u>	<u>0.960</u>	<u>0.978</u>	<u>0.988</u>	<u>0.970</u>
D-AEN	<b>0.443</b>	<b>0.500</b>	<b>0.561</b>	<b>0.484</b>	<b>0.968</b>	<b>0.984</b>	<b>0.990</b>	<b>0.977</b>

从实验结果可以看出: (1) RemoveR 在 7 个指标上优于 RemoveBR, 表明用于实体表示学习的双向注意力机制对模型性能有很大的影响, 说明 D-AEN 在实体的表示学习中引入关系的方向信息评估邻域的重要性起了相当大的作用。

(2) RemoveRA 相较 RemoveR 在 6 个指标上取得了较大的提升, 说明在关系的表示学习中融入邻域信息能学习到更有效的关系表示, 使实体和关系进行了更充分的语义交互。(3) 相对 RemoveRA, D-AEN 在所有指标上均表现出了更好的性能, 说明在关系的表示学习中, 引入注意力机制评估邻域的重要性而后以不

同的权重融合邻域信息更新关系的向量表示,有益于关系的表示学习。从以上分析可以看出 D-AEN 中的每个组件都设计得合理。

#### 4.3.10 案例分析

本节在 FB15k-237 数据集上进行案例分析对 D-AEN 的预测能力进行直观地验证。如表 4.9 所示,给出几个三元组的头实体和关系,预测得分最高的 4 个尾实体,粗体表示预测集中的正确尾实体,下划线表示不在测试集中但是在训练集和验证集中的正确尾实体。实验结果证明 D-AEN 可以成功的预测测试集中的尾实体,尽管测试集中的尾实体不总是取得最好的排名。

表 4.9 在 FB15k-237 数据集上进行案例分析的实验结果

头实体和关系	预测的尾实体
(James Madison, organization founder)	(1) <b>Democratic Party</b> ; (2) United States Military Academy; (3) Democratic-Republican Party; (4) Episcopal Church.
(National Football League, team)	(1) <b>Los Angeles Chargers</b> ; (2) <u>Carolina Panthers</u> ; (3) <u>New York Jets</u> ; (4) <u>Detroit Lions</u> .
(The X-Files, actor)	(1) <u>William B. Davis</u> ; (2) <b>Gillian Anderson</b> ; (3) <u>Cary Elwes</u> ; (4) <u>Robert Patrick</u> .
(marriage, location of ceremony)	(1) <u>Paris</u> ; (2) <u>Sydney</u> ; (3) <b>Las Vegas</b> ; (4) <u>London</u> .

#### 4.4 本章小结

本章提出 D-AEN 模型用于解决大多数基于图神经网络的知识表示学习方法

忽略的邻域同时影响实体和关系的表示学习这一问题，通过在实体的知识表示学习中引入基于关系方向的双边注意力机制评估邻域的重要性，同时引入关系注意力机制评估邻域重要性以将邻域信息以恰当的重要性融入关系的向量表示中。D-AEN 充分的考虑了知识图谱的结构信息，使学到的实体和关系表示尽可能地保存知识图谱的本质结构信息，极大促进了实体与关系之间和实体与实体之间的语义交互。在经典的链接预测数据集上表现出了先进的性能。

## 5 总结与展望

### 5.1 本文工作总结

知识表示学习在低维稠密的向量空间中对知识图谱进行建模，学习知识图谱中实体和关系的向量表示以保存知识图谱的结构信息和语义信息。主流的知识表示学习模型基于单个三元组对知识图谱进行建模，然而，这些模型只关注了三元组内部的交互，忽略了实体和关系存在的邻域结构，因此在整体上忽略了知识图谱的结构信息。本文从知识图谱的结构信息出发，探究了在建模知识图谱中融入实体和关系所在邻域的信息的知识表示学习方法。本文的研究成果总结如下：

(1) 提出基于聚合的图卷积知识表示学习方法，融合实体的部分邻域信息。实体与不同的邻居实体通过不同的关系相连体现了实体所在的邻域结构，表示实体同时处于不同的语义环境。针对大部分基于三元组的模型在知识表示学习过程中无法有效融入实体所在邻域的信息这一问题，提出基于聚合的图卷积知识表示学习方法，旨在为实体学习语义更丰富的向量表示。该方法设计基于聚合的图卷积网络编码器对中心实体的邻居实体及相连关系的向量表示进行聚合，而后将聚合后的表示用于更新中心实体的表示。针对基因本体进行建模用于分析基因功能相似度，实验结果表明该方法能有效捕获实体的部分邻域信息，从而学习到有效的实体表示，性能明显优于传统基因功能相似度分析方法。

(2) 提出基于双注意力的图卷积知识表示学习方法，融合实体和关系的全部邻域信息。实体和关系所处的不同三元组表示了它们所处的不同邻域，说明实体和关系在不同邻域中具有的不同语义，反过来，实体和关系所在的不同邻域对它们的表示学习也有一定的影响。针对邻域同时影响实体和关系的表示学习这一问题，提出基于双注意力的图卷积知识表示学习方法，力求在实体和关系的表示学习过程中充分融入邻域信息。首先，该方法在图卷积网络编码器中设计两个注意力机制同时评估邻域的重要性，其次，根据不同重要性融合邻域信息，最后，根据融合的邻域信息更新实体和关系的向量表示。在标准的链接预测数据集上对模型性能进行评估，实验结果表明该方法能有效捕获实体和关系的全部邻域信息，从而促进实体与实体之间和实体与关系之间的语义交互，性能明显优于已



有的基于三元组的方法和基于图神经网络的方法。

## 5.2 研究展望

针对当前大部分知识表示学习方法无法有效捕获知识图谱结构信息的问题,本文探究了在知识表示学习过程中通过图卷积网络融合部分邻域信息和整个邻域信息的知识表示学习方法,虽然取得了一些阶段性的成果,但是本文工作仍有可以改进的空间。

(1) 在融合部分邻域信息的知识表示学习中,基于聚合的图卷积知识表示学习方法总以相同的重要性融合中心实体的邻居实体及相连关系聚合后的表示,并且关系的表示学习是通过自动更新来实现。因此,尝试在融入部分邻域信息时引入注意力机制动态地评估它们的重要性能有效建模实体所在邻域的差异,同时设计合理的关系表示学习方式也具有重要意义。

(2) 在融合整个邻域信息的知识表示学习中,基于双注意力的图卷积知识表示学习方法只针对当前知识图谱中实体和关系的一阶邻域结构进行建模。然而,知识图谱以图结构的形式存储实体和关系,包含实体和关系的高阶邻域结构以及实体间的多步路径信息。因此,下一步工作可以尝试利用这些信息改进现有的方法。

(3) 近年来,随着时代发展,知识图谱不断更新和扩大,其结构变得越来越复杂,导致无法充分捕获其结构信息。因此,从多视角对知识图谱进行表示学习能尽可能充分地挖掘知识图谱的结构信息,从而最大化保存其包含的语义信息,更好地服务下游知识驱动任务。

## 参考文献

- [1] Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase[J]. *Communications of the ACM*, 2014, 57(10): 78~85
- [2] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]. In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008. 1247~1250
- [3] Bizer C, Lehmann J, Kobilarov G, et al. Dbpedia-a crystallization point for the web of data[J]. *Journal of web semantics*, 2009, 7(3): 154~65
- [4] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]. In: *Proceedings of the 16th international conference on World Wide Web*, 2007. 697~706
- [5] Fellbaum C. WordNet[M]. *Theory and applications of ontology: computer applications*. Springer. 2010: 231~43
- [6] Zhang F, Yuan N J, Lian D, et al. Collaborative knowledge base embedding for recommender systems[C]. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016. 353~362
- [7] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs[C]. In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013. 1533~1544
- [8] Bordes A, Chopra S, Weston J. Question Answering with Subgraph Embeddings[C]. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014. 615~620
- [9] Miller E. An Introduction to the Resource Description Framework[J]. *Bulletin of the American Society for Information Science*, 1998, 25(1): 15~19
- [10] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]. In: *Proceedings of the 26th International Conference on Neural Information Processing System*, 2013. 3111~3119
- [11] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 2013. 2787~2795
- [12] Nickel M, Tresp V, Kriegel H-P. A three-way model for collective learning on multi-relational data[C]. In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011. 809~816
- [13] Weston J, Bordes A, Yakhnenko O, et al. Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction[C]. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013. 1366~1371
- [14] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C].

- In: Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014. 1112~1119
- [15] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015. 2181~2187
- [16] Ji G, He S, Xu L, et al. Knowledge graph embedding via dynamic mapping matrix[C]. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015. 687~696
- [17] Sun Z, Deng Z, Nie J, et al. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space[C]. In: Proceedings of the 7th International Conference on Learning Representations, 2018.
- [18] Zhang Z, Cai J, Zhang Y, et al. Learning hierarchy-aware knowledge graph embeddings for link prediction[C]. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020. 3065~3072
- [19] Yang B, Yih W, He X, et al. Embedding Entities and Relations for Learning and Inference in Knowledge Bases[C]. In: Proceedings of the 3rd International Conference on Learning Representations, 2015.
- [20] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[C]. In: Proceedings of the 33rd International Conference on Machine Learning, 2016. 2071~2080
- [21] Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs[C]. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016. 1955~1961
- [22] Liu H, Wu Y, Yang Y. Analogical inference for multi-relational embeddings[C]. In: Proceedings of the 34th International Conference on Machine Learning, 2017. 2168~2178
- [23] Balažević I, Allen C, Hospedales T. TuckER: Tensor Factorization for Knowledge Graph Completion[C]. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019. 5185~5194
- [24] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion completion[C]. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013. 926~934
- [25] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2d knowledge graph embeddings[C]. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018. 1811~1818
- [26] Nguyen T D, Nguyen D Q, Phung D Q. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network[C]. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018. 327~333
- [27] Jiang X, Wang Q, Wang B. Adaptive convolution for multi-relational learning[C]. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 978~987
- [28] Vashishth S, Sanyal S, Nitin V, et al. InteractE: Improving convolution-based knowledge graph

- embeddings by increasing feature interactions[C]. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020. 8986~8990
- [29] Vu T, Nguyen T D, Nguyen D Q, et al. A capsule network-based embedding model for knowledge graph completion and search personalization[C]. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 2180~2189
- [30] Cai L, Wang W Y. KBGAN: Adversarial Learning for Knowledge Graph Embeddings[C]. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018. 1470~1480
- [31] Guo S, Wang Q, Wang B, et al. SSE: Semantically smooth embedding for knowledge graphs[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 29(4): 884~97
- [32] Xie R, Liu Z, Sun M. Representation learning of knowledge graphs with hierarchical types[C]. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016. 2965~2971
- [33] Xie R, Liu Z, Jia J, et al. Representation learning of knowledge graphs with entity descriptions[C]. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016. 2659~2665
- [34] Wang Z, Li J. Text-enhanced representation learning for knowledge graph[C]. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016. 1293~1299
- [35] Xie R, Liu Z, Luan H, et al. Image-embodied knowledge representation learning[C]. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017. 3140~3146
- [36] Feng J, Huang M, Yang Y, et al. GAKE: Graph aware knowledge Embedding[C]. In: Proceedings of the 26th International Conference on Computational Linguistics, 2016. 641~651
- [37] Schlichtkrull M, Kipf T N, Bloem P, et al. Modeling relational data with graph convolutional networks[C]. In: Proceedings of the 15th European Semantic Web Conference, 2018. 593~607
- [38] Shang C, Tang Y, Huang J, et al. End-to-end structure-aware convolutional networks for knowledge base completion[C]. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019. 3060~3067
- [39] Nathani D, Chauhan J, Sharma C, et al. Learning Attention-based Embeddings for Relation Prediction in Knowledge Graphs[C]. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. 4710~4723
- [40] Zhang Z, Zhuang F, Zhu H, et al. Relational graph neural network with hierarchical attention for knowledge graph completion[C]. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020. 9612~9619
- [41] Bruna J, Zaremba W, Szlam A, et al. Spectral networks and deep locally connected networks on graphs[C]. In: Proceedings of the 2nd International Conference on Learning Representations, 2014.

- [42] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering[C]. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016. 3844~3852
- [43] Xu B, Shen H, Cao Q, et al. Graph Wavelet Neural Network[C]. In: Proceedings of the 7th International Conference on Learning Representations, 2019.
- [44] Niepert M, Ahmed M, Kutzkov K. Learning convolutional neural networks for graphs[C]. In: Proceedings of the International Conference on Machine Learning, 2016. 2014~2023
- [45] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs[C]. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. 1025~1035
- [46] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[C]. In: Proceedings of the 5th International Conference on Learning Representations, 2017.
- [47] Veličković P, Cucurull G, Casanova A, et al. Graph Attention Networks[C]. In: Proceedings of the 6th International Conference on Learning Representations, 2018.
- [48] Vashishth S, Sanyal S, Nitin V, et al. Composition-based Multi-Relational Graph Convolutional Networks[C]. In: Proceedings of the 7th International Conference on Learning Representations, 2019.
- [49] Yu D, Yang Y, Zhang R, et al. Knowledge embedding based graph convolutional network[C]. In: Proceedings of the 2021 Web Conference, 2021. 1619~1628
- [50] Zhao Y, Zhou H, Xie R, et al. Incorporating Global Information in Local Attention for Knowledge Representation Learning[C]. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 13th International Joint Conference on Natural Language Processing, 2021. 1341~1351
- [51] Consortium G O. The Gene Ontology (GO) database and informatics resource[J]. Nucleic acids research, 2004, 32: 258~261
- [52] Camon E, Magrane M, Barrell D, et al. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology[J]. Nucleic acids research, 2004, 32: 262~266
- [53] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language[J]. Journal of artificial intelligence research, 1999, 11: 95~130
- [54] Jiang J, Conrath D. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy[C]. In: Proceedings of the 10th Research on Computational Linguistics International Conference, 1997. 19~33
- [55] Lin D. An Information-Theoretic Definition of Similarity[C]. In: Proceedings of the 15th International Conference on Machine Learning, 1998. 296~304
- [56] Pesaranghader A, Matwin S, Sokolova M, et al. simDEF: definition-based semantic similarity measure of gene ontology terms for functional similarity analysis of genes[J]. Bioinformatics, 2016, 32(9): 1380~1387
- [57] Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association[J].

- Bioinformatics, 2007, 23(2): 257~258
- [58] Pesquita C, Faria D, Bastos H, et al. Metrics for GO based protein semantic similarity: a systematic evaluation[J]. BMC bioinformatics, 2008, 9(5): 1~16
- [59] Sánchez D, Batet M. A semantic similarity method based on information content exploiting multiple ontologies[J]. Expert Systems with Applications, 2013, 40(4): 1393~1399
- [60] Teng Z, Guo M, Liu X, et al. Measuring gene functional similarity based on group-wise comparison of GO terms[J]. Bioinformatics, 2013, 29(11): 1424~1432
- [61] Tian Z, Wang C, Guo M, et al. An improved method for functional similarity analysis of genes based on Gene Ontology[J]. BMC systems biology, 2016, 10(4): 465~484
- [62] Tian Z, Fang H, Ye Y, et al. A novel gene functional similarity calculation model by utilizing the specificity of terms and relationships in gene ontology[J]. BMC bioinformatics, 2022, 23(1): 1~14
- [63] Benabderrahmane S, Smail-Tabbone M, Poch O, et al. IntelliGO: a new vector-based semantic similarity measure including annotation origin[J]. BMC bioinformatics, 2010, 11(1): 1~16
- [64] Zhang J, Jia K, Jia J, et al. An improved approach to infer protein-protein interaction based on a hierarchical vector space model[J]. BMC bioinformatics, 2018, 19(1): 1~14
- [65] Jain S, Bader G D. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology[J]. BMC bioinformatics, 2010, 11(1): 1~14
- [66] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, 2010. 249~256
- [67] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The journal of machine learning research, 2014, 15(1): 1929~1958
- [68] Kingma D, Ba J. Adam: A Method for Stochastic Optimization[C]. In: Proceedings of the 3rd International Conference on Learning Representations, 2015.
- [69] Wang J, Du Z, Payattakool R, et al. A new method to measure the semantic similarity of GO terms[J]. Bioinformatics, 2007, 23(10): 1274~1281
- [70] Ye R, Li X, Fang Y, et al. A Vectorized Relational Graph Convolutional Network for Multi-Relational Network Alignment[C]. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019. 4135~4141
- [71] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016. 2818~2826
- [72] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]. In: Proceedings of the 32nd International Conference on Machine Learning, 2015. 448~456
- [73] Lin X V, Socher R, Xiong C. Multi-Hop Knowledge Graph Reasoning with Reward Shaping[C]. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018. 3243~3285

参考文献

---

- [74] Li Q, Wang D, Feng S, et al. Global Graph Attention Embedding Network for Relation Prediction in Knowledge Graphs[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.

## 个人简历、在学期间发表的学术论文与研究成果

### 个人简历

方海川，男，重庆市忠县人，1997年9月25日生。

2015年9月——2019年6月，大连海洋大学，计算机科学与技术专业，获得工学学士学位；

2019年9月——至今，郑州大学，软件工程专业，攻读工学硕士学位。

### 研究成果

- [1] **Fang H**, Wang Y, Tian Z, Ye Y. Learning knowledge graph embedding with a dual-attention embedding network[J]. Expert Systems With Applications, 2022. (中科院 1 区, 二审)
- [2] Tian Z\*, **Fang H\***, Teng Z, Ye Y. GOGCN: Graph Convolutional Network on Gene Ontology for functional similarity analysis of genes[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022. (中科院 2 区, 二审, \*代表同等贡献)
- [3] Tian Z, **Fang H**, Ye Y, Zhu Z. A novel gene functional similarity calculation model by utilizing the specificity of terms and relationships in gene ontology[J]. BMC bioinformatics, 2022, 23(1): 1~14 (中科院 2 区, 已发表)
- [4] Tian Z, **Fang H**, Ye Y, Zhu Z. SWE: a novel method with semantic-weighted edge for measuring gene functional similarity[C]. In: Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine, 2020. 1672~1678. (CCF B 类会议, 已发表)

### 参与项目

- [1] 国家自然科学基金：图传播 IB 方法的模型及传播机制研究（62176239）
- [2] 国家自然科学基金：基于多元数据融合的药物重定位方法研究（61801432）
- [3] 国家自然科学基金：传播 IB 方法的研究（61772475）

### 荣誉奖励

- [1] 2019 年研究生一等奖学金
- [2] 2020 年研究生二等奖学金
- [3] 2021 年研究生一等奖学金
- [4] 2021 年郑州大学三好研究生
- [5] 2022 年郑州大学优秀毕业生



## 致谢

行文至此，仿佛重新经历了一遍备战考研直到硕士即将毕业，心中感受颇多。三年以来，我自觉“换了一副新面孔”，收获良多，值此论文完成之际，心中不自觉浮现感恩二字。

感谢我的导师叶阳东教授。老师每一次对我的教导，对我来说都是一次阅读内心的契机，能让我反省自己、提升自己。老师常说，科研是不断内求、不断修心的过程，在这个过程中，老师给我树立了榜样，教会了我方法，教导我严谨、踏实，这将使我受益终生。

感谢我的副导师田侦老师。田老师就像我的大哥哥一样，无论在生活中还是学习中，他总能给我很多帮助，很多启发。一次次分享，一次次交流，都能让我从内心感到他对我的关心和期望。

感谢我的师兄师姐，胡世哲、张明明、王有为、李辉、孙中川、毛奕桥、郭强，曾鑫和夏春管等对我学习或生活上的帮助，与他们交流，总能让我受益匪浅。感谢与我同届的张麒、史凯远和钟李红同学，与他们一起努力是我的幸运。感谢我的师妹师弟，彭祥余和余跃，与他们一起成长让我感到很有成就感。感谢实验室的李鹏翔和李祥瑞同学，与他们同桌是我快乐的源泉。感谢实验室的娄铮铮老师，吴云鹏老师，闫小强老师，他们对工作和学习的认真态度深深感染了我。

感谢我的室友于芳星和李垒昂，三年朝夕相处，我们结下了浓厚的友情。人生苦短，愿君历尽千帆，归来仍是少年。

感谢我的家人对我无微不至的关怀与支持，无以为报，惟愿身心健康。

感谢杜钰琳的一路相伴。

感谢参与本论文评审的各位专家及学者，你们的意见和建议使我受益匪浅。感谢在我生命中来了又去的人们。

方海川  
2022年3月