李君涛, 欧阳智, 杜逆索. 基于多阶段推理的多跳机器阅读理解[J]. 智能计算机与应用,2024,14(8):1-10. DOI:10.20169/j.issn.2095-2163.240801

# 基于多阶段推理的多跳机器阅读理解

李君涛[1], 欧阳智[2], 杜逆索[2]

(1 贵州大学 数学与统计学院,贵阳 550025; 2 贵州大学 贵州省大数据产业发展应用研究院, 贵阳 550025)

**摘 要**: 多跳机器阅读理解需要从相关文档中提取关键线索,并以此推理出问题的答案。随着自然语言技术的进步,该任务吸引了越来越多研究者的关注,但很少注重关键线索之间直接的信息交互。因此,本文提出了一种基于多阶段推理的多跳阅读理解模型结构,基于逐步精炼信息的思路,通过基于 PLM 的深度学习模型,依次提取与问题有关的相关文档集、支持句集以及最终的答案片段,并在答案预测模型中提出多级关系图卷积网络,以实现线索信息的充分交互。在 HotpotQA 数据集上进行的实验表明,与基准模型相比答案片段 *EM* (Exact Match)值提升了 24.26%, *F*1 值提升了 23.86%。

**关键词**: 多跳阅读理解; 多阶段推理; 图卷积网络; 自然语言处理

**中图分类号**: TP391.1 **文献标志码**: A **文章编号**: 2095-2163(2024)08-0001-10

## Multi-hop machine reading comprehension based on multi-stage reasoning

LI Juntao[1], OUYANG Zhi[2], DU Nisuo[2]

(1 School of Mathematics and Statistics, Guizhou University, Guiyang 550025, China;
2 Guizhou Big Data Academy, Guizhou University, Guiyang 550025, China)

**Abstract**: Multi-hop machine reading comprehension involves extracting key clues from relevant documents and inferring the answer to a given question based on those clues. With the advancement of natural language processing technology, this task has attracted increasing attention from researchers. However, little attention has been paid to direct interaction between key clues. Therefore, this paper proposes a multi-stage reasoning-based multi-hop reading comprehension model structure. Following the idea of progressively refining information, it sequentially extracts relevant document sets, supporting sentence sets, and final answer fragments based on a PLM-based deep learning model that is related to the question. In addition, a multi-level relational graph convolution network is proposed in the answer prediction model to facilitate sufficient interaction between key information. Experiments conducted on the HotpotQA dataset demonstrate that, compared to the baseline model, the proposed model achieves an improvement of 24.26% in Exact Match (*EM*) value and 23.86% in *F*1 value for answer fragment prediction.

**Key words**: multi-hop reading comprehension; multi-stage rreasoning; graph convolutional network; natural language processing

## 0 引 言

机器阅读理解(Machine Reading Comprehension, MRC)是一项使机器能根据给定的文本信息回答人类相关问题的技术,该技术可以帮助人类从文本数据中快速提取重点信息,理解相关内容并进行拓展,显著提升人类获取和处理信息的效率。因此,如何让机器更好地理解自然语言以提升其智能化水平问题,吸引了不少国内外学者的关注。随着预训练语言模型的发展,模型在 SQuAD[1-2]、RACE[3]、CoQA[4]等单跳阅读理解数据集上的表现已取得显著提升。然而,这些模型仍缺少整合多段落中的线索信息并进行多跳推理能力,难以处理更加复杂的多跳阅读理解问题。

不同于只需要定位文档中的某一片段,并从中提取与问题匹配的信息作为答案的单跳阅读理解任务,多跳阅读理解需要模型提取多个文档中的逻辑信息,并进行多次推理得出最终的答案。为了获取多句子、

**基金项目**: 国家自然科学基金(72261004); 贵州大学培育项目(贵大培育[2020]41 号)。

**作者简介**: 李君涛(1995 - ),男,硕士研究生,主要研究方向:自然语言处理,数据挖掘; 杜逆索(1986- ), 男,博士,讲师,主要研究方向:人工智能,仿真模拟。

**通讯作者**: 欧阳智(1987-),男,博士,讲师,主要研究方向:知识管理,机器学习。Email: zouyang@gzu.edu.cn

**收稿日期**: 2023-05-22

多段落之间的交互信息,需要更有针对性和跳跃性地提取相关文档的逻辑推理信息。针对非结构化文本数据的多跳阅读理解任务,现有的方法中基于图结构神经网络的方法是最受关注的研究框架之一。

在阅读理解任务中,图神经网络的方法主要是通过将语义信息和推理链的逻辑关系表示成图结构化数据,然后以迭代的方式更新图数据中的节点表示,使得每个节点能够融合其邻居节点的信息,并逐渐传播和聚合整个图的信息。这种方式生成的节点信息更具代表性,可以很好地表示多跳的推理逻辑,还可以捕获节点间的依赖关系。图神经网络的方法在处理多跳阅读理解任务中取得了相当不错的效果,如 Qiu 等[5]提出的 DFGN 模型,Tu 等[6]提出的 SAE 模型。然而,这些模型却很少关注关键线索间的信息交互,进而导致图神经网络进行推理时,存在信息冗余和关键特征提取不充分的问题。

为解决上述图神经网络模型的不足,本文提出了基于多阶段推理的模型结构来强化阅读理解中关键信息之间的交互。首先,通过两阶段的文档提取模型,筛选出与回答问题有关的相关文档;然后通过基于图网络的支持句选择模型,标记相关文档中的线索支持句集;最后将相关文档和被标记的支持句集作为输入,通过答案预测模型预测最终的答案。

## 1 相关工作

随着 BERT[7]、AlBERT[8]等预训练语言模型(Pre-Trained Language Model, PLM)的提出,自然语言处理技术也随之进入了一个全新的时代,仅通过微调就能满足大部分下游任务的需求。而以 HotpotQA[9]、WikiMultiHopQA[10]等抽取式多跳阅读理解数据集为自然语言处理领域提供了更具挑战性的任务。现有针对多跳阅读理解任务的研究工作主要有以下几种研究思路:

(1)改进检索方法。该方法使用信息检索(Information Retrieval, IR)模型对开放领域语料进行初步过滤,并利用单跳阅读理解(Reading Comprehension, RC)模型从这些相关文档中提取问题的答案。例如,Chen 等[11]提出了维基百科问答系统 RrQA,通过 TF-IDF 的文本相识度计算,检索出维基百科语料库中的相关文档,再从文档中筛选出问题的答案。Grail 等[12]提出了名为 LQR-net 的潜在问题重构网络,该网络通过在潜在空间中重新构建问题表示并积累信息,使模型能够以更有效的方式改写问题和积累信息。然而,这些方法为非端到端的方法,检索阶段的结果好坏会影响到阅读阶段,造成错误传递进而影响整个训练结果。

(2)基于问题分解的研究。其基本思想是将复杂问题分解为若干简单的子问题,训练模型整合子问题的答案生成最终答案。例如,Jiang 等[13]提出了一种可解释的、基于控制器的自组装神经模块网络用于多跳推理。其中的模块可以自动组装成一个可解释的网络,以便更好地理解模型的推理过程。Perez 等[14]提出了一种无监督的问题分解方法。该方法使用了一个基于掩码的自编码器,该自编码器可以将问题分解成子问题,并且可以在不使用任何标签或人工注释的情况下进行训练。但是,在处理复杂问题时此类方法会出现答案类型覆盖不全、问题无法分解等情况。

(3)通过记忆网络和注意力机制对文本信息进行交互。例如,Seo 等[15]在处理阅读理解任务时采用了一种双向注意力流(BiDAF)网络,该网络可以同时对问题和文本进行建模,并使用了一种新的方法来计算注意力权重,从而更好地捕捉问题和文本之间的交互信息。如,朱斯琪等[16]设计了一种深度交互融合网络模型,在段落级别通过多层自注意力机制实现不同段落间信息的交互融合,从而筛选出与问题相关的段落,在推理阶段通过段落间的交叉注意力对相关段落进行信息交互。这些基于深度学习的模型,虽然实现了对文章和问题之间信息的深度交互,但是无法有效强化特征信息间的交互,难以引入额外的细粒度信息。

此外,图神经网络也是处理阅读理解问题的有利工具。结合命名实体识别技术 Tu 等[17]、Cao 等[18]、刘啸等[19]都通过 Spacy 或 StanfordCoreNLP 等工具包识别上下文命名实体并以此构建实体图,之后利用 GCN[20]、GAT[21]及其变体对实体图中的信息进行多跳推理。Huang 等[22]提出的 BFR 模型同样是将文档中的每个句子作为节点。与 SAE 模型不同的是 BFR 模型构建的节点图中边的类型只有一种,但是同一种边包含不同的权重关系,在图神经推理交互时信息交互的多少根据边的权重决定。

尽管上述针对多跳阅读理解任务的模型方法取得了不错的效果,然而在针对关键线索的提取和推理上仍然存在着不足,很少有研究工作关注对支持线索直接提取和其中关键特征信息的推理交互。因此,本文提出了基于多阶段推理的多跳阅读理解模型,通过文本分类模型提取问题的相关文档,并通过逐级提取信息的两阶段推理模型,解决现有模型进

行多跳推理时存在的信息冗余,以及关键特征提取不充分的问题。

## 2　模型结构

多跳阅读理解任务首先需要从多个候选文档中提取出与问题相关的文档,从这些相关文档中进行推理得出问题最终的答案。因此,本文将任务整体划分为两个子任务:文档提取任务和阅读推理任务。针对每个子任务的模型都包含两个阶段,共 4 个阶段,如图 1 所示。

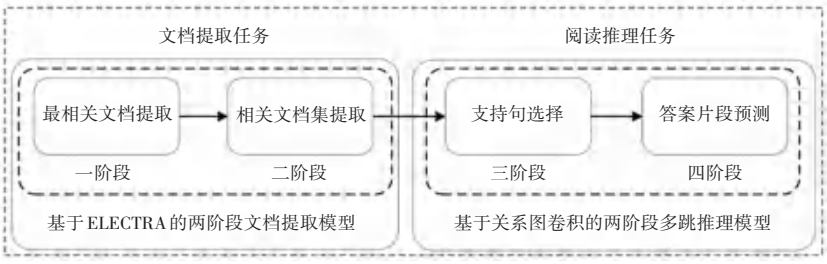


**图 1　多跳阅读理解任务多阶段执行流程**

**Fig. 1　Multi-hop reading comprehension tasks involve multi-stage execution processes**

第一、二阶段的模型用于处理文档提取任务,对应基于 ELECTRA 的两阶段文档提取模型,三、四阶段的模型用于执行阅读推理任务,对应基于关系图卷积的两阶段多跳推理模型。第一阶段模型将提取与问题最相关的文档;第二阶段在此基础上进一步提取剩余文档,得到所有与问题相关的文档集合;第三阶段将第二阶段得到相关文档集作为输入,训练模型获取与推理问题答案相关的支持句集;第四阶段模型将结合第三阶段得出的支持句集预测问题最终的答案。

### 2.1　文档提取

借鉴 FE2H 模型[23]中的相关文档提取思路,本文文档提取模型结构如图 2 所示。通过对候选文档的两阶段提取,模型可以有效地捕捉组合文档与问题的相关关系,进一步考虑到将判断文档与问题是否相关的二分类任务转化为预测文档组合与问题相关程度的任务,让模型以更细粒度的角度判断文档的组合类别,进而提高模型提取相关文档的能力。因此,在第二阶段构建了基于文档组合与问题相关程度的多分类任务。任务目标是预测出与问题完全相关的文档组合,该文档组合将作为问题最终的相关文档集。

多分类标签生成规则为:与问题完全不相关的文档组归为一类,标签为 0;组合中存在某一文档与问题相关的归为一类,标签为 1;组合中所有文档都与问题相关的归为一类,标签为 2。对所有的文档集处理后,根据总样本数据的分布通过一定策略平衡数据样本,如按一定比例对占比高的样本进行欠采样以及生成一部分占比低的样本。

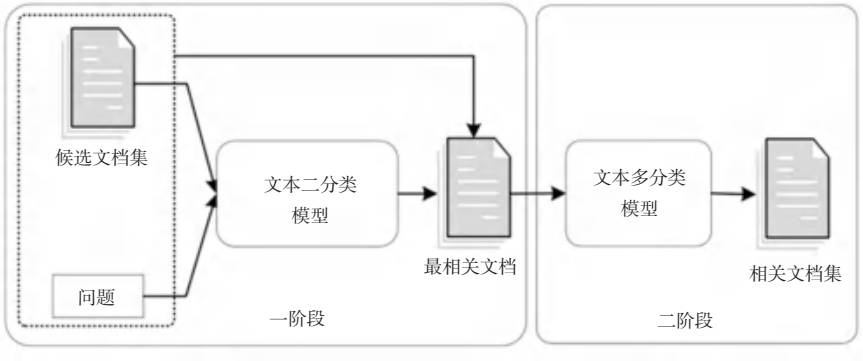


**图 2　两阶段文档提取模型结构图**

**Fig. 2　Two-stage document extraction model architecture diagram**

### 2.2　多跳推理

在文档提取任务中获取了与回答问题有关的相关文档集,结合此文档集进行最终答案预测相关的多跳推理,模型结构如图 3 所示,其中包含两个阶段。一阶段是支持句选择模型,目标是使模型预测出与回答问题正确答案相关的关键支持句;二阶段是答案预测模型,目标是结合一阶段预测的支持句集推理出问题最终的答案。

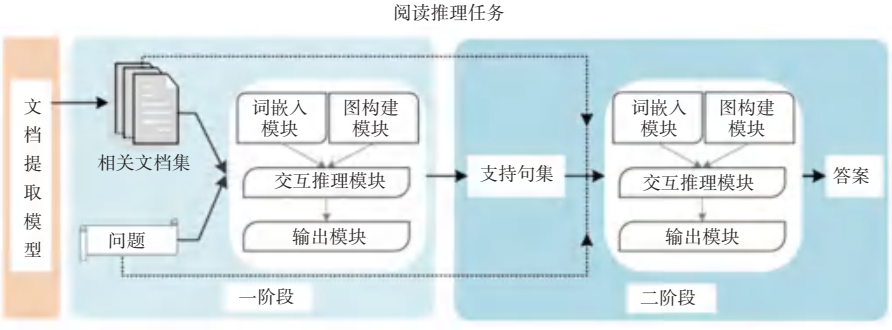


**图 3 两阶段多跳推理模型架构图**

**Fig. 3 Two-stage multi-hop reasoning model architecture diagram**

为了更好地融入图结构数据中的边类型信息，基于关系图卷积网络的两阶段多跳推理模型，在第一阶段采用了关系图卷积网络进行多跳推理。而为了有效利用第一阶段模型得到支持句信息以推理问题最终的答案，在关系图卷积网络的基础上提出了多级关系图卷积网络来执行第二阶段的交互推理。

关系图卷积网络（Relational Graph Convolution Network，RGCN）是 Schlichtkrull 等[24]在图卷积网络基础上提出的网络模型，用以解决图卷积网络不能融合边类型信息的问题。关系图卷积网络模型的算法如下：

给定图信息 $G=(N,E,R)$，其中 $N$ 为节点集合，$R$ 为边类型集合，$r\in R$，$E$ 为边的集合，$(n_i,r,n_j)\in E$。对于输入的图信息，每一层的节点信息根据公式(1)的 RGCN 算法进行更新。

$$h_i^{(l+1)}=\sigma\left(\sum_{r\in R}\sum_{j\in N_i^r}\frac{1}{c_{i,r}}\boldsymbol{W}_r^{(l)}h_j^{(l)}+\boldsymbol{W}_0^{(l)}h_j^{(l)}\right)\quad(1)$$

其中，$\boldsymbol{h}^l$ 为第 $l$ 层的节点表示向量；$\sigma(\cdot)$ 为激活函数；$c_{i,r}$ 为正则化常量；$N_i^r$ 为与节点 $i$ 为 $r$ 类型关系的邻居节点的集合；$\boldsymbol{W}_0^l$ 为第 $l$ 层自连接的训练权重矩阵；$\boldsymbol{W}_r^l$ 为第 $l$ 层第 $r$ 种边类型的权重矩阵。

多级关系图卷积网络在全局相关文档和问题节点信息交互的基础上，加强了支持句节点信息的传递。首先，全局节点交互层通过关系图卷积网络对全局的相关文档和问题节点信息进行交互，得到各节点的全局表示向量 $\boldsymbol{gl-node}_i^l$；随后，在支持句节点交互层基于预测支持句的图信息，对支持句的全局表示信息再次进行关系图卷积运算，得到更新的支持句节点 $\boldsymbol{sup-node}_j^l$；最后，通过公式(2)所示方式对全局节点信息和支持句节点信息进行融合，得到最终的节点表示 Node。

$$\text{Node}=\text{Attention}(\text{Concat}(\boldsymbol{gl-node},\boldsymbol{sup-node}))\quad(2)$$

### 2.2.1 支持句选择模型

支持句选择模型包含 4 个模块，对于输入的相关文档集和问题先通过图构建模块生成相应的图结构数据，同时通过词嵌入模块对文本数据向量化，随后通过交互推理模块对问题和文档信息进行交互，最后通过输出模块得出支撑推理链的支持句集。具体框架如图 4 所示：

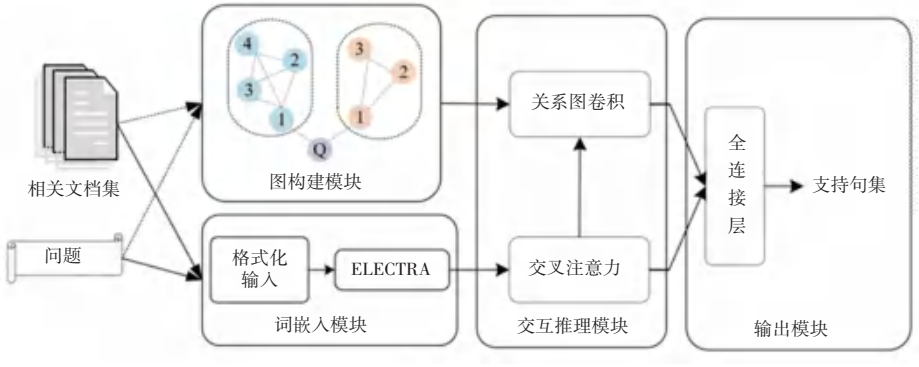


**图 4 支持句选择模型架构图**

**Fig. 4 Supporting sentence selection model architecture diagram**

各模块的详细说明如下：

1）图构建模块。该模块主要是根据文挡选择模型得到的相关文档集通过设定的规则构建节点图，该节点图将用于关系图卷积网络中的信息交互。

节点图构建规则如下:

(1)为弱化同一文档中前后句子词信息存在距离依赖的问题,同一文档中的句子节点之间建立一种边的连接,生成邻接矩阵 $M_d$。

(2)基于文档首句往往是总结句特征这一事实,将问题节点和每一个相关文档的首句节点相连,建立另一种边的连接,生成领接矩阵 $M_q$。

2)词嵌入模块。该模块主要是将问题 $Q$ 和根据文挡选择模型得到的相关文档集 $D_{\text{relevant}}$ 转化为本任务所需的输入格式,并通过预训练语言模型生成文本的表示向量。首先将问题和相关文档集拼接为 [CLS] + $Q$ + [SEP] + $D_{\text{relevant}}$ + [SEP] 的形式作为输入 *seq*, 接着使用 ELECTRA 的 Large 模型对 *sep* 进行编码处理,得到公式(3)所示的向量表示 $QD \in R^{L\times d}$。其中,$L$ 为文本输入的最大长度, $d$ 为词向量的表示维度。

$$QD = \text{ELECTRA} - \text{Large}(seq) \tag{3}$$

3)交互推理模块。该模块主要是根据构建好的节点图以及向量表示 $d$, 通过交叉注意力和图卷积网络,对问题和相关文档信息进行交互推理。

首先,从向量表示 $QD$ 拆分出问题表示 $R_q$ 和文档表示 $R_c$, 随后通过交叉注意力网络将问题信息融入到文档中,计算过程如公式(4)-公式(6)所示。

$$Q = R_c\, W^Q \tag{4}$$

$$K = R_q\, W^K \tag{5}$$

$$V = R_q\, W^V \tag{6}$$

根据公式(7)-公式(9)计算多头交叉注意力。

$$\text{cross} - \text{Attention}\ (Q,K,V) = \text{Softmax}\left(\frac{Q\cdot K^T}{\sqrt{d_k}}\right)V \tag{7}$$

$$head_i = \text{cross} - \text{Attention}\ (Q_i,K_i,V_i) \tag{8}$$

$$\text{MH} - \text{CrossAttention}\ (Q,K,V) = \text{Concat}\ (head_1, \cdots, head_i)\ W^o \tag{9}$$

将多头交叉注意力的表示结果与预训练模型得到的原始表示 $QD$ 融合,得到更新后的整合文本表示,如公式(10)所示。

$$QD' = QD + w * \text{MH} - \text{CrossAttention}(Q,K,V) \tag{10}$$

通过关系图卷积网络进行节点信息更新的具体步骤如下:

**步骤 1**　生成节点表示。首先根据问题和文档中句子的开始位置 $Pos^i_{\text{start}}$ 和结束位置 $Pos^i_{\text{end}}$, 得到每个句子的矩阵表示 $M^i_{\text{sent}} \in R^{len_{\text{sent}} * d}$。其中, $len_{\text{sent}}$ 表示最长的句子长度。然后通过一层神经元网络层计算句子中每个 *token* 的注意力得分,根据注意力得分聚合句子信息得到句子的节点表示,如公式(11)所示。其中, $f(\cdot)$ 表示包装了线性层和非线性激活函数的神经单元网络。

$$Node_i = M^i_{\text{sent}} \otimes f(QD'[Pos^i_{\text{start}}:Pos^i_{\text{end}}]) \tag{11}$$

**步骤 2**　利用带有门控机制的关系图卷积网络(Gate-RGCN)更新节点表示。首先按照公式(12)对节点进行关系图卷积运算,获取邻居节点的语义信息,然后按照公式(13)、公式(14)的门控机制进行节点信息流量控制,最后按照公式(15)进行节点图同种特征的信息融合,经过一定次数的图运算后,得到最终的节点信息表示 $Node_{\text{end}}$。

$$Node_i^{l+1} = \sigma\Big(\sum_{r\in R}\sum_{j\in N_i^r} W_r^{(l)}\, Node_j^{(l)} + W_0^{(l)}\, Node_j^{(l)}\Big) \tag{12}$$

$$w^i = \sigma(f_{\text{gate}}(Concat(Node_i^{l+1}, Node_i^l))) \tag{13}$$

$$Node_i^{l+1} = w^i \otimes \tanh(Node_i^{l+1}) + (1 - w^i) \otimes Node_i^l \tag{14}$$

$$\text{Node}_i^{l+1} = \text{mean}\Big(\sum_{r\in M_d, M_q} Node_i^{l+1}\Big) \tag{15}$$

其中, $Node_i^{l+1}$ 表示第 $i$ 个节点在 $l$ 次更新后的节点表示, $\boldsymbol{R}$ 表示邻接矩阵; $N_i^r$ 表示第 $r$ 种边类型下第 $i$ 个节点的所有邻居节点的集合; $\boldsymbol{W}_r^{(l)}$ 表示在第 $l$ 层图运算中第 $r$ 种边类型下权重矩阵; $\boldsymbol{W}_0^{(l)}$ 表示固定的初始化关系矩阵; $\boldsymbol{f}_{\text{gate}}(\cdot)$ 表示一层线性转化函数; $\sigma(\cdot)$ 为 sigmoid 激活函数;Ä表示哈达马积。

4)输出模块。该模块主要是对交互推理模块得到的节点信息进行最后的关于问题答案的支持句预测。首先,按照公式(16)通过两层全连接层网络 $f_{\text{sent}}(\cdot)$ 分别对交叉注意力层得到的整合文档表示 $QD'$ 和卷积网络层得到的全局节点表示 $Node_{\text{end}}$ 计算支持句概率预测值,最后按照公式(17)通过二分类交叉熵损失函数(Binary Cross Entropy, BCE)计算支持句预测的损失 $L_{\text{sent}}$。在预测时,选择预测概率大于 threshold 的句子作为最终支持句预测结果。

$$\hat{y}_{\text{sent}} = f_{\text{sent}}(QD') + w \cdot f_{\text{sent}}(Node_{\text{end}}) \tag{16}$$

$$L = \text{BCE}(\hat{y}_{\text{sent}}, y_{\text{sent}}) \tag{17}$$

### 2.2.2　答案预测模型

答案预测模型包含 4 个模块,框架如图 5 所示。

(1)图构建模块。该模块除了根据文挡选择模型得到的相关文档集通过设定的规则构建节点图之外,还通过支持句信息构建支持句节点图。支持句节点图通过支持句节点之间两两相连进行构建,得到关于相关文档的邻接矩阵 $\boldsymbol{M}_d$、关于问题的邻接矩阵 $\boldsymbol{M}_q$ 以及关于支持句的邻接矩阵 $\boldsymbol{M}_s$。

(2)词嵌入模块。该模块是将问题 $Q$ 和相关文

档集转化为如下格式:[*CLS*] + “yes” + “no” + *Q* + [SEP] + $D_{\text{relevant}}$ + [SEP],接着使用 ELECTRA 模型对 *sep* 进行编码处理,生成整合后的文档表示 $QD \in R^{L\times d}$。

(3)交互推理模块。该模块根据上述构建好的节点图以及向量表示 *d* 通过交叉注意力和多级图卷积网络,对问题和相关文档信息进行交互推理。交叉注意力网络的信息更新过程与支持句选择模型中交互推理模块的交叉注意力计算过程一致,得到更新后的整合文本表示 $QD'$,然后通过多级关系图卷积网络进行节点信息更新。具体过程为:首先根据公式(11)~公式(15)得到全局节点信息表示 $Node^{\text{global}}$,然后将支持句节点表示和支持句邻接矩阵 $\boldsymbol{M}_s$ 作为 Gate-RGCN 的输入,按照公式(18)-(20)对支持句节点进行更新,最后按照公式(21)将全局节点表示和支持句表示进行拼接通过注意力网络得到最终的节点表示 $Node_{\text{fuse}}$。

$$Node_i^{\sup(l+1)} = \sigma\Big(\sum_{r\in M_s}\sum_{j\in N_i} W^{\sup} Node_j^{\sup(l)} + W_0^{\sup} Node_j^{\sup(l)}\Big) \quad (18)$$

$$w^i = \sigma(f_{\text{gate}}(Node_i^{\sup(l+1)})) \quad (19)$$

$$Node_i^{\sup} = w^i \otimes \tanh(Node_i^{\sup(l+1)}) + (1-w^i) \otimes Node_i^{\sup(l)} \quad (20)$$

$$Node_{\text{fuse}} = \text{Attention}(\text{Concat}(Node^{\text{global}}, Node^{\sup(l+1)})) \quad (21)$$

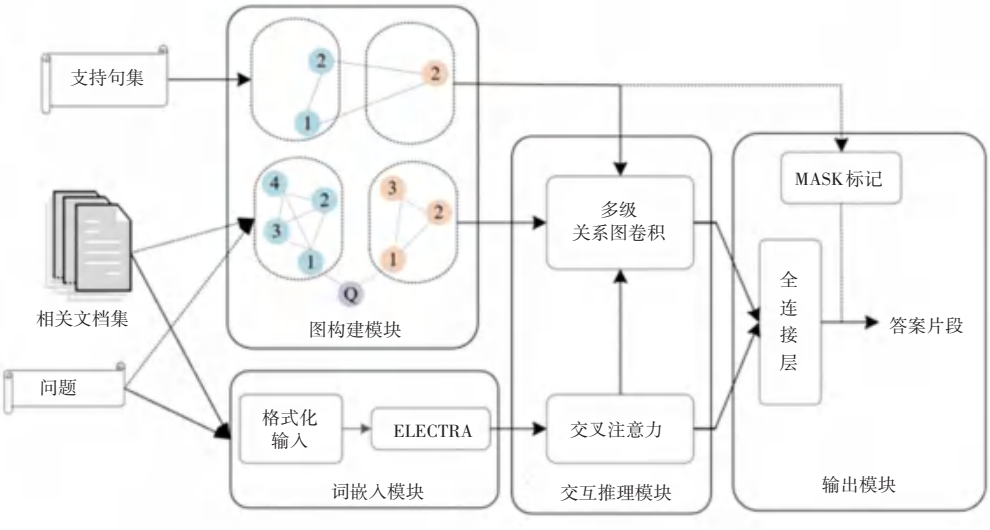


**图 5 答案预测模型架构图**

**Fig. 5 Answer prediction model architecture diagram**

(4)输出模块。该模块通过交互推理模块得到更新后的整合文本表示 $QD'$,以及更新后的节点表示 $Node_{\text{fuse}}$ 进行问题答案的最终预测。首先根据预测的支持句集可以得到支持句集的标记 $MASK_{\text{sup}}$,该标记将限定预测答案片段的始末位置只在支持句集的范围内。通过注意力网络对融合节点表示 $Node_{\text{fuse}}$ 的信息进行交互拓展,随后通过两层全连接层网络 $f_{\text{span}}(\cdot)$ 得到相关文档集中每一个 *token* 基于节点信息的得分 $y_{\text{sent}}$,最后结合支持句的 MASK 标记得到最终的答案开始位置得分 $\hat{y}_{\text{start}}$ 和结束位置得分 $\hat{y}_{\text{end}}$。如公式(22)~公式(24)所示。

$$y_{\text{sent}} = f_{\text{span}}(\text{Attention}(Node_{\text{fuse}}, QD')) \quad (22)$$

$$\hat{y}_{\text{start}} = f_{\text{span}}(QD') + y_{\text{sent}} + MASK_{\text{sup}} \quad (23)$$

$$\hat{y}_{\text{end}} = f_{\text{span}}(QD') + y_{\text{sent}} + MASK_{\text{sup}} \quad (24)$$

在训练时,通过公式(25)计算答案范围预测的损失 $L_{\text{span}}$。其中,$CE(\cdot)$ 是交叉熵损失函数。在预测时,通过 sigmoid 激活函数得到答案起始位置的预测概率值,如公式(26)所示。

$$L_{\text{span}} = CE(\hat{y}_{\text{start}}, y_{\text{start}}) + CE(\hat{y}_{\text{end}}, y_{\text{end}}) \quad (25)$$

$$P_{\text{start}}, P_{\text{end}} = \sigma(\hat{y}_{\text{start}}, \hat{y}_{\text{end}}) \quad (26)$$

## 3 实验结果与分析

### 3.1 实验环境

实验使用的编程语言是 Python3.8,基于 CUDA10.2,Pytorch 1.6.0 框架搭建模型的实验环境;服务器操作系统为 Ubuntu18.04,内存为 128 GB,GPU 为 12 GB Nvidia Titan XP,模型在 8 块 GPU 上运行。模型中的具体参数见表 1。

**表 1　阅读推理模型参数**

**Table 1　Reading comprehension model parameters**

| 参数 | 参数值 |
|---|---|
| 预训练模型 | ELECTRA-Large |
| 词嵌入维度 | 1 024 |
| 拼接文本最大长度 | 512 |
| 问题最大长度 | 64 |
| 一般隐藏层维度 | 1 024 |
| 训练 batch_size | 3 |
| 验证、测试 batch_size | 12 |
| 初始学习率 | $1 \times 10^{-5}$ |
| 优化器 | Adam |
| adam_epsilon | $1 \times 10^{-8}$ |
| noise_lambda | 0.15 |
| 支持句选择模型图卷积层数 | 2 |
| 支持句选择模型预测阈值 | 0.50 |
| 答案预测模型图卷积层数 | 3 |
| 训练轮次 | 8 |

## 3.2　实验数据

本模型采用了 HotpotQA 数据集 distractor 版本进行实验设计。数据集包含来自维基百科的超过 10 万个问答样本。数据集中每个样例都是由(候选文档,问题,支持事实,答案)四元组组成,与简单的阅读理解数据集不同,该数据集不仅给出了答案标签,还提供了支持句作为线索标签。为此,还引入了新的评价指标 sup_*F*1 和 sup_*EM*, 用于评估多跳推理模型找到正确支持证据的能力。目前,大多数多跳推理模型都是针对该数据集进行研究的。表 2 为 HotpotQA 数据集的数据样例。其中斜体字体为推理链上的关键句子,加粗下划线字体为正确答案。

## 3.3　实验结果

为了验证整体模型的性能,选择了 4 个有代表性的基于图神经网络的阅读推理模型(DFGN[5]、SAE[6]、HGN[25]、AMGN[26])以及基线模型(Baseline Model[7]),在上述数据集的验证集上进行比较,整体模型与其他模型在上述评价指标上的结果见表 3。

**表 2　HotpotQA 数据集数据样例**

**Table 2　HotpotQA dataset data sample**

| | |
|---|---|
| 文档 1 | [1] *Allison Beth "Allie" Goertz (born March 2, 1991) is an American musician.*<br>[2] *Goertz is known for her satirical songs based on various pop culture topics.*<br>[3] *Her videos are posted on YouTube under the name of Cossbysweater.*<br>[4] Subjects of her songs have included the film "The Room", the character Milhouse from the television show "The Simpsons", and the game Dungeons & Dragons.<br>… |
| 文档 2 | [1] Bartholomew JoJo "Bart" Simpson is a fictional character in the American animated television series "The Simpsons" and part of the Simpson family.<br>[2] He is voiced by Nancy Cartwright and first appeared on television in "The Tracey Ullman Show" short "Good Night" on April 19, 1987.<br>[3] Cartoonist Matt Groening created and designed Bart while waiting in the lobby of James L. Brooks´ office.<br>… |
| 文档 3 | [1] *Milhouse Mussolini van Houten is a fictional character featured in the animated television series "The Simpsons", voiced by Pamela Hayden, and created by Matt Groening who named the character after President Richard Nixon's middle name.*<br>[2] Later in the series, it is revealed that Milhouse´s middle name is "Mussolini." |
| … | … |
| 问题 | Musician and satirist Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who? |
| 支持事实 | 文档 1:1、2、3;文档 3:1 |
| 答案 | President Richard Nixon |

由表 3 可见,在支持句的预测能力上,本模型与目前基于图神经网络的模型中表现最好的 AMGN 模型相比,在 *EM* 值和 *F*1 值上分别提升了 1.78%和 0.55%,表明相比于需要兼顾支持句选择和答案预测的多任务模型,以提升支持句选择的精度作为模型目标在支持句选择上更有优势。而在答案片段的预测能力上,本模型除了比 AMGN 模型在 *EM* 值和 *F*1 值上分别降低了 0.67%和 0.49%外,相比于其他模型都有一定的提升,充分说明了先提取支持句再预测答案片段的分阶段模型存在可行性。最后对比

模型整体的联合评价效果,本模型相比于效果最好的 AMGN 模型在 *EM* 值和 *F*1 值上分别提升了 1.16%和 0.26%,整体结果来看,本模型可以为处理多跳阅读理解任务提供另一种解决思路。

**表 3 模型各指标结果对比**

**Table 3 Model metric comparison results**

| 模型 | 支持句 | | 答案片段 | | 联合 | |
|---|---|---|---|---|---|---|
| | *EM*/ % | *F*1 /% | *EM*/ % | *F*1/ % | *EM*/ % | *F*1 /% |
| Baseline | 20.32 | 64.49 | 45.60 | 59.02 | 10.83 | 40.16 |
| DFGN | 51.50 | 81.62 | 56.31 | 69.69 | 33.62 | 59.82 |
| SAE-large | 61.53 | 86.86 | 66.92 | 79.62 | 45.36 | 71.45 |
| HGN-large | 62.76 | 88.47 | 69.22 | 82.19 | 47.11 | 74.21 |
| AMGN+ | 63.57 | 88.83 | 70.53 | 83.37 | 47.77 | 75.24 |
| **本模型** | **65.35** | **89.38** | **69.86** | **82.88** | **48.83** | **75.50** |

### 3.4 消融实验

为了验证支持句选择模型中阅读推理部分的网络结构是否能够有效提升实验效果,在验证集上进行消融实验,来验证交叉注意力网络和关系图卷积网络是否能够有效提升模型的效果,实验结果见表 4。由此可见,当实验去除了支持句选择模型交互推理模块的交叉注意力网络,模型在验证集上支持句的 *EM* 值和 *F*1 值分别降低了 0.85%和0.65%;当实验去除了支持句选择模型交互推理模块的关系图卷积网络,在验证集上支持句的 *EM* 值和 *F*1 值分别降低了 1.14%和 0.97%,表明通过关系图卷积网络加深文档中句子与句子之间、句子与问题之间的信息融合,可以有效提升支持句预测模型的精度。

为了验证答案预测模型中输出、推理和输出中的关键结构是否能够有效提升实验效果,在验证集上进行消融实验来验证各结构的效果,实验结果见表 5。可见,当实验去除了答案预测模型交互推理模块的交叉注意力网络,模型在验证集上答案片段的 *EM* 值和 *F*1 值分别降低了 0.77%和 0.67%;当实验去除了答案预测模型交互推理模块的多级关系图卷积网络,模型在验证集上答案片段的 *EM* 值和 *F*1 值都呈现显著下降;当实验将 MASK 标记操作替换为直接将支持句集作为答案预测模型的初始输入,即将任务转化为单跳阅读理解任务时,模型在验证集上答案片段的 *EM* 值和 *F*1 值分别下降了1.95%和 1.69%,这表明直接将第一阶段支持句选择模型预测得到的支持句集作为第二阶段答案预测模型的原始输入并不可取。原因在于文档中的前后句子间有一定的承接关系,并且可能存在使用代词指代前面出现过的名词的情况,损失语义信息,而通过 MASK 标记的方式是在模型输出时对答案片段截取的范围进行限制,不会存在上述问题。

**表 4 支持句选择模型消融实验结果**

**Table 4 Results of supporting sentence selection model ablation experiment**

| | *EM*/ % | 浮动 | *F*1/ % | 浮动 |
|---|---|---|---|---|
| 支持句模型 | **65.35** | - | **89.38** | - |
| a)去除交叉注意力 | 64.50 | −0.85 | 88.73 | −0.65 |
| b)去除关系图卷积 | 64.21 | −1.14 | 88.41 | −0.97 |

**表 5 答案预测模型消融和对比实验结果**

**Table 5 Results of answer prediction model ablation and comparative experiments**

| | *EM*/ % | 浮动 | *F*1/ % | 浮动 |
|---|---|---|---|---|
| 答案预测模型 | **69.86** | | **82.88** | |
| a)去除交叉注意力 | 69.08 | −0.78 | 82.21 | −0.67 |
| b)去除多级关系图卷积网络 | 68.77 | −1.09 | 81.93 | −0.95 |
| c)替换 MASK 标记为输入时提取支持句 | 67.90 | −1.96 | 81.17 | −1.71 |

### 3.5 结果分析

本节对于模型在验证集上不同类别(comparison, bridge)和不同答案种类 (yes/no,span)的数据,进行了评价指标的比对,结果见表 6。

从 Yang 等[7]的分析可知,distract 版本 HotpotQA 数据集中问题类别为 comparison 的数据样例占比 20%,bridge 类型的占 80%。而从实验结果来看,模型对样本量较少的 comparison 类型的问题表现更好,而针对问题需要在不同文档中找到桥接词并进行推导时,模型效果在各方面都表现一定程度的下降,这说明模型还需要在更有效地提取文档中的承接线索,以此对不同段落文档的数据中建立更准确的联系。从答案种类的角度看,对于 yes/no 类别问题,模型在答案片段预测的 *EM* 值和 *F*1 值上分别达到了 93.04%和 93.15%的高水平表现;而 span 类型数据答案始末位置不确定,导致预测难度增大,模型表现相较于 yes/no 类别的问题存在较大差距。

**表 6　不同类别问题实验结果**

**Table 6　Results of experiments on different categories of questions**

| 类别 | | 支持句 | | 答案片段 | | 联合 | |
|---|---|---|---|---|---|---|---|
| | | *EM*/ % | *F*1 /% | *EM*/ % | *F*1 /% | *EM*/ % | *F*1 /% |
| 问题类别 | comparison | 77.31 | 92.87 | 80.02 | 85.26 | 63.04 | 79.88 |
| | bridge | 62.33 | 88.49 | 67.32 | 82.27 | 45.25 | 74.40 |
| 答案种类 | yes/no | 82.34 | 94.08 | 93.04 | 93.15 | 77.73 | 87.45 |
| | span | 64.24 | 89.07 | 68.38 | 82.21 | 46.96 | 74.74 |

此外,本文还探究了模型预测答案的错误样例,错误类型可以分为答案不全类、是否错误类、实体错误类这 3 种,表 7 中列举了其统计结果和部分样例。统计结果显示在数据集所有的错误结果中,答案预测不全的错误类型占比最大,然而其大部分预测的内容并不影响答案所表达的意思。如样例所示,问题的标准答案为“Greenwich Village, New York City”,而模型的预测结果为“Greenwich Village”,线索支持句预测准确,在非严格限制回答格式的情况下可以认为是正确答案。但由于不完全一致,使得 EM 值降低,这也解释了评价指标中 *EM* 值和 *F*1 值相差较大的原因。实体错误类占 25.42%。从样例中可以看出虽然结果回答错误,但能预测到与标准答案同属性的政府职位,而线索支持句的预测上 3 个关键句子提取了其中 2 个,支撑线索不完全导致最终预测结果出现偏差。

**表 7　错误样例**

**Table 7　Examples of Error**

| 类型 | 样例 |
|---|---|
| 答案不全类(73.14%) | 问题:The director of the romantic comedy "Big Stone Gap" is based in what New York city? |
| | 预测答案:Greenwich Village　　标准答案:Greenwich Village, New York City |
| | 支持句 EM 比:2/2 |
| 是否错误类(1.44%) | 问题:Are Random House Tower and 888 7th Avenue both used for real estate? |
| | 预测答案:yes　　标准答案:no |
| | 支持句 EM 比:3/2 |
| 实体错误类(25.42%) | 问题:What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell? |
| | 预测答案:United States ambassador　　标准答案:Chief of Protocol |
| | 支持句 EM 比:2/3 |

## 4　结束语

本文提出一种处理多跳阅读理解任务的多阶段模型结构。基于逐步精炼信息的思路,通过基于 PLM 的深度学习模型依次提取与问题有关的相关文档集、支持句集以及最终的答案片段,并在答案预测模型中提出多级关系图卷积网络对关键信息进行充分交互。在 Distract 版本的 HotpotQA 数据集上进行的实验表明,本文模型与基于图神经网络的其他模型相比综合表现具有一定优势。对最后的实验结果进行分析,发现支撑线索不完全或冗余会影响最终预测结果的准确性,针对此部分的问题将是后续模型改进的重点方向。在未来的工作中,本文将专注于答案为片段式以及推理线索为桥连类型的问题进行研究,使模型具备更加综合且全面的推理能力。

## 参考文献


[1] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100000+ questions for machine comprehension of text[C]//Proceedings of the 2016 Conference of the Empirical Methods in Natural Language Processing. IEEE, 2016: 2383 - 2392.

[2] RAJPURKAR P, JIA R, LIANG P. Know what you don't know: Unanswerable questions for SQuAD[C] //Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. IEEE, 2018: 784-789.

[3] LAI G, XIE Q, LIU H, et al. Race: Large - scale reading comprehension dataset from examinations[J]. arXiv preprint arXiv:1704.04683, 2017.

[4] REDDY S, CHEN D, MANNING C. Coqa: A conversational question answering challenge[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 249-266.

[5] QIU L, XIAO Y, QU Y, et al. Dynamically fused graph network for multi - hop reasoning[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. IEEE,

2019：6140-6150.
[6] TU M，HUANG K，WANG G，et al. Select，answer and explain：Interpretable multi - hop reading comprehension over multiple documents [C]//Proceedings of the Thirty - Fourth AAAI Conference on Artificial Intelligence. 2020：9073-9080.
[7] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. 2019：4171-4186.
[8] LAN Z，CHEN M，GOODMAN S，et al. Albert：A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv：1909.11942，2019.
[9] YANG Z，QI P，ZHANG S，et al. Hotpotqa：A dataset for diverse，explainable multi - hop question answering [C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. IEEE，2018：2369-2380.
[10] HO X，NGUYEN A K D，SUGAWARA S，et al. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps[J]. arXiv preprint arXiv：2011.01060，2020.
[11] CHEN D，FISCH A，WESTON J，et al. Reading Wikipedia to answer open - domain questions [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017：1870-1879.
[12] GRAIL Q，PEREZ J，GAUSSIER E J G. Latent question reformulation and information accumulation for multi-hop machine [P]. U.S.：17/015，501，2021-8-19.
[13] JIANG Y，BANSAL M. Self-assembling modular networks for interpretable multi - hop reasoning [C] //Proceedings of International Joint Conference on Natural Language Processing. Association for Computational Linguistics. IEEE，2019：4473 - 4483.
[14] PEREZ E，LEWIS P，CHO K，et al. Unsupervised question decomposition for question answering[C] //Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. IEEE，2020：8864 - 8880.
[15] SEO M，KEMBHAVI A，FARHADI A，et al. Bidirectional attention flow for machine comprehension [J]. arXiv preprint arXiv：1611.01603，2016.
[16] 朱斯琪，过弋，王业相. 基于深度交互融合网络的多跳机器阅读理解[J]. 中文信息学报，2022，36(5)：67-75.
[17] TU M，WANG G，HUANG J，et al. Multi - hop reading comprehension across multiple documents by reasoning over heterogeneous graphs [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019：2704-2713.
[18] CAO Y，FANG M，TAO D，et al. BAG：Bi-directional attention entity graph convolutional network for multi - hop reasoning question answering[C]//Proceedings of 2018 Conference of North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Association for Computational Linguistics. IEEE，2019：357-362.
[19] 刘啸，杨敏. 基于动态图神经网络的会话式机器阅读理解研究[J]. 集成技术，2022，11(2)：67-78.
[20] KIPF T. WELLING M. Semi-supervised classification with graph convolutional networks[C]//Proceedings of the 5th International Conference on Learning Representations. IEEE，2017：1-8.
[21] VELIČKOVIĆ P，CUCURULL G，CASANOVA A，et al. Graph attention networks[J]. arXiv preprint arXiv：1710.10903，2017.
[22] HUANG Y，YANG M. Breadth first reasoning graph for multi-hop question answering[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. IEEE，2021：5810-5821.
[23] LI X Y，LEI W J，YANG Y B. From easy to hard：Two-stage selector and reader for multi - hop question answering [C]// Proceedings of 2023 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2023：1-5.
[24] SCHLICHTKRULL M，KIPF T N，BLOEM P，et al. Modeling relational data with graph convolutional networks [C]// Proceedings of European Semantic Web Conference. IEEE，2018：593-607.
[25] FANG Y，SUN S，GAN Z，et al. Hierarchical graph network for multi - hop question answering [C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. IEEE，2020：8823-8838.
[26] LI R，WANG L，WANG S，et al. Asynchronous multi-grained graph network for interpretable multi-hop reading comprehension [C]// Proceedings of the 2021 Conference on International Joint Conference on Artificial Intelligence. IEEE，2021：3857-3863.