计算机科学
COMPUTER SCIENCE

解耦知识蒸馏在文档级关系抽取中的应用

刘乐, 肖蓉, 杨肖

引用本文

刘乐, 肖蓉, 杨肖. 解耦知识蒸馏在文档级关系抽取中的应用[J]. 计算机科学, 2025, 52(8): 277-287.

LIU Le, XIAO Rong, YANG Xiao. Application of Decoupled Knowledge Distillation Method in Document-level RelationExtraction [J]. Computer Science, 2025, 52(8): 277-287.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240600050

解耦知识蒸馏在文档级关系抽取中的应用

刘 乐 肖 蓉 杨 肖
湖北大学计算机与信息工程学院 武汉 430062
(202131116020073@stu.hubu.edu.cn)

摘 要 文档级关系抽取是自然语言处理领域中的一个重要研究方向,旨在从无结构或半结构的自然语言文档中提取实体之间的语义关系。提出了结合使用解耦知识蒸馏方法和交叉多头注意力机制来解决文档级关系抽取任务。首先,交叉多头注意机制不仅能够并行关注不同注意力头中的元素,使模型在不同粒度和层级上进行信息的交流和整合,而且允许模型在计算头实体与尾实体之间的注意力时,同时考虑它们与关系之间的相关性,从而提升模型对复杂关系的理解能力,增强模型对实体特征表示的学习。此外,为了进一步优化模型性能,还引入了解耦知识蒸馏方法去适应远程监督数据。该方法将原始KL散度损失中的目标类别知识蒸馏损失TCKDL和非目标类别知识蒸馏损失NCKDL解耦为了两个可以通过超参数调整其权重重要性的独立部分,提高了知识蒸馏过程的灵活性和有效性,特别是在处理DocRED远程监督数据中的噪声时,能够更精准地进行知识迁移和学习。实验结果表明,所提模型在DocRED数据集上能够更有效地提取实体对之间的关系。

关键词:自然语言处理;文档级关系抽取;DocRED;交叉多头注意力;解耦知识蒸馏;远程监督数据;KL散度
中图分类号 TP391

Application of Decoupled Knowledge Distillation Method in Document-level Relation Extraction

LIU Le,XIAO Rong and YANG Xiao
School of Computer Science and Information Engineering,Hubei University,Wuhan 430062,China

Abstract Document-level relation extraction is an important research direction in the field of natural language processing,aiming to extract semantic relationships between entities from unstructured or semi-structured natural language documents. This paper proposes a solution combining decoupled knowledge distillation and cross multi-head attention mechanisms to address the DocRE task. Firstly,the cross multi-head attention mechanism can not only simultaneously focus on elements in different attention heads,enabling the model to exchange and integrate information at different granularities and levels but also allow the model to consider the correlation between head and tail entities and their relations when calculating attention,thereby enhancing the model's understanding of complex relationships and improving the learning of entity feature representations. Additionally,to further optimize the model's performance,this paper introduces a decoupled knowledge distillation method to adapt to distantly supervised data. This method decouples the original KL divergence loss into target class knowledge distillation loss(TCKDL) and non-target class knowledge distillation loss(NCKDL),which can adjust their weight importance through hyperparameters,increasing the flexibility and effectiveness of the knowledge distillation process. Particularly,it enables more precise knowledge transfer and learning when dealing with noise in the DocRED distantly supervised data. Experimental results show that the proposed model can more effectively extract relationships between entity pairs on the DocRED dataset.

Keywords Natural language processing,Document-Level relation extraction,DocRED,Cross Multi-head attention,Decoupled knowledge distillation,Distantly supervised data,Kullback-Leibler divergence

1 引言

文档级关系抽取(Document-level Relation Extraction,DocRE)是自然语言处理(Natural Language Processing,NLP)领域的重要研究方向。其目标是从无结构或半结构的自然语言文档中提取实体之间的语义关系,然后预测文档中所有实体对的关系标签,从而形成结构化的关系三元组(头实体,关系,尾实体),这些提取出来的关系三元组可以广泛应用于智能问答[1]、知识图谱构建[2]和信息提取[3]等多个下游任务。

到稿日期:2024-06-05 返修日期:2024-11-16
基金项目:湖北省自然科学基金(E1KF291005);云南省自然科学基金(2022KZ00125)
This work was supported by the Hubei Provincial Natural Science Foundation(E1KF291005) and Yunnan Provincial Natural Science Foundation (2022KZ00125).
通信作者:肖蓉(20040363@hubu.edu.cn)

在 DocRE 任务中，利用预训练语言模型[4-5]（Pretrained Language Model，PLM）生成文本向量表示是一种普遍采用的方法。预训练语言模型通常在大规模文本数据上进行无监督学习，这使得它能够初步捕捉文本中丰富的语义信息。然而，预训练语言模型往往无法充分捕捉实体之间的复杂关系，这是因为实体通常在文本中具有特定的上下文与语境，而传统的注意力机制一般无法有效地处理实体之间的复杂交互。为了解决这一问题，本文引入了交叉多头注意力机制，该机制不仅能够使模型并行关注不同注意力头中的元素，使其在不同粒度和层级上进行信息的交流和整合，而且允许模型在计算头实体与尾实体之间的注意力时，同时考虑它们与关系之间的相关性，从而帮助模型更全面地理解关系的复杂性。

除此之外，本文还提出了一种基于 Zhao 等[6]的解耦知识蒸馏思想所实现的方法，旨在更有效地利用 DocRED[7]的远程监督数据集。具体而言，该方法通过改进传统知识蒸馏损失 KL 散度（Kullback-Leibler Divergence），设计了一种更为精细的损失：解耦 KL 散度（Decoupled Kullback-Leibler Divergence，DKL）。DKL 实现了对原始 KL 散度中目标类型知识蒸馏损失（Target Class Knowledge Distillation Loss，TCKDL）和非目标类型知识蒸馏损失（Non-Target Class Knowledge Distillation Loss，NCKDL）两大部分的解耦。其中，TCKDL 专注于传递目标类别的精确知识，而 NCKDL 则处理非目标类别之间的相对关系。这种设计使得模型在处理不同类别的知识时能够独立优化，避免了传统耦合方式带来的相互干扰。具体而言，DKL 在 DocRE 任务实际应用中的作用表现为：它通过提高学生模型的判断力和灵活性，使其在面对 DocRED 远程监督数据中的噪声时，能够更加精准地学习和应用教师模型的有效知识，从而显著提升模型的训练效果。因此，通过解除 TCK-DL 和 NCKDL 之间的不合理耦合，DKL 使得学生模型可以对教师模型传递的知识进行更细致地区分和处理，这种独立优化的机制不仅提高了模型对噪声的鲁棒性，还增强了模型的总体表现和泛化能力。

综上所述，本文的主要创新点在于提出了一种结合使用交叉多头注意力机制和解耦知识蒸馏方法的文档级关系抽取模型。其中，交叉多头注意力机制允许模型在计算注意力时考虑实体之间的复杂关系，而解耦知识蒸馏方法则通过解耦原始 KL 散度损失中的目标类别知识蒸馏损失 TCKDL 和非目标类别知识蒸馏损失 NCKDL，提高了模型对远程监督数据中噪声的鲁棒性和学习效率。这些创新显著提升了文档级关系抽取任务的性能。

2 相关工作

2.1 基于图神经网络的模型

基于图神经网络的文档级关系抽取模型利用依赖关系构建文档级图，然后使用图神经网络进行推理。具体而言，该方法一般是将实体及其在文档中的上下文建模为节点，关系建模为图上的边。其关键在于通过图的结构，将实体之间的关系进行建模，从而在全局范围内捕捉实体之间的关联。

时至今日，已有不少利用基于图神经网络的模型来解决文档级关系抽取任务的研究工作。最初的图神经网络模型由 Scarselli 等[8]于 2008 年提出，为文档级关系抽取任务奠定了基础。随后的研究不断完善和拓展了这一模型。2020 年，Nan 等[9]提出了一种基于潜在结构细化的推理方法，该方法通过动态生成文档级图结构并在其上进行多跳推理，从而显著提升了文档级关系抽取的性能。同年，Zeng 等[10]提出了一种双图谱推理方法，该方法通过实体提及节点和句子边构建文档级图，在关系抽取任务中取得了令人瞩目的效果。2021 年，名为基于特征组合的图卷积神经网络模型 FC-GCN[11]被提出，该模型避免了解析错误对模型性能的影响，并通过结合先验知识和经验，有效地初始化了对称邻接矩阵，在 DocRE 任务中取得了显著的性能提升。2023 年，Wang 等[12]提出了一种基于异构图注意力网络的模型 NEHGAN。与传统方法不同，该模型通过结合节点类型和边类型，构建单词级、提及级和实体级 3 个图，用于捕捉文档中实体之间的复杂关联，提高了关系抽取性能。总的来说，以上方法都为 DocRE 任务提供了新的思路和方法，丰富了图神经网络在该领域的应用场景。

2.2 基于 Transformer 架构的模型

基于 Transformer[13]架构的文档级关系抽取模型主要利用注意力机制，通过对文档中不同位置的实体赋予不同的注意力权重来捕捉实体之间的关联信息。

该架构在文档级关系抽取领域自 2017 年以来蓬勃发展。在这个发展过程中，Vaswani 等[14]提出的 Transformer 模型为自然语言处理领域带来了革命性的变革，其基于注意力机制的架构为后续研究奠定了基础。随着这一框架的引入，研究者开始探索如何将其应用于文档级关系抽取任务。2018 年，Verga 等[15]首次将自注意机制应用于完整摘要的生物关系抽取，为使用 Transformer 进行关系抽取任务开辟了新的方向。同年，Devlin 等[4]提出了 BERT 这种基于双向 Transformer 的预训练语言模型。此后，2019 年 Liu 等[5]提出的 RoBERTa 模型进一步优化了 BERT 的预训练过程和微调策略，为文档级关系抽取任务提供了更加精细的模型。随着时间的推移，研究者还在不断探索新的技术手段以提高模型性能。2021 年，Zhou 等[16]提出了自适应阈值和本地上下文池化的方法，结合局部和全局信息，进一步提升了文档级关系抽取的准确性。与此同时，Xu 等[17]研究了如何更好地建模实体结构和提及依赖关系，优化了文档级关系抽取任务的处理方法。2022 年，Xie 等[18]提出了 Eider 模型，通过高效的证据抽取和推理阶段融合，推动了文档级关系抽取技术的发展。同年，Tan 等[19]提出了使用轴向注意力机制增强实体对特征向量表示的方法，该方法帮助模型捕获了关系三元组之间的相互依赖关系，从而显著提升了文档级关系抽取的性能，为 DocRE 任务中的特征工程提供了一种创新的解决方案。

2.3 基于远程监督学习的模型

基于远程监督学习的文档级关系抽取模型解决了标注数据不足的问题，该方法将大规模未标记的文本数据和已有的知识库生成标注数据。这一方法的关键在于，通过远程监督生成的标注数据进行模型训练。

2009 年，Mintz 等[20]提出了一种远程监督算法，利用知识库中的信息自动生成标记数据，用于训练关系抽取模型。

2021 年，随着研究的逐步深入，Xu 等[17]在 DocRED 数据集[7]上采用一种朴素的自适应方法 SSAN-Adapt 来进行实体关系抽取。具体而言，该方法首先用关系抽取损失在远程监督数据上进行模型预训练，然后用相同的目标在人工标注数据上进行模型微调。

随着远程监督学习方法的发展，研究逐渐转向知识蒸馏相关的方法，以进一步提升模型的性能和适应性。2022 年，Tan 等[19]利用 Hinto 等[21]提出的知识蒸馏的思想，提出了一种用于文档级关系提取的半监督学习框架，通过迁移教师模型的知识来提高学生模型的性能。具体而言，先通过教师模型在 DocRED[7]远程监督数据上生成预测结果即软标签，然后结合真实标签即硬标签来联合监督训练学生模型。2023 年，Zhang 等[22]提出了一种基于推理多头自注意力单元和自蒸馏训练框架的文档级关系抽取模型，显著提升了性能。该模型通过 4 个注意力头分别建模 4 种推理模式，以覆盖更多关系三元组，并通过自蒸馏训练框架来解决训练和测试间的输入差距问题，从而更好地推理文档中的隐含关系。同年，Ma 等[23]提出的 DREEAM 模型，通过引导注意力机制聚焦于包含关系线索的证据句，结合自训练策略，显著提升了文档级关系抽取的性能。

3　任务定义

本文的文档级关系抽取任务的形式化如下。给定一个文档 D，它由一组实体$\{e_i\}_{i=1}^{n}$组成，其中 n 表示文档中的实体个数。文档级关系抽取任务的目标是预测文档中各个实体对$(e_h,e_t)_{h,t\in\{1,\cdots,n\},h\neq t}$之间的关系类型，其中 e_h 和 e_t 分别表示头实体与尾实体。关系集定义为 $R\cup\{NR\}$，其中 NR(Null Relation)代表无关系。由于一个实体可能在文档中以提及的形式出现多次，因此对于每个实体 e_i，它都可以有多个提及$\{m_j^i\}_{j=1}^{N_{e_i}}$，其中 N_{e_i} 表示某个实体 e_i 在文档中的提及次数。如果存在一对实体对(e_h,e_t)的任何一对提及存在某种关系，那么说明该实体对(e_h,e_t)也存在该种关系。而对于不存在任何关系的实体对，则将其标记为 NR。在测试阶段，模型将预测出所有实体对$(e_h,e_t)_{h,t\in\{1,\cdots,n\},h\neq t}$的关系标签。

4　模型结构

如图 1 所示，本文模型结构大体分为两个部分，即特征工程和软硬标签联合监督学习。

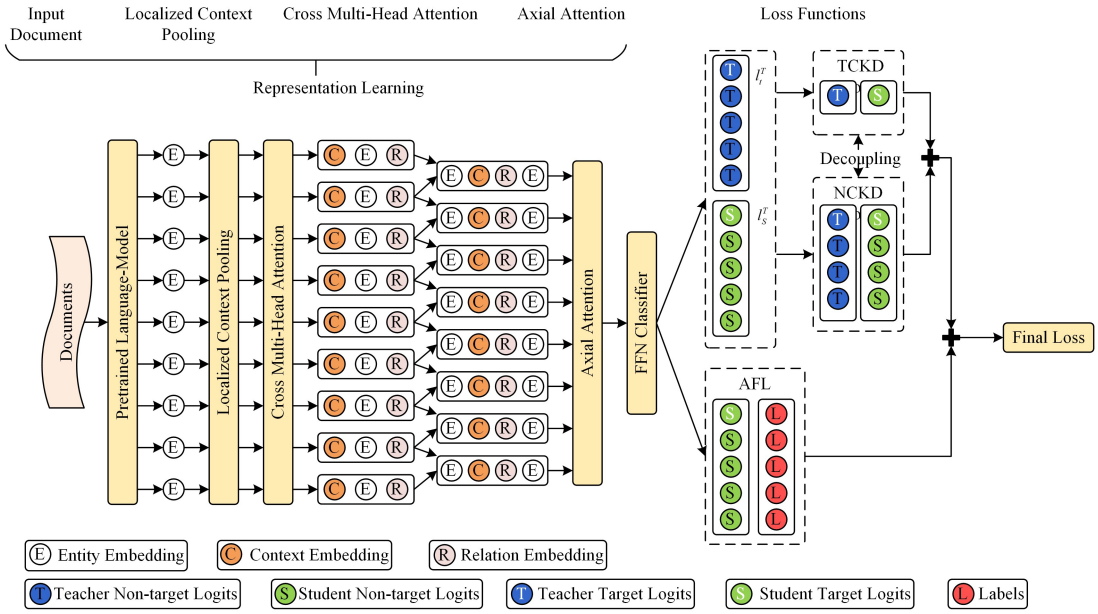

图 1　模型架构

Fig. 1　Model architecture

在特征工程部分，首先通过预训练语言模型生成最初的实体特征向量表示，其次使用本地上下文池[16]方法来获取每个实体的上下文增强特征向量表示，然后利用交叉多头注意力机制引入关系作为参考实体，从而考虑头尾实体与关系之间的相关性，最终进一步增强实体的特征向量表示，待实体与实体之间组成头尾实体对后，再使用轴向注意力机制[19]捕获实体对之间的相互依赖信息，从而增强实体对的特征向量表示。

在软硬标签联合监督学习部分，学生模型的特征工程输出结果将输入至前馈神经网络(Feed-Forward Network，FFN)层进行关系分类，然后输出关系的预测结果。学生模型的关系预测结果将用于其在知识蒸馏阶段的软硬标签联合监督学习中：硬标签监督学习采用自适应焦点损失(Adaptive Focal Loss，AFL)作为关系分类损失，以衡量学生模型的关系预测结果与真实标签之间的损失。该关系分类损失的优势在于，AFL 能够让模型更好地从长尾类中学习。而软标签监督学习则采用解耦 KL 散度 DKL 作为知识蒸馏损失，以衡量学生模型预测结果与老师模型预测结果之间的损失。该知识蒸馏损失的优势在于，DKL 分为目标类别知识蒸馏损失 TCKDL 和非目标类别知识蒸馏损失 NCKDL 两个彼此解耦的部分，其精细化的设计允许学生模型以一种更为灵活与有效的方式去利用远程监督数据进行训练，优化了知识迁移的过程。因此，知识蒸馏阶段最终输出的总损失是 AFL，TCKDL 和 NCKDL 这 3 个部分的加权和，最后模型将利用总损失进行

梯度反向传播及模型参数更新。

需要注意的是，学生模型与教师模型共享相同的网络架构，但它们在训练过程中的角色和使用方式有所不同。具体而言，首先使用 DocRED 人工标注数据训练教师模型，确保其性能达到最佳。然后，使用训练好的教师模型在 DocRED 远程监督数据上进行预测，生成软标签（老师模型的预测结果）。对于学生模型，预训练阶段在远程监督数据集上结合使用教师模型生成的软标签和数据集中自带的硬标签（真实标签）来进行联合监督学习。这一步骤便是知识蒸馏的过程。在知识蒸馏过程中，学生模型通过学习软标签获取教师模型的知识；同时通过学习硬标签，保持对真实数据的准确性。在完成远程监督数据的预训练后，再使用人工标注的数据集对学生模型进行微调，以进一步提升模型在特定任务上的性能。最后，对学生模型进行全面的性能评估，以验证其在实际应用中的效果。这种训练策略的好处在于，通过联合利用软标签和硬标签，学生模型能够更有效地学习和泛化。

4.1 特征表示学习

4.1.1 头尾实体特征表示

对于给定文档 $D=[token_p]_{p=1}^{l}$，$token_p$ 表示位置 p 上的标记，l 表示文档 D 的总标记个数，即文档长度。与先前的文档级关系抽取工作一样，本文使用一个预训练语言模型作为编码器，并通过在实体提及的起止位置插入符号“*”来表示实体提及的位置。随后，向预训练语言模型 PLM 输入文档 D，则可获得该文档中标记的嵌入矩阵表示 $\boldsymbol{H}$，具体计算式如式(1)所示：

$$\boldsymbol{H}=PLM([token_1,\cdots,token_l])=[h_1,\cdots,h_2] \tag{1}$$

对于实体 e_i，它都有提及 $\{m_j^i\}_{j=1}^{N_{e_i}}$，其中 N_{e_i} 为实体 e_i 的提及次数。通过应用 LogSumExp Pooling 技术[24]可以获得其全局的实体嵌入表示，这种表示能够捕捉实体的语义信息。具体计算式如式(2)所示：

$$h_{e_i}=\log\sum_{j=1}^{N_{e_i}}\exp(h_{m_j}) \tag{2}$$

其中，h_{e_i} 为 e_i 的聚合特征向量表示。

4.1.2 本地上下文池化增强头尾实体特征表示

以往研究已经强调上下文信息对于预测不同实体对之间关系的重要性。因此，本文采用 Zhou 等[16]提出的本地上下文池方法，以关注实体在文档中相关上下文的本地化表示。

首先，对于每个实体 e_i，通过平均池化来整合其提及所在上下文中的注意力输出。具体计算式如式(3)所示：

$$A_{e_i}=\sum_{j=1}^{N_{e_i}}a_{m_j} \tag{3}$$

其中，a_{m_j} 为提及 m_j 位置的自注意力权值。

其次，计算上下文特征向量表示。整体计算式如式(4)和式(5)所示：

$$q^{(h,t)}=\sum_{i=1}^{n}(A_{e_h}^i\cdot A_{e_t}^i) \tag{4}$$

$$c^{(h,t)}=H^{\mathrm{T}}q^{(h,t)} \tag{5}$$

其中，n 表示注意力头数，“·”表示 Hadamard 乘积，A_{e_h} 是头实体 e_h 的聚合注意力输出，A_{e_t} 则是尾实体 e_t 的聚合注意力输出，$q^{(h,t)}$ 是实体对 (e_h,e_t) 平均池化后的注意力权值，$\boldsymbol{H}$ 是整个文档的上下文向量表示。

最后，将上下文特征向量表示 $c^{(h,t)}$ 与实体特征向量表示 h_{e_i} 融合，得到上下文增强特征向量表示。具体计算式如式(6)所示：

$$z_h=\tanh(W_h h_{e_h}+W_c c^{(h,t)}) \tag{6}$$

其中，z_h 是实体对 (e_h,e_t) 中头实体 e_h 的上下文增强特征向量表示。然后，以同样的计算方式获得尾实体上下文增强特征向量表示 z_t。

需要注意的是，本节提到的上下文特征向量表示 $c^{(h,t)}$ 也称为关系实体特征向量表示，在后文中简称为关系实体，其相关数学符号的下标记为 r。

4.1.3 交叉多头注意力增强头尾实体特征表示

传统的多头注意力机制能够并行关注不同注意力头中的元素，并允许模型在不同粒度和层级上进行信息的交流和整合，从而丰富与增强对实体特征表示的学习。而对于交叉多头注意力机制，通过引入关系作为参考实体，该机制能够允许模型在计算头实体与尾实体之间的注意力时，同时考虑它们与关系之间的相关性，从而使得目标实体可以更好地进行表示与建模。因此，该机制有助于在多头注意力的基础上进一步增强实体的特征向量表示，并捕捉到实体之间的复杂关系。交叉多头注意力模型如图 2 所示。

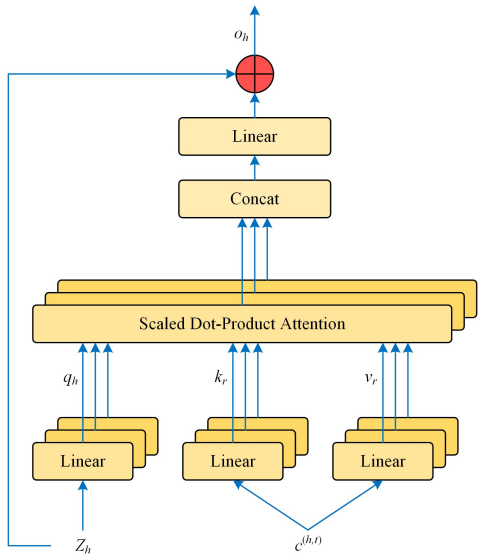

图 2 交叉多头注意力模型

Fig. 2 Cross multi-head model

具体而言，交叉多头注意力机制首先通过交叉注意力机制计算头实体与关系实体之间的注意力分数，并加权关系实体的值向量得到交叉注意力输出 $CrossAttention(q_h,k_r,v_r)$。然后，通过汇集所有交叉注意力头的输出 $Head_i$，并经线性变换后得到最终的交叉多头注意力输出 $CrossMultiHeads(z_h)$。最后，将交叉多头注意力的输出与原始输入进行残差连接得到最终的输出结果 o_h。在这里，残差连接的作用不仅可以提高模型的稳定性和鲁棒性，还能够充分融合原始输入数据特征和经交叉多头注意力机制处理后的输出数据特征，从而提供更为丰富的特征表示。整体计算式如式(7)—式(10)所示：

$$CrossAttention(q_h,k_r,v_r)=\mathrm{Softmax}\left(\frac{q_h k_r^{\mathrm{T}}}{\sqrt{d/h}}\right)v_r \tag{7}$$

$$Head_i = CrossAttention(z_h, c^{(h,t)}), i=1,2,\cdots,n \quad (8)$$

$$CrossMultiHeads(z_h) = \boldsymbol{W}_o cat(Head_1, Head_2, \cdots, Head_n) + b_o \quad (9)$$

$$o_h = z_h + CrossMultiHeads(z_h) \quad (10)$$

其中，$q_h = \boldsymbol{W}_q z_h$ 表示头实体查询特征表示；$k_r = \boldsymbol{W}_k c^{(h,t)}$ 表示关系键特征表示；$v_r = \boldsymbol{W}_v c^{(h,t)}$ 表示关系值特征表示；$\boldsymbol{W}_q$，$\boldsymbol{W}_k$，$\boldsymbol{W}_v$，$\boldsymbol{W}_o$ 表示可学习的权重矩阵；b_o 表示偏置项；n 表示设置的交叉注意力头数；cat 表示对不同的交叉注意力头计算的结果进行拼接；o_h 表示经残差连接后的最终输出。然后，通过采取同样的计算方式获得尾实体的交叉多头注意力增强特征向量表示 o_t。

4.1.4　实体对特征表示

继 Zhou 等[16]之后，Tan 等[19]使用分组双线性函数进行特征的组合。嵌入 o_h 的实体首先被分成 k 个大小相等的组，即 $o_h = [z_h^1, z_h^2, \cdots, z_h^k]$。然后对 o_t 做同样的拆分。具体的计算式如式(11)和式(12)所示：

$$y_i^{(h,t)} = \sum_{j=1}^{k} (o_h^{jT} \boldsymbol{W}_{y_i}^j o_t^j) + b_i \quad (11)$$

$$y^{(h,t)} = [y_1^{(h,t)}, y_2^{(h,t)}, \cdots, y_d^{(h,t)}] \quad (12)$$

其中，$y_i^{(h,t)}$ 表示维度 i 上的实体对特征向量表示，$\boldsymbol{W}_{y_i}^j$（$i=1,\cdots,d, j=1,\cdots,k$）表示维度 i 上的权重矩阵，b_i 表示维度 i 上的偏置项，$y^{(h,t)}$ 表示全部实体对特征矩阵表示。

需要注意的是，对于给定文档 D 中的 n 个实体，需要对 $n(n-1)$ 个实体对排列进行分类。为了编码所有实体对及其位置，Tan 等[19]选择使用一个维度为 $n \times n \times d$ 的矩阵 $\boldsymbol{G}$ 来表示文档 D 的所有实体对（d 表示隐藏层维度），并在训练和推理过程中忽略 $n \times n$ 索引的对角线。

4.1.5　轴向注意力增强实体对特征表示

Tan 等[19]提出使用两跳关注来编码每个实体对 (e_h, e_t) 表示的轴向相邻信息，而不是仅使用头尾嵌入来进行关系分类。他们认为，关注轴心因素是更为有效和符合直觉的。后续实验也确实证明了轴向注意力机制的有效性，该机制允许模型能够捕捉到与实体对共享相邻关系的语义信息最为丰富的邻居，从而提升关系抽取效果。具体来说，轴向注意力模块依次沿高度轴和宽度轴方向独立执行自注意力计算，并在每次计算后通过残差连接融合原始输入和计算结果。对于某对实体对 (e_h, e_t)，轴向注意力模块的输出的具体计算式如式(13)和式(14)所示：

$$x_{\text{wide}}^{(h,t)} = x_{\text{high}}^{(h,t)} + \sum_{p \in 1,\cdots,n} \text{softmax}_p(q_{(h,t)}^{\mathrm{T}} k_{(h,p)}) v_{(h,p)} \quad (13)$$

$$x_{\text{high}}^{(h,t)} = y^{(h,t)} + \sum_{p \in 1,\cdots,n} \text{softmax}_p(q_{(h,t)}^{\mathrm{T}} k_{(p,t)}) v_{(p,t)} \quad (14)$$

其中，$q_{(i,j)} = \boldsymbol{W}_Q y^{(i,j)}$，$k_{(i,j)} = \boldsymbol{W}_k y^{(i,j)}$，$v_{(i,j)} = \boldsymbol{W}_V y^{(i,j)}$ 表示实体对特征向量表示 y 在位置 (i,j) 上的线性投影；$\boldsymbol{W}_Q$，$\boldsymbol{W}_K$，$\boldsymbol{W}_v$ 表示可学习的权重矩阵；λ 表示轴向注意力模块的输出；$softmax_p$ 函数表示一个适用于所有可能的 $p=(i,j)$ 位置的 Softmax 函数，该函数的作用是关注两跳关系三元组的一跳邻居。

4.2　自适应焦点损失

最后有一个用于预测关系的线性层，如式(15)所示：

$$l^{(h,t)} = \boldsymbol{W}_l x_{\text{wide}}^{(h,t)} + b_l \quad (15)$$

其中，$l^{(h,t)}$ 表示对所有关系的预测输出 logit，$\boldsymbol{W}_l$ 表示将关系嵌入映射到每个类的 logit 的权重矩阵，b_l 表示偏置项。

本文采用的关系抽取损失函数是由 Tan 等[19]提出的自适应焦点损失 AFL。AFL 作为 Zhou 等[16]提出的自适应阈值损失（Adaptive Thresholding Loss，ATL）的增强版本而引入，旨在解决多标签分类任务中的长尾类问题。

与 ATL 类似，AFL 也由两部分组成，分别针对正类和负类。具体来说，在训练过程中标签空间被分为两个子集：正类子集 P_T 和负类子集 N_T。正类子集 P_T 包含实体对 (e_h, e_t) 中存在的关系，如果 (e_h, e_t) 之间不存在关系，则 P_T 为空（$P_T = \varnothing$）。另一方面，负子集 N_T 包含不属于正类 $N_T = R \backslash N_T$ 的关系类。对于正类，其概率计算式如式(16)所示：

$$\mathrm{P}(x_i | e_h, e_t) = \frac{\exp(l_{x_i}^{(h,t)})}{\exp(l_{x_i}^{(h,t)}) + \exp(l_{\text{TH}}^{(h,t)})} \quad (16)$$

其中，x_i 的 logit 分别与阈值类 TH 的 logit 排序。对于负类，其概率计算式如式(17)所示：

$$P(x_{\text{TH}} | e_h, e_t) = \frac{\exp(l_{x_{\text{TH}}}^{(h,t)})}{\sum_{x_j \in N_T \cup \{TH\}} \exp(l_{x_j}^{(h,t)})} \quad (17)$$

AFL 利用焦点损失的思想来平衡正面类的概率分布。最终的 AFL 损失函数如式(18)所示：

$$L_{\text{RE}} = \sum_{x_i \in P_T} (1 - P(x_i))^{\gamma} \log(P(x_i)) + \log(P(x_{\text{TH}})) \quad (18)$$

其中，γ 是超参数，$P(x_i | e_h, e_t)$ 简记为 $P(x_i)$，$P(x_{\text{TH}} | e_h, e_t)$ 简记为 (x_{TH})。该损失函数的设计目的是使得损失主要集中在低置信阶层，以更好地优化长尾类。

4.3　解耦知识蒸馏

先前已有研究表明，从远程监督数据中进行预训练有利于文档级关系的提取[17]。其中，知识蒸馏就是 DocRE 任务远程监督学习中的一种常用方法。一般而言，选取 KL 散度（Kullback-Leibler Div-ergence）、交叉熵（Cross-Entropy，CE）、均方误差（Mean Squared Error，MSE）作为知识蒸馏损失。然而，为了进一步利用远程监督数据，本文采用了一种新的解耦知识蒸馏方法来解决 DocRE 任务。该方法选用了解耦 KL 散度（Decoupled Kullback-Leibler Divergence，DKL）作为知识蒸馏损失，以衡量老师模型与学生模型预测结果之间的差异。DKL 的数学符号表示为 L_{DKL}，其具体推导过程如式(19)—式(25)所示。

首先，对于第 i 类的分类概率 p_i，其具体计算式如式(19)所示：

$$p_i = \frac{\exp(l_i)}{\sum_{j=1}^{C} \exp(l_j)} \quad (19)$$

其中，l 表示 logit，c 表示类的个数。

其次，根据式(19)构造两种新的概率分布，具体如下。

第一种概率分布：目标类与非目标类的二分类概率分布 $b = [p_t, p_{\backslash t}]$。该概率分布和分类监督信号高度耦合。该分布包含全部目标类概率和全部非目标类概率两个元素，具体的计算分别如式(20)和式(21)所示：

$$p_t = \frac{\sum_{k=1}^{C} \exp(l_k)}{\sum_{j=1}^{C} \exp(l_j)} (k=t) \quad (20)$$

$$p_{\backslash t} = \frac{\sum_{k=1}^{C} \exp(l_k)}{\sum_{j=1}^{C} \exp(l_j)} (k \neq t) \quad (21)$$

特别地，对于式(20)，需要注意的是：由于 L_{DKL} 此前仅被用于单标签分类问题中，因此简单地将原本的单目标类概率改为多目标类概率之和的形式，以适配 DocRE 这一多分类任务。

第二种概率分布：非目标类内部竞争的多分类概率分布 $\hat{p}=[\hat{p}_1,\hat{p}_2,\cdots,\hat{p}_{C-1}]$，即在预测样本为非目标类的前提下每个类各自的概率(总和为 1)。该概率分布和分类监督信号是不相关的，换句话说，从这个概率分布中无法得知目标类上的预测置信度。该分布包含在预测样本为非目标类的前提下每个类各自的概率，其具体计算式如式(22)所示：

$$\hat{p}_i=\frac{\exp(z_i)}{\sum_{C_{j=1,j\neq t}}\exp(z_j)} \tag{22}$$

随后，在构造完两种新的概率分布后，可以根据式(20)—式(22)，换一种形式具体描述 KL 散度，如式(23)所示：

$$\begin{aligned}L_{KL}&=p_t^T\log\left(\frac{p_t^T}{p_t^S}\right)+p_{\setminus t}^T\sum_{i=1,i\neq t}^{C}\hat{p}\left(\log\left(\frac{\hat{p}_i^T}{\hat{p}_i^S}\right)+\log\left(\frac{p_{\setminus t}^T}{p_{\setminus t}^S}\right)\right)\\&=p_t^T\log\left(\frac{p_t^T}{p_t^S}\right)+p_{\setminus t}^T\log\left(\frac{p_{\setminus t}^T}{p_{\setminus t}^S}\right)+p_{\setminus t}^T\sum_{i=1,i\neq t}^{C}\hat{p}_i^T\log\left(\frac{\hat{p}_i^T}{\hat{p}_i^S}\right)\\&=\mathrm{KL}(b^T\parallel b^S)+(1-p_t^T)\mathrm{KL}(\hat{p}^T\parallel\hat{p}^S)\end{aligned} \tag{23}$$

最后，利用 Zhao 等[6]提出的解耦 KL 散度损失的想法，来解除它们之间不合理的耦合，以灵活地调整各个部分的重要性。具体而言，将限制 L_{NCKD} 的权重 $(1-p_t^T)$ 替换为 β，并给 L_{TCKD} 设置一个权重 α。由此得到解耦 KL 散度损失 L_{DKL}。其具体计算式如式(24)和式(25)所示：

$$L_{DKL}=\alpha L_{TCKD}+\beta L_{NCKD} \tag{24}$$

$$L_{DKL}=\alpha\mathrm{KL}(b^T\parallel b^S)+\beta\mathrm{KL}(\hat{p}^T\parallel\hat{p}^S) \tag{25}$$

至此，完成了对 L_{DKL} 的推导，从而可以得到模型在远程监督数据上预训练时的总损失 L_{Total}。总损失的具体计算式如式(26)所示：

$$L_{Total}=\lambda L_{RE}+L_{DKD}=\lambda L_{RE}+\alpha L_{TCKD}+\beta L_{NCKD} \tag{26}$$

其中，α，β，λ 都是可调节的超参数。

5 实验设置与结果分析

5.1 实验数据集构建

在文档级关系抽取公共数据集 DocRED[7]上评估了本文模型。DocRED 数据集是一个众包的大型文档级关系抽取数据集，它分为人工标注 DocRED 数据集(Human-annotated DocRED)和远程监督 DocRED 数据集(Distantly Supervised DocRED)。其中，人工标注 DocRED 数据集经过了大规模的人工标注，共包含 5053 篇文档。而远程监督 DocRED 数据集则是通过将 Wikidata[25]知识库中的关系信息自动映射到维基百科文档中生成的，共包含 101873 篇文档，常作为远程监督学习方法中的远程监督数据。具体而言，首先通过在文档中识别实体并将其链接到知识库中的对应实体，然后根据知识库中的关系将这些关系标注到包含这些实体的文本段落中。然而，虽然这样能够生成大量标注数据，但是也引入了噪声，如错误标注或遗漏正确的关系等。

人工标注 DocRED 数据集与远程监督 DocRED 数据集的详细信息如表 1 所列。

表 1 数据集的详细信息

Table 1 Dataset details

数据集	数据集配置	值
Human-annotated DocRED	文档数	5053
	句子数	40276
	词语数	1002000
	实体数	132375
	关系类型数	96
	关系实例数	63427
	关系事实数	56354
Distantly Supervised DocRED	文档数	101873
	句子数	828115
	词语数	21368000
	实体数	2558350
	关系类型数	96
	关系实例数	1508320
	关系事实数	881298

5.2 实验参数设置与评价指标

采用了 PyTorch 框架实现所提方法，在武汉大学超算资源分区平台上进行本研究的实验。

实验选取两种流行的预训练语言模型作为文档编码器，分别是 BERT-base[4]和 RoBERTa-large[5]。这些预训练模型在大规模文本语料上进行了训练，能够捕捉丰富的语义信息。此外，为了进一步提升模型性能，采用了 AdamW[26]作为优化器，这是一种被广泛使用且在自然语言处理任务中表现出色的优化算法。在模型的两个阶段的训练中，主要的训练参数设置如下：在知识蒸馏阶段，学习率设置为 3×10^{-5} 或 1×10^{-5}，并在训练初期的 6%步骤中使用学习率热身[27]的策略，让模型更好地适应新的任务。此外，Transfomer 层之间的 dropout 率设置为 0.1，以防过拟合；最大梯度规范设置为 1.0，以避免梯度爆炸问题的发生。而在微调阶段，则进一步调整学习率为 1×10^{-6}，并加载已经完成知识蒸馏的学生模型作为训练起点。具体的实验参数设置如表 2 和表 3 所列。

表 2 知识蒸馏阶段实验参数设置

Table 2 Experimental parameter settings for the knowledge distillation phase

编码器	参数	值
BERT-base	TrainBatchSize	2
	TestBatchSize	2
	学习率	3×10^{-5}
	epoch	50
RoBERTa-large	TrainBatchSize	1
	TestBatchSize	1
	学习率	1×10^{-5}
	epoch	40

表 3 微调阶段实验参数设置

Table 3 Experimental parameter settings for the fine-tuning phase

编码器	参数	值
BERT-base	TrainBatchSize	1
	TestBatchSize	1
	学习率	1×10^{-6}
	epoch	10
RoBERTa-large	TrainBatchSize	1
	TestBatchSize	1
	学习率	1×10^{-6}
	epoch	20

本文在验证集和测试集上的主要评估指标是 F1 分数和 Ign_F1 分数。其中，Ign_F1 分数表示忽略标注训练数据中已知关系事实的 F1 分数。

5.3　实验结果分析

5.3.1　基线模型实验对比分析

为验证本文模型的有效性，将其与近三年基于 Transformer 的 DocRE 模型进行对比。对比实验结果如表 4 和表 5 所列。

表 4　基于 BERT-base 编码的对比模型实验结果

Table 4　Results of comparative model experiments based on BERT-base encoding

(%)

模型	Dev		Test	
	F1	Ign_F1	F1	Ign_F1
SSAN	58.95	56.68	58.41	56.06
ATLOP	61.09	59.22	61.30	59.31
AFKD*	61.70	59.74	61.65	59.59
EIDER*	62.23	60.24	62.19	60.07
DREEAM	62.55	60.51	62.49	60.03
Ours	62.72	60.78	62.58	60.48

表 5　基于 RoBERTa-large 编码的对比模型实验结果

Table 5　Results of comparative model experiments based on RoBERTa-large encoding

(%)

模型	Dev		Test	
	F1	Ign_F1	F1	Ign_F1
SSAN	62.08	60.25	61.42	59.47
ATLOP	63.18	61.32	63.40	61.39
AFKD*	63.76	61.83	63.72	61.89
EIDER*	63.91	61.89	64.02	61.85
DREEAM	64.20	62.29	64.27	62.21
Ours	64.46	62.50	64.85	62.76

本研究基于 AFKD 模型实现 DocRE 任务，并对验证集和测试集进行多次实验。其中，验证集的 F1 分数和 Ign_F1 分数是基于 5 次实验结果的平均值，而测试集的 F1 分数和 Ign_F1 分数则是基于 Codalab 平台上得到的最佳实验结果。此外，本文模型采用 5.2 节的参数设置和评价指标。其中，未带“*”号的对比模型均采用原文实验结果，而带“*”号的对比模型则是在本文实验环境下得到的实验结果。对比实验结果的具体分析如下。

1）SSAN-Adapt[17]。该模型将实体提及之间的独特依赖结构嵌入标准注意力机制中，从而得到一种新的实体编码结构注意力网络，旨在提升文档级关系抽取中的实体关系表示能力。与该模型相比，本文模型在使用 BERT-base 编码时，在整体上分别提升了 3.77 个百分点、4.1 个百分点、4.17 个百分点、4.42 个百分点；而当使用 RoBERTa-large 编码时，在整体上分别提升了 2.38 个百分点、2.25 个百分点、3.42 个百分点、3.29 个百分点。

2）ATLOP[16]。该模型通过自适应阈值处理实体对的多标签分类问题，并且使用上下文信息增强实体对表示，来提高关系抽取的准确性和鲁棒性。与该模型相比，本文模型在使用 BERT-base 编码时，在整体上分别提升了 1.63 个百分点、1.56 个百分点、1.28 个百分点、1.17 个百分点；而当使用 RoBERTa-large 编码时，在整体上分别提升了 1.28 个百分点、1.18 个百分点、1.45 个百分点、1.37 个百分点。

3）AFKD[19]。首先，通过应用轴向注意力来学习实体对之间的相互依赖关系，从而增强实体对的表示能力；然后，采用自适应焦点损失解决 DocRE 类别不平衡问题；最后，利用知识蒸馏的方法来减小人工标注数据和远程监督数据之间的差异。与该模型相比，本文模型在使用 BERT-base 编码时，在整体上分别提升了 1.02 个百分点、1.04 个百分点、0.93 个百分点、0.89 个百分点；而当使用 RoBERTa_large 编码时，在整体上分别提升了 0.7 个百分点、0.67 个百分点、1.13 个百分点、0.87 个百分点。

4）EIDER[18]。首先，训练标记证据句子；然后，将这些证据句子形成伪文档表示，与原文档共同抽取实体对关系；最后，在推理阶段融合提取的关系并获取最终的实体对关系预测值。与该模型相比，本文模型在使用 BERT-base 编码时，在整体上分别提升了 0.49 个百分点、0.54 个百分点、0.39 个百分点、0.41 个百分点；而当使用 RoBERTa-large 编码时，在整体上分别提升了 0.55 个百分点、0.61 个百分点、0.83 个百分点、0.91 个百分点。

5）DREEAM[23]。该模型采用证据句子作为监督信号，引导模型在自我训练策略下，从大量数据中自动学习提取证据句子，并赋予其较高权重，从而准确获取实体对的关系。与该模型相比，本文模型在使用 BERT-base 编码时，验证集上的 F1 分数和 Ign_F1 分数均有小幅提升，测试集上的 F1 分数也有小幅提升，Ign_F1 分数则提升了 0.45 个百分点；而当使用 RoBERTa-large 编码时，验证集上的 F1 分数和 Ign_F1 分数均有小幅度提升，测试集的 F1 分数和 Ign_F1 分数则分别提升了 0.58 个百分点和 0.55 个百分点。

由表 4 和表 5 可知，本文模型在 DocRED 数据集上实现了全局整体最优，说明了本文模型结合使用交叉多头注意力机制与解耦知识蒸馏方法在 DocRE 任务中的有效性。

本文模型性能较优的原因主要有两点：在特征表示上，本文模型增加的交叉多头注意机制不仅能够并行关注不同注意力头中的元素，允许模型在不同粒度和层级上进行信息的交流和整合，而且可以在计算头实体与尾实体之间的注意力时，同时考虑它们与关系之间的相关性，以帮助模型更全面地理解关系的复杂性，更好地捕获实体之间的相互依赖关系，从而丰富与增强模型对实体特征向量表示的学习。在远程监督学习上，远程监督 DocRED 数据集虽然是通过将知识库中的关系信息自动映射到维基百科文档中而生成的，提供了大量标注数据，但也引入了噪声，如错误标注和遗漏正确关系等。因此，老师模型在本身硬标签就存在噪声样本的远程监督 DocRED 数据集上训练后，所预测生成的软标签（老师模型的预测结果）自然也有噪声问题。对此，本文应用解耦知识蒸馏方法来解决该问题。该方法通过解耦原始 KL 散度损失中的目标类别知识蒸馏损失 TCKDL 与非目标类别知识蒸馏损失 NCKDL 之间的不合理耦合，使模型能够独立优化不同类别的知识，避免了传统耦合方式所带来的相互干扰。其中，TCKDL 专注于传递目标类别的精确知识，确保学生模型能

够准确捕捉关键类别的特性；NCKDL 则处理非目标类别之间的相对关系，确保学生模型理解类别之间的复杂交互。通过这种双重的独立优化机制，学生模型的判断力和灵活性显著提高，能够更细致且独立地评估目标类别和非目标类别的置信度，从而过滤掉老师模型预测结果中可能的错误标签，确保了学生模型能够更准确地学习与应用教师模型传递过来的有效知识。综上所述，基于 DKL 的解耦知识蒸馏方法通过减少耦合干扰、精细化知识传递、提高模型判断力和灵活调整权重，使得学生模型能够更精准地学习教师模型的软标签，从而在 DocRE 任务中取得更好的结果。

综上所述，本文模型结合使用交叉多头注意力机制和解耦知识蒸馏方法，在两种编码方法下的开发集和测试集上均表现出色，有力证明了该方法在提升文档级关系抽取任务中的有效性。

5.3.2 交叉注意力头数实验对比分析

根据表 6 中的实验数据，可以明显观察到，交叉注意力头数设置为 8 时，在验证集和测试集上获得了较好的结果。具体来说，验证集上 8 个交叉注意头的 F1 分数最高，Ign_F1 分数则稳居第二的位置；而在测试集上，F1 分数最高，Ign_F1 分数同样稳居第二的位置。同时，8 个交叉注意力头的结果在整体上表现较为稳定。选择 8 个交叉注意力头的原因是：在这个设置下，模型能够充分捕获实体之间的相互关系，同时避免过多的计算复杂性。过少的交叉注意力头数可能限制模型获取全局语义信息的能力，而过多的交叉注意力头则可能导致过渡拟合和增加计算开销。综上所述，从实验数据来看，选择 8 个交叉注意力头在验证集和测试集上都表现出较好的效果，这使得其成为较为合理的交叉注意力头数设置，可以在性能和计算开销之间取得平衡。

表 6 交叉多头注意力头数对比分析

Table 6 Comparison analysis of cross-multihead attention heads

模型	Heads	Dev		Test	
		F1/%	Ign_F1/%	F1/%	Ign_F1/%
cross multi-head attention	4	62.68	60.73	62.50	60.39
	8	**62.72**	**60.78**	**62.58**	**60.48**
	12	62.59	60.81	62.41	60.45
	16	62.70	60.68	62.54	60.48

5.3.3 各种知识蒸馏损失函数实验对比分析

在本节实验中，知识蒸馏这种远程监督学习方法被应用到 DocRE 任务中，以探究其对模型性能的影响。DocRE 任务是一项复杂的关系抽取任务，需要模型能够准确地识别文本中实体之间的关系。然而，人工标注 DocRE 数据集通常规模较小且标注成本较高，模型常常会因为训练数据量限制而造成模型训练效果不佳。知识蒸馏方法则通过利用教师模型的知识来指导学生模型的训练，在远程监督的情况下，可以帮助学生模型从教师模型的预测中学习到更多的信息，从而提升模型的性能。因此，将知识蒸馏方法应用到 DocRE 任务中，可以有效地解决数据稀缺性和标注成本高昂的问题，提高模型的性能。具体而言，采用了 KL 散度（Kullback-Leibler Divergence）、交叉熵（Cross-Entropy，CE）、均方误差（Mean Squared Error，MSE）和解耦 KL 散度（Decoupled Kullback-Leibler Divergence，DKL）4 种不同的知识蒸馏损失函数进行对比实验，以比较它们在关系抽取任务中的效果。通过这种细致的对比分析，以帮助理解不同损失函数在 Doc-RE 任务的知识蒸馏过程中的作用和效果，为未来的研究和应用提供指导。总的来说，该研究扩展了知识蒸馏损失函数的选择，为文档级关系抽取任务的研究提供了新的视角。

根据表 7 的实验结果可知，在 DocRE 任务中选用 DKL 作为知识蒸馏损失函数的模型表现最佳。分析认为，选 DKL 作为知识蒸馏损失时的蒸馏效果与模型性能优于其他 3 种常见知识蒸馏损失函数 KL，CE 和 MSE 的主要可能原因在于：首先，对于 KL，DocRE 任务中的实体关系通常是复杂多样的，而 KL 散度损失仅简单地衡量两个概率分布之间的差异，可能无法充分表达实体关系之间的复杂性；其次，对于 CE 和 MSE，这两种损失函数对样本标签中存在的噪声样本数据比较敏感，而 D-ocRE 远程监督数据中又存在规模较大的噪声样本数据，因此 CE 和 MSE 可能引入较大的损失，从而使模型训练受到干扰，最终影响模型的性能；最后，对于 DKL，它通过对原始 KL 散度损失进行解耦，实现了对目标类别知识蒸馏损失 TCKDL 和非目标类别知识蒸馏损失 NCKDL 的分别处理和权重灵活调整。因此，对于远程监督 DocRED 数据集存在的噪声问题，DKL 能够通过改变超参数来调整各部分权重，由此使得模型更加关注那些具备更高准确性与高可信度的老师模型预测的标签样本，从而减小噪声标签样本对模型训练的影响，进而提升模型性能。综上所述，DKL 能够更灵活地根据实际 Doc-RE 任务的需求和 DocRED 数据集的特点，调整优化知识蒸馏过程，从而促使教师模型与学生模型之间进行更为准确且充分的知识迁移，进而提升知识蒸馏效果，提高模型性能。因此，它被选为本文首选的知识蒸馏损失函数，以优化模型在关系抽取任务中的效果。

表 7 知识蒸馏损失函数对比分析

Table 7 Comparison analysis of knowledge distillation loss functions

(%)

模型	损失函数	Dev		Test	
		F1	Ign_F1	F1	Ign_F1
Ours	KL	61.61	59.69	61.54	59.53
	CE	61.77	59.85	61.66	59.68
	MSE	62.12	60.20	62.09	60.06
	DKL	**62.72**	**60.78**	**62.58**	**60.48**

5.3.4 解耦知识蒸馏超参数调优实验分析

对于式(26)中的解耦知识蒸馏总损失的权重超参数 α，β，λ 的调整采用逐步优化的方法，即先固定两个超参数，在给定的网格空间内，调整一数，找到其最优值后再固定这个最优值，接着调整下一个超参数，以此类推。具体过程如下：首先，在固定 $\lambda=1$，$\alpha=1$ 的情况下，调整 $\beta\in\{1-p_t^T,1,3,5,7,9\}$（其中，当 $\alpha=1$，$\beta=1-p_t^T$ 时，便是没有进行解耦的原始 KL 散度的形式），记录每次实验在验证集上的 F1 分数和 Ign_F1 分数，发现当 $\beta=7$ 时，模型性能最佳，因此将 β 固定为 7。其次，在固定 $\lambda=1$，$\beta=7$ 的情况下，调整 $\alpha\in\{0.1,0.5,1.0,1.5,2.0\}$，记录每次实验在验证集上的 F1 分数和 Ign_F1 分数，发现当 $\alpha=1.5$ 时，模型性能最佳，因此将 α 固定为 1.5。最后，

在固定 $\alpha=1.5,\beta=7$ 的情况下，调整 $\lambda\in\{0.5,1.0,1.4,1.5,2.0\}$，记录每次实验在验证集上的 F1 分数和 Ign_F1 分数，发现当 $\lambda=1.4$ 时，模型性能最佳，因此将 λ 固定为 1.4。至此，可以得到一组超参数组合 $\alpha=1.5,\beta=7,\lambda=1.4$，但因为是单独地将最优的参数组合在一起，其整体效果可能有所变化，所以在确定所有超参数的最优值后，再进行一次完整的验证，以确保组合后的参数仍然能达到最优效果，具体的实验结果如表 8 所列。

表 8　解耦知识蒸馏中超参数调优化分析

Table 8　Analysis of hyperparameter tuning in decoupled knowledge distillation

α	β	λ	Dev		Test	
			F1/%	Ign_F1/%	F1/%	Ign_F1/%
1.0	7	1.4	62.50	60.47	62.43	60.27
2.0	7	1.4	62.64	60.58	62.29	60.11
1.5	5	1.4	62.41	60.34	62.30	60.25
1.5	9	1.4	62.57	60.62	62.49	60.36
1.5	7	1.0	62.42	60.51	62.34	60.19
1.5	7	2.0	62.35	60.23	62.21	60.04
1.5	**7**	**1.4**	**62.72**	**60.78**	**62.58**	**60.48**

由表 8 可知，对于目标类别知识蒸馏损失 TCKDL 的权重超参数 α，当 α 值过低时，会导致目标类别知识传递不足，模型可能无法充分学习到目标类别的特征，从而影响分类性能；当 α 值过高时，则会导致模型过度关注目标类别，忽略其他重要特征，从而导致过拟合，模型在训练数据上表现很好，但在验证数据上表现不佳。对于非目标类别知识蒸馏损失 NCKDL 的权重超参数 β，当 β 值过低时，模型在非目标类别上的知识传递不充分，无法有效过滤噪声数据，降低了模型的鲁棒性和准确性；而当 β 值过高时，会导致模型过度关注非目标类别，忽视了目标类别的重要性，影响目标类别的分类准确性。对于关系分类损失 AFL 的权重超参数 λ，当 λ 值过低时，关系分类损失的权重较低，导致模型在训练过程中对关系分类的关注度不足。这意味着模型在训练过程中很难充分地学习关系的特征，从而降低了关系分类的准确性。而当 λ 值过高时，关系分类损失的权重较大，过高的关系分类损失权重会掩盖 TCKDL 和 NCKDL 的作用，使得模型无法有效利用教师模型传递的知识。这会削弱解耦知识蒸馏方法的优势，从而降低整体性能。

5.3.5　结合交叉多头注意力机制与解耦知识蒸馏方法在各种模型上应用的实验与分析

根据表 9 和表 10 的实验数据可以看出，整体而言，所有模型在结合使用交叉多头注意力机制和解耦知识蒸馏方法后，F1 和 Ign_F1 分数均有所提升，表明结合使用交叉多头注意力机制和解耦知识蒸馏方法在提升模型性能方面是有效的。

表 9　本实验环境下未应用本文方法的模型对比实验结果

Table 9　Comparative experimental results of models without applying the proposed method in this environment

(%)

模型	Dev		Test	
	F1	Ign_F1	F1	Ign_F1
ATLOP*	60.78	58.72	60.94	58.77
EIDER*	62.23	60.24	62.19	60.07
AFKD*	61.70	59.74	61.65	59.59

表 10　本实验环境下应用本文方法的模型对比实验结果

Table 10　Comparative experimental results of models applying the proposed method in this environment

(%)

模型	Dev		Test	
	F1	Ign_F1	F1	Ign_F1
(ATLOP+CrossHead-Attention+DKD)*	61.53	59.51	61.62	59.65
(EIDER+CrossHead-Attention+DKD)*	62.65	60.70	62.56	60.51
(AFKD+CrossHead-Attention+DKD)*	62.72	60.78	62.58	60.48

具体而言，ATLOP 和 AFKD 模型在结合使用交叉多头注意力机制和解耦知识蒸馏方法后提升较大。其中，AFKD 的提升最为显著：在验证集和测试集上，F1 分数和 Ign_F1 分数分别提升了 1.02 个百分点和 1.04 个百分点，以及 0.93 个百分点和 0.89 个百分点。这表明，AFKD 模型在结合新的方法后，能够更有效地提升特征表示和知识传递的效果。而 ATLOP 也表现出较为满意的提升：在验证集和测试集上，F1 分数和 Ign_F1 分数分别提升了 0.75 个百分点和 0.79 个百分点，以及 0.68 个百分点和 0.88 个百分点。这进一步验证了提出的结合使用交叉多头注意力机制和解耦知识蒸馏的方法在解决 DocRE 任务时的广泛适用性和有效性，为未来在类似任务中的应用提供了重要参考。ATLOP 和 AFKD 提升较大的原因主要在于：这些模型在原始特征表示和知识传递机制上存在更多的优化空间，因此当结合使用交叉多头注意力机制和解耦知识蒸馏方法后，能够进一步增强特征表示能力，使模型能够更好地捕捉复杂关系。并且，通过优化知识传递过程，提高了模型对不同类别关系的处理能力，增强了噪声过滤能力，特别是在 DocRE 远程监督数据集上的表现更为突出。

然而，EIDER 模型结合使用交叉多头注意力机制和解耦知识蒸馏方法后提升幅度相对较小。

EIDER 模型在验证集和测试集上的 F1 分数和 Ign_F1 分数分别提升了 0.42 个百分点和 0.46 个百分点，以及 0.37 个百分点和 0.44 个百分点。EIDER 模型提升不大的原因主要在于，其核心优化领域与结合使用交叉多头注意力机制和解耦知识蒸馏方法的优化重点有重叠；并且，EIDER 模型已有的优化措施使得其在信息利用和知识传递方面已经相当高效，因此引入新的方法虽然增加了一些性能，但模型的复杂性增加和边际收益递减。

EIDER 模型相比于 ATLOP 与 AFKD 模型，其提升幅度相对较小。

5.4　消融实验

本文对基线模型主要进行了两个方面的改进：首先引入交叉多头注意力机制帮助模型更全面地理解关系的复杂性和捕捉实体之间的相互依赖关系，从而增强模型对实体特征向量的学习能力；然后，还使用了解耦知识蒸馏的方法，以一种更准确且灵活的方式来学习教师模型传递过来的知识，从而更为有效地利用远程监督数据，进而提高了蒸馏效果。为了验证以上两个改进点对模型性能的影响，本文使用 BERT-base 编码进行消融实验，并比较了 4 个不同模型在验证集和测试集上的性能。由表 11 的实验结果可知，本文模型的各个

模块都对 DocR-E 起到了一定作用。具体来看：首先，基线模型在使用交叉多头注意力机制进行实体特征增强后，整体的提升分别为 0.42 个百分点、0.46 个百分点、0.44 个百分点、0.47 个百分点。可以看出，增加交叉多头注意力机制后，模型在验证集和测试集上的表现均有所提升，说明交叉多头注意力机制有助于模型更好地理解和学习实体之间的关系。其次，基线模型引入解耦知识蒸馏方法来适应远程监督数据后，整体的提升分别为 0.89 个百分点、0.77 个百分点、0.89 个百分点、0.79 个百分点。可以看出，引入解耦知识蒸馏方法后，在验证集和测试集上，模型性能都得到了显著提升，表明 DKD 方法在优化知识迁移过程和减小噪声影响方面具有显著作用。最后，当基线模型结合使用交叉多头注意力机制和解耦知识蒸馏方法后，整体的提升分别为 1.02 个百分点、1.04 个百分点、0.93 个百分点、0.89 个百分点。此时的模型在开发集和测试集上的表现均达到最佳，进一步验证了这两种方法的协同作用对模型性能的提升有显著贡献。

综上所述，单独引入交叉多头注意力机制或解耦知识蒸馏方法均能提升模型性能，而两者结合使用时，效果最为显著。这表明，交叉多头注意力机制和解耦知识蒸馏方法在文档级关系抽取任务中的结合使用能够有效增强模型对实体的特征表示，并帮助其更充分且准确地学习到远程监督学习的有效知识，从而显著提升模型的整体性能和鲁棒性。

表 11 消融实验结果分析

Table 11 Ablation experiment results analysis

(%)

模型	预训练语言模型	Dev		Test	
		F1	Ign_F1	F1	Ign_F1
AFKD	BERT-base	61.70	59.74	61.65	59.59
AFKD+Attention	BERT-base	62.12	60.20	62.09	60.06
AFKD+DKD	BERT-base	62.59	60.51	62.54	60.38
AFKD+Attention+DKD	BERT-base	62.72	60.78	62.58	60.48

结束语 文档级关系抽取在自然语言处理领域具有重要研究意义和广泛应用。相比于句子级关系抽取，文档级关系抽取需要在更大范围的上下文中识别实体间的关系，这对于理解复杂的文本语义、信息抽取以及知识图谱的构建等方面具有重要作用。在法律文书、医学报告、科学文献等领域，准确的文档级关系抽取可以极大地提高信息检索的精度和效率，促进自动化信息处理和决策支持系统的发展。然而，文档级关系抽取相较于句子级抽取，也面临着更大的挑战。因为文档级关系抽取需要处理更大数量级的文本数据，涉及更复杂的上下文语境。对此，本文提出了一种结合使用交叉多头注意力机制和解耦知识蒸馏方法的模型，旨在解决文档级关系抽取任务中的关键挑战。具体来说，交叉多头注意力机制通过并行关注不同注意力头中的元素，以及在计算注意力时同时考虑实体与关系之间的相关性，显著增强了模型对复杂关系的理解与实体特征的表示能力。解耦知识蒸馏方法则通过解除原始 KL 散度损失中目标类别知识蒸馏损失 TCKDL 和非目标类别知识蒸馏损 NCKDL 之间存在的不合理耦合，提高了知识蒸馏过程的灵活性和有效性，特别是在处理远程监督数据中的噪声时，能够更精准地进行知识迁移和学习。

在 DocRED 数据集上进行实验，本文方法在验证集和测试集上均取得了显著的性能提升，证明了其有效性和鲁棒性。因此，结合使用交叉多头注意力机制和解耦知识蒸馏方法能显著提高文档级关系抽取任务中的模型性能。这为未来的研究提供了新的思路和方法，具有重要的应用价值和理论意义。

尽管提出的基于交叉多头注意力机制和解耦知识蒸馏方法的模型在文档级关系抽取任务中取得了显著成果，但仍存在一些需要进一步改进与完善的地方。首先，对于特征表示学习方面，可以尝试引入更复杂的注意力机制或其他最新研究成果，以进一步提升全局语义信息的获取和表达能力。此外，对于远程监督学习，当前的解耦知识蒸馏方法还依赖于手工调整超参数值，希望在未来工作中能探索出一种能够自适应调整超参数的方法或类似技术，以提高模型的适应性。

参 考 文 献

[1] YANG Z, WANG Y, GAN J, et al. Design and research of intelligent question-answering(Q&A) system based on high school course knowledge graph[J]. Mobile Networks and Applications, 2021, 26(5): 1884-1890.

[2] YU H, LI H, MAO D, et al. A relationship extracti-onmethod for domain knowledge graph construction[J]. World Wide Web, 2020, 23: 735-753.

[3] XUW, CHEN K, ZHAO T. Document-level relation extraction with reconstruction[C]// Proceedings of the AAAI Conference on Artificial Inteligence. 2021: 14167-14175.

[4] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pretraining of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[5] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv: 1907.11692, 2019.

[6] ZHAO B, CUI Q, SONG R, et al. Decoupled knowledge distillation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11953-11962.

[7] YAO Y, YE D, LI P, et al. DocRED: A large-scale document-level relation extraction dataset[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 764-777.

[8] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80.

[9] NAN G, GUO Z, SEKULIĆ I, et al. Reasoning with Latent Structure Refinement for Document-Level Relation Extraction [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020: 1546-1557.

[10] ZENG S, XU R, CHANG B, et al. Double Graph Based Reasoning for Document-level Relation Extraction[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020: 1630-1640.

[11] XU J, CHEN Y, QIN Y, et al. A feature combination-based graph convolutional neural network model for relation extraction [J]. Symmetry, 2021, 13(8): 1458.

[12] WANG N，CHEN T，REN C，et al. Document-level relation extraction with multi-layer heterogeneous graph attention network [J]. Engineering Applications of Artificial Intelligence，2023，123：1-10.

[13] WOLF T，DEBUT L，SANH V，et al. Transformers：State-of-the-Art Natural Language Processing[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020：38-45.

[14] VASWANI A，SHAZEER N，PARMARN，et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017.

[15] VERGA P，STRUBELL E，MCCALLUMA. Sim-ultaneously self-attending to all mentions for full-abstract biological relation extraction[J]. arXiv：1802. 10569，2018.

[16] ZHOU W，HUANG K，MAT，et al. Document-level relation extraction with adaptive thresholding andlocalized context pooling [C]// Proceedings of the AAAI Conferenceon Artificial Intelligence. 2021：14612-14620.

[17] XU B，WANG Q，LYU Y，et al. Entity structure within and throughout：Modeling mention dependencies for document-level relation extraction[C]// Proceedings of the AAAI Conference on Artificial Intelligence，2021：14149-14157.

[18] XIE Y，SHEN J，LI S，et al. Eider：empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion[C]// Proceedings of the Association for Computational Linguistics. 2022：257-268.

[19] TAN Q，HE R，BING L，et al. Document-level relation extraction with adaptive focal loss and knowledge distillation[C]// Proceedings of Findings of the Association for Computational Linguistics. ACL，2022：1672-1681.

[20] MINTZ M，BILLS S，SNOW R，et al. Distant sup-ervision for relation extraction without labeled data[C]// Proceedings of the Joint Conference ofthe 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. ACL，2009：1003-1011.

[21] HINTO N G，VINYALS O，DEAN J. Distilling the knowledge in a neural network[J]. arXiv：1503. 02531，2015.

[22] ZHANG L，SU J，MIN Z，et al. Exploring self-distillation based relational reasoning training for document-level relation extraction[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023：13967-13975.

[23] MA Y，WANG A，OKAZAKIN. DREEAM：Guiding attention with evidence for improving doc-ument-level relation extraction [J]. arXiv：2302. 08675，2023.

[24] JIA R，WONG C，POON H. Document-Level Nary Relation Extraction with Multiscale Representation Learning[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. 2019：3693-3704.

[25] VRANDEČIĆD，KRÖTZSCHM. Wikidata：a free collaborative knowledgebase [J]. Communica-tions of the ACM，2014，57(10)：78-85.

[26] LOSHCHILOV I，HUTTERF. Decoupled weight decay regularization[J]. arXiv：1711. 05101，2017.

[27] GOYALP，DOLLÁR P，GIRSHICK R，et al. Accurate，large minibatch sgd：Training imagenet in 1 hour[J]. arXiv：1706. 02677，2017.

LIU Le，born in 2003，bachelor. His main research interests include natural language processing and information extraction.

XIAO Rong，born in 1980，Ph.D，lecturer. Her main research interests include natural language processing and information extraction.

（责任编辑：喻藜）