

基于提示调优的少样本多模态关系抽取分层网络

李维淦^{1,2}, 桂宁³, 钱育蓉^{2,4}, 陈嘉颖^{1,2}

(1. 新疆大学软件学院, 新疆 乌鲁木齐 830046; 2. 新疆大学软件工程重点实验室, 新疆 乌鲁木齐 830091;

3. 中南大学计算机学院, 湖南 长沙 410083; 4. 新疆大学计算机科学与技术学院, 新疆 乌鲁木齐 830091)

摘要: 多模态关系抽取(MRE)方法通过利用多模态信息辅助关系抽取任务。现有 MRE 模型要达到良好的关系抽取效果,通常需要在大量标注数据的基础上进行有效训练,然而在少样本情况下却表现不佳。针对上述问题,本文充分利用关系标签中丰富的语义知识和先验知识,提出一种新的提示调优分层网络。首先,设计实现基于知识注入的文本提示模块,具体来说,通过利用关系标签中隐含的实体信息,引入虚拟实体类型词构建提示模板,使模型可以感知样本中潜在的实体类型范围,并利用上下文对引入的虚拟关系答案词进行不断优化,以表达最佳的语义信息,从而提升模型在少样本情况下的性能。其次,利用实体对与关系之间的相互制约关系,设计实现实体-关系协同优化模块,以进一步提升关系抽取效果。最后,在文本编码器的每个自注意力层中,通过引入基于视觉前缀的注意力机制,将分层的多尺度视觉特征与文本信息深度融合,从而生成更加有效且鲁棒的文本表示,显著降低模型对错误的敏感性。在多模态神经关系抽取数据集(MNRE)上的实验结果表明,该模型的精确率、召回率、F1 值分别达到 84.97%、83.91%与 84.43%,较基准模型均有一定提升,特别是在少样本情况下,本文所提模型显著优于基准模型,展现出良好的关系抽取效果。

关键词: 关系抽取;少样本;提示调优;先验知识;多模态信息

中图分类号: TP391.1

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0070493

Hierarchical Network for Few-shot Multimodal Relation Extraction Based on Prompt-tuning

LI Weigan^{1,2}, GUI Ning³, QIAN Yurong^{2,4}, CHEN Jiaying^{1,2}

(1. School of Software, Xinjiang University, Urumqi 830046, Xinjiang, China;

2. Key Laboratory of Software Engineering, Xinjiang University, Urumqi 830091, Xinjiang, China;

3. School of Computer Science and Engineering, Central South University, Changsha 410083, Hunan, China;

4. School of Computer Science and Technology, Xinjiang University, Urumqi 830091, Xinjiang, China;)

【Abstract】 The Multimodal Relation Extraction (MRE) method assists in relation extraction tasks by using multimodal information. To achieve a good relation extraction performance, existing MRE models usually need to be effectively trained on a large amount of labeled data. However, they perform poorly in the case of few-shot samples. To address this issue, this study utilizes the rich semantic and prior knowledge of relation labels and proposes a new prompt-tuning hierarchical network. First, a text-prompt module based on knowledge injection is designed and implemented. Specifically, by utilizing the entity information hidden in relation labels and introducing virtual entity type words to construct prompt templates, the model can perceive the potential range of entity types in the sample and continuously optimize the introduced virtual relation answer words using context to express the best semantic information, thereby improving the performance of the model in the case of few-shot samples. Second, by utilizing the mutual constraint relationship between entity pairs and relations, an entity-relation collaborative optimization module is designed and implemented to further improve the effect of relation extraction. Finally, in each self-attention layer of the text encoder, a visual prefix-based attention mechanism is introduced to deeply integrate the layered multiscale visual features with the text information, thereby generating more effective and robust text representations and significantly reducing the sensitivity of the model to errors. The experimental results on the multimodal neural relation extraction dataset (MNRE) show that the precision, recall, and F1 score of the model reach 84.97%, 83.91%, and 84.43%, respectively, which are all improved compared with the benchmark model. Especially in the case of few-shot samples, the model proposed in this study is significantly better than the benchmark model, demonstrating good relation extraction performance.

【Key words】 relation extraction; few-shot; prompt-tuning; prior knowledge; multimodal information

基金项目: 国家自然科学基金(62266043);新疆维吾尔自治区重点研发专项(2023B01029-1, 2023B01029-2);天山创新团队项目(2023D14012);新疆维吾尔自治区杰出青年科学基金(2023D01E01);青年拔尖人才项目(2023TSYCCX0043)。

作者简介: 李维淦(CCF 学生会员),男,硕士研究生,主研方向为多模态信息抽取;桂宁、钱育蓉(CCF 高级会员、通信作者),教授、博士;陈嘉颖,副教授、博士。

收稿日期: 2024-10-16

修回日期: 2025-02-25

E-mail: qyr@xju.edu.cn

0 引言

关系抽取在信息抽取中发挥着重要作用,也是知识图谱构建的关键环节之一。关系抽取是从非结构化文本中抽取结构化知识的一种方法,旨在从给定文本数据中抽取出实体之间的特定关系。以往关于关系抽取的研究以纯文本为主,近年来,随着社交媒体上多模态数据的急剧增多以及多模态学习的快速发展,多模态关系抽取(MRE)任务被提出,该任务使用多模态信息作为附加输入辅助关系抽取。目前 MRE 的主流研究通过借助图像的视觉信息为文本提供缺失的上下文,以提升关系抽取效果。

在多模态领域的应用中,大部分数据未经标注,可用的已标注数据通常有限,并且为 MRE 任务收集和注释大量的多模态数据需要投入巨大的人力和时间成本,因此,针对少样本 MRE 的研究变得至关重要。现有关于 MRE 的研究虽然已取得一些成果,但是这些研究均依赖于大量的训练数据,而在少样本情况下表现效果不佳。此外,以往的 MRE 研究主要关注如何充分利用文本信息和视觉信息,却忽略关系标签中丰富的先验知识。在提示调优领域,现有提示调优策略通过融合原始输入和提示模板来预测[MASK]标记处的词汇,并将预测的标记词映射到相应的任务标签集合,从而使预训练语言模型(PLM)在少样本任务上具有更好的性能,其挑战在于构造适当的提示模板 $T(\cdot)$ 和答案集合 Z ,两者合称为提示。

针对上述问题,本文提出一种新的提示调优分层网络用于少样本多模态关系抽取任务,并且充分利用关系标签中丰富的先验知识和语义知识。使用提示调优进行关系抽取时,不需要在 PLM 上添加额外的分类器,因此可以消除 PLM 的预训练任务与关系抽取任务之间的差距。本文使用提示调优方法,在构造提示模板时,利用关系标签中潜在实体的平均聚合嵌入对特定标记词进行初始化,生成虚拟实体类型词,使得模型可以感知数据集的潜在实体类型范围,并将关系标签拆分词的平均聚合嵌入作为虚拟关系答案词,充分利用关系标签中的隐含语义信息,供语言模型填充[MASK]标记,并通过上下文对虚拟关系答案词进行不断优化,从而提升模型在 MRE 少样本情况下的性能。此外,本文还考虑到实体对和关系之间的相互制约关系,这是以往 MRE 研究中所未关注到的,例如,样本中标记实体类型分别为“person”和“organization”,那么所属关系类型一定不会是“per/per/peer”,相反,如果样本的关系类型是“per/per/couple”,那么该样本中一定

不会出现“person”之外的标记实体类型。使用实体对预测和关系预测协同优化,可以利用实体对和关系的相互制约,限制预测范围,从而进一步提升抽取结果的准确性。本文在多模态关系抽取数据集(MNRE)^[1]及其少样本情况下进行了实验验证。

1 相关工作

1.1 多模态关系抽取

作为信息抽取和知识图谱构建的核心步骤^[2-3],关系抽取是学者们长期研究的重要领域。早期方法包括基于模式的方法^[4-5]、基于卷积神经网络(CNN)/循环神经网络(RNN)的方法^[6]和基于图^[7]的方法。随着 PLM^[8]的发展,使用 PLM 作为骨干网络的模型来进行纯文本关系抽取已经成为主流方法^[9],一些研究表明,基于来自 Transformer 的双向编码器表示(BERT)^[8]的模型效果显著优于基于 RNN 和基于图的模型。ZENG 等^[10]利用分段最大池化和多实例学习,有效地解决了抽取过程中的噪声问题。SOARES 等^[11]提出基于 BERT 的空白匹配方法(MTB),这是一种面向关系抽取的预训练方法,用于从文本中学习关系模式。CHEN 等^[12]将提示工程应用于关系抽取任务,提出一种知识感知的协同优化方法(KnowPrompt)。

尽管上述方法在纯文本关系抽取任务上表现良好,但无法胜任社交媒体等充满多模态数据的应用场景,而且多模态关系抽取数据集也相对匮乏。基于上述情况,ZHENG 等^[1]提出多模态关系抽取任务,并制作完成该任务的首个多模态神经关系抽取数据集(MNRE);ZHENG 等^[13]提出高效图对齐模型(MEGA),使用图结构化的视觉信息来指导文本关系的提取,并且考虑结构图和语义图的相似性;CHEN 等^[14]提出一种层次化视觉前缀融合网络(HVPPeT),通过在 BERT 的每个自注意力层中使用基于视觉前缀的注意力机制,将分层多尺度的视觉特征结合起来,生成有效的文本表示,以降低错误敏感性;吴海鹏等^[15]将双向注意力引入多模态关系抽取任务;XU 等^[16]提出一种基于混合专家系统的统一视觉提示调优框架(VisualPT-MoE),用于融合不同类型的图像特征进行多模态关系抽取;CUI 等^[17]提出具有信息瓶颈的多模态表示学习(MMIB),通过基于信息论的表示学习来解决多模态关系抽取中的模态噪声和模态间隙问题。尽管上述研究在具备充足训练数据时均展现出不错的性能,但在训练数据不足的情况下未进行验证或表现不佳,而本文提出的提示调优分层网络在少样本情

况下依然能保持出色的关系抽取效果。

1.2 提示调优

随着生成式预训练 Transformer-3 (GPT-3)^[18] 被提出,提示调优在许多自然语言处理任务中已取得出色表现。一些研究通过手动构建提示的方法^[19],已验证提示调优的优势。DING 等^[20]在实体类型相关的任务中应用提示调优,通过构建面向实体的文本化器和模板进行提示学习,其中文本化器本质上是一个映射函数,能够将模型的预测结果或隐藏表示映射为自然语言中的词汇、短语或句子等表达形式。HU 等^[21]使用校准策略将外部知识纳入文本化器中。为避免繁琐的提示设计工作,学者们已经积极探索自动搜索离散提示的方法:GAO 等^[22]、SCHICK 等^[23]首先探索提示模板和答案词的自动化生成方法;SHIN 等^[24]进一步提出梯度引导搜索技术,自动生成模板和词汇表中的标签词。此外,一些学者提出连续提示的方法^[25],专注于利用可学习的连续嵌入作为提示模板。以上研究表明提示调优的有效性,然而在面对关系抽取任务时,由于需要设计适用于该任务的特定提示模板和答案集合,因此无法直接应用提示调优策略。

2 任务定义

2.1 多模态关系抽取

MRE 数据集可以表示为 $\mathcal{D} = \{\mathcal{X}, \mathcal{I}, \mathcal{R}\}$, 其中,

$\mathcal{X}$ 是文本序列集合, $\mathcal{I}$ 是图像集合, $\mathcal{R}$ 是关系标签集合。给定文本序列 $X = \{x_1, x_2, \dots, x_s, \dots, x_o, \dots, x_n\} \in \mathcal{X}$ 及其关联图像 $I \in \mathcal{I}$, MRE 任务的目标是利用 X 及 I 的信息, 预测头实体 x_s 和尾实体 x_o 之间的关系 $r \in \mathcal{R}$ 。

2.2 关系抽取的提示调优

使用 [MASK] 标记和额外的附加词构建提示模板 $T(\cdot)$, 其中 [MASK] 标记供语言模型 $\mathcal{L}$ 填充标签词。将文本序列 X 映射到 $T(\cdot)$, 获取提示输入 $X_{\text{prompt}} = T(X)$ 。因为 $\mathcal{L}$ 可以预测 [MASK] 位置的正确标记, 可以通过 $\mathcal{L}$ 编码 X_{prompt} 来获得 [MASK] 的隐藏向量, 并产生概率分布 [如式 (1) 所示], 来描述 $\mathcal{Z}$ 中的哪些标记适合替换 [MASK] 标记。

$$p(r \mid X) = p([\text{MASK}] = \mathcal{M}(r) \mid X_{\text{prompt}}) \quad (1)$$

式中: $\mathcal{M}(\cdot)$ 是将关系标签集合 $\mathcal{R}$ 连接到答案集合 $\mathcal{Z}$ 的映射。

3 模型

提示调优分层网络由分层视觉前缀提示获取模块、基于知识注入的文本提示模块、多模态特征融合模块和实体-关系协同优化模块 4 个部分组成, 如图 1 所示 (彩色效果见《计算机工程》官网 HTML 版, 下同)。视觉前缀提示获取模块以预训练的移位窗口 Transformer (Swin Transformer)^[26] 作为骨干

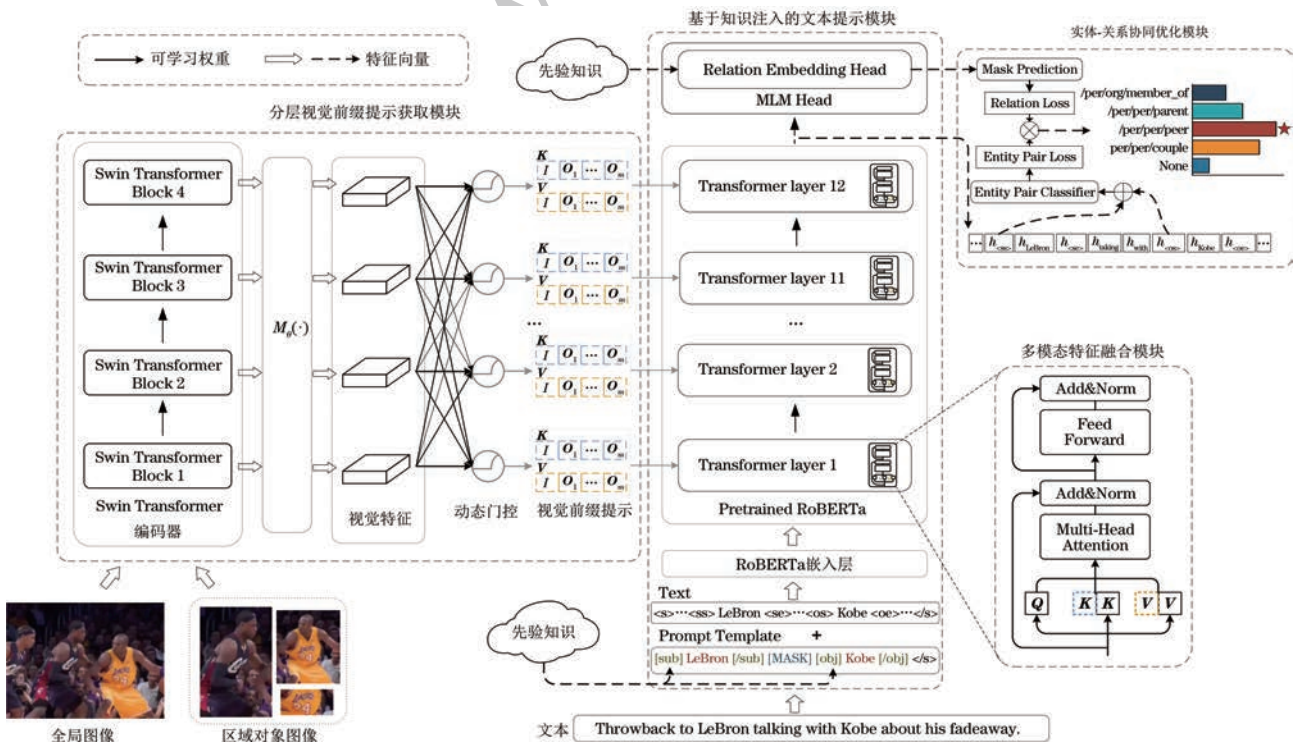


图 1 提示调优分层网络的模型结构

Fig.1 Structure of prompt-tuning hierarchical network

网络提取分层视觉特征,并通过动态门控生成视觉前缀提示。基于知识注入的文本提示模块使用关系标签中的先验知识,设计基于虚拟实体类型词和虚拟关系答案词的提示,并使用预训练的健壮型 BERT 预训练模型 (RoBERTa)^[27] 提取文本特征。多模态融合特征模块则是在 RoBERTa 的自注意力层中进行视觉前缀提示与文本表示融合。实体-关系协同优化模块采用预训练的掩码语言模型输出 (MLM Head) 层来预测关系类型,并加入实体对分类器,两者协同优化模型。

3.1 分层视觉前缀提示获取模块

CHEN 等^[14]提出的分层视觉前缀融合网络 (HVPNeT),首次在多模态领域的任务中使用金字塔分层视觉特征,表明使用分层网络结构可以获得多尺度视觉信息,使得不同大小的视觉对象可以在相应的尺度上具有适当的特征表示。

视觉接地工具^[28]可以将关于图像的自然语言查询对应到图像的正确区域。首先,使用该工具获取原始图像 I 中前 m 个显著的局部视觉对象,并将原始图像和局部对象图像重新缩放为 224×224 像素,分别作为全局图像 $\tilde{I}$ 和区域对象图像 $O = \{o_1, o_2, \dots, o_m\}$ 。

然后,使用预训练的 Swin Transformer 对给定图像进行编码,生成多尺度视觉特征列表 $\{F_1, F_2, \dots, F_c\}$,并将其映射到 $M_\theta(\cdot)$,计算方法如式(2)和式(3)所示:

$$V_c = \text{Conv}_{1 \times 1}(F_c) \quad (2)$$

$$V_i = \text{Conv}_{1 \times 1}(\text{Pool}(F_i)), i = 1, 2, \dots, c-1 \quad (3)$$

式中: i 表示骨干网络的第 i 个块; c 表示骨干网络的块数(对于 Swin Transformer, $c = 4$);Pool 表示池化操作,目的是将多尺度视觉特征变为相同尺寸的分层视觉特征。为匹配 RoBERTa 中 Transformer 的嵌入尺寸,对分层视觉特征进行 1×1 的卷积操作。

使用 CHEN 等^[14]提出的动态门控机制来聚合分层视觉特征,旨在预测一个归一化向量,以表示每个块对视觉特征的贡献。在动态门控中, $g_i^{(l)} \in [0, 1]$ 表示从视觉骨干网络第 i 块到 RoBERTa 第 l 层的路径概率,计算方法如式(4)所示:

$$g^{(l)} = \mathbb{G}^{(l)}(V) \in \mathbb{R}^c \quad (4)$$

式中: $\mathbb{G}^{(l)}(\cdot)$ 表示根据 RoBERTa 中第 l 层的门控函数。具体操作为:1)对来自第 i 块的视觉特征 V_i 进行全局平均池化操作,实现特征压缩;2)将 c 个块的视觉特征相加,并生成平均向量;3)利用多层感知机(MLP)层 W_l 进一步降低特征维度;4)生成第

l 层 Transformer 的概率向量,计算方法如式(5)所示:

$$g^{(l)} = \text{Softmax}\left(f\left(W_l\left(\frac{1}{c} \sum_{i=1}^c P(V_i)\right)\right)\right) \quad (5)$$

式中: $f(\cdot)$ 表示激活函数 Leaky ReLU; $P(\cdot)$ 表示全局平均池化操作。为匹配第 l 层的 Transformer,对分层视觉特征进行聚合,得到 $V_{\text{gated}}^{(l)}$,计算方法如式(6)所示:

$$V_{\text{gated}}^{(l)} = g^{(l)} V^{(l)} \quad (6)$$

第 l 层 Transformer 获得的最终视觉特征 $\tilde{V}_{\text{gated}}^{(l)}$,通过式(7)的连接操作获得:

$$\tilde{V}_{\text{gated}}^{(l)} = [V_{\text{gated}}^{(l,1)}; V_{\text{gated}}^{(l,o_1)}; \dots; V_{\text{gated}}^{(l,o_m)}] \quad (7)$$

使用一组线性变换 $W_k^l \in \mathbb{R}^{d \times 2 \times d}$,将第 l 层的聚合分层视觉特征投影到自注意力模块中,使其与文本表示处于相同的嵌入空间,获取视觉前缀提示 $\{\phi_k^l, \phi_v^l\}$ 的方法如式(8)所示:

$$\{\phi_k^l, \phi_v^l\} = \tilde{V}_{\text{gated}}^{(l)} W_k^l \quad (8)$$

3.2 基于知识注入的文本提示模块

3.2.1 实体知识注入

自动构建提示的方法通常需要额外的计算成本,于是根据关系抽取任务的特点,本文手动设计提示模板 $T(\cdot)$ 如式(9)所示:

$$T(\cdot) = \text{"}\cdot\text{[sub]\{SUB\[/sub][MASK][obj]\{OBJ\[/obj]."} \quad (9)$$

式中: $\{SUB\}$ 和 $\{OBJ\}$ 分别为文本序列 X 中头实体 x_s 和尾实体 x_o 预留的位置; [sub] 和 [obj] 是虚拟实体类型词标记,并且分别代表 x_s 和 x_o 的起始位置; [/sub] 和 [/obj] 分别代表 x_s 和 x_o 的结束位置。

对 MNRE 的关系标签集合进行分析,得到该数据集的潜在实体集合 $C_E = \{\text{per}, \text{loc}, \text{org}, \text{misc}\}$ 。使用集合 C_E 的聚合嵌入表示,为 $T(\cdot)$ 中 [sub] 和 [obj] 标记的权重进行初始化[如式(10)所示],并得到最终的提示模板 $\tilde{T}(\cdot)$ 。

$$e_{\text{[sub]}} = e_{\text{[obj]}} = \frac{1}{|C_E|} \sum_{i=1}^{|C_E|} e(C_E^{(i)}) \quad (10)$$

式中: $e_{\text{[sub]}}$ 和 $e_{\text{[obj]}}$ 分别表示对 [sub] 和 [obj] 进行初始化所用的平均聚合嵌入表示; $e(\cdot)$ 表示 $\mathcal{L}$ 的嵌入层。

本文使用预训练语言模型 RoBERTa 提取文本特征。首先在文本序列 $X = \{x_1, x_2, x_s, \dots, x_o, \dots, x_n\}$ 的首尾分别添加“<s>”和“</s>”标记,然后标记实体的起始位置和结束位置,在 x_s 的两侧添加“<ss>”“<se>”标记,在 x_o 的两侧添加“<os>”

“<oe>”标记,得到文本序列 $\tilde{X}$ 。将其映射到 $\tilde{T}(\cdot)$,得到 RoBERTa 最终的提示输入序列 X_{prompt} ,处理方式如式(11)所示:

$$X_{\text{prompt}} = \tilde{T}(\tilde{X}) \quad (11)$$

在实体周围分配虚拟实体类型词,并用潜在实体集合的平均聚合嵌入对其进行初始化,可以使得模型感知数据集的潜在实体类型范围,从而在训练数据不足时依旧表现出色,由于初始化的虚拟实体类型词并不是特定实体的精确类型,因此并不存在作弊行为。

3.2.2 关系知识注入

在以往的研究中,使用提示调优进行关系抽取,一般会采用自动生成的方式,将语言模型词汇表中的一个标签词与一个关系标签建立一对一映射。然而,这种方法在搜索过程中计算复杂度较高,且未充分利用关系标签中丰富的语义知识。

关系标签中存在特定实体类型信息,例如关系标签“per/per/peer”,表明该关系类型是在两个实体类型均为“per”基础上的特定关系“peer”,这是一种隐含的语义信息。本文假设在 PLM 词汇空间中存在一个虚拟关系答案词 $Z' \in \mathcal{Z}'$,它可以表示关系的这种隐含语义。使用可学习的关系嵌入构建虚拟关系答案词,来完全表示对应的关系标签 r ,并使用 MLM Head 层进行关系类型预测。因此,可以将 $p(r|X)$ 重新定义为在虚拟关系答案词集合 $\mathcal{Z}'$ 上 [MASK] 位置的概率分布。具体操作为:将关系标签 r 拆分并对缩写词进行补充,得到 r 的拆分词集 $C_r = \{r_1, r_2, \dots, r_N\}$ (其中, N 表示拆分得到的单词数量),例如,关系标签“per/per/peer”的拆分词集为 {person, person, peer}。使用加权平均函数对拆分词的嵌入表示进行权重平均化处理,以初始化关系嵌入,如式(12)所示:

$$e_{[\text{rel}]}(Z') = \frac{1}{N} \sum_{i=1}^N e(r_i) \quad (12)$$

式中: $e_{[\text{rel}]}(Z')$ 是虚拟关系答案词 Z' 的嵌入。

3.3 多模态特征融合模块

将获取的视觉前缀提示插入 RoBERTa 每个自注意力层的文本序列之前,可以使 RoBERTa 的每一层都能感知分层多尺度视觉特征,便于做出更准确和更全面的预测决策。具体来说,给定输入文本的序列,首先将预处理序列第 $l-1$ 层的上下文表示 $H^{l-1} \in \mathbb{R}^{n \times d}$ (其中, n 为 token 序列的长度, d 表示使用 RoBERTa 提取文本特征时所需的特征维度)映射到查询(query)、键(key)、值(value)向量中,如

式(13)所示:

$$Q^l = H^{l-1} W_l^q, K^l = H^{l-1} W_l^k, V^l = H^{l-1} W_l^v \quad (13)$$

式中: Q^l, K^l, V^l 分别表示文本序列在 RoBERTa 第 l 层的 query、key 和 value 向量; W 为可学习权重。第 l 层基于视觉前缀提示的注意力的计算方法如式(14)所示:

$$A_{\text{Prefix_Attention}}^l = \text{Softmax}\left(\frac{Q^l [\phi_k^l; K^l]^T}{\sqrt{d}}\right) [\phi_v^l; V^l] \quad (14)$$

在此基础上,得到样本在 RoBERTa 上的最终表示 H^L ,如式(15)所示:

$$H^L = U(X_{\text{prompt}}, \hat{V}_{\text{gated}}^{(l)}) \quad (15)$$

式中: $U(\cdot)$ 表示基于视觉前缀的注意力操作。

3.4 实体-关系协同优化模块

3.4.1 实体对预测

通过进一步分析 MNRE 数据集的关系标签集合,发现数据集中预定义的 23 种关系类型共涉及 10 种实体对类型,例如“per/per”“per/loc”等。基于实体对与关系之间的相互制约关系,本文为关系抽取任务添加实体对预测,实体对类型可以限定关系类型的检索范围,从而提升关系抽取效果。

MNRE 数据集用于实体对预测可表示为 $\mathcal{D}_{\text{ep}} = \{(X^{(i)}, I^{(i)}, E^{(i)})\}_{i=1}^M$ (其中, M 表示样本数量, $E \in \mathcal{E}$ 表示实体对标签, $\mathcal{E}$ 表示实体对标签集合)。通过“<ss>”和“<os>”标记对应的输出隐藏嵌入,使用 Softmax 函数计算 $\mathcal{E}$ 上的概率分布,计算方法如式(16)所示:

$$p(E|X) = \text{Softmax}(W[H_{<\text{ss}>}^L, H_{<\text{os}>}^L]) \quad (16)$$

通过最小化 $p(E|X)$ 在整个 $\mathcal{X}$ 上的交叉熵损失对 $\mathcal{L}_{\text{ep}}$ 和 W 进行微调, $\mathcal{L}_{\text{ep}}$ 的计算方法如式(17)所示:

$$\mathcal{L}_{\text{ep}} = - \sum_{i=1}^M E^{(i)} \ln(p(E^{(i)} | U(X_{\text{prompt}}^{(i)}, \hat{V}_{\text{gated}}))) \quad (17)$$

3.4.2 关系预测

将关系知识注入方法构建的虚拟关系答案词用于填充 [MASK] 标记,并根据上下文对其不断优化,以提升关系类型的预测效果。对于关系抽取任务, MNRE 数据集可表示为 $\mathcal{D}_{\text{re}} = \{(X^{(i)}, I^{(i)}, r^{(i)})\}_{i=1}^M$,根据提示调优方法,将概率分布 $p(r|X)$ 重新定义为在虚拟关系答案词集合 $\mathcal{Z}'$ 上 [MASK] 位置的概率分布,如式(18)所示:

$$p(r|X) = p([\text{MASK}] = Z' | X_{\text{prompt}}) \quad (18)$$

通过最小化 $p(r|X)$ 在整个 $\mathcal{X}$ 上的交叉熵损失,来对 $\mathcal{L}_{[\text{MASK}]}$ 进行微调。 $\mathcal{L}_{[\text{MASK}]}$ 的计算方法如

式(19)所示:

$$\mathcal{L}_{[\text{MASK}]} = - \sum_{i=1}^M r^{(i)} \ln(p([\text{MASK}] = \mathcal{Z}'^{(i)} | U(X_{\text{prompt}}^{(i)}, \hat{V}_{\text{gated}}))) \quad (19)$$

3.4.3 协同优化

由于实体对和关系之间存在相互制约关系,因此根据实体对预测和关系预测,确定最终的损失函数如式(20)所示:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{cpr}} + \beta \mathcal{L}_{[\text{MASK}]} \quad (20)$$

式中: α 和 β 为超参数。

4 实验与分析

4.1 数据集

本文所有实验均在 MNRE 数据集^[1]上进行。该数据集涵盖音乐、运动、社会事件等各种主题,包含 15 848 个样本、9 201 张图片和 23 种预定义的关系,其原始数据来自多模态命名实体识别数据集 Twitter-2015^[29]和 Twitter-2017^[30],以及部分从推特上爬取的数据。ZHENG 等^[1]通过人工标记的方法对原始数据进行标注,并滤除错误样本。

4.2 实验设置

实验通过 RoBERTa-base 来提取文本特征,特征维度 $d=768$,局部视觉对象数量 m 的最大值设置为 3, Swin Transformer 使用基于 ImageNet-1K 训练所得的预训练权重,并在训练时固定参数,超参数 α 和 β 均设置为 0.5。模型采用 AdamW 优化器训练目标函数,在前 10% 的梯度更新中,学习率线性预热到最大值,然后在剩余的训练中线性衰减。所有非偏置参数的权重衰减设置为 0.01。将批量大小固定为 16,初始学习率设置为 3×10^{-5} 。对模型进行 10 个 epoch 的训练,并在第 3 个 epoch 之后进行评估。本文模型在 GeForce RTX 3090 显卡上进行训练。

4.3 评价指标

本文采用精确率(P)、召回率(R)及 F1 值(F_1)作为衡量指标,计算公式如式(21)~式(23)所示:

$$P = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \quad (21)$$

$$R = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \quad (22)$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (23)$$

式中: N_{TP} 表示被正确预测为关系 r 的样本数; N_{FP} 表示被错误预测为关系 r 的样本数; N_{FN} 表示被错误预测为其他关系的样本数。

4.4 对比模型

为验证本文所提模型的有效性,将其与一些经典的单模态关系抽取模型和主流多模态关系抽取模型进行对比实验,对比模型如下:

1) PCNN^[10]: 结合分段最大池化和多实例学习,是关系抽取中最常用的基线之一。

2) MTB^[11]: 借助远程监督的思想,通过设置一个比较不同关系向量的预训练任务,以更专注于学习关系特征。

3) PURE^[31]: 提供一种基于实体类型标记的简单有效的方法来提高关系抽取任务的性能。

4) UMT^[32]: 使用多模态交互模块来获得图像感知的单词表示和单词感知的视觉表示,来指导最终的预测。

5) UMGF^[33]: 采用图像对象检测器作为图像模态和图神经网络的交互单元来实现多模态交互。

6) BERT+SG^[13]: 利用 BERT 获取文本特征,利用场景图提取视觉特征,将两者连接起来预测多模态关系。

7) BERT+SG+Att^[13]: 考虑视觉场景图与文本之间的语义相似度,并使用注意力机制来计算语义相似度。

8) MEGA^[13]: 利用视觉场景图和文本依赖图之间的结构对齐和语义对齐,找到与文本最相关的视觉关系,从而提高多模态关系抽取的性能。

9) VisualBERT^[34]: 在 Transformer 层的每个块中,通过自注意力隐式地将输入文本和与输入图像相关的区域进行对齐。

10) HVPNeT^[14]: 在每一层 Transformer 的注意力计算过程中增加视觉信息,提高多模态关系抽取的性能。

11) VisualPT-MoE^[16]: 将不同的图像表示投影到提示中以消除语义差距,并从伪图像表示中引入提示以减轻不匹配图像引起的干扰。

12) MMIB^[17]: 提出一种基于信息瓶颈原理的精细化正则化器,用于提取文本图像表示,去除冗余噪声。

4.5 对比实验

在 MNRE 数据集上,将所提模型与基准模型进行对比实验,实验结果汇总如表 1 所示,最优指标以加粗字体标示。

从表 1 的实验结果可看出,所提模型较纯文本关系抽取模型 PURE 的 F1 值提升 23.21 百分点,相比于多模态关系抽取模型 MMIB, F1 值提升 1.2 百分点。所提模型优于文本单模态模型,主要

表 1 在 MNRE 数据集上的实验结果

Table 1 The results of experiments on MNRE dataset %

模型	精确率	召回率	F1 值
PCNN	62.85	49.69	55.49
MTB	64.46	57.81	60.86
PURE	64.50	58.19	61.22
VisualBERT	57.15	59.48	58.30
UMT	62.93	63.88	63.46
UMGF	64.38	66.23	65.29
BERT+SG	62.95	62.65	62.80
BERT+SG+Att	60.97	66.56	63.64
MEGA	64.51	68.44	66.41
HVPNeT	83.64	80.78	81.85
VisualPT-MoE	82.44	81.41	81.92
MMIB	83.49	82.97	83.23
所提模型	84.97	83.91	84.43

是因为多模态模型能够利用图像信息,为关系抽取提供更为丰富的语义内容。同时,本文所提模型优于主流的 MRE 模型,这一提升得益于所提模型采用的基于先验知识注入的文本提示与实体关系协同优化,使得模型可以感知到任务中潜在实体类型的范围,并且可以根据实体对类型对于关系类型的限制,来对[MASK]的预测进一步优化。

为验证模型在训练数据不足时的可用性,对 MNRE 原始训练集使用两种方式进行采样,得到低资源训练集用于训练以模拟少样本实验情况:第 1 种方式是在 MNRE 原始训练集中对每个类的 K 个实例进行采样, K 分别取 8、16 和 32,以形成低资源训练集,实验结果如表 2 所示;第 2 种方式是在 MNRE 原始训练集中随机抽取 2%~40% 的数据,形成低资源训练集,随后在低资源情况下进行对比实验,实验结果如图 2 所示。在表 2 和图 2 中,w/o Text Prompt 表示在提取文本特征的时候不使用提示调优,而是使用常见的微调策略,具体操作为:不再对输入序列增加提示,并且将[MASK]预测改为使用 MLP 作为分类器预测关系类型。两种实验方式的结果均显示,对比主流的 MRE 模型,本文所提出的方法在低资源情况下表现出明显优势,尤其在采样数据越少时,优势更为显著。与去除文本提示模块的模型版本相比,原模型表现出明显领先,以此证明在少样本情况下基于知识注入的文本提示模块发挥着重要作用。

表 2 K -shot 训练数据的实验结果Table 2 The results of experiments with K -shot training data %

采样	模型	F1 值
8-shot	HVPNeT	18.92
	VisualPT-MoE	27.90
	所提模型 w/o Text Prompt	20.49
	所提模型	34.54
16-shot	HVPNeT	22.06
	VisualPT-MoE	33.45
	所提模型 w/o Text Prompt	32.93
	所提模型	40.06
32-shot	HVPNeT	39.98
	VisualPT-MoE	39.70
	所提模型 w/o Text Prompt	43.13
	所提模型	44.48

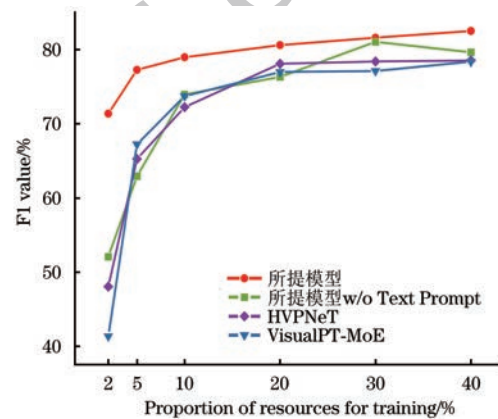


图 2 百分比随机采样的实验结果

Fig. 2 The results of experiments randomly sampled by percentage

4.6 消融实验

为验证本文所提模块的有效性,在 MNRE 数据集上进行消融实验,以 w/o Entity Pair Classifier 表示不对实体对类型进行预测,只保留[MASK]预测。消融实验结果如表 3 所示,可以看出:不使用提示调优,模型的 F1 值下降 1.25 百分点;去除实体对预测,模型的 F1 值下降 2.21 百分点。实验结果证明本文所提模块在 MRE 任务中的有效性,并且显示在训练数据充足的情况下,由于实体对预测可以限制关系类型的预测范围,其对最终关系抽取的影响更为显著。

表 3 消融实验结果

Table 3 The results of the ablation experiment %

模型	精确率	召回率	F1 值
所提模型	84.97	83.91	84.43
w/o Text Prompt	83.70	82.66	83.18
w/o Entity Pair Classifier	82.42	82.03	82.22

为进一步验证本文所提模块在低资源情况下的有效性,选取采样率为 5% 的训练数据集进行消融实验,实验结果如表 4 所示。从实验结果可以看出:不使用提示调优时,模型的 F1 值下降 1.14 百分点;去除实体对预测时,模型的 F1 值下降 14.34 百分点。实验结果表明,基于知识注入的文本提示模块在少样本情况下相比于实体-关系协同优化模块发挥着更为关键的作用。

表 4 5%训练数据的消融实验结果

Table 4 The results of the ablation experiment with 5% training data

模型	精确率	召回率	F1 值
所提模型	77.81	76.72	77.26
w/o Text Prompt	63.68	62.19	62.92
w/o Entity Pair Classifier	76.79	75.47	76.12

4.7 超参数对比实验

为验证实体-关系协同优化模块中实体预测与关系预测的有效性,本研究使用 MNRE 全部训练数据进行超参数对比实验,实验结果如图 3 所示,其中, $\alpha + \beta = 1$ 。从实验结果可以得知:当 α 和 β 均取 0.5 时,模型效果最佳;随着 α 取值的增大或减小, F1 值均有所下降,尤其是 α 取值大于 0.5 时, F1 值下降较为显著,以此证明为关系抽取任务增加实体对预测步骤是可行的,但应合理选取实体对预测与关系预测在最终目标函数中的合理权重,当实体对预测与关系预测的权重均为 0.5 时,模型能够发挥出最佳性能。

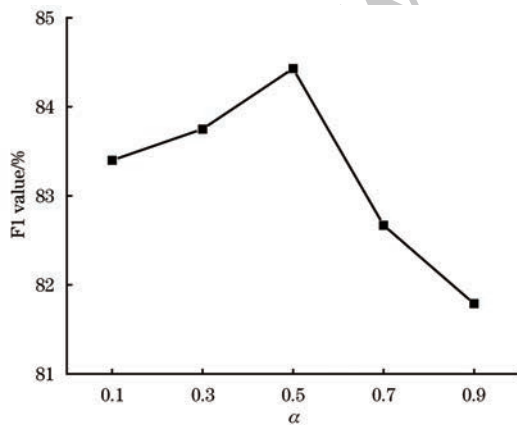


图 3 超参数对比实验结果

Fig.3 The results of the hyperparameter comparison experiment

5 结束语

本文提出一种提示调优分层网络用于多模态关系抽取任务,充分利用关系标签中丰富的语义知识和先验知识,并考虑到实体对和关系之间的相互制

约关系,进一步提高关系抽取的效果。实验结果表明,相较于以往的纯文本关系抽取方法与多模态关系抽取方法,本文所提模型在精确率、召回率和 F1 值 3 个指标上均有一定提升,尤其在少样本情况下提升更为显著。受到 HE 等^[35]的启发,在未来的工作中,将考虑利用大语言模型生成额外的附加信息,来提升关系抽取准确性,或者利用大语言模型合成训练数据,从数据增强的角度提升模型在少样本情况下的表现效果。

参考文献

- [1] ZHENG C M, WU Z W, FENG J H, et al. MNRE: a challenge multimodal dataset for neural relation extraction with visual evidence in social media posts[C]//Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Shenzhen, China: IEEE Press, 2021: 1-6.
- [2] 贺泽华, 夏方毅, 王健, 等. 原型对比学习驱动的鲁棒性关系抽取方法[J]. 重庆邮电大学学报(自然科学版), 2025, 37(1): 17-28.
- [3] HE Z H, XIA F Y, WANG J, et al. Robust relation extraction method based on prototypical contrastive learning[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2025, 37(1): 17-28. (in Chinese)
- [4] 李冬梅, 张扬, 李东远, 等. 实体关系抽取方法研究综述[J]. 计算机研究与发展, 2020, 57(7): 1424-1448.
- [5] LI D M, ZHANG Y, LI D Y, et al. Review of entity relation extraction methods[J]. Journal of Computer Research and Development, 2020, 57(7): 1424-1448. (in Chinese)
- [6] CALIFF M E, MOONEY R J. Relational learning of pattern-match rules for information extraction[C]//Proceedings of the 16th National Conference on Artificial Intelligence. Berlin, Germany: Springer, 1999: 328-334.
- [7] HUFFMAN S B. Learning information extraction patterns from examples[C]//Proceedings of International Joint Conference on Artificial Intelligence. Berlin, Germany: Springer, 1995: 246-260.
- [8] ZHANG Y H, ZHONG V, CHEN D Q, et al. Position-aware attention and supervised data improve slot filling[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 35-45.
- [9] GUO Z J, NAN G S, LU W, et al. Learning latent forests for medical relation extraction[C]//Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, 2020: 3651-3657.
- [10] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of Conference on the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S. l.]: Association for Computational Linguistics, 2019: 3498-4195.
- [11] LI J, WANG R X, ZHANG N Y, et al. Logic-guided semantic representation learning for zero-shot relation classification[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: Association for Computational Linguistics, 2020: 2967-2978.
- [12] ZENG D J, LIU K, CHEN Y B, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Proceedings of the 2015 Conference on

- Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015: 1753-1762.
- [11] SOARES L B, FITZGERALD N, LING J, et al. Matching the blanks: distributional similarity for relation learning[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 2895-2905.
- [12] CHEN X, ZHANG N Y, XIE X, et al. KnowPrompt: knowledge-aware prompt-tuning with synergistic optimization for relation extraction[C]//Proceedings of the ACM Web Conference 2022. New York, USA: ACM Press, 2022: 2823-2833.
- [13] ZHENG C M, FENG J H, FU Z, et al. Multimodal relation extraction with efficient graph alignment[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York, USA: ACM Press, 2021: 5298-5306.
- [14] CHEN X, ZHANG N Y, LI L, et al. Good visual guidance make a better extractor: hierarchical visual prefix for multimodal entity and relation extraction[C]//Proceedings of the Findings of the Association for Computational Linguistics: NAACL, 2022. Seattle, USA: Association for Computational Linguistics, 2022: 1607-1618.
- [15] 吴海鹏, 钱育蓉, 冷洪勇. 基于双向注意力机制的多模态关系抽取[J]. 计算机工程, 2024, 50(4): 160-167.
WU H P, QIAN Y R, LENG H Y. Multimodal relation extraction based on bidirectional attention mechanism[J]. Computer Engineering, 2024, 50(4): 160-167. (in Chinese)
- [16] XU B, HUANG S, DU M, et al. A unified visual prompt tuning framework with mixture-of-experts for multimodal information extraction[C]//Proceedings of International Conference on Database Systems for Advanced Applications. Berlin, Germany: Springer, 2023: 544-554.
- [17] CUI S Y, CAO J X, CONG X, et al. Enhancing multimodal entity and relation extraction with variational information bottleneck[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 1274-1285.
- [18] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Washington D.C., USA: IEEE Press, 2020: 1877-1901.
- [19] BEN-DAVID E, OVED N, REICHART R. PADA: example-based prompt learning for on-the-fly adaptation to unseen domains[C]//Proceedings of the Transactions of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2022: 414-433.
- [20] DING N, CHEN Y L, HAN X, et al. Prompt-learning for fine-grained entity typing[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 6888-6901.
- [21] HU S D, DING N, WANG H D, et al. Knowledgeable prompt-tuning: incorporating knowledge into prompt verbalizer for text classification[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 2225-2240.
- [22] GAO T Y, FISCH A, CHEN D Q. Making pre-trained language models better few-shot learners[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). [S. l.]: Association for Computational Linguistics, 2021: 3816-3830.
- [23] SCHICK T, SCHMID H, SCHÜTZE H. Automatically identifying words that can serve as labels for few-shot text classification[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: Association for Computational Linguistics, 2020: 5569-5578.
- [24] SHIN T, RAZEGHI Y, LOGAN IV R L, et al. AutoPrompt: eliciting knowledge from language models with automatically generated prompts[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S. l.]: Association for Computational Linguistics, 2020: 4222-4235.
- [25] HAMBARDZUMYAN K, KHACHATRIAN H, MAY J. WARP: word-level adversarial reprogramming[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg, USA: Association for Computational Linguistics, 2021: 4921-4933.
- [26] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE Press, 2022: 9992-10002.
- [27] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. [2024-09-01]. <https://arxiv.org/abs/1907.11692>.
- [28] YANG Z Y, GONG B Q, WANG L W, et al. A fast and accurate one-stage approach to visual grounding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE Press, 2019: 4682-4692.
- [29] LU D, NEVES L, CARVALHO V, et al. Visual attention model for name tagging in multimodal social media[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, USA: Association for Computational Linguistics, 2018: 1990-1999.
- [30] LI J, SUN A X, HAN J L, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 50-70.
- [31] ZHONG Z X, CHEN D Q. A frustratingly easy approach for entity and relation extraction[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S. l.]: Association for Computational Linguistics, 2021: 50-61.
- [32] YU J F, JIANG J, YANG L, et al. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2020: 3342-3352.
- [33] ZHANG D, WEI S, LI S, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2021: 14347-14355.
- [34] LI L H, YATSKAR M, YIN D, et al. VisualBERT: a simple and performant baseline for vision and language[EB/OL]. [2024-09-01]. <https://arxiv.org/abs/1908.03557>.
- [35] HE X Y, LI S X, ZHANG Y N, et al. The more quality information the better: hierarchical generation of multi-evidence alignment and fusion model for multimodal entity and relation extraction[J]. Information Processing & Management, 2025, 62(1): 103875.