**引用格式：**侯长波, 王睿奇, 韩吉南, 等. 跨模态自适应融合网络用于多模态关系抽取 [J]. 电子科技大学学报, 2026, 55(4): 590-599.
Hou Changbo, Wang Ruiqi, Han Jinan, et al. Feature-adaptive fusion network for multimodal relation extraction[J]. Journal of University of Electronic Science and Technology of China, 2026, 55(4): 590-599. (in Chinese)

# 跨模态自适应融合网络用于多模态关系抽取

侯长波[1]，王睿奇[1*]，韩吉南[2]，汤曜辉[1]

（1. 哈尔滨工程大学信息与通信工程学院，哈尔滨 150001；2. 北京宇航系统工程研究所，北京 100076）

**摘要：**针对现有多模态关系抽取（MRE）模型难以有效解决多模态特征融合不充分和实际应用下多模态数据不相关的问题，提出一种跨模态自适应融合网络（CAFNeT），旨在提升多模态关系抽取的性能。首先基于扩散生成模型生成新的图像数据，并与原始文本数据组成新的图文对作为网络的辅助信息；其次在多层次跨模态交互模块中设计了一种自适应多头交叉注意力机制，分别从句子层面和词汇层面实现粗细粒度的跨模态特征交互，深入地捕捉模态间的语义依赖关系；最后，在多模态自适应融合模块中设计了一种基于辅助信息的图文不相关修正机制，通过自适应调节机制动态调整原始图文对信息与辅助信息在最终表示中的贡献比例，有效抑制图文不相关数据信息的影响，同时避免关键信息丢失。在公开多模态关系抽取数据集上的实验结果表明，CAFNeT 在精确率、召回率和 F1 值上分别达到 91.30%、90.16% 与 90.72%，相较于基准模型均有一定提升，进一步的消融研究证实了所提方法的有效性。

**关键词：**关系抽取；多模态信息；交叉注意力机制；跨模态融合；自适应权重；扩散模型
**中图分类号：**TP391.1 **文献标志码：**A DOI: 10.12178/1001-0548.2025165

# Feature-adaptive fusion network for multimodal relation extraction

HOU Changbo[1], WANG Ruiqi[1*], HAN Jinan[2], and TANG Yaohui[1]

(1. School of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China;
2. Beijing Institute of Astronautical Systems Engineering, Beijing 100076, China)

**Abstract:** To address the challenges of insufficient multimodal feature fusion and irrelevant multimodal data in practical applications faced by existing multimodal relation extraction (MRE) models, this paper proposes a cross-modal adaptive fusion network (CAFNeT), aiming to enhance MRE performance. Firstly, novel image data is generated using diffusion generative models and combined with the original text data to form new image-text pairs serving as auxiliary information for the network. Secondly, a multi-level cross-modal interaction module incorporating an adaptive multi-head cross-attention mechanism is designed. This mechanism facilitates coarse- and fine-grained cross-modal feature interactions at both the sentence and word levels, enabling the deep capture of semantic dependencies between modalities. Finally, a multimodal adaptive fusion module is introduced, featuring an image-text irrelevance correction mechanism based on auxiliary information. This mechanism dynamically adjusts the contribution proportions of the original image-text pair information and the auxiliary information in the final representation through an adaptive regulation mechanism. This effectively suppresses the influence of irrelevant image-text data while avoiding the loss of critical information. Experimental results on public MRE datasets demonstrate that CAFNeT achieves precision, recall, and F1-scores of 91.30%, 90.16%, and 90.72%, respectively, showing improvements over baseline models. Further ablation studies confirm the effectiveness of the proposed method.

**Key words:** relation extraction; multimodal information; cross-attention mechanism; cross-modal fusion; adaptive weight; diffusion model

关系抽取作为信息抽取领域的核心任务之一，是自然语言处理中的关键研究方向，同时也是知识图谱构建的关键环节，其目的是从文本中识别出实体之间的语义关系。近年来，随着社交网络迅速发展，社交媒体的内容通常呈现多种形式和模态。其中不同形式、模态的内容之间呈现出互补、替代等

收稿日期：2025−08−05
基金项目：中央高校基本科研业务费专项资金（3072025ZN0801）
作者简介：侯长波，博士，副教授，主要从事多模态信号智能处理方面的研究。
*通信作者 E-mail: 2019081022@hrbeu.edu.cn

关系，传统关系抽取方法主要依赖文本特征，在处理缺乏文本上下文的社交媒体上的多模态数据时，仅靠纯文本信息难以确定实体间的关系。因此，多模态关系抽取（multimodal relation extraction, MRE）任务被提出，旨在利用社交媒体中与文本相关的图片作为附加信息辅助关系抽取。

MRE 通过融合不同模态的信息，能够有效缓解实体关系歧义的问题。然而，文本数据和图像数据在表达方式和信息密度上都存在较大差异，这种文本特征与视觉特征的不完全匹配，给 MRE 任务带来了极大的挑战。为解决以上问题，研究人员对该领域进行了深入探索。文献 [1] 提出一种高效图对齐模型，首先分别构建文本表示图和图像表示图，之后通过语义和结构对齐机制学习文本与图像特征，用于实体关系预测。文献 [2] 提出一种层次化视觉前缀融合网络，将视觉特征作为每一层注意力计算的前缀，并通过基于 Transformer 的双向编码器表示（bidirectional encoder representations from Transformers, BERT）融合文本与视觉特征。文献 [3] 提出一种基于混合专家系统的统一视觉提示调优框架。文献 [4] 提出了一种多模态网络框架，采用多尺度视觉特征提取模块获取视觉特征，并提出多模态特征融合模块学习多模态表示。

虽然现有关于 MRE 的研究已取得一些成果，但仍存在以下不足之处。

其一是模态间特征融合不充分。在社交媒体领域，用户提供的文本通常为非结构化的短文本，存在信息量少、词语模糊性大、上下文信息缺失等特点，仅单独使用文本的句子级特征或词级别特征难以挖掘文本的全部信息。图像在社交媒体上通常作为文本的补充信息，包含多个实体对象，其中大部分对于 MRE 任务属于不相关对象，容易导致模态间特征交互过程中引发信息冗余。现有的研究方法以整个文本和图像作为对象学习语义关联，忽略了更深层次的细粒度语义关系，同时图像的全局信息并非能够对所有文本中的实体语义关系的预测起到正向促进作用。

其二是多模态数据并非完全相关。在实际社交媒体的推文中，不同用户的图文对数据质量并不一致，不同推文的图文对数据分别呈现完全相关、部分相关和完全不相关的情况。然而现有的多模态关系抽取网络模型默认输入的图文对数据呈现某种相关关系，没有考虑到实际场景下不相关的多模态数据会对多模态学习带来歧义的现象。

为解决上述问题，本文提出一种跨模态自适应融合网络 CAFNeT 用于 MRE 任务。首先对每个原始文本数据基于扩散生成模型生成对应的高质量图像数据，并与原始文本数据组成新的图文对数据作为网络模型的第二输入源；在模态对齐阶段通过引入文本词级别与句子级别特征，并设计出一种自适应多头交叉注意力机制，与图像特征进行跨模态交互，充分挖掘和利用不同模态间的互补信息；同时设计了一种多模态自适应融合模块，动态调整原始图文对信息与辅助信息在最终表示的贡献比例，对最终的识别结果进行辅助修正，有效抑制图文不相关信息的影响，同时防止关键信息丢失。在 MNRE 数据集[5]上的全面实验评估表明，本文提出的方法显著优于对比方法。

## 1　相关工作

关系抽取作为信息抽取和知识图谱的关键步骤[6-7]，最近受到广泛关注。早期关系抽取方法包括基于规则的方法[8]、基于词典驱动的方法[9]、基于传统机器学习的方法[10-12]、基于卷积神经网络/循环神经网络的方法[13-14] 和基于 BERT 的方法[15]。

虽然上述方法在传统关系抽取任务中展现出良好的效果，但社交媒体上的文本呈现出简短、包含特殊字符、表达偏口语化和未收录的网络流行词语爆发等特点，针对这样具有高噪声的文本语料，传统基于文本信息抽取的模型面临巨大的挑战[16]。针对现有关系抽取方法在社交媒体上表现一般的问题，研究人员开始聚焦多模态领域，构建多模态架构的关系抽取模型。

文献 [17] 提出一种基于多模态检索的网络框架，其中文本和图像的检索模块分别从知识库中检索相关知识，并通过专家融合模块完成关系预测。文献 [18] 提出一种用于实体−对象和关系−图像对齐的预训练任务，能够从大量未标记的图像−字幕对中提取自监督信号，提高模型性能。文献 [19] 提出一种基于信息瓶颈原理的多模态表示学习方法，来解决模态间噪声带来的负面影响。文献 [20] 提出将文本与图像建模为交叉模态图的方法，利用基于图的信息瓶颈的原理来细化图结构。文献 [21] 提出一种基于双通道视觉语义信息协同增强的方法，用于消除无用视觉关系的影响。文献 [22] 提出一种多证据对齐融合模型，将大语言模型生成的文本证据与图像证据进行融合，提高模型性能。

尽管上述方法具有不错的性能，但模态间细粒度特征交互不充分、特征融合不够高效，导致文本

在与图像建立关联的过程中，容易受到无关对象的影响，对识别结果造成干扰；同时上述方法没有考虑到图文完全不相关的情况，导致模型在学习不相关的图文对数据的过程中，会导致模型的识别性能下降。为解决上述问题，本文通过生成辅助信息和从不同层面进行粗细粒度的跨模态特征交互来提升模型的关系抽取效果。

## 2 跨模态自适应融合网络模型

CAFNeT 由多模态特征提取模块、多层次跨模态交互模块、多模态自适应融合模块和输出层 4 部分组成，如图 1 所示。

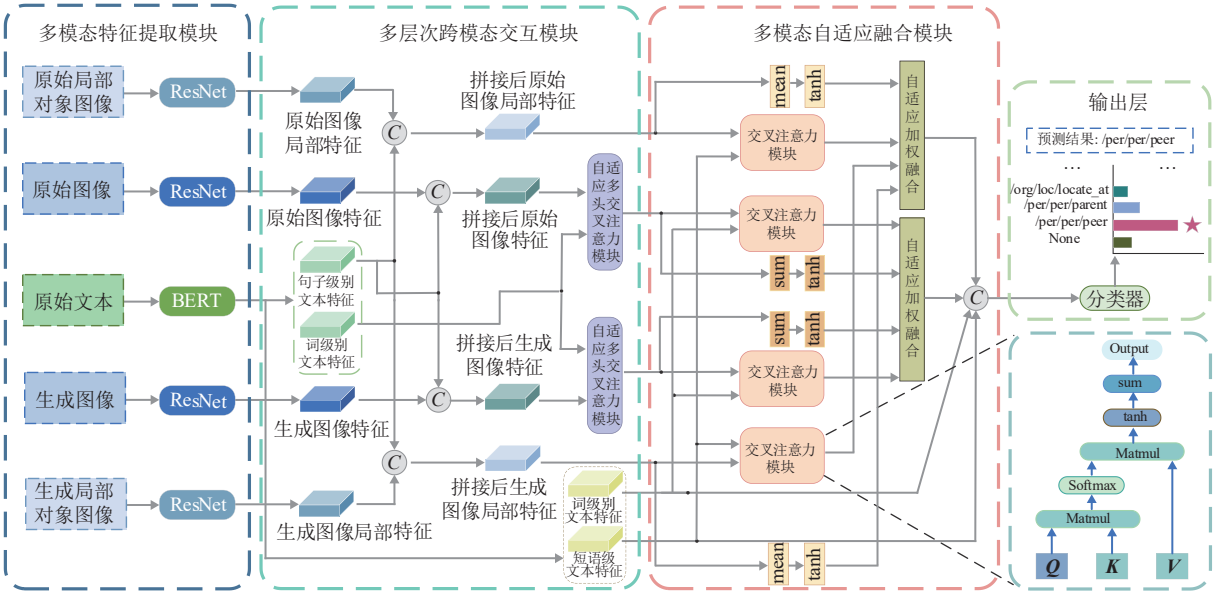



图 1 跨模态自适应融合网络模型整体结构

### 2.1 多模态特征提取模块

#### 2.1.1 辅助信息生成

文献 [23] 使用扩散生成模型，在 LAION-5B 数据集的子集[24] 上训练稳定扩散生成模型[25]，将文本看作是图像的翻译，通过反向翻译生成翻译图像数据，作为补充信息。文献 [2] 使用视觉工具[26-27] 提取具有前$m$个显著性的局部视觉对象，作为全局图像的补充。这两种方法分别通过扩散生成模型和视觉工具生成额外的图像数据，作为原本数据的补充，给本研究很大的启发。

#### 2.1.2 文本特征提取

文本作为显示信息，具有更强的概括能力，在 MRE 任务中作为主导模态，提供主要的逻辑结构与语义线索。

本文使用预训练模型 BERT[28] 对文本进行编码，得到文本特征向量，计算方法如式 (1) 和式 (2) 所示：

$$\boldsymbol{H} = \mathrm{BERT}(\boldsymbol{T}, \boldsymbol{A}) \in \mathbb{R}^{B\times L\times d} \tag{1}$$

$$\boldsymbol{S} = \frac{1}{L}\sum_{i=1}^{L} \boldsymbol{H}_{[i;i]} \in \mathbb{R}^{B\times d} \tag{2}$$

式中，输入 token 序列为$\boldsymbol{T} \in \mathbb{R}^{B\times L}$；$B$表示训练过程中每次迭代所使用的样本数；$L$表示截断的最大长度；$d$表示隐藏维度；$\boldsymbol{A} \in \{0,1\}^{B\times L}$表示 Transformer 的掩码；$\boldsymbol{H}$表示通过预训练模型 BERT 得到的词级别文本特征；$\boldsymbol{S} \in \mathbb{R}^{B\times d}$表示经过平均池化后得到的句子级文本特征。

#### 2.1.3 图像特征提取

图像作为隐式信息，语义模糊度较高，在 MRE 任务中通常作为补充模态，用于增强文中表达不清的信息。

本文基于文献 [23] 使用的扩散生成模型生成翻译图像数据$I_{\mathrm{diff}}$和文献 [2] 利用视觉工具提取具有前$m$个显著性的局部视觉对象的方法，得到原始图像$I_{\mathrm{ori}}$、原始局部对象图像$I_{\mathrm{ori}}^{\mathrm{obj}}$、生成图像$I_{\mathrm{diff}}$和生成局部对象图像$I_{\mathrm{diff}}^{\mathrm{obj}}$。

使用预训练模型 ResNet50[29] 对$I_{\mathrm{ori}}$、$I_{\mathrm{ori}}^{\mathrm{obj}}$、$I_{\mathrm{diff}}$和$I_{\mathrm{diff}}^{\mathrm{obj}}$进行编码，得到图像特征向量。计算方法如式 (3) 所示：

$$\begin{cases} \boldsymbol{v}_{\mathrm{ori}} = f_{\mathrm{ResNet}}(I_{\mathrm{ori}}) \\ \boldsymbol{v}_{\mathrm{ori}}^{\mathrm{obj}} = f_{\mathrm{ResNet}}(I_{\mathrm{ori}}^{\mathrm{obj}}) \\ \boldsymbol{v}_{\mathrm{diff}} = f_{\mathrm{ResNet}}(I_{\mathrm{diff}}) \\ \boldsymbol{v}_{\mathrm{diff}}^{\mathrm{obj}} = f_{\mathrm{ResNet}}(I_{\mathrm{diff}}^{\mathrm{obj}}) \end{cases} \tag{3}$$

式中，$\boldsymbol{v}_{\mathrm{ori}}$、$\boldsymbol{v}_{\mathrm{ori}}^{\mathrm{obj}}$、$\boldsymbol{v}_{\mathrm{diff}}$和$\boldsymbol{v}_{\mathrm{diff}}^{\mathrm{obj}}$分别表示$I_{\mathrm{ori}}$、$I_{\mathrm{ori}}^{\mathrm{obj}}$、$I_{\mathrm{diff}}$和$I_{\mathrm{diff}}^{\mathrm{obj}}$经过预训练模型 ResNet50[29] 编码后的图

像特征。

### 2.2　多层次跨模态交互模块

针对当前模型模态间特征交互不充分的问题，本文设计了一种多层次跨模态交互模块，将句子级文本特征与图像特征进行交互实现粗粒度对齐，词级别文本特征与拼接后的原始图像和拼接后的生成图像进行交互实现细粒度对齐，充分挖掘和利用不同模态间的交互信息。

粗粒度对齐和细粒度对齐在多层次跨模态交互模块中分层协作。粗粒度对齐关注整体语义关联，通过句子级特征与图像特征的拼接，提供模态间全局上下文的对齐；细粒度对齐捕捉局部细节依赖，从词级别和图像局部对象层面，深入挖掘词汇与视觉对象间的语义对应，聚焦于与文本语义高度相关的视觉区域，抑制无关特征。两者通过自适应机制分层协作，粗粒度对齐为细粒度对齐提供上下文约束，减少冗余搜索空间，而细粒度对齐通过注意力权重动态聚焦关键信息，避免无关特征干扰，从而实现高效协同。

#### 2.2.1　粗粒度对齐

通过将图像特征$\boldsymbol{I}$和句子级文本特征$\boldsymbol{S}$在特征层面进行拼接来构建一个多模态表示，实现图像特征$\boldsymbol{I}$与句子级文本特征$\boldsymbol{S}$粗粒度对齐。粗粒度对齐提供了模态间全局语义的初步融合，为后续细粒度交互奠定基础。

将$\boldsymbol{I}$进行线性映射得到$\boldsymbol{I}'$，在降低维度的同时将$\boldsymbol{I}$映射到一个语义空间，为后续更容易与$\boldsymbol{S}$进行对齐与融合，之后将$\boldsymbol{S}$进行扩展得到$\boldsymbol{S}'$，使其与图像每个位置进行对齐，为后续的拼接提供结构上的一致性，将$\boldsymbol{I}'$与$\boldsymbol{S}'$进行拼接后的结果经过多层感知机（multi-layer perceptron, MLP），进一步提纯特征：

$$\boldsymbol{I}' = \text{Linear}(\boldsymbol{I}) \tag{4}$$

$$\boldsymbol{S}' = \text{Repeat}(\boldsymbol{S}, \boldsymbol{K}) \tag{5}$$

$$\boldsymbol{O} = \text{MLP}(C(\boldsymbol{I}', \boldsymbol{S}')) \tag{6}$$

式中，Linear表示线性映射；Repeat表示扩展操作；$\boldsymbol{K}$表示扩展的大小；$C$表示拼接操作；$\boldsymbol{O}$表示拼接融合后的特征。

#### 2.2.2　细粒度对齐

现有的研究方法仅关注文本和图像的整体关联，忽略了更深层次的细粒度语义关系，因此本文设计了一种自适应多头交叉注意力模块（adaptive multi-head cross-attention module, AMCM），如图 2 所示。通过将词级别文本特征$\boldsymbol{H}$经过MLP提取文本词级语义得到$\boldsymbol{H}'$，之后分别与拼接融合后的原始图像特征$\boldsymbol{O}_{\text{ori}}$和拼接融合后的生成图像特征$\boldsymbol{O}_{\text{diff}}$进行深层交互融合。

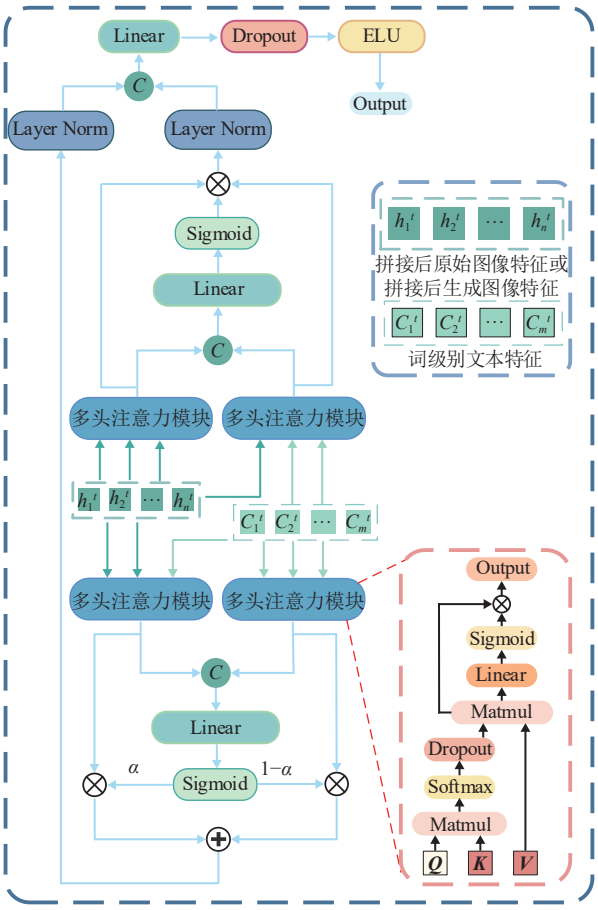



图 2　自适应多头交叉注意力模块结构图

如图 2 所示，首先将拼接后的图像特征$\boldsymbol{O}' \in \{\boldsymbol{O}_{\text{ori}}, \boldsymbol{O}_{\text{diff}}\}$与$\boldsymbol{H}'$进行线性投影，得到相应模态的查询（query, $\boldsymbol{Q}$）、键（key, $\boldsymbol{K}$）和值（value, $\boldsymbol{V}$），计算过程如式 (7) 和式 (8) 所示：

$$\begin{cases} \boldsymbol{Q}_{O'} = \boldsymbol{W}_{O'}^{Q} \cdot \boldsymbol{O}' \\ \boldsymbol{K}_{O'} = \boldsymbol{W}_{O'}^{K} \cdot \boldsymbol{O}' \\ \boldsymbol{V}_{O'} = \boldsymbol{W}_{O'}^{V} \cdot \boldsymbol{O}' \end{cases} \tag{7}$$

$$\begin{cases} \boldsymbol{Q}_{H'} = \boldsymbol{W}_{H'}^{Q} \cdot \boldsymbol{H}' \\ \boldsymbol{K}_{H'} = \boldsymbol{W}_{H'}^{K} \cdot \boldsymbol{H}' \\ \boldsymbol{V}_{H'} = \boldsymbol{W}_{H'}^{V} \cdot \boldsymbol{H}' \end{cases} \tag{8}$$

式中，$\boldsymbol{W}_{O'}^{Q}, \boldsymbol{W}_{O'}^{K}, \boldsymbol{W}_{O'}^{V}, \boldsymbol{W}_{H'}^{Q}, \boldsymbol{W}_{H'}^{K}, \boldsymbol{W}_{H'}^{V} \in \mathbb{R}^{d \times d}$，表示权重矩阵。

之后将$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分为$h$个注意力头，每个头维度为$d_h = d/h$，使模型从多个子空间学习不同的注意力，注意力计算过程如式 (9) 所示：

$$\boldsymbol{A}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_h}}\right)\boldsymbol{V} \tag{9}$$

式中，$\boldsymbol{A}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$表示计算出的注意力。

对每个头的注意力进行逐点加权，动态保留重要特征，抑制噪声，如式 (10) 和式 (11) 所示：

$$\boldsymbol{G}=\sigma(\text{Linear}(\boldsymbol{A}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}))) \tag{10}$$

$$\boldsymbol{A}'(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\boldsymbol{A}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})\odot\boldsymbol{G} \tag{11}$$

式中，$\sigma$表示Sigmoid函数；$\boldsymbol{G}$表示值域在 [0,1] 之间的权重因子；$\odot$表示逐元素乘法；$\boldsymbol{A}'(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$表示经过每个头的最终输出。

将每个头的输出拼接后形成多头注意力，如式 (12) 和式 (13) 所示：

$$\boldsymbol{n}_i=\boldsymbol{A}'(\boldsymbol{Q}_i,\boldsymbol{K}_i,\boldsymbol{V}_i) \tag{12}$$

$$\boldsymbol{M}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=C(n_1,n_2,\cdots,n_h)\boldsymbol{W}^o \tag{13}$$

式中，$\boldsymbol{M}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$表示每个头的输出拼接后形成的多头注意力。

分别从图像内部、图像关注文本、文本关注图像和文本内部 4 种角度计算多头注意力，得到 4 种注意力：

$$\begin{cases}\boldsymbol{o}_1=\boldsymbol{M}(\boldsymbol{Q}_{O'},\boldsymbol{K}_{O'},\boldsymbol{V}_{O'})\\ \boldsymbol{o}_2=\boldsymbol{M}(\boldsymbol{Q}_{O'},\boldsymbol{K}_{H'},\boldsymbol{V}_{H'})\\ \boldsymbol{o}_3=\boldsymbol{M}(\boldsymbol{Q}_{H'},\boldsymbol{K}_{O'},\boldsymbol{V}_{O'})\\ \boldsymbol{o}_4=\boldsymbol{M}(\boldsymbol{Q}_{H'},\boldsymbol{K}_{H'},\boldsymbol{V}_{H'})\end{cases} \tag{14}$$

式中，$\boldsymbol{o}_1$表示图像自注意力；$\boldsymbol{o}_2$表示图像−文本交叉注意力；$\boldsymbol{o}_3$表示文本−图像交叉注意力；$\boldsymbol{o}_4$表示文本自注意力。

相较于直接对不同注意力来源的特征信息进行拼接，产生信息冗余，通过自适应融合机制，模型能够自动学习对不同注意力的偏好，并平滑地结合各自的输出，提升模型鲁棒性，保留重要信息的同时过滤冗余信息。计算过程如式 (15) 和式 (16) 所示：

$$\begin{cases}\alpha_{12}=\sigma(\text{Linear}([\boldsymbol{o}_1;\boldsymbol{o}_2]))\in[0,1]\\ \alpha_{34}=\sigma(\text{Linear}([\boldsymbol{o}_3;\boldsymbol{o}_4]))\in[0,1]\end{cases} \tag{15}$$

$$\begin{cases}\boldsymbol{F}_{12}=\alpha_{12}\boldsymbol{o}_1+(1-\alpha_{12})\boldsymbol{o}_2\\ \boldsymbol{F}_{34}=\alpha_{34}\boldsymbol{o}_3+(1-\alpha_{34})\boldsymbol{o}_4\end{cases} \tag{16}$$

式中，$[\boldsymbol{o}_1;\boldsymbol{o}_2]$和$[\boldsymbol{o}_3;\boldsymbol{o}_4]$表示在最后一个维度拼接；$\alpha_{12}$和$\alpha_{34}$表示加权权重；$\boldsymbol{F}_{12}$和$\boldsymbol{F}_{34}$表示加权融合后的特征。

最后将融合结果进行归一化和非线性激活，保证模型梯度稳定，增强表达能力：

$$\begin{cases}\tilde{\boldsymbol{F}}_{12}=\text{LayerNorm}(\boldsymbol{F}_{12})\\ \tilde{\boldsymbol{F}}_{34}=\text{LayerNorm}(\boldsymbol{F}_{34})\end{cases} \tag{17}$$

$$\tilde{\boldsymbol{F}}'=C(\tilde{\boldsymbol{F}}_{12},\tilde{\boldsymbol{F}}_{34}) \tag{18}$$

$$\boldsymbol{F}=\text{ELU}(\text{Dropout}(\text{Linear}(\tilde{\boldsymbol{F}}'))) \tag{19}$$

式中，LayerNorm表示归一化操作；Dropout表示Dropout层；ELU表示激活函数；$\tilde{\boldsymbol{F}}_{12}$和$\tilde{\boldsymbol{F}}_{34}$表示归一化后的特征；$\tilde{\boldsymbol{F}}'$表示$\tilde{\boldsymbol{F}}_{12}$与$\tilde{\boldsymbol{F}}_{34}$进行拼接后的特征；$\boldsymbol{F}$表示经过多层次跨模态交互后的输出特征。

### 2.3 多模态自适应融合模块

针对当前模型没有考虑到图文完全不相关的情况。本文设计了一种多模态自适应融合模块（multimodal adaptive fusion, MAF），通过动态调整原始图文对信息与辅助信息的加权比例，强化模态间关键特征的交互作用，有效抑制图文不相关数据信息的影响，防止信息丢失，提升整体性能。

首先将图像特征与文本特征进行线性投影，得到相应模态的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，之后将图像模态的$\boldsymbol{Q}$与文本模态的$\boldsymbol{K}$、$\boldsymbol{V}$放入交叉注意力模块中，计算注意力，计算过程如式 (20) 所示：

$$\begin{cases}\boldsymbol{f}_1=\sum(\tanh(\boldsymbol{A}(\boldsymbol{Q}_1,\boldsymbol{K}_1,\boldsymbol{V}_1)))\\ \boldsymbol{f}_2=\sum(\tanh(\boldsymbol{A}(\boldsymbol{Q}_2,\boldsymbol{K}_2,\boldsymbol{V}_2)))\\ \boldsymbol{f}_3=\sum(\tanh(\boldsymbol{A}(\boldsymbol{Q}_3,\boldsymbol{K}_3,\boldsymbol{V}_3)))\\ \boldsymbol{f}_4=\sum(\tanh(\boldsymbol{A}(\boldsymbol{Q}_4,\boldsymbol{K}_4,\boldsymbol{V}_4)))\end{cases} \tag{20}$$

式中，$\boldsymbol{Q}_1$表示拼接后原始图像局部特征的线性投影；$\boldsymbol{Q}_2$表示交互后的原始图像特征的线性投影；$\boldsymbol{Q}_3$表示交互后的生成图像特征的线性投影；$\boldsymbol{Q}_4$表示拼接后生成图像局部特征的线性投影；$\boldsymbol{K}_1$、$\boldsymbol{V}_1$、$\boldsymbol{K}_4$、$\boldsymbol{V}_4$表示短语级文本特征的线性投影；$\boldsymbol{K}_2$、$\boldsymbol{V}_2$、$\boldsymbol{K}_3$、$\boldsymbol{V}_3$表示词级别文本特征的线性投影；$\boldsymbol{f}_1$、$\boldsymbol{f}_2$、$\boldsymbol{f}_3$、$\boldsymbol{f}_4$表示进行交叉注意力计算后，各自获得的融合特征。

随后将融合特征与经过交叉注意力模块前的特征输入自适应加权融合模块中进行自适应融合，动态调整各模态特征的加权比例，如图 3 所示。

通过对输入特征生成自适应权重，最后得到加权融合后的特征$\boldsymbol{f}_1^{\text{final}}$和$\boldsymbol{f}_2^{\text{final}}$。

$$\begin{cases}\boldsymbol{M}=[\boldsymbol{f}_1;\boldsymbol{m}_1;\boldsymbol{f}_4;\boldsymbol{m}_4]\\ \boldsymbol{N}=[\boldsymbol{f}_2;\boldsymbol{m}_2;\boldsymbol{f}_3;\boldsymbol{m}_3]\end{cases} \tag{21}$$

$$\begin{cases}\boldsymbol{a}=\text{Softmax}(\text{MLP}(\boldsymbol{M}))\\ \boldsymbol{b}=\text{Softmax}(\text{MLP}(\boldsymbol{N}))\end{cases} \tag{22}$$

$$\begin{cases}\boldsymbol{f}_1^{\text{final}}=a_1\boldsymbol{f}_1+a_2\boldsymbol{m}_1+a_3\boldsymbol{f}_4+a_4\boldsymbol{m}_4\\ \boldsymbol{f}_2^{\text{final}}=b_1\boldsymbol{f}_2+b_2\boldsymbol{m}_2+b_3\boldsymbol{f}_3+b_4\boldsymbol{m}_3\end{cases} \tag{23}$$

式中，$\boldsymbol{m}_1$和$\boldsymbol{m}_4$分别表示拼接后原始图像局部特征和拼接后生成图像局部特征经过归一化后的特征；$\boldsymbol{m}_2$和$\boldsymbol{m}_3$分别表示原始图像和生成图像经过多层次跨模态交互后的输出特征经过归一化后的特征；$\boldsymbol{a}=[a_1,a_2,a_3,a_4]$和$\boldsymbol{b}=[b_1,b_2,b_3,b_4]$分别表示图像整体特征和图像目标特征的自适应权重；$\boldsymbol{f}_1^{\text{final}}$和$\boldsymbol{f}_2^{\text{final}}$分别表示经过自适应加权融合后得到的图像整体融合特征和图像目标融合特征。

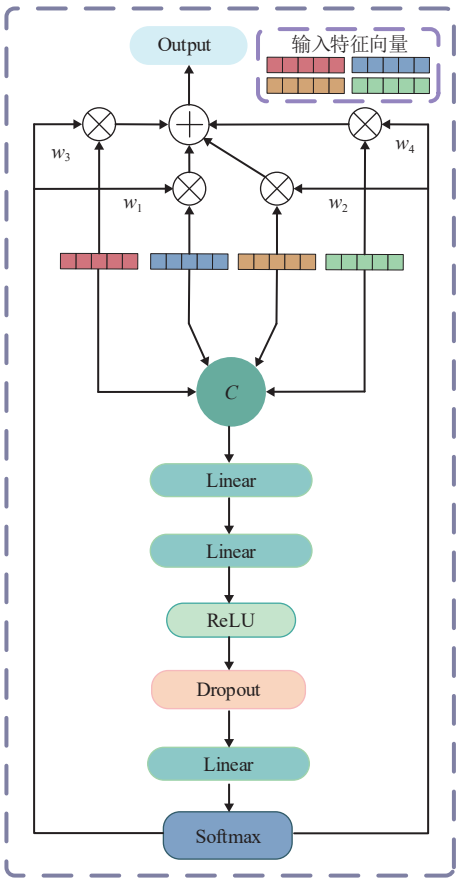



图 3　自适应加权融合模块结构图

最后对加权融合后的特征与文本特征进行拼接，得到最终的输出特征$\boldsymbol{F}_{\text{final}}$。

### 2.4　输出层

关系提取任务的目的是预测实体与实体之间存在的语义关系，如式 (24) 所示，输出层使用MLP和Softmax作为分类器来预测实体间存在的关系，并输出每个关系对应的概率：

$$p_{\text{output}}=\text{Softmax}(\text{MLP}(\boldsymbol{F}_{\text{final}})) \tag{24}$$

式中，$p_{\text{output}}\in n$表示$n$个预定义关系的分类概率。

## 3　实验与结果分析

### 3.1　数据集

本文所有实验均在 MNRE 数据集[5] 上进行，该数据集包含 15 848 个样本、9 201 张图片和 23 种预定义关系。该数据集的原始数据主要来自多模态命名实体识别中 Twitter-2015[30] 和 Twitter-2017[31] 的数据。文献 [5] 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。

### 3.2　实验设置

本实验采用 Hugging Face 库中的 BERT-base-uncased 模型默认的分词器作为文本编码器，并将最大长度设为 128 来预处理数据。采用 ResNet50[29] 作为视觉编码器，对原始图像、文献 [2] 的目标图像和扩散生成模型生成的翻译图像进行编码。模型采用 AdamW 优化器进行训练，在训练的前 10% 的梯度更新中，学习率预热达到最大值，之后在剩余的训练过程中线性衰减。所有非偏置参数的权重设置为 0.01。将批量大小固定为 16，初始学习率为 $2\times10^{-5}$。对模型进行 30 个 epoch 的训练，并在第 3 个 epoch 之后进行评估。本文模型在 GeForce RTX 4090 显卡上进行训练。

### 3.3　评价指标

本文采用精确率（precision, P）、召回率（recall, R）及 F1 值（F1 score, F1）作为评价指标。

### 3.4　基准模型

为验证所提模型的有效性，本文分别与一些经典的单模态关系抽取模型和当前主流多模态关系抽取模型进行比较。

#### 3.4.1　单模态关系抽取模型

PCNN[32]：设计了一种分段最大池化，以捕捉不同实体之间的结构信息。

BERT[28]：第一个基于微调的表示模型，减少了任务特定体系结构的需求。

MTB[15]：利用 BERT 使用分布假设拓展，构建与任务无关的关系表示。

PURE[33]：提出一种基于 pipeline 的端到端关系抽取方法。

#### 3.4.2　多模态关系抽取模型

MEGA[1]：对文本和视觉场景图采取有效的对齐策略。

MoRe[17]：提出多模态检索的框架，检索与输入文本和图像相关的知识，并通过专家混合模块来组合预测结果。

HVPNeT[2]：提出一种视觉增强方法，通过基于 Transformer 架构的分层视觉前缀融合网络，将视觉线索集成到文本分析中。

MKGformer[34]：通过使用一种混合型变换器将视觉−文本知识进行整合，增强了信息检索、问答

和推荐等任务。

MMIB[19]：提出一种基于变分信息瓶颈原理的网络框架，预测相关信息与噪声信息。通过基于互信息的对比学习方法，增强文本表示和图像表示的一致性。

RSRNeT[4]：通过引入多尺度视觉特征提取，将视觉特征作为前缀信息输入每一层 Transformer 的自注意力机制中，实现视觉和文本信息的深度融合。

VisualPT-MoE[3]：引入视觉提示调优机制，将视觉提示作为科学系的前缀嵌入文本编码器中，引导文本表示学习。采用专家混合策略融合多种图像表示，提升模型对多样化图像信息的利用效率。

AMNet[35]：通过图像−文本对齐和图像−文本匹配模块来增强表示，并辨别语义对应关系。

Prompt me up[18]：提出实体−对象和关系−图像对齐的预训练任务，利用大量未标记的图像−文本对进行自监督预训练，挖掘图像和文本之间的潜在对齐关系。

CE-DCVSI[21]：提出一种双通道视觉语义信息协同增强机制，通过引入双通道结构，分别处理图像的全局特征和局部特征，通过协同增强策略，充分融合两种视觉信息，提升模型对图像语义的理解能力。

HGMAF[22]：提出一种多证据对齐融合机制，将大语言模型生成的文本证据与图像证据进行融合，充分利用不同来源的信息，提高模型的准确性。

### 3.5 对比实验

在 MNRE 数据集[5] 上，将本文提出的 CAFNeT 模型与基准模型进行对比实验，验证本文模型在 MRE 任务中的有效性。实验结果如表 1 所示，本文模型与最优指标皆以加粗字体标示。

从表 1 的实验结果可以看出，本文所提出的模型 CAFNeT 相较于纯文本关系抽取模型 PURE 的 F1 值提升了 29.5 个百分点，相较于多模态关系抽取模型 HGMAF，F1 值提升了 3.38 个百分点。

**表 1 在 MNRE 数据集上的实验结果**

%

| 模态 | 模型 | 精确率 | 召回率 | F1值 |
|---|---|---|---|---|
| Text | PCNN | 62.85 | 49.69 | 55.49 |
| | BERT | 58.58 | 60.25 | 59.40 |
| | MTB | 64.46 | 57.81 | 60.86 |
| | PURE | 64.50 | 58.19 | 61.22 |
| Text+Image | MEGA[1]（2021） | 64.51 | 68.44 | 66.41 |
| | MoRe[17]（2022） | 68.23 | 68.79 | 68.51 |
| | HVPNeT[2]（2022） | 83.64 | 80.78 | 81.85 |
| | MKGformer[34]（2022） | 82.67 | 81.25 | 81.95 |
| | VisualPT-MoE[3]（2023） | 84.81 | 83.75 | 84.28 |
| | AMNet[35]（2023） | 85.24 | 83.91 | 84.57 |
| | Prompt me up[18]（2023） | 84.95 | 85.76 | 84.86 |
| | MMIB[19]（2024） | 83.49 | 82.97 | 83.23 |
| | RSRNeT[4]（2024） | 84.78 | 83.06 | 83.89 |
| | CE-DCVSI[21]（2025） | 86.59 | 85.56 | 86.08 |
| | HGMAF[22]（2025） | 88.35 | 86.36 | 87.34 |
| | **CAFNeT** | **91.30** | **90.16** | **90.72** |

本文提出的 CAFNeT 模型优于传统的单模态关系抽取模型和主流的多模态关系抽取模型，这一提升得益于所提模型的多层次跨模态交互模块和多模态自适应融合模块，使得模型能够充分挖掘和利用不同模态间的互补信息，动态调整不同特征的贡献比例，增强模态间的关键特征信息，防止关键信息丢失。

为验证模型在低资源下的可用性，从 MNRE 原始训练集中随机以 5% ~ 50% 进行采样，形成低资源训练集，并以此训练集进行对比实验，结果如图 4 所示。

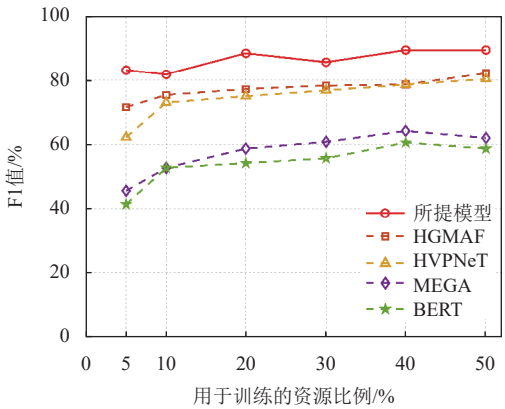



图 4 随机采样实验结果

实验结果表明，相较于主流的 MRE 模型，本文提出的 CAFNeT 模型在低资源情况下表现出显著优势。

### 3.6　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，其中 w/o AMCM 表示在多层次跨模态交互模块中，去除 AMCM，不对模态间进行细粒度视觉和文本对齐；w/o MAF 表示去除多模态自适应融合模块中辅助信息对结果修正的影响；w/o MAF 和 w/o AMCM 表示同时去除 MAF 和 AMCM。

消融实验结果如表 2 所示，去除 AMCM，模型的 F1 值下降 1.3 个百分点；去除 MAF，模型的 F1 值下降 8.71 个百分点；同时去除 AMCM 和 MAF，模型的 F1 值下降 9.08 个百分点。实验结果证明本文所提模型在 MRE 任务中的有效性。

**表 2　消融实验结果**

| 模型 | 精确率/% | 召回率/% | F1值/% | 每个epoch训练+验证消耗时间/min | 显存消耗/MB |
|---|---|---|---|---|---|
| **CAFNeT** | **91.30** | **90.16** | **90.72** | 5.78 | 18 648 |
| w/o AMCM | 90.42 | 88.44 | 89.42 | 5.53 | 18 429 |
| w/o MAF | 82.57 | 81.46 | 82.01 | 5.58 | 18 504 |
| w/o AMCM 和 w/o MAF | 81.97 | 81.31 | 81.64 | **5.50** | **18 166** |

### 3.7　不同扩散模型

本文使用的扩散模型为潜空间扩散模型（latent diffusion model, LDM）[25]。为探究不同扩散模型生成的翻译图像对实验结果的影响，与其他具有文本生成图像功能的扩散模型 CogView[36] 和 LafitE[37] 进行比较，结果如表 3 所示。本文模型与最优指标皆以加粗字体标示。

**表 3　基于不同扩散模型的实验结果**　　%

| 模型 | 精确率 | 召回率 | F1值 |
|---|---|---|---|
| **LDM** | **91.30** | **90.16** | **90.72** |
| LafitE | 90.88 | 89.27 | 90.07 |
| CogView | 89.76 | 87.81 | 88.77 |

从表 3 的实验结果可以看出，本文使用的扩散模型 LDM 相较于扩散模型 LafitE 的 F1 值提升了 0.65 个百分点，相较于扩散模型 CogView 的 F1 值提升了 1.95 个百分点。

实验结果表明，相较于经典的扩散模型 LafitE 和 CogView，本文提出的 CAFNeT 使用的 LDM 模型在 MRE 任务中具有一定的优势。

为确认不同扩散模型的生成性能，采用弗雷歇初始距离（fréchet inception distance, FID）[38]、初始分数（inception score, IS）[39] 和生成图片平均消耗时间作为评价指标。FID 和 IS 的计算过程为：

$$\mathrm{FID}=\left\|\boldsymbol{\mu}_r-\boldsymbol{\mu}_g\right\|^2+\mathrm{Tr}\left(\boldsymbol{\Sigma}_r+\boldsymbol{\Sigma}_g-2\left(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g\right)^{1/2}\right) \tag{25}$$

$$\mathrm{IS}=\exp\left(E_{x\sim p(g)}\left[D_{\mathrm{KL}}(p(y\mid x)\|p(y))\right]\right) \tag{26}$$

式中，$\boldsymbol{\mu}_r$和$\boldsymbol{\mu}_g$分别代表原始图像和生成图像经过预训练 Inception V3[40] 提取特征后的均值向量；$\boldsymbol{\Sigma}_r$和$\boldsymbol{\Sigma}_g$分别表示原始图像特征向量和生成图像特征向量的协方差矩阵；Tr(·)表示矩阵的迹；$x$表示扩散模型生成的图像；$p(g)$表示扩散模型所产生的图像分布；$D_{\mathrm{KL}}(\cdot\|\cdot)$表示 KL 散度；$p(y|x)$表示在给定图像$x$时，预训练 Inception V3 输出的类别概率分布；$p(y)$表示生成图像的类别概率分布均值。

与其他经典扩散模型 LafitE 和 CogView 进行比较，生成性能结果如表 4 所示。

**表 4　不同扩散模型基于文本条件的图像生成结果**

| 模型 | FID | IS | 生成图片平均消耗时间/s |
|---|---|---|---|
| **LDM** | **13.54** | 24.37 | 1.52 |
| LafitE | 25.36 | **25.04** | **0.12** |
| CogView | 26.82 | 19.14 | 3.21 |

从表 4 的实验结果可以看出，本文使用的扩散模型 LDM 虽然在 IS 上略低于扩散模型 LafitE，在计算效率上与 LafitE 存在差距，但是在 FID 上显著优于其他模型，生成模型的鲁棒性更好，生成的图片质量更高、多样性更好。

### 3.8　案例分析

为了更加直观地展示 CAFNeT 模型的优势，本文从 MNRE 数据集中选择了一组代表性的案例样本，并将 CAFNeT 模型与传统的关系抽取模型 PURE 和主流的 MRE 模型 MEGA、HVPNeT 模型进行比较，如图 5 所示。

基于纯文本的单模态关系抽取模型 PURE 难以确定实体之间的正确关系，而主流 MRE 模型 MEGA 和 HVPNeT 尽管利用了图像信息，但往往聚焦于无关的视觉对象，反而在后续融合过程中引入了噪声。相比之下，CAFNeT 通过多层次跨模态交互模块，实现了文本和目标对象之间的细粒度对齐，并通过多模态自适应融合模块，保留重要信息，抑制不相关多模态数据信息的负面影响，最终正确预测

了实体间的关系。



文本: Registration is rocking for the 2018 OMFPOA Symposium in Windsor !!!

头实体: OMFPOA

尾实体: Windsor

正确输出: None

PURE: misc/loc/present_in

MEGA: misc/loc/locate_at

HVPNeT: misc/loc/locate_at

CAFNeT: None

图 5 预测结果图

## 4 结束语

本文提出一种跨模态自适应融合网络用于多模态关系抽取任务，充分挖掘和利用不同模态之间的互补信息，增强模态间的特征交互，并改进特征融合策略，考虑多模态数据不相关的情况，通过扩散模型生成图像组成新的图文对，作为辅助信息修正最终的识别结果，保留重要信息的同时抑制不相关多模态信息的影响，进一步提升关系抽取的效果。实验结果表明，相较于传统的单模态关系抽取方法与现有的多模态关系抽取方法，本文提出的模型在精确率、召回率和 F1 值 3 个方面均有一定的提升，同时在低资源情况下也具有有效性。在未来的工作中，将考虑利用大语言模型生成额外的提示信息来提升关系抽取的准确性。

## 参 考 文 献


[1] Zheng Changmeng, Feng Junhao, Fu Ze, et al. Multimodal relation extraction with efficient graph alignment[C]// Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 5298-5306.

[2] Chen Xiang, Zhang Ningyu, Li Lei, et al. Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction[C]// Proceedings of the Association for Computational Linguistics: NAACL 2022. Seattle: Association for Computational Linguistics, 2022: 1607-1618.

[3] Xu Bo, Huang Shizhou, Du Ming, et al. A unified visual prompt tuning framework with mixture-of-experts for multimodal information extraction[C]//Proceedings of the 28th International Conference on Database Systems for Advanced Applications. Tianjin: Springer, 2023: 544-554.

[4] Wang Min, Chen Hongbin, Shen Dingcai, et al. RSRNeT: A novel multi-modal network framework for named entity recognition and relation extraction[J]. PeerJ Computer Science, 2024, 10: e1856.

[5] Zheng Changmeng, Wu Zhiwei, Feng Junhao, 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. Shenzhen: IEEE, 2021: 1-6.

[6] 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.
Li Dongmei, Zhang Yang, Li Dongyuan, et al. Review of entity relation extraction methods[J]. Journal of Computer Research and Development, 2020, 57(7): 1424-1448. (in Chinese)

[7] 曾泽凡, 胡星辰, 成清, 等. 基于预训练语言模型的知识图谱研究综述[J]. 计算机科学, 2025, 52(1): 1-33.
Zeng Zefan, Hu Xingchen, Cheng Qing, et al. Survey of research on knowledge graph based on pre-trained language models[J]. Computer Science, 2025, 52(1): 1-33. (in Chinese)

[8] Humphreys K, Gaizauskas R, Azzam S, et al. University of Sheffield: Description of the LaSIE-II system as used for MUC-7[C]//Proceedings of the 7th Message Understanding Conference (MUC-7). Fairfax, Virginia: Association for Computational Linguistics, 1998: 84-89.

[9] Neelakantan A, Collins M. Learning dictionaries for named entity recognition using minimal supervision[C]// Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg: Association for Computational Linguistics, 2014: 452-461.

[10] Sun Xia, Dong Lehong. Feature-based approach to Chinese term relation extraction[C]//Proceedings of the International Conference on Signal Processing Systems. Singapore: IEEE, 2009: 410-414.

[11] 郭喜跃, 何婷婷, 胡小华, 等. 基于句法语义特征的中文实体关系抽取[J]. 中文信息学报, 2014, 28(6): 183-189.
Guo Xiyue, He Tingting, Hu Xiaohua, et al. Chinese named entity relation extraction based on syntactic and semantic features[J]. Journal of Chinese Information Processing, 2014, 28(6): 183-189. (in Chinese)

[12] 甘丽新, 万常选, 刘德喜, 等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展, 2016, 53(2): 284-302.
Gan Lixin, Wan Changxuan, Liu Dexi, et al. Chinese named entity relation extraction based on syntactic and semantic features[J]. Journal of Computer Research and Development, 2016, 53(2): 284-302. (in Chinese)

[13] Nguyen T H, Grishman R. Relation extraction: Perspective from convolutional neural networks[C]//Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Denver: Association for Computational Linguistics, 2015: 39-48.

[14] Zeng Daojian, Liu Kang, Lai Siwei, et al. Relation classification via convolutional deep neural network[C]// Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin: Association for Computational Linguistics, 2014: 2335-2344.

[15] 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: Association for Computational Linguistics, 2019: 2895-2905.

[16] 王永胜, 李培峰, 王中卿, 等. 多模态信息抽取研究综述[J]. 软件学报, 2025, 36(4): 1665-1691.
Wang Yongsheng, Li Peifeng, Wang Zhongqing, et al. Survey on multimodal information extraction research[J]. Journal of Software, 2025, 36(4): 1665-1691. (in Chinese)

[17] Wang Xinyu, Cai Jiong, Jiang Yong, et al. Named entity and relation extraction with multi-modal retrieval[C]// Proceedings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi: Association for Computational Linguistics, 2022: 5925-5936.

[18] Hu Xuming, Chen Junzhe, Liu Aiwei, et al. Prompt me up:

Unleashing the power of alignments for multimodal entity and relation extraction[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023: 5185-5194.
[19] Cui Shiyao, Cao Jiangxia, Cong Xin, 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.
[20] Wu Shengqiong, Fei Hao, Cao Yixin, et al. Information screening whilst exploiting! Multimodal relation extraction with feature denoising and multimodal topic modeling[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto: Association for Computational Linguistics, 2023: 14734-14751.
[21] Gong Yunchao, Lv Xueqiang, Yuan Zhu, et al. CE-DCVSI: Multimodal relational extraction based on collaborative enhancement of dual-channel visual semantic information[J]. Expert Systems with Applications, 2025, 262: 125608.
[22] He Xinyu, Li Shixin, Zhang Yuning, 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.
[23] Zheng Changmeng, Feng Junhao, Cai Yi, et al. Rethinking multimodal entity and relation extraction from a translation point of view[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: Association for Computational Linguistics, 2023: 6810-6824.
[24] Schuhmann C, Vencu R, Beaumont R, et al. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Sydney: Curran Associates Inc., 2021.
[25] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 10674-10685.
[26] Zhang Dong, Wei Suzhong, Li Shoushan, et al. Multi-modal graph fusion for named entity recognition with targeted visual guidance[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [S.l.]: AAAI Press, 2021: 14347-14355.
[27] Yang Zhengyuan, Gong Boqing, Wang Liwei, et al. A fast and accurate one-stage approach to visual grounding [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 4682-4692.
[28] Devlin J, Chang Mingwei, 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, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.
[29] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.
[30] Lu Di, 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. Melbourne: Association for Computational Linguistics, 2018: 1990-1999.
[31] Zhang Qi, Fu Jinlan, Liu Xiaoyu, et al. Adaptive co-attention network for named entity recognition in tweets [C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 5674-5681.
[32] Zeng Daojian, Liu Kang, Chen Yubo, 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: Association for Computational Linguistics, 2015: 1753-1762.
[33] Zhong Zexuan, Chen Danqi. 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 (NAACL-HLT 2021). [S.l.]: Association for Computational Linguistics, 2021: 50-61.
[34] Chen Xiang, Zhang Ningyu, Li Lei, et al. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid: ACM, 2022: 904-915.
[35] Dai Yinglong, Gao Feng, Zeng Daojian. An alignment and matching network with hierarchical visual features for multimodal named entity and relation extraction[C]// Proceedings of the 30th International Conference on Neural Information Processing. Changsha: Springer, 2023: 298-310.
[36] Ding Ming, Yang Zhuoyi, Hong Wenyi, et al. CogView: Mastering text-to-image generation via transformers[C]// Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2021: 19822-19835.
[37] Zhou Yufan, Zhang Ruiyi, Chen Changyou, et al. Towards language-free training for text-to-image generation[C]// Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022: 17886-17896.
[38] Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017: 6629-6640.
[39] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona: Curran Associates Inc., 2016: 2234-2242.
[40] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2818-2826.

责任编辑　张　莉