

事件结构—内容协同驱动的多模态虚假新闻检测

摘要：在基于事件图的多模态虚假新闻检测中，论元省略和谓词缺失容易导致事件要素组织不一致，标题语义、视觉场景和外部知识等内容信息也未得到充分利用。针对上述问题，本文提出事件结构—内容协同检测方法（Dual-Perspective Collaboration for Multimodal Fake News Detection, DPC）。该方法协同建模事件结构与图文内容，包括以下要素：1）依据谓词存在性对事件要素进行差异化组织并映射至统一的关系建模空间；2）基于多视角图注意力构建事件关系，通过相对模态门控调节文本与视觉事件表示的贡献，以增强结构一致性判别；3）利用百科知识增强标题语义表示，通过跨模态注意力机制融合文本语义与对象级视觉信息，补充内容层面的判别证据。在 Twitter、Weibo 和 PHEME 三个公开数据集上，DPC 的准确率为 0.912~0.938，且在无谓词结构子集上的宏 F1 均高于所比较的各变体。实验结果表明，DPC 可提高事件信息不完整条件下的图文一致性判别能力。

关键词：虚假新闻检测；多模态；多视角图注意力网络；事件图；相对模态门控

中图分类号：TP391

文献标志码：A

Multimodal Fake News Detection Driven by Event Structure—Content Collaboration

Abstract: In event-graph-based multimodal fake news detection, argument omission and predicate absence can lead to inconsistent organization of event elements, while content information, including headline semantics, visual scenes, and external knowledge, remains underutilized. To address these issues, this paper proposes Dual-Perspective Collaboration for Multimodal Fake News Detection (DPC). The proposed method jointly models event structures and textual-visual content through three components: 1) organizing event elements adaptively according to predicate presence and mapping them into a unified relation modeling space; 2) modeling event relations using multi-view graph attention and employing relative modality gating to regulate the contributions of textual and visual event representations, thereby enhancing structural consistency discrimination; and 3) enriching headline semantic representations with encyclopedic knowledge and integrating textual semantics with object-level visual information through cross-modal attention to supplement content-level discriminative evidence. Experiments on three public datasets, Twitter, Weibo, and PHEME, show that DPC achieves accuracy scores ranging from 0.912 to 0.938 and consistently outperforms all compared variants in macro-F1 on predicate-free structure subsets. The results demonstrate that DPC can improve text-image consistency discrimination under conditions of incomplete event information.

Key words: fake news detection, multimodal, multi-view graph attention network, event graph, relative modality gating

0 引言

社交媒体已成为公众获取和分享新闻信息的重要渠道^[1]。社交媒体新闻多以图文形式呈现，错误配图、旧图复用等现象增加了真实性判别的难度，推动了联合分析文本与图像的多模态虚假新闻检测（multimodal fake news detection）研究^[2-3]。

近年来，部分研究通过事件图建模图文内容：以谓词（描述事件动作或状态的核心成分）、论元

（围绕谓词承担施事、受事等语义角色的成分）等事件要素为节点、要素间的语义关系为边构建事件图，并通过比较图文两侧的事件要素及其关系判断跨模态一致性。CLIP-Event^[4]在视觉语言预训练中引入事件及论元角色对齐，为事件关系的跨模态建模提供了基础；Event-Radar^[5]进一步将事件图用于多模态虚假新闻检测。这类方法能够从事件要素及其关系层面刻画新闻语义，为分析图文内容之间的细粒度一致性提供了新的思路。

然而,真实社交媒体中的短标题常省略论元,甚至不含明确谓词,使事件级多模态检测面临两个问题。其一,不同标题中可显式获得的事件要素不同。如图1所示,含有明确谓词的标题即使部分论元未明示,仍可依据已有谓词和论元形成事件关系;而部分标题缺少明确谓词,无法建立谓词—角色结构。若采用固定的事件角色形式,不同表达下的事件结构难以得到一致组织。其二,新闻真实性判断依赖多来源信息的共同作用。事件关系能够刻画被抽取事件要素之间的联系,但标题整体语义、视觉场景以及外部知识等信息并不完全包含于显式事件关系中。不同来源的信息具有不同的表达特点,如何在事件结构建模基础上充分利用这些互补信息,仍是事件级多模态检测需要解决的问题。

现有方法难以充分应对上述问题。对于前者,已有事件结构方法通常以可显式抽取的事件要素及其关系作为建模基础,能够利用事件关系进行细粒度图文一致性分析,但对于不同标题表达下事件要素组织形式的差异仍缺少进一步考虑。对于后者,已有研究进一步挖掘不同模态和不同粒度的内容线索,例如EM-FEND^[6]利用视觉实体和图像内嵌文本补充图文语义,BMR^[7]从文本、图像模式和图像语义等多个视角学习新闻表示。这些方法表明,不同来源的信息均能够为真实性判断提供补充依据,但其主要关注多模态内容证据增强与融合,尚未进一步考虑事件结构表示与整体图文内容之间的互补关系。

针对上述不足,本文提出事件结构—内容协同检测方法DPC(Dual-Perspective Collaboration for Multimodal Fake News Detection)。该方法以事件表达形式为基础,对不同标题中的事件信息进行统一组织:对于具有明确谓词的标题,采用谓词—论元形式表示事件要素;对于缺少明确谓词的标题,则采用主题—实体形式,并通过事件索引映射到统一关系建模空间,实现不同表达形式下事件结构的一致表示。在此基础上,DPC联合建模事件结构信息与图文内容信息,通过事件结构推理学习跨模态关系,并利用标题语义、视觉场景和百科描述增强内容证据,最终实现不完整事件条件下的图文一致性判别。

本文的主要贡献如下:

(1) 提出事件结构—内容协同检测方法DPC。



图1 不同标题表达下的事件要素与内容线索示例
标题用于说明结构类型, $\emptyset$ 仅表示文本未提及相应论元。

该方法针对不同标题表达下事件要素组织不一致及内容信息利用不足的问题,将事件结构信息与图文内容信息进行协同建模,实现不完整事件条件下的图文一致性判别。

(2) 设计面向异构事件表达的统一关系建模方法。针对谓词缺失和论元省略导致的事件结构不一致问题,采用谓词—论元与主题—实体两种形式组织事件输入,并通过事件索引映射到统一关系空间,实现不同表达下事件结构的一致建模。

(3) 在 Twitter、Weibo 和 PHEME 三个公开基准数据集上验证了所提方法的有效性: DPC 在 Twitter 上的检测性能优于对比方法,在 Weibo 上接近但略低于最优方法,在 PHEME 上高于公开文献参照准确率;消融实验与基于大语言模型的对照测试验证了各组成部分的作用,在事件结构证据受限的无谓词结构子集上,完整模型的 F1 仍高于所比较的各变体。

1 相关工作

1.1 基于整体表征的检测方法

早期多模态虚假新闻检测方法主要通过联合编码文本与图像特征学习新闻整体表示,以实现跨模态信息融合与真实性判断。EANN^[8]通过事件对抗训练学习跨事件共享表示, MVAE^[2]利用变分自编码器约束多模态特征融合, SpotFake^[3]通过融合文本和图像表示完成分类。随着研究深入, SAFE^[9]、

CAFE^[10]和MCAN^[11]进一步利用跨模态注意力机制增强文本与图像之间的信息交互, FND-CLIP^[12]借助视觉语言预训练模型提升图文语义对齐能力, BMR^[7]则从多个视角学习新闻表示并进行融合。该类方法推动了多模态信息融合的发展, 但其主要依赖文本与图像的整体语义表示, 对新闻中具体事件要素及其关系的刻画仍然有限。

1.2 基于细粒度证据的检测方法

为提升模型对细粒度语义线索的利用能力, 部分研究进一步从实体、对象及事件结构等层面对新闻证据进行建模。EM-FEND^[6]利用视觉实体与文本实体之间的不一致关系进行检测, 并结合图像内嵌文本补充语义信息; KMGCN^[13]将文本词、知识概念和视觉词组织到统一图结构中, 通过知识关联增强多模态表示。这类方法能够利用实体级语义线索辅助真实性判断, 但实体对应关系主要反映对象层面的关联, 对于新闻事件中的动作及角色关系描述仍然有限。针对上述不足, 事件结构建模方法进一步利用谓词及其语义角色关系刻画新闻事件。CLIP-Event^[4]将事件及论元角色对齐引入视觉语言预训练过程, 为跨模态事件关系建模提供了基础; Event-Radar^[5]进一步构建事件图, 并通过多视角学习利用事件关系进行真实性判断。相比实体级关联, 事件结构能够更加细粒度地描述参与对象、行为及语义角色之间的关系, 为图文一致性判断提供结构化依据。然而, 现有事件结构方法通常依赖显式抽取的事件要素, 并采用相对固定的事件角色表示, 当标题存在谓词缺失或论元省略时, 不同表达形式下的事件结构组织仍面临挑战。

1.3 大语言模型驱动的检测方法

近年来, 大语言模型 (large language model, LLM) 逐渐应用于虚假新闻检测中的语义理解、理由生成和结构信息抽取等任务。Hu 等人^[14]探索利用 LLM 生成判别理由辅助虚假新闻检测; FACTGUARD^[15]结合新闻内容与常识理由增强文本真实性分析; GLPN-LLM^[16]进一步将 LLM 生成的伪标签融入跨模态图结构学习, 以辅助多模态虚假新闻检测。除作为判别辅助模型外, LLM 在信息抽取和结构化知识生成方面展现出较强能力, 可以为下游任务提供实体、关系及事件层面的语义表示。现有研究表明, LLM 能够为新闻语义理解和辅助信息生成提供支持, 但其主要作用集中于判别

辅助、知识增强或标签传播等过程, 如何利用 LLM 获取事件结构信息, 并进一步与多模态检测模型结合仍值得研究。

1.4 知识增强与模态融合方法

除事件结构建模外, 外部知识增强和模态融合机制也是提升多模态虚假新闻检测性能的重要方向。TagMe^[17]通过实体链接将短文本映射至 Wikipedia, 为新闻提供额外知识信息; QMF^[18]研究质量感知的动态融合机制, 根据不同模态信息调整融合过程; Event-Radar^[5]从事件不一致、文本情绪和视觉模式等多个证据视角学习新闻表示, 并调节不同视角的信息贡献。这些方法能够通过外部知识补充或模态交互优化提升检测性能, 但主要关注知识增强或模态融合过程本身, 对于事件结构信息与图文内容信息之间的协同关系仍缺少深入研究。

此外, 文本对抗攻击领域已有面向硬标签黑盒文本的 TextSwindler^[19]和基于双重语义过滤的 DLSF^[20]等方法, 但其研究目标与本文的多模态虚假新闻检测有所不同。

综上, 现有多模态虚假新闻检测方法分别从整体语义表示、细粒度证据建模、知识增强和模态融合等角度推动了该领域的发展。然而, 在事件表达不完整条件下, 如何有效组织不同形式的事件信息, 并进一步联合利用事件结构信息与图文内容信息进行真实性判别, 仍是该领域需要解决的问题。针对上述不足, 本文从事件表达适应性和结构—内容协同两个方面展开研究。

2 方法设计

2.1 问题定义

给定多模态新闻集 $\mathcal{D} = \{(P_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, 第 i 条样本由新闻 P_i 与标签 y_i 构成, 其中 $P_i = \{T_i, I_i, \mathcal{M}_i\}$, T_i 为标题文本, I_i 为新闻图像, $\mathcal{M}_i$ 为外部工具获得的实体链接、百科描述、目标检测及事件抽取信息。标签 $y_i \in \{0, 1\}$, 0 表示真实新闻, 1 表示虚假新闻。为表述简洁, 下文省略样本下标。目标是学习判别函数 $f_\theta: P \mapsto \hat{y}$, 联合事件结构一致性与图文内容语义输出真伪预测。

2.2 DPC 总体框架

本文提出事件结构—内容协同检测方法 DPC (Dual-Perspective Collaboration for Multimodal Fake News Detection), 总体框架如图 2 所示。DPC 由多

模态事件抽取、内容证据增强、事件结构推理和协同判别组成。

给定标题 T 与图像 I ，DPC 首先依据谓词存在性选择事件表示方式，将不同标题表达下的事件要素组织为统一结构输入，并获得用于内容建模的图文表示。随后，内容证据增强分支利用标题语义、百科知识与视觉对象信息学习内容一致性，事件结构推理分支基于事件节点及其关系建模跨模态结构一致性。两个分支分别产生内容得分与结构得分，经 Softmax 转换为概率后，通过全局可学习权重 α 在概率空间进行协同融合，得到最终预测结果。

2.3 多模态事件抽取

本模块根据标题中的谓词存在性选择对应的事件表示方式，并组织文本与视觉侧事件要素，为后续结构推理提供统一输入。针对不同标题表达形式，分别构建谓词结构与无谓词结构，并记录事件要素对应的位置索引及跨模态关联状态。其中，文本侧根据谓词存在性抽取事件要素并构建对应结构，视觉侧负责对象检测及跨模态关联建立。

该模块输出内容表示与结构表示两类互补信息。内容表示用于刻画新闻整体语义，结构表示用于描述事件要素之间的关系，并分别输入内容证据增强和事件结构推理模块。

2.3.1 文本事件要素抽取

固定主谓宾模板在论元省略时容易产生空字段，而在标题缺少明确谓词时难以形成有效事件结构。本文使用 DeepSeek-V3^[21] 判断标题中的谓词存在性，并根据结构判定结果将文本事件表示分为谓词结构和无谓词结构。对应地，本文设计谓词抽取模式和主题—实体抽取模式，并采用温度为 0 的受约束 JSON 格式输出：

$$\mathcal{S}(T) = \begin{cases} (s^?, v, o^?, p^?), & T \in \mathcal{C}_{\text{pred}}, \\ (\tau, \{e_i\}_{i=1}^{N_e}), & T \in \mathcal{C}_{\text{te}}. \end{cases} \quad (1)$$

其中， $\mathcal{C}_{\text{pred}}$ 与 $\mathcal{C}_{\text{te}}$ 分别表示谓词结构与无谓词结构， N_e 为关键实体数量。谓词抽取模式输出谓词 v 以及允许缺失的主语 s 、宾语 o 和介词短语 p 。当标题包含明确谓词但部分论元缺失时，仍采用谓词结构表示；仅在无法获得明确谓词时采用主题—实体结构。主题—实体抽取模式通过主题词 τ 和关键实体集合保留无谓词标题中的有效语义线索，为后续结构传播提供输入。谓词标签 $\mathcal{C}_{\text{pred}}$ 根据结构判定结果确定，用于选择后续事件输入组织形式。抽取结果均保持标题中的原文表述，事件抽取配置及异常响应处理规则见第 3.1.2 节。

2.3.2 文本特征编码

标题经 CLIP 文本编码器得到词元矩阵 T^K 与句向量 t_{cls} 。参与跨模态匹配的事件要素表示由对应位

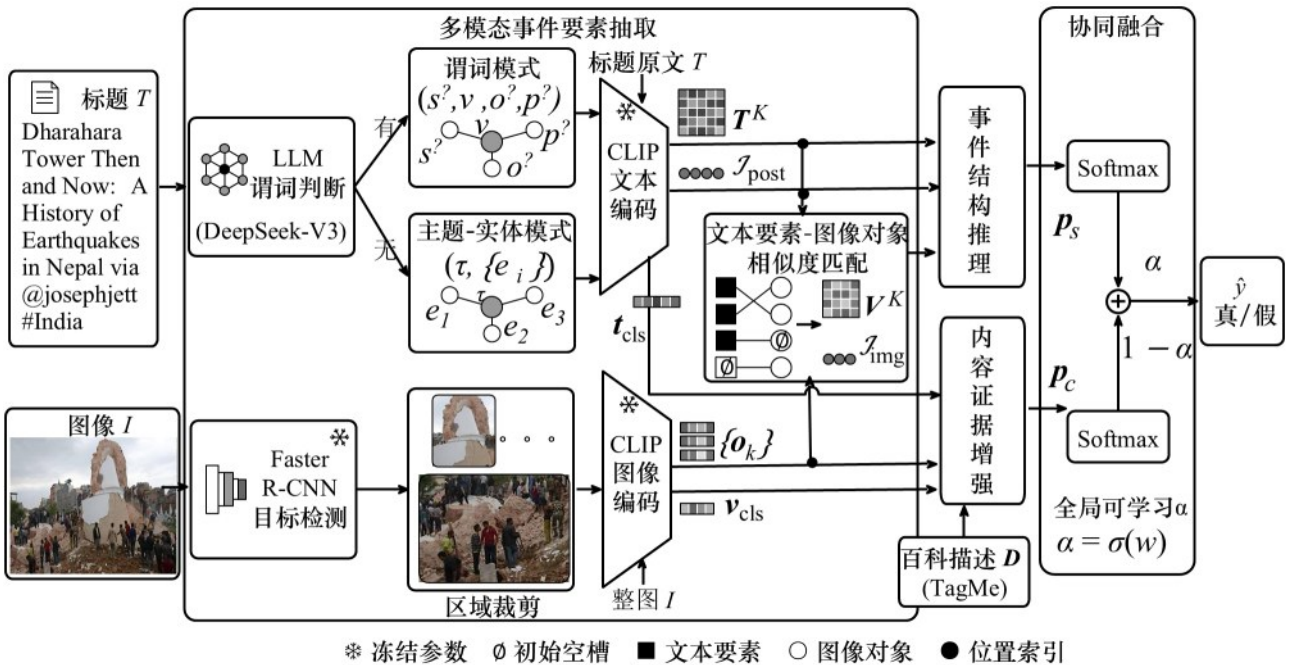


图 2 DPC 总体框架 (内容证据增强与事件结构推理分别见图 3 和图 4)

置的词元池化获得, 并经 CLIP 文本投影头映射和 L2 归一化后, 与图像区域表示投影到同一语义空间。图像区域表示同样经过图像投影头映射和 L2 归一化处理。

为保持事件要素与原文位置之间的对应关系, 本文记录各要素的位置索引。对于能够定位到原文的要素, 采用对应词元表示; 对于无法定位的非空要素, 则进行独立编码; 对于缺失角色, 则保留空位信息。

文本事件索引 $\mathcal{I}_{\text{post}}$ 用于记录事件要素在标题中的位置关系, 包括句级表示位置以及可定位谓词、主题和外围要素对应的词元位置。索引按照原文位置去重并排序。参与跨模态匹配的外围要素根据结构类型取 $\{s, o, p\}$ 或实体集合, 谓词 v 与主题 τ 参与区域配对。上述位置索引与跨模态匹配结果用于构建后续事件结构表示。

2.3.3 视觉特征编码与事件要素匹配

图像经 CLIP 得到整图向量 $\mathbf{v}_{\text{cls}}$, Faster R-CNN^[22]检测区域经裁剪后输入 CLIP 编码, 得到对象集合 $\{\mathbf{o}_k\}$ 。在归一化后的文本外围要素表示与对象表示之间计算余弦相似度矩阵 $\mathbf{S}$, 并采用带阈值约束的互斥贪心策略完成跨模态匹配, 即优先选择相似度最高的未匹配文本—视觉对, 仅当该相似度不低于阈值 τ_{match} 时建立关联。空文本角色不参与匹配过程。本文依据三个数据集上的匹配率与拒配率确定 τ_{match} (见第 3.2.3 机制验证分析), 取值为 0.2。

由于事件结构推理需要同时利用已对齐信息与

未对齐信息, 本文保留匹配过程中的槽位状态、候选相似度及对象索引, 并将其作为结构输入的一部分。视觉节点矩阵 $\mathbf{V}^K$ 由整图表示、有效对象表示以及未匹配槽位的占位表示组成。对于已匹配槽位, 直接引用对应对象表示以避免重复计算; 对于未匹配槽位, 则保留其对应的匹配信息, 不因占位处理而丢失。

2.4 内容证据增强

社交媒体标题通常较短, 事件要素还可能因论元省略或对象漏检而不完整。仅依赖事件结构难以充分利用新闻整体语义及外部知识等补充信息。因此, 本文设计内容证据增强模块, 结合标题语义、实体百科描述和视觉对象信息, 从整体内容层面学习图文一致性表示, 为事件结构推理提供互补证据, 该模块框架图如图 3 所示。

2.4.1 知识增强的文本表征

为补充短标题中可能缺失的背景信息, 本文利用实体关联知识增强文本表示。本文使用 TagMe^[17]将标题 T 中的实体链接至维基百科, 并将词条首句描述编码为知识表示矩阵 $\mathbf{D} \in \mathbb{R}^{M \times d}$, 使其与 $\mathbf{t}_{\text{cls}}$ 映射至同一语义空间, 其中 M 表示有效百科描述数量。随后, 句向量 $\mathbf{t}_{\text{cls}}$ 与知识描述经共享投影后, 以标题语义作为查询、百科描述作为键和值进行缩放点积注意力, 并通过残差连接保留标题自身语义:

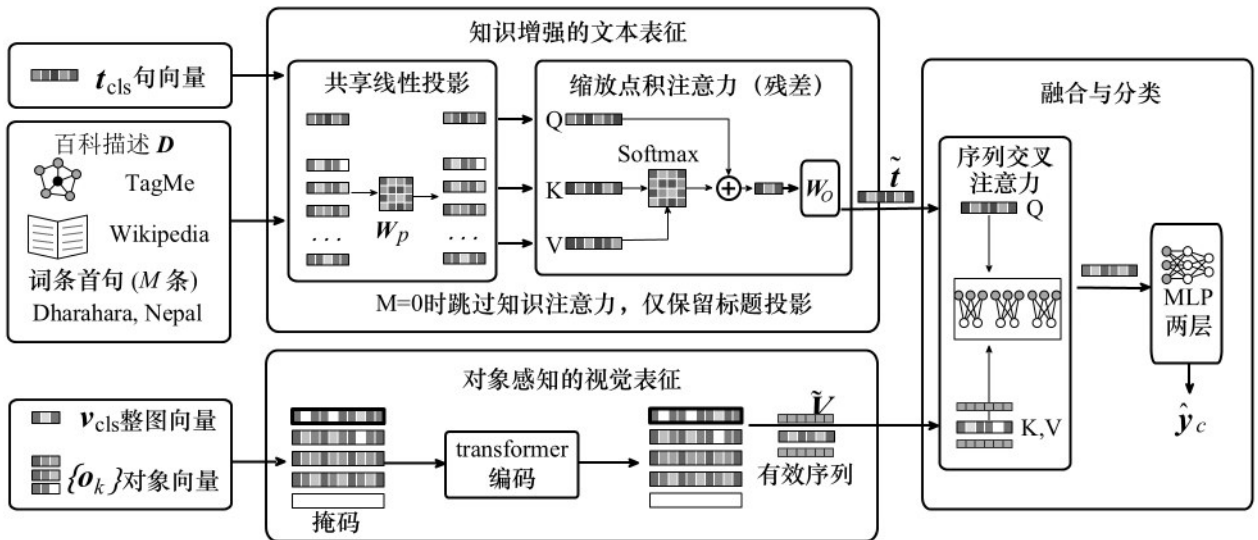


图3 内容证据增强框架

$$\tilde{t} = W_o \left(q + \text{Softmax} \left(\frac{qK^T}{\sqrt{d'}} \right) V \right), \quad (2)$$

其中, q 来自投影后的句向量, K 、 V 来自有效百科描述, 投影维度 $d'=256$ 。当 $M=0$ 时以有效性掩码跳过知识注意力, 直接输出投影后的句向量。

2.4.2 对象感知的视觉表征与跨模态融合

为充分利用图像中的对象级视觉证据, 本文将整图信息与检测对象统一组织为视觉序列。视觉侧以整图向量与对象向量组成序列 $[v_{\text{cls}}; o_1; \dots; o_K]$, 其中整图仅出现一次, 真实对象按有效长度保留, 批填充位置由掩码 m 屏蔽。该序列经 Transformer 编码后返回完整输出序列 $\tilde{V} \in \mathbb{R}^{(K+1) \times d'}$ 。随后, 以知识增强后的文本向量 $\tilde{t}$ 作为查询、 $\tilde{V}$ 作为键和值进行多头交叉注意力 (MHA), 使文本能够在整图与各对象之间进行选择性地聚合。由于键和值采用对象级视觉序列, 注意力权重能够根据文本查询在不同视觉对象之间自适应分配, 从而实现对象级视觉证据选择, 而非仅依赖单一视觉表示进行融合。MHA 在 Softmax 之前应用键填充掩码 m 。当 $K=0$ 时序列仅保留整图项, 避免所有键均被屏蔽。融合过程定义为式 (3):

$$\begin{aligned} u &= \text{LN}(\tilde{t} + \text{MHA}(\tilde{t}, \tilde{V}, \tilde{V})), \\ z_c &= \text{LN}(u + \text{FFN}(u)). \end{aligned} \quad (3)$$

式 (3) 中的 $\tilde{V}$ 保留序列维度, FFN 为前馈网络, 内容得分 $\hat{y}_c$ 由 z_c 经两层 MLP 得到。

内容表征过程主要关注图文内容层面的语义一

致性。然而, 仅依赖整体内容匹配难以充分刻画事件要素之间的结构关系差异。因此, 本文进一步利用事件节点及其关系进行结构传播, 学习跨模态结构表示。

2.5 事件结构推理

为进一步建模细粒度事件关系, 本文构建事件结构推理模块, 从事件节点及其关系出发学习跨模态结构表示。该模块以节点索引诱导三类边集, 在共享节点矩阵和图注意力参数下进行多视角传播, 并通过相对模态门控调节跨模态交互项的贡献, 框架图如图 4 所示。

2.5.1 统一节点图与三类视图

将文本节点矩阵与视觉节点矩阵沿节点维拼接, 得到统一节点矩阵 $H^{(0)} = [T^K; V^K] \in \mathbb{R}^{N \times d}$, 其中 N 为节点总数。第 2.3 节的要素匹配仅标记事件语义的承载位置, 未参与对齐的词元与未被配对的对象仍需保留在传播空间中。由于传播初始阶段难以准确确定节点间关联强度, 预设边权或依据阈值裁剪边连接均可能损失潜在有效的信息路径。因此, 本文不预设边权, 而是在全部节点对之间建立无向连接并保留自环, 将节点关联强度交由图注意力网络端到端学习。

在此基础上, 根据文本事件索引和图像事件索引诱导文本事件边集与图像事件边集, 对应形成文本事件视图、图像事件视图和全图视图。三个视图共享同一节点矩阵, 仅通过边索引和传播掩码限定信息传播范围。两类事件节点仅通过索引标记, 不

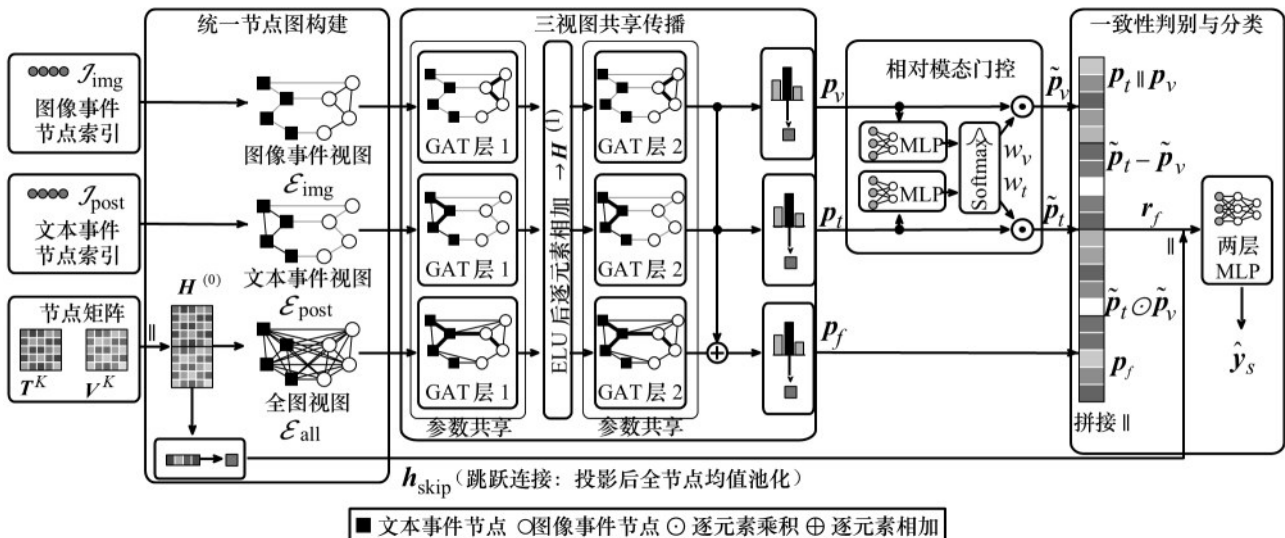


图4 事件结构推理框架(图中连边为结构示意)

物理构造独立子图，为后续参数共享和边掩码并行传播提供基础。

2.5.2 多视角图注意力传播与相对模态门控

为同时保留全局节点关联与事件局部关系，本文在统一节点矩阵上构建三个传播视图。多视角图注意力传播网络在同一节点矩阵上使用共享的图注意力参数，并通过不同边掩码在上述三个视图上并行传播：

$$H_{\star}^{(l+1)} = \text{LN}\left(\text{GAT}_{w,a}^{(l)}(H^{(l)}; \mathcal{E}_{\star})\right), \quad (4)$$

$$\star \in \{\text{all}, \text{post}, \text{img}\}.$$

本文设置两层图注意力网络。第一层三个视图的输出各经指数线性单元 (ELU) 激活后逐元素相加，形成共享中间表示 $H^{(1)} = \sum_{\star} \text{ELU}(H_{\star}^{(1)})$ ，以实现不同传播范围的信息交互。第二层再以该表示为共同输入并按三类边集传播，使跨模态对齐信号持续注入各视图演化。第二层输出分别用于局部事件一致性建模和全局事件表示保留。文本事件视图与图像事件视图的结果各自保留，分别在文本事件节点和图像事件节点上做最大池化，得到供跨模态比较的 p_t 和 p_v 。三个视图的结果同时逐元素相加，在全部节点上做最大池化，得到全局结构表示 $p_f \in \mathbb{R}^{d_o}$ ，其中 d_o 为池化后视图表示的维度。

两个轻量 MLP 分别由 p_t 与 p_v 产生相对分配分数 g_t 、 g_v ，经 $[w_p, w_v] = \text{Softmax}([g_t, g_v])$ 得到权重，并定义 $\tilde{p}_t = w_t p_t$ 、 $\tilde{p}_v = w_v p_v$ 。其中，该权重由分类目标端到端学习，用于调节文本与视觉事件表示在一致性判别中的相对贡献，而非作为独立的模态质量估计。为避免门控完全抑制任一模态的信息，模型保留未加权的局部事件表示和全图表示作为残差信息通路。全图表示 p_f 保留为全局结构上下文，不参与门控调整。

在此基础上，利用门控后的模态表示构造差异项和乘积项，并与未加权的局部事件表示及全图表示共同映射为一致性判别向量：

$$r_f = W_r [p_t \parallel p_v \parallel \tilde{p}_t - \tilde{p}_v \parallel \tilde{p}_t \odot \tilde{p}_v \parallel p_f]. \quad (5)$$

其中 $\parallel$ 为拼接、 $\odot$ 为逐元素乘积， $r_f \in \mathbb{R}^{d_r}$ ， $W_r \in \mathbb{R}^{d_r \times 5d_o}$ 。式 (5) 保留未加权的 p_t 、 p_v 和 p_f 作为残差信息通路，使门控仅调节交互项贡献而不完全抑制任一模态。最终，将 r_f 与投影后全节点均值池化的 h_{skip} 拼接，经两层 MLP 得到结构得分 $\hat{y}_s$ 。

2.6 协同判别与联合优化

内容分支与结构分支分别建模内容语义与事件结构证据，两类输出从不同视角提供真实性判别依据。DPC 选择在概率层融合二者，以避免直接合并两个分类头的得分时受到输出尺度差异影响。两分支分类得分分别经 Softmax 归一化得到 p_c 和 p_s ，再引入可学习参数 w ，并通过 $\alpha = \sigma(w)$ 将融合权重约束到 (0,1)，最终概率为

$$p_{\text{final}} = \alpha p_s + (1 - \alpha) p_c. \quad (6)$$

取 p_{final} 中概率较大的类别作为预测结果 $\hat{y}$ 。DPC 采用负对数似然损失进行端到端训练：

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \log(p_{\text{final}}^{(i)}[y_i] + \epsilon), \quad (7)$$

其中 B 为批大小， $\epsilon=10^{-8}$ 用于防止 $\log 0$ 。

在概率空间融合后，训练目标直接作用于与推理阶段一致的融合概率。由于融合权重 α 与 $1-\alpha$ 分别作用于两分支概率，模型可在联合训练过程中同步调整分支表示与融合比例。

训练过程中，多模态事件要素抽取中的 CLIP 与 Faster R-CNN 保持冻结，内容增强与结构推理模块及融合参数 w 由同一优化器联合更新。 w 初始化为 1.38 ($\alpha_0 \approx 0.8$)，用于设置训练初期的融合比例。

3 实验

本节在三个公开的多模态虚假新闻检测基准上评估 DPC 方法，分别从整体检测性能、模型组件贡献、关键机制验证、抽取方式与架构设计影响以及无谓词结构适应能力等方面开展实验分析。

3.1 实验设置

3.1.1 数据集

实验采用 Twitter^[23]、Weibo^[24] 和 PHEME^[25] 三个公开多模态虚假新闻数据集。其中，Twitter 包含 MediaEval 任务中的图文推文，Weibo 由微博辟谣平台与权威媒体新闻构成，PHEME 包含多个事件中的谣言与非谣言样本。

实验采用各数据集官方发布版本及划分方式。预处理阶段仅保留同时包含文本和图像的样本，去除视频及模态缺失数据，Weibo 多图像样本仅保留首张图像，PHEME 删除重复且无图像的文本样本。处理后的数据规模及谓词结构分布如表 1 所示。

表 1 中的谓词结构划分由 DeepSeek-V3 判定，用于抽取模式选择及后续分组分析。为验证结构判

表1 数据集统计

数据集	训练	测试	谓词结构	无谓词结构	总计
Twitter	14 483	1 934	6 743 (41.1%)	9 674 (58.9%)	16 417
Weibo	7 531	1 880	8 024 (85.3%)	1 387 (14.7%)	9 411
PHEME	1 367	589	1 428 (73.0%)	528 (27.0%)	1 956

谓词结构允许不及物、被动和论元省略,但必须存在明确谓词;无谓词结构以主题、实体或名词短语为主。

定可靠性,从测试集中随机抽取100条样本进行人工复核,结果与LLM判定一致(100/100)。

3.1.2 实验配置

本文采用DeepSeek-V3完成事件结构判定与要素抽取。模型根据谓词存在性选择SVOP或TOPIC_ENTITY结构,并生成与原文本一致的JSON结果。其中,谓词结构保留已有角色信息,无谓词结构记录主题及实体集合。模型温度设置为0,异常空响应重新请求,抽取过程不使用新闻真假标签。

Twitter和PHEME采用CLIP ViT-B/16^[26], Weibo采用Chinese-CLIP ViT-B/16^[27],所有骨干网络均冻结。文本编码器生成维度为512的词元特征和句向量,其中CLIP采用EOS位置,Chinese-CLIP采用[CLS]位置作为句级表示;图像编码器生成全局特征,并对Faster R-CNN检测对象进行裁剪后重新编码。TagMe与维基百科用于实体链接和百科描述获取。训练阶段仅更新内容证据增强、事件结构推理及融合参数。

实验在单张NVIDIA A40 GPU上完成,训练20轮,批大小设置为64。主实验采用随机种子42、100和2024进行三次独立训练,结果报告均值和样本标准差;表2和表3中DPC结果取三次运行均值,其余实验采用单次运行结果。

3.1.3 对比方法

本文选取多类代表性方法进行比较,包括经典多模态检测方法EANN^[8]、att-RNN^[24]、MVAE^[2]、SAFE^[9]和SpotFake^[3],以及基于预训练视觉语言模型的方法CLIP+MLP、CAFE-CLIP^[10]、MCAN-CLIP^[11]、FND-CLIP^[12]、BMR^[7]、Event-Radar^[5]和GLPN-LLM^[16]。上述方法覆盖整体语义表示、细粒度证据建模、视觉语言预训练和事件结构建模等不同技术路线,能够从多个角度评价本文方法的检测能力。

其中,CLIP+MLP用于评估基础图文表示的检

测性能;Event-Radar作为事件结构建模方法,用于比较事件关系信息在多模态虚假新闻检测中的作用。GLPN-LLM为2025年提出的方法,本文采用其CLIP编码器变体进行比较,以减少不同视觉语言骨干带来的性能差异。除部分方法外,对比结果取自公开文献报告,缺失结果记为“—”。

3.1.4 评估指标

本文采用准确率(Accuracy)和宏F1(Macro-F1)评价模型性能。其中,Macro-F1分别计算真实新闻和虚假新闻两类样本的F1值,并取其算术平均。考虑到多模态虚假新闻检测数据集中可能存在类别分布差异,本文以Macro-F1作为主要评价指标,以减少类别比例对模型评价的影响。

3.2 实验结果

3.2.1 检测性能分析

为验证DPC在多模态虚假新闻检测任务中的有效性,本文将其与不同类型代表性方法进行比较。表2给出了DPC与对比方法在Twitter、Weibo和PHEME三个数据集上的整体检测性能,表3进一步报告DPC对真实新闻和虚假新闻两类样本的分类结果。

在Twitter数据集上,DPC取得0.938的准确率和0.938的宏F1,优于其他比较方法。其中,相比性能较高的Event-Radar,准确率和宏F1分别提升1.00和1.49个百分点;相比近期方法GLPN-LLM,准确率和宏F1分别提升5.0和4.8个百分点。进一步观察分类指标可知,真实类和虚假类F1分别达到0.939和0.937,表明模型能够较均衡地识别两类样本。此外,Twitter数据集中存在较多缺少明确谓词的标题,DPC通过自适应事件表示和内容证据增强缓解事件结构不完整带来的影响。

在Weibo数据集上,DPC的准确率为0.912,宏F1为0.911,均低于Event-Radar,分别相差0.7和0.8个百分点。与GLPN-LLM相比,DPC的准确率高4.5个百分点,但宏F1低0.4个百分点。三种随机种子的实验结果波动较小,说明DPC在不同初始化条件下具有稳定的检测表现。

在PHEME数据集上,DPC的准确率为0.912,较Event-Radar的0.901高1.1个百分点,较GLPN-LLM的0.865高4.7个百分点。但DPC的宏F1为0.897,低于GLPN-LLM的0.907。分类结果显示,DPC对虚假新闻的召回率达到0.963,但精确率为

表2 在三个数据集上的检测性能对比

方法	Twitter		Weibo		PHEME	
	准确率	F1	准确率	F1	准确率	F1
EANN	0.648	0.639	0.782	0.780	0.681	0.721
att-RNN	0.664	0.664	0.772	0.769	—	—
MVAE	0.745	0.744	0.824	0.823	0.852	0.827
SAFE	0.762	0.761	0.763	0.761	0.811	0.767
SpotFake	0.778	0.761	0.892	0.836	—	—
CLIP+MLP	0.857	0.853	0.887	0.886	0.870	0.845
CAFE-CLIP	0.879	0.857	0.897	0.896	0.882	0.856
FND-CLIP	0.902	0.896	0.907	0.907	0.875	0.857
MCAN-CLIP	0.917	0.911	0.900	0.899	0.882	0.861
BMR	0.883	0.870	0.889	0.889	0.863	0.830
Event-Radar	<u>0.928</u>	<u>0.923</u>	0.919	0.919	<u>0.901</u>	0.880
GLPN-LLM	0.888	0.890	0.867	<u>0.915</u>	0.865	0.907
DPC（本文）	0.938 ± 0.002	0.938 ± 0.002	<u>0.912 ± 0.001</u>	0.911 ± 0.001	0.912 ± 0.010	<u>0.897 ± 0.010</u>

加黑和下划线分别表示最高和次高报告值。DPC 结果报告三个随机种子的均值±样本标准差。对比方法结果来源于公开文献,缺失结果记为“—”。

表3 DPC 的分类别指标(三种子均值)

数据集	准确率	宏 F1	真实新闻			虚假新闻		
			准确率	召回率	F1	准确率	召回率	F1
Twitter	0.938	0.938	0.929	0.948	0.939	0.947	0.927	0.937
Weibo	0.912	0.911	0.892	0.919	0.905	0.930	0.905	0.917
PHEME	0.912	0.897	0.984	0.892	0.936	0.775	0.963	0.859

宏 F1 为逐种子计算后取均值,与由表中已舍入的两类 F1 直接平均所得结果可能存在 0.001 量级的差异。

0.775, 说明模型在检出较多虚假新闻的同时, 仍存在一定的误报问题。

总体而言, DPC 在三个公开数据集上均取得了具有竞争力的检测性能, 表明事件结构信息能够提供细粒度关系约束, 而内容信息能够补充整体语义证据, 二者协同建模有助于提升不完整事件条件下的多模态虚假新闻检测能力。

3.2.2 消融实验

表 4 给出了 DPC 各组成部分的消融结果, 包括结构推理分支、内容证据增强分支、交互机制以及融合策略的影响。

两条单分支模型均低于完整 DPC。其中, 仅内容证据增强模型在 Twitter、Weibo 和 PHEME 上的宏 F1 分别下降 2.5、2.1 和 2.4 个百分点, 说明标题语义、视觉对象以及外部知识能够提供事件结构之

外的重要判别依据; 仅事件结构推理模型分别下降 1.4、1.5 和 0.9 个百分点, 表明事件关系建模能够提供细粒度一致性约束。两类信息具有互补性, 联合建模能够覆盖不同粒度的判别线索。

在结构推理分支中, 去除三视图分离后, Twitter 和 Weibo 上的宏 F1 分别下降 5.9 和 3.3 个百分点, 说明分别建模文本事件关系、视觉事件关系和跨模态全局关系能够提升结构表示能力。去除显式交互和局部向量后, PHEME 上的宏 F1 分别下降 6.6 和 7.5 个百分点, 表明关系交互与局部残差信息对于复杂事件场景下的结构判别具有重要作用。

在内容证据增强分支中, 将融合视觉、文本侧特征的交叉注意力依次替换为拼接融合后, Twitter、Weibo 和 PHEME 上的宏 F1 都出现下降, 说明对象级选择性交互能够更有效地利用与文本相关的视觉证据以及百科知识。去除百科知识注意力后, 三个数据集上的宏 F1 分别下降 2.5、1.9 和 6.8 个百分点, 表明外部知识能够补充短标题中缺失的背景语义。

将可学习融合权重替换为静态权重后, 三个数据集上的宏 F1 分别下降 2.4、0.8 和 1.7 个百分点, 说明根据训练目标自动调整结构信息与内容信息贡献比例, 有助于实现两类信息的有效协同。

表4 消融实验结果(宏F1)

分组	变体	Twitter		Weibo		PHEME	
		Acc	F1	Acc	F1	Acc	F1
分支	仅内容证据增强	0.915	0.915	0.891	0.891	0.896	0.866
	仅事件结构推理	0.926	0.926	0.897	0.897	0.893	0.881
事件结构推理	去三视图分离	0.885	0.881	0.882	0.879	0.890	0.851
	去显式交互	0.910	0.907	0.894	0.893	0.887	0.824
	去局部向量	0.907	0.903	0.890	0.890	0.874	0.815
	去相对模态门控	0.912	0.907	0.891	0.887	0.893	0.850
	视觉侧拼接融合	0.913	0.910	0.882	0.885	0.892	0.873
内容证据增强	文本侧拼接融合	0.910	0.907	0.892	0.892	0.874	0.831
	去百科知识注意力	0.919	0.915	0.891	0.893	0.891	0.822
融合	可学习→静态	0.916	0.916	0.904	0.904	0.894	0.873
	完整DPC	0.940	0.940	0.912	0.912	0.905	0.890

各变体与完整DPC均为单次运行结果,完整DPC一行与表2的三种子均值存在细小差异。

3.2.3 机制验证分析

为验证DPC中相对模态门控和文本事件要素一视觉对象匹配机制的有效性,本文分别开展受控模态退化实验和匹配阈值分析。

在门控机制验证中,本文固定训练完成的DPC模型,仅在测试阶段对文本外围要素、视觉对象以及两侧信息分别进行50%随机删除,并以冻结干净输入门控权重的模型作为对照,结果如表5所示。

表5 受控模态退化实验结果

数据集	扰动	w _t	Δw_t	DPC 宏F1	对照F1
Twitter	文本要素	0.521	-0.002	0.916	0.917
Twitter	视觉对象	0.523	-0.0002	0.915	0.915
Twitter	两侧信息	0.520	-0.003	0.913	0.912
Weibo	文本要素	0.542	-0.004	0.885	0.886
Weibo	视觉对象	0.541	-0.005	0.887	0.886
Weibo	两侧信息	0.539	-0.007	0.881	0.881
PHEME	文本要素	0.529	-0.004	0.872	0.872
PHEME	视觉对象	0.531	-0.002	0.873	0.871
PHEME	两侧信息	0.530	-0.003	0.869	0.871

w_t表示相对模态门控中文本侧融合权重, Δw_t 表示退化前后该权重变化量。对照组固定干净输入下的门控权重。表5中的DPC与表4中的完整DPC使用同一模型。

由表5可知,三类退化均导致宏F1下降,双侧退化影响更明显。与固定干净输入下门控权重的对照组相比,各退化条件下DPC的宏F1差异均较小(绝对差值 ≤ 0.002),同时文本侧门控权重变化较小,未呈现明显的权重迁移趋势。因此,在本文实验设置下,相对模态门控主要体现为调节文本与视觉事件表示在结构一致性判别交互项中的相对贡献。现有结果不足以支持将该权重解释为模态质量或可靠性的直接估计。该结果说明,本文门控机制的作用主要在于协调不同信息来源在联合判别过程中的贡献,而非对单一模态质量进行显式预测。

除相对模态门控外,文本事件要素一视觉对象匹配策略也是结构传播的重要组成部分。为分析匹配阈值 τ_{match} 对候选对象筛选的影响,本文统计不同阈值下的匹配率与拒配率。

表6 不同匹配阈值下的匹配率与拒配率(%)

阈值 τ_{match}	Twitter		Weibo		PHEME	
	匹配率	拒配率	匹配率	拒配率	匹配率	拒配率
无阈值	100.00	0.00	100.00	0.00	100.00	0.00
0.20	90.23	9.77	96.29	3.71	89.78	10.22
0.30	7.12	92.88	5.39	94.61	8.67	91.33

匹配率表示被接受配对的要素占参与匹配要素的比例,拒配率表示最大余弦相似度低于阈值而未建立匹配关系的比例。

如表6所示,无阈值策略会将所有候选要素强制匹配,三个数据集匹配率均达到100%,但无法

过滤低相似度候选, 容易引入错误跨模态关联。当 $\tau_{\text{match}}=0.2$ 时, 三个数据集仍保持较高匹配覆盖率 (89.78%~96.29%), 同时能够拒绝 3.71%~10.22% 的低相似度候选, 在保证有效结构连接的同时降低噪声关联影响。当阈值提高至 0.3 时, 接受匹配比例仅为 5.39%~8.67%, 超过 90% 的候选关系被拒绝, 导致大量潜在事件关系无法参与后续结构传播。因此, 综合考虑跨模态关联覆盖与低质量匹配过滤能力, 本文采用 $\tau_{\text{match}}=0.2$ 。

3.2.4 事件结构建模与协同判别对照实验

为分析事件表示构建方式、结构传播过程以及结构与内容协同判别机制对检测性能的影响, 本文设置 E1~E4 四组对照。E1 采用依存句法规则替代 LLM 抽取, 其余保持不变; E2 仅利用 LLM 抽取结果进行分类, E3 在此基础上引入全图 GAT 传播, E4 采用三视图共享传播网络。各设置采用相同数据划分和训练预算, 结果如表 7 所示。

表 7 抽取方式、结构传播与协同判别对照实验(宏 F1)

设置	Twitter	Weibo	PHEME
E1: 依存规则抽取+DPC	0.913	0.887	0.841
E2: LLM 抽取+线性分类器	0.734	0.764	0.626
E3: LLM 抽取+全图 GAT+MLP	0.841	0.803	0.817
E4: LLM 抽取+三视图传播+MLP	0.875	0.846	0.861
完整 DPC	0.940	0.912	0.890

由表 7 可知, 不同结构建模设置之间存在明显性能差异。E2 仅利用 LLM 抽取结果进行分类, 三个数据集上的性能最低, 说明事件要素抽取结果仍

需要通过关系建模转化为有效结构表示。E3 引入全图 GAT 传播后, 宏 F1 均有所提升, 表明节点间关系传播能够增强事件关联建模能力。进一步地, E4 采用三视图共享传播网络后性能继续提升, 说明分别建模不同传播范围的事件关系能够提升结构表示能力。

相比 E4, 完整 DPC 进一步引入内容证据增强、相对模态门控和协同判别机制后, 三个数据集上的性能均进一步提升, 说明在事件结构表示基础上融合图文内容证据能够提供互补判别信息。此外, E1 相比完整 DPC 在 Twitter、Weibo 和 PHEME 上的宏 F1 分别低 2.7、2.5 和 4.9 个百分点, 表明在本实验设置下, 采用 LLM 进行事件结构判定与要素抽取的方案优于依存句法规则替代方案。

3.2.5 无谓词结构鲁棒性分析

根据 LLM 判定得到的谓词标签 C_{pred} , 本文按事件表达形式统计测试样本, 分别评估谓词结构和无谓词结构子集上的检测性能, 以分析 DPC 对不同事件表达形式的适应能力及无谓词回退策略和内容分支的作用。V1 为完整 DPC, 根据 C_{pred} 自适应选择谓词结构或主题—实体结构; V2 移除无谓词回退策略; V3 始终采用主题—实体结构, 并由 LLM 重新抽取全部样本的主题—实体表示; V4 仅保留结构分支。

如图 5 所示, 灰色柱和斜线柱分别表示谓词结构与无谓词结构子集上的宏 F1。V1 在 Twitter、Weibo 和 PHEME 的无谓词子集上分别取得 0.923、0.903 和 0.875 的宏 F1, 较 V2 分别提升 20.2、15.9 和 7.3 个百分点, 表明自适应选择事件表示形式能

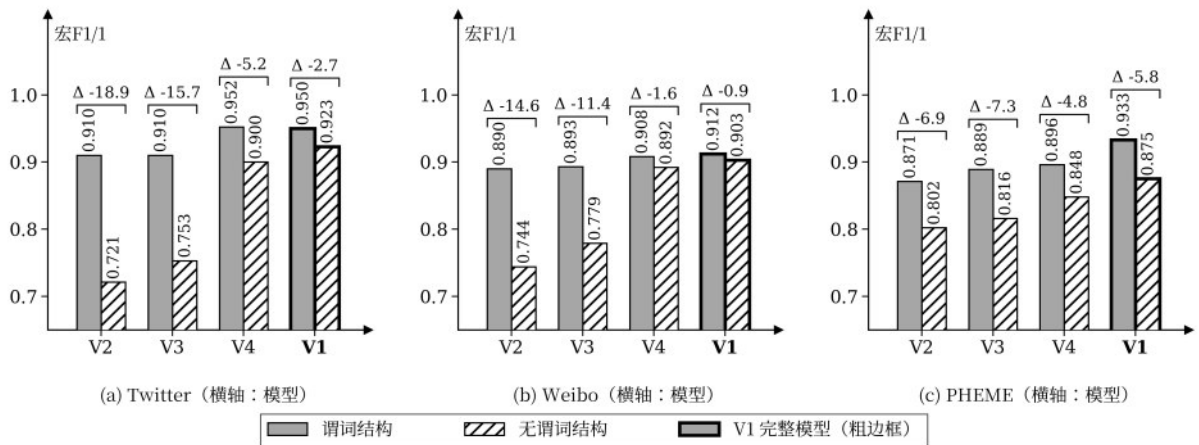


图 5 谓词结构与无谓词结构子集上的宏 F1 及组间差距

Δ为“无谓词结构宏 F1-谓词结构宏 F1”, 单位为百分点; V1 为完整 DPC (粗边框)。子集划分依据 LLM 结构标签。

够缓解谓词缺失对结构建模的影响。相比 V3, V1 在两类子集上均取得更高性能, 说明固定事件表示难以适应不同文本结构。

V1 在三个数据集上的组间差距 Δ 分别为 -2.7、-0.9 和 -5.8 个百分点。其差距绝对值均小于 V2, 但与 V3 相比, 仅在 Twitter 和 Weibo 上更小, 在 PHEME 上略大, 表明自适应事件表示的改善效果因数据集而异。此外, 相比 V4, V1 在三个无谓词子集上的宏 F1 分别提升 2.3、1.1 和 2.7 个百分点, 说明内容分支能够在结构证据不足时提供有效补充。

4 结束语

本文针对社交媒体短标题中事件要素表达不完整及事件结构与图文内容协同不足的问题, 提出了事件结构—内容协同检测方法 DPC。该方法根据谓词存在性自适应选择事件表示形式, 通过统一关系建模空间和多视角图注意力传播学习事件关系, 并利用相对模态门控调节文本与视觉事件表示的相对贡献, 结合标题语义、对象级视觉信息和百科描述补充内容判别依据。在 Twitter、Weibo 和 PHEME 三个公开数据集上的实验表明, DPC 取得了较好的检测效果, 尤其在无谓词结构子集上优于所比较的各变体, 验证了该方法对事件要素不完整情况的适应性。消融实验、机制验证实验及无谓词结构分析进一步验证了各组成模块的作用。但当前相对模态门控尚缺乏显式模态质量监督, 事件结构抽取的可靠性也有待进一步验证。未来将探索引入显式模态质量监督的融合方法, 扩大事件结构抽取的人工核验范围, 并开展更充分的跨数据集验证和统一实验比较。

参考文献:

- [1] Kumpel A S, Karnowski V, Keyling T. News sharing in social media: a review of current research on news sharing users, content, and networks [J]. *Social Media + Society*, 2015, 1(2): 1-14.
- [2] Khattar D, Goud J S, Gupta M, et al. MVAE: multimodal variational autoencoder for fake news detection[C]//The World Wide Web Conference. New York: ACM Press, 2019: 2915-2921.
- [3] Singhal S, Shah R R, Chakraborty T, et al. SpotFake: a multi-modal framework for fake news detection[C]//2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). Piscataway: IEEE Press, 2019: 39-47.
- [4] Li M, Xu R, Wang S, et al. CLIP-Event: connecting text and images with event structures[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2022: 16420-16429.
- [5] Ma Z, Luo M, Guo H, et al. Event-Radar: event-driven multi-view learning for multimodal fake news detection[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok: Association for Computational Linguistics, 2024: 5809-5821.
- [6] Qi P, Cao J, Li X, et al. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM Press, 2021: 1212-1220.
- [7] Ying Q, Hu X, Zhou Y, et al. Bootstrapping multi-view representations for fake news detection[C]//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2023: 5384-5392.
- [8] Wang Y, Ma F, Jin Z, et al. EANN: event adversarial neural networks for multi-modal fake news detection[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2018: 849-857.
- [9] Zhou X, Wu J, Zafarani R. SAFE: similarity-aware multi-modal fake news detection[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer, 2020: 354-367.
- [10] Chen Y, Li D, Zhang P, et al. Cross-modal ambiguity learning for multimodal fake news detection[C]//Proceedings of the ACM Web Conference 2022. New York: ACM Press, 2022: 2897-2905.
- [11] Wu Y, Zhan P, Zhang Y, et al. Multimodal fusion with co-attention networks for fake news detection[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online: Association for Computational Linguistics, 2021: 2560-2569.
- [12] Zhou Y, Yang Y, Ying Q, et al. Multimodal fake news detection via CLIP-guided learning[C]//2023 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE Press, 2023: 2825-2830.
- [13] Wang Y, Qian S, Hu J, et al. Fake news detection via knowledge-driven multimodal graph convolutional networks[C]//Proceedings of the 2020 International Conference on Multimedia Retrieval. New York: ACM Press, 2020: 540-547.
- [14] Hu B, Sheng Q, Cao J, et al. Bad actor, good advisor: exploring the role of large language models in fake news detection[C]//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2024: 22105-22113.
- [15] He J, Zhang H, Xiao Y, et al. FACTGUARD: event-centric and commonsense-guided fake news detection[C]//Proceedings of the 40th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2026: 363-371.
- [16] Hu S, Hu J, Zhang H. Synergizing LLMs with global label propagation for multimodal fake news detection[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vienna: Association for Computational Linguistics, 2025: 1426-1440.
- [17] Ferragina P, Scaiella U. TAGME: on-the-fly annotation of short text fragments (by Wikipedia entities)[C]//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2010: 1625-1628.
- [18] Zhang Q, Wu H, Zhang C, et al. Provable dynamic fusion for low-quality multimodal data[C]//Proceedings of the 40th International Con-

- ference on Machine Learning. Honolulu: PMLR, 2023: 41753-41769.
- [19] 熊熙, 刘钊荣, 张帅, 余艳. TextSwindler: 面向硬标签黑盒文本的对抗攻击算法[J]. 中文信息学报, 2024, 38(12): 18-29.
- Xiong X, Liu Z R, Zhang S, Yu Y. TextSwindler: hard-label black-box textual adversarial attack[J]. Journal of Chinese Information Processing, 2024, 38(12): 18-29.
- [20] 熊熙, 丁广政, 王娟, 张帅. DLSF: 基于双重语义过滤的文本对抗攻击方法[J]. 计算机科学, 2025, 52(10): 423-432.
- Xiong X, Ding G Z, Wang J, Zhang S. DLSF: a textual adversarial attack method based on dual-level semantic filtering[J]. Computer Science, 2025, 52(10): 423-432.
- [21] DeepSeek-AI. DeepSeek-V3 technical report[EB/OL]. (2024-12-27) [2026-09-11]. <https://arxiv.org/abs/2412.19437>.
- [22] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems 28. Red Hook: Curran Associates, 2015: 91-99.
- [23] Boididou C, Andreadou K, Papadopoulos S, et al. Verifying multimedia use at MediaEval 2015[C]//Working Notes Proceedings of the MediaEval 2015 Workshop. CEUR Workshop Proceedings, Vol. 1436. Aachen: CEUR-WS.org, 2015.
- [24] Jin Z, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM Press, 2017: 795-816.
- [25] Zubiaga A, Liakata M, Procter R. Exploiting context for rumour detection in social media[C]//Social Informatics. Cham: Springer, 2017: 109-123.
- [26] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the 38th International Conference on Machine Learning. [S.l.]: PMLR, 2021: 8748-8763.
- [27] Yang A, Pan J, Lin J, et al. Chinese CLIP: contrastive vision-language pretraining in Chinese[EB/OL]. (2022-11-02) [2026-09-11]. <https://arxiv.org/abs/2211.01335>.