联系电话：13121614763
联系邮件：zhangpengpengwsh@163.com

基于大语言模型与视觉语言模型的多模态事实核查

张芃芃[1]，彭勃[2*]，董晶[2]，程皓楠[3]
（1. 北华航天工业学院遥感信息工程学院，廊坊 065000；2. 中国科学院自动化研究所模式识别实验室，北京 100190；3. 中国传媒大学媒体融合与传播国家重点实验室，北京 100024）

摘要：多模态事实核查旨在联合多种模态的媒体内容以抽取有效信息来检测社交媒体背景下的虚假信息。针对已有研究对事实核查领域专用数据集过于依赖以及在图像理解和语义相似度计算方面可解释性弱的问题，提出了一种全新的基于预训练大模型的多模态事实核查自动化方法，并在公开数据集 COSMOS 上进行了实验。结果表明该方法达到了 0.859 的正确率，且在每次核查时都能提供清晰的理由，相较于传统的基线方法具有更高的准确性和更强的可解释性。此外，还深入分析了不同的方法变体，以及数据集中各种虚假信息的判别场景，验证了本方法凭借在多模态信息语义理解方面的强大能力，可以灵活应对不同情境下的脱离上下文(OOC, out-of-context)检测。本文方法为社交网络中多模态媒体内容的事实核查工作提供有力的技术支持和新的思考方向。

关键词：深度学习；大语言模型；视觉语言模型；多模态；事实核查
中图分类号：TP37　　**文献标识码：**A

Multimodal fact-checking based on large language models and vision language models

ZHANG Pengpeng[1], PENG Bo[2*], DONG Jing[2], CHENG Haonan[3]
(1. School of Remote Sensing Information Engineering, North China Institute of Aerospace Engineering, Langfang 065000, China; 2. New Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; 3. State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China)

Abstract: Multimodal fact-checking aims to combine multimodal media content to extract valid information to detect false information in the context of social media. Aiming at the problems of over-reliance on domain-specific datasets for fact verification and weak interpretability in terms of image understanding and semantic similarity comparison in

基金项目：国家重点研发计划（2021YFC3320103）；媒体融合与传播国家重点实验室（中国传媒大学）开放课题（SKLMCC2022KF002）；国家自然科学基金（62272460）；北京市自然科学基金（4232037）
作者简介(*为通讯作者)：张芃芃(1998-)，男，主要从事图像取证研究。Email:zhangpengpengwsh@163.com；彭勃(1991-)，男，博士，副研究员，主要从事计算机视觉、图像取证研究。Email:bo.peng@nlpr.ia.ac.cn；董晶(1983-)，女，博士，研究员，主要从事图像处理、模式识别、多媒体内容安全研究。Email:jdong@nlpr.ia.ac.cn；程皓楠(1994-)，女，博士，副研究员，主要从事音频合成、处理与鉴伪研究。Email: haonancheng@cuc.edu.cn

existing studies, this paper proposed a novel automated multimodal fact-checking method based on a pre-trained large model and conducted exhaustive experiments on the publicly available dataset COSMOS. The results show that the method achieves an accuracy of 0.859 and provides clear justifications in every verification, which provides higher accuracy and stronger interpretability compared to traditional baseline methods. In addition, this paper also deeply analyzed different method variants and various false information discrimination scenarios in the dataset, verifying that this method can flexibly cope with out-of-context (OOC) detection in different contexts by the strong capability in semantic understanding of multimodal information. With the continuous progress of large model technology in the future, the method proposed in this paper will show more excellent performance in the field of fact-checking, which provides strong technical support and a new idea for fact-checking of multimodal media content in social networks.
Keywords: deep learning; large language models (LLM); vision language models (VLM); multimodal; fact-checking

1 引言

互联网和社交媒体的快速发展带来更加便捷的信息获取和消费方式的同时，也为虚假信息的传播提供了便利。虚假的信息会错误地引导网民的认知，甚至制造恐慌，不利于舆论走向和社会安定。面对互联网中海量的媒体内容，受限于主要依赖于人工校验的传统虚假信息检测方法，极高的成本使错误的消息难以被及时发现与纠正。另一方面，在社交网络环境下，图像与文本两种媒体模态被大众广泛应用。文本内容能够全面地描述语义信息，而图像内容相对文本具有直观视觉信息和更强吸引力，因此联合不同模态的媒体内容以抽取更多有效信息逐渐受到国内外研究团队的关注。随着大数据与人工智能领域中大量关键技术的突破，诸多借助深度学习的多模态虚假信息检测方法被相继提出[1-3]，目前的主流研究可大致被分为基于内容模式的检测方法和事实核查(fact-checking)方法两个方向[4]。

基于内容模式的检测方法旨在利用深度学习强大的特征提取能力，分别从图像与文本中获取虚假信息所共有的模式(如写作风格、语言特点等)，然后进行信息真伪的二分类判别[5-7]。针对仅有的少数研究只是简单拼接多模态内容，难以实现特征深层次融合的问题，Jin 等[8]提出了一种具有注意力机制的 RNN 模型，以融合来自文本、图像和社会背景的特征来完成谣言检测任务，并在由微博和推特收集的两个多模态数据集上进行评估，验证了方法的有效性。为了挖掘模态之间的相关性，Khattar 等[9]利用多模式变分自动编码器来学习图文的共享表示，从而发现推特中不同模态数据之间的相关性，并将其与分类器耦合以检测虚假信息。尽管这类方法能够端到端地识别出新闻的真伪，但受限于训练数据集规模与质量和特定来源的信息特征，其判别泛化能力较弱。同时，新闻信息具有较强的时效性，仅利用历史数据训练的方法难以准确检测新出现的新闻。

为了应对上述问题，事实核查方法则首先识别出信息中的关键内容，然后在参考数据源中检索相关事实，再根据这些事实和细节来核验信息的真伪[10-12]。相对基于内容模式的检测方法，这可以显式地描述出判别阶段中的逻辑，并且减轻对信息中惯用词汇和行文风格的依赖，使检测过程更加符合实际。针对基于内容模式的检测方法数据集收集困难的问题，Müller-Budack 等[13]提出了一个不依赖于训练数据的多模态实体关系一致性检验系统。该系统首先通过命名实体链接从文本中提取实体信息，接着在搜索引擎中检索这些实体对应的参考图像，然后通过计算原信息中的图像与参考图像之间的相似度完成一致性判别。

本文所关注的脱离上下文(OOC, out-of-context)检测是一类常见的事实核查场景，在OOC 情况下同一幅真实图像被两个或更多网站解释成不同或相反的含义。由于图像无需经过编辑或篡改，造假成本低廉，在社交网络中更易产生误导作用。Aneja 等[14-16]收集了新闻频道和 Fact-checking 网站中由图像和标题文本组成的新闻，构建了 OOC 检测的公开数据集 COSMOS。在该数据集中，C1、C2 分别表示不同来源的两段文本，I 表示出现在这两个来源中的同一幅图像，例如图 1 中案例 1，虽然两段文本描述的目标一致，但语义上互斥排他，因此符合 OOC 情况；而案例 2 中的两段文本描述的目标并不一致，无法判断出信息是否虚假，因此这种情况被判定为非脱离上下文（NOOC, not out-of-context)。同时，该团队还针对 COSMOS 提出了一个自监督学习的检测方法。该方法分为训练和测试两个阶段，在训练阶段利用神经网络将数据集中每幅图像中典型目标与其对应的一段语义相关文本和一段语义无关文本分别编码提取特征，并在最大化目标与其语义相关文本的相似度的同时最小化目标与其语义无关文本的相似度。而在测试阶段，则将 C1、C2 同时编码后匹配图像 I 中相似度最大的目标，若两目标交并比（IoU）低于阈值 t_i，这意味着两段文本描述的对象不一致，因此直接将其判定为 NOOC；若两目标 IoU 高或等于 t_i，则需要进一步判断 C1、C2 语义相似度是否低于 t_s，若低则说明两段文本描述对象一致但语义不同，故判定为 OOC，反之则为 NOOC。值得注意的是，OOC 检测方法能够广泛应用和推广到其他设定的虚假信息检测任务中以识别图像错误解读（如旧图新用）的情况，如已知一组真实参考图文情况下，对另一使用了同一张图像的待测报道的文本真伪进行判别。

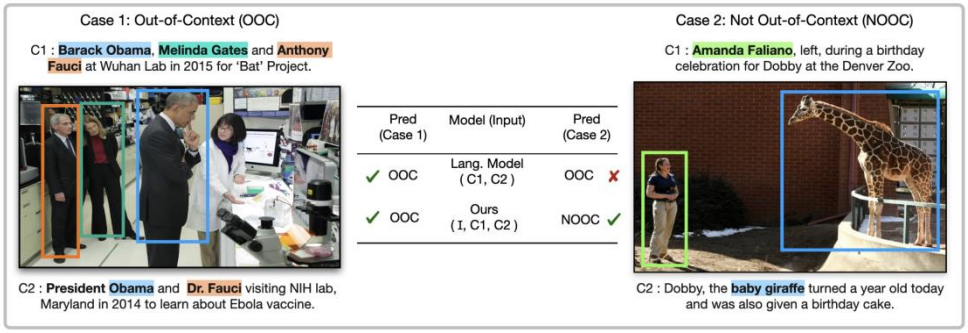

图 1 OOC/NOOC 示意图(图源自文献[14])

尽管一些学者[17-20]在图像特征提取和语义推理方面改进了 OOC 的检测方法使判别精度得到提升，该类方法尚需要依赖事实核查领域专用数据集，并且在图像理解和语义相似度计算方面可解释性仍然较弱。近年来，随着高性能计算和通用人工智能的发展，OpenAI[21-22]开发了一款出色的大语言模型 GPT-3.5 并引起轰动，其卓越的文本生成能力和深度理解能力在人工智能领域引起了广泛关注。该模型基于深度学习技术，通过训练海量的文本数据，能够捕捉上下文信息来理解深层次语义，并生成流畅的文本内容，从而实现对自然语言的高效处理，但大语言模型通常仅能处理文本。另一方面，针对现有不同种类媒体的单模态大模型难以实现有效对齐的问题， Li 等[23]提出了一个轻量级模型 Q-Former，该模型通过有效整合现有的图像与文本预训练单模态大模型，成功构建了多模态的视觉语言大模型 BLIP-2。这搭建起了多模态信息间的桥梁，为不同种类的预训练模型提供了灵活应用的基础。以上模型具有广泛的应用场景，这为准确、高效的事实核查提供了新的方式和可能。

基于上述调研与思考，本文提出一种基于预训练大模型的多模态事实核查方法，利用视觉语言模型 BLIP-2 和大语言模型 GPT3.5 在公开数据集 COSMOS 上进行了实验，详尽分析了数据集中出现错误解读的模式，经过系统测试不同图像描述算法、提示词(Prompt)设计方案和判断方式对预测结果的影响，最终取得了 0.859 的正确率(Accuracy)且能够在每次核查时给出理由，证明了其准确性和可解释性。随着大模型能力的不断增强，本文所提事

实核查方法的效果有望在未来进一步提升，这为社交网络中多模态媒体内容核查的实际应用提供了新思路。

2 方法设计

本文提出的多模态事实核查方法如图 2 所示，包括文本推理、图像语义理解和图文冲突检测三个模块。该方法能够在无可用训练数据集情况下，借助预训练大模型对给定的一幅图像 I 和两段文本 C1、C2 进行 OOC 检测。具体地，首先在文本推理模块中判断 C1、C2 在语义上是否矛盾以及是否一致，若确定矛盾且不一致则判定为 OOC，若确定一致且不矛盾则判定为 NOOC。对其他情况则进一步利用图像语义理解模块提取图像 I 的语义特征并表达成文本 C0，然后在图文冲突检测模块中结合 C0、C1 和 C2 最终确定判定结果，并给出判定过程的原因解释。

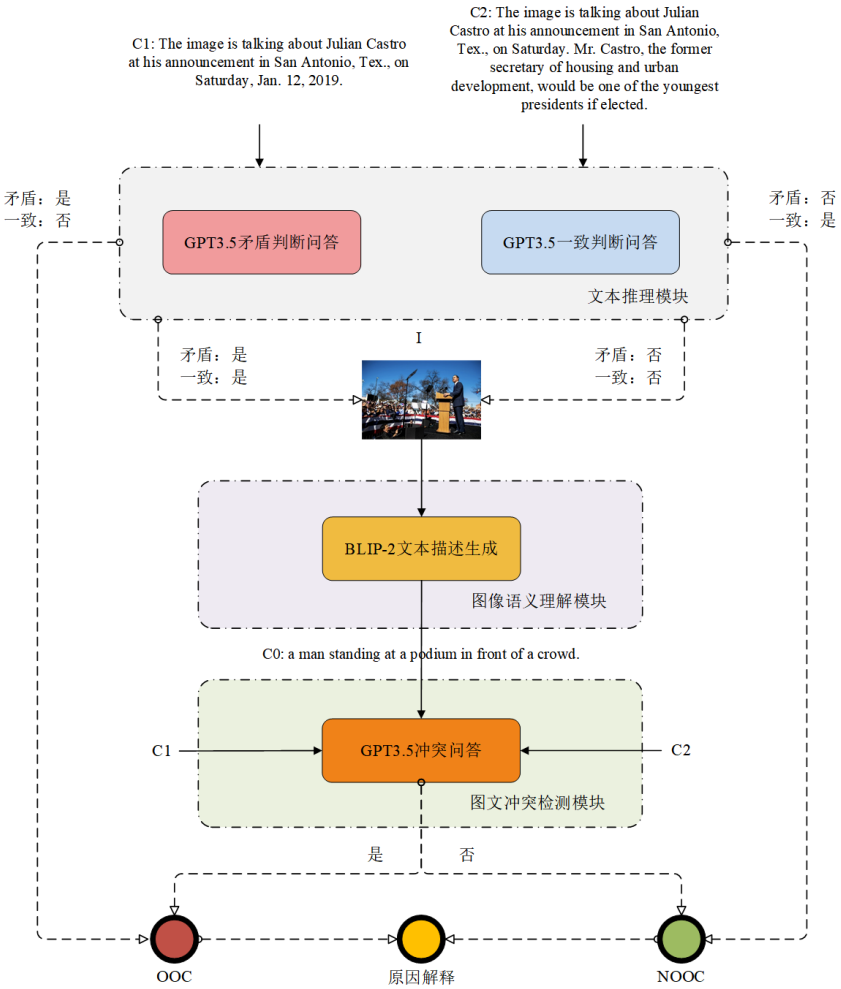

图 2 本文方法工作流图（实线表示数据流，虚线表示逻辑流）

2.1 文本推理模块

OOC 检测旨在判别 C1、C2 是否对同一目标解释成了不同或相反的含义，因此该过程主要依赖于文本信息。如果 C1、C2 在语义上的内容是矛盾且不一致的，则可以在无需利用图像的情况下直接将其判定为 OOC；同理，当 C1、C2 所描述的内容一致且不矛盾，则直接将其判断为 NOOC。为了满足语义关系推断的需求，文本推理模块针对目前最流行的大语言模型 GPT3.5 设计提示词，以获得两段文本间的矛盾和一致关系。为了让 GPT3.5 更精准地理解推理任务，引导其回复实际所需要的答案，提示词开头通常需要为模型设定身份扮演，然

后再提出具体问题，本方法采用的提示词设计如下：

a. 文本矛盾判断提示词：

“You are a detective now, I will tell two propositions and you will respond with whether the two propositions are contradictory.

Proposition A: [C1]

Proposition B: [C2]”

(“现在你是一名侦探，我将告诉你两个命题，你要回复这两个命题是否矛盾。

命题 A：[C1]

命题 B：[C2]”)

b. 文本一致判断提示词：

“You are a detective now. I will tell two propositions and you will respond with whether two propositions have the same semantics.

Proposition A: [C1]

Proposition B: [C2]”

(“现在你是一名侦探，我将告诉你两个命题，你要回复这两个命题是否一致。

命题 A：[C1]

命题 B：[C2]”)

其中[C1]、[C2]将分别被对应的文本所替代，在 GPT3.5 接收到提示词后会首先回复“Yes”或“No”字样以表达判断结果，接着解释作出该判断的理由。

2.2 图像语义理解模块

当文本内容无法直接判断出是否存在 OOC 情况时，则需要借助图像内容以获得判定依据。视觉语言模型 BLIP-2 不但在视觉问答领域表现优异，同时具备为图像生成文本描述的功能。因此为了充分利用图像信息，图像语义理解模块凭借 BLIP-2 预训练模型生成对图像 I 的文字描述 C0。

2.3 图文冲突检测模块

根据 OOC 检测的实际意义可知，其本质是在判定当两段文本解释同一幅图像时是否发生冲突。图文冲突检测模块借助 GPT3.5 强大的自然语言理解和逻辑推理能力，对仅利用文本信息未能辨别的情况进行 OOC 检测,提示词如下：

“I tell you the content of an image and two paragraphs of text. These two pieces of text are used to explain the content of the same image, and you have to judge whether the two interpretations of the same image conflict. {"Conflict": "<Yes/No>", "Explanation": "<>"}

Image: [C0]

Text 1: The image is talking about [C1]

Text 2: The image is talking about [C2]”

(“我告诉你一幅图像和两段文本，这两段文本被用来解释同一幅图像的内容，你需要判断对这张图像的两种解释是否冲突。 {“冲突”：“ < 是/否 > ”，“理由”：“<>”}

图像：[C0]

文本 1：[C1]

文本 2：[C2]”)

其中[C0]将被图像语义理解模块中获得的文本所替代，[C1]、[C2]将被原始输入的两段文本所替代。同样，当 GPT3.5 接收到提示词后会首先回复“Yes”或“No”字样以表达输入图文是否冲突的判断结果（若冲突则判定为 OOC，反之则判定为 NOOC），而后会自动解释作出该判断的理由。

3 实验与讨论

3.1 实验设置与评价指标

COSMOS 是一个用于 OOC 检测任务的公开数据集，由于本文方法无需训练，实验仅使用测试集 1000 幅图像与其对应的 2000 段文本。检测过程中视觉语言模型 BLIP-2 选用 blip2-flan-t5-xl 版本，大语言模型 GPT-3.5 使用 gpt-3.5-turbo 版本，其它参数在文本推理模块中均为默认，在图文冲突检测模块中设置 GPT-3.5 参数 temperature 为 0。OOC 检测是一种二分类任务，本文使用该研究领域常用 Accuracy 指标评估检测性能。本文实验基于 Python3.8 配置版本为 0.27.2 的 openai 库，并调用 ChatCompletion 类在线使用 ChatGPT 功能，能够全自动地对待测数据进行核查。

3.2 实验结果

在 COSMOS 测试集上对比了本文方法的两个变体和 Aneja 团队所提出的基线方法[14]，其中本文方法的变体方法包括：

（1）仅文本推理：为了使仅文本推理方法直接判断出输入文本是否存在冲突，实验采用以下提示词：

“I tell you two pieces of text. You have to judge whether the two texts conflict. Please answer in the following json format only. {"Conflict": "<Yes/No>", "Explanation": "<>"}

Text 1: [C1]

Text 2: [C2]”

(“我告诉你两段文本。你需要判断这两段文本是否冲突。请仅用以下 json 格式回答。{“冲突”：“<是/否>”，“理由”：“<>”}

文本 1: [C1]

文本 2: [C2]”)

其中[C1]、[C2]将被原始输入的两段文本所替代，当 GPT3.5 接收到提示词后会首先回复“Yes”或“No”字样以表达输入文本是否冲突的判断结果（若冲突则判定为 OOC，反之则判定为 NOOC），而后会自动解释作出该判断的理由。

（2）图像语义理解+图文冲突检测：为充分利用图像信息来判断输入信息是否存在冲突，实验先利用 BLIP-2 预训练模型生成对图像 I 的文字描述 C0，然后将 C0 和原始输入的两段文本同时输入 2.3 节所述的图文冲突检测模块中，以进行 OOC/NOOC 的判别与原因解释。

从表 1 的结果可以发现，基于 GPT3.5 的仅文本推理方法由于未利用图像视觉信息，其正确率略低于基线方法；先利用 BLIP-2 进行图像语义理解然后展开图文冲突检测的正确率与基线方法效果基本持平；而本文完整方法在正确率指标下优于基线方法和上述两个变体方法。这说明相较于基线方法，经过预训练的大模型能够更加深入地理解输入图像与文本间的语义关系，并且本文结合文本推理、图像语义理解、图文冲突检测的方法流程比变体方法具有更好效果，提示词设计方案达到较好水平，能够准确实现 OOC 检测。

表 1 不同方法在 COSMOS 数据集上的结果

方法 \ 评价指标	正确率
基线方法[14]	0.836
仅文本推理	0.797
图像语义理解+图文冲突检测	0.830
本文完整方法	**0.859**

3.3 实例分析

通过总结归纳 COSMOS 测试集样本，发现 OOC 检测可分为以下两种情况，分别是两段文本内容对应图像中的不同目标和两段文本内容对应图像中的相同目标, 其中后者两段文本的关系又可分为一致、否定和补充三类。本文方法凭借强大的多模态信息语义理解能力，能够在不同复杂情境下完成 OOC 检测任务。

3.3.1两段文本内容对应图像中的不同目标

COSMOS 测试集中编号 960 样本（见图 3），真值标签为 NOOC：

I:

图 3 COSMOS 测试集中编号 960 样本图像

C1: Women stand in line wearing jewellery, blue dresses and multicoloured headwraps in Mali

C2: On the same day, gold miners wait backstage at an event in Mali

(C1：在马里，妇女们戴着珠宝、身穿蓝色连衣裙、头戴五颜六色的头饰排队

C2：同一天，在马里的一场活动中，金矿工人在后台等待)

本文方法首先在文本推理模块中利用文本信息判断样本文字的语义关系，在文本矛盾判断时得到回答“No, these two propositions are not contradictory. They simply describe different groups of people and their activities in Mali.”（“不，这两个命题并不矛盾。它们只是描述了马里的不同人群及其活动。”），在文本一致判断时得到回答“No, Proposition A and Proposition B do not have the same semantics.”（“不，命题 A 和命题 B 的语义不相同。”），这说明 C1、C2 不矛盾且不一致。因此需要在图像语义理解模块中生成对图像的文本解释“women in blue turbans standing in a tent”（“戴着蓝色头巾的妇女站在帐篷里”），然后利用图文冲突检测模块获得最终 OOC 判别结果“{"Conflict": "No", "Explanation": "Both texts mention Mali and the image shows women wearing blue turbans, which could be interpreted as headwraps. The difference is that text 1 mentions gold miners waiting backstage, while text 2 mentions women wearing jewelry and dresses. However, these details do not necessarily conflict with each other as they could both be happening at the same event or in the same location."}”（“{“冲突”：“否”，“理由”：“两段文本都提到马里，图像显示的是戴着蓝色头巾的妇女，这可以被解释为头饰。不同的是，文本 1 提到了在后台等待的金矿工人，而文本 2 提到了戴着珠宝、穿着连衣裙的妇女。然而，这些细节并不一定相互冲突，因为它们可能发生在同一活动或同一地点中。”}”）。

由结果可知，本文方法能够完成非同一目标描述场景下的 NOOC 判定，并指出 C1、C2 描述重点以及 NOOC 判定原因。

3.3.2 两段文本描述相同目标

① 两文本描述具有一致关系

COSMOS 测试集中编号 11 样本（见图 4），真值标签为 NOOC：

I:

图 4 曼谷被闲置的用来运送游客的嘟嘟车

C1: Tuk-tuk car parked in Bangkok because there are no tourists

C2: In Bangkok, tuk-tuks that used to transport tourists around the city are seen idle because of travel bans and border closures during the pandemic.

(C1：由于没有游客，嘟嘟车停在曼谷

C2：在曼谷，由于疫情期间的旅行禁令和边境关闭，过去用来运送游客的嘟嘟车被闲置。)

本文方法首先在文本推理模块中利用文本信息判断样本文字的语义关系，在文本矛盾判断时得到回答“The two propositions are not contradictory, but they convey the same information in different ways. Both propositions suggest that tuk-tuk cars in Bangkok are not being used for tourism purposes due to the pandemic.”（“这两个命题并不矛盾，但它们以不同的方式传达了相同的信息。这两个命题都表明，由于疫情原因，曼谷的嘟嘟车没有被用于旅游目的。”），在文本一致判断时得到回答“Yes, both propositions have the same semantics. They convey the same meaning that tuk-tuk cars in Bangkok are parked and not being used due to the absence of tourists caused by the travel bans and border closures related to the pandemic.”（“是的，这两个命题具有相同的语义。它们传达了相同的含义，即由于与疫情相关的旅行禁令和边境关闭，导致没有游客，曼谷的嘟嘟车停着，无人使用。”），这说明 C1、C2 不矛盾且一致，因此可直接将该样本分为 NOOC。

可见，本文方法在两文本具有相同语义的情况下对一致关系作出了准确判断。

② 两文本描述具有否定关系

COSMOS 测试集中编号 3 的样本（见图 5），真值标签为 OOC：

I:

图 5 COSMOS 测试集中编号 3 样本图像

C1: A shocking report about the former child actor Kirk Cameron was just another hoax from a well-known purveyor of satire.

C2: Actor, musician, director and devoted follower of Christ, Kirk Cameron, has been stricken with an extremely rare form of what can only be described as a biblical disease.

(C1：一篇关于前儿童演员柯克·卡梅隆的令人震惊的报道只是一家知名讽刺作家的又一场骗局。

C2：演员、音乐家、导演和基督的忠实追随者柯克·卡梅伦患上了一种极其罕见的疾病，这种疾病只能被描述为圣经中的疾病。)

本文方法首先在文本推理模块中利用文本信息判断样本文字的语义关系，在文本矛盾判断时得到回答“These two propositions are contradictory. In Proposition A, it is stated that the report about Kirk Cameron was a hoax, while Proposition B claims that he has a rare disease.”（“这两个命题是矛盾的。在命题 A 中，关于柯克·卡梅伦的报道是一场骗局，而 B 命题则声称他患有罕见疾病。”），在文本一致判断时得到回答“No, these two propositions do not have the same semantics as they are talking about different situations. Proposition A talks about a hoax report involving Kirk Cameron, while Proposition B talks about Kirk Cameron being affected by a rare biblical disease.”（“不，这两个命题的语义不同，因为它们谈论的是不同的情况。命题 A 谈论的是一份涉及柯克·卡梅伦的恶作剧报告，而命题 B 谈论的是柯克·卡梅伦受到一种罕见的圣经疾病的影响。”），这说明 C1、C2 矛盾并且 C1 是对 C2 的否定，此时可直接将该样本判定为 OOC。

因此，本文方法能够理解文本间的否定语义，从而指出 C1、C2 中存在的 OOC 问题。

③ 两文本描述具有补充关系

COSMOS 测试集中编号 0 样本（见图 6），真值标签为 NOOC：

I:

图 6 COSMOS 测试集中编号 0 样本图像

C1: Julian Castro at his announcement in San Antonio, Tex., on Saturday, Jan. 12, 2019.

C2: Julian Castro at his announcement in San Antonio, Tex., on Saturday. Mr. Castro, the former secretary of housing and urban development, would be one of the youngest presidents if elected.

(C1：朱利安·卡斯特罗于 2019 年 1 月 12 日星期六在德克萨斯州圣安东尼奥市发表声明。

C2：周六，朱利安·卡斯特罗在德克萨斯州圣安东尼奥市发表声明。如果当选，前住房和城市发展部长卡斯特罗将成为最年轻的总统之一。)

本文方法首先在文本推理模块中利用文本信息判断样本文字的语义关系，在文本矛盾判断时得到回答“The two propositions are not contradictory.”（“这两个命题并不矛盾。”），

在文本一致判断时得到回答“No, Proposition A is simply stating the time and location of Julian Castro's announcement, while Proposition B is making a statement about Julian Castro's potential presidency. These are two different propositions with different meanings.” (“不，命题 A 只是简单地说明朱利安·卡斯特罗发表声明的时间和地点，而命题 B 则是关于朱利安·卡斯特罗可能担任总统的陈述。这是两个不同的命题，有着不同的含义。”)，这说明 C1、C2 不矛盾且不一致。因此需要在图像语义理解模块中生成对图像的文本解释“a man standing at a podium in front of a crowd”（“一个站在人群面前的讲台上的男人”），然后利用图文冲突检测模块获得最终 OOC 判别结果“{"Conflict": "No", "Explanation": "Both texts are talking about the same event and person, and provide additional details such as the location and date of the announcement. They do not conflict with each other."}” (“{“冲突”：“否”，“理由”：“两个文本都在谈论同一事件和同一个人，并提供了其他细节，如发布的地点和日期。它们彼此不冲突。”}”)。因此在本例中，本文方法能够发现输入文本间的补充关系，并完成准确的 NOOC 判定。

此外，互为补充关系的两文本除了可能出现上述例子中语义上的并存情况外，若补充的内容不符合事实，则可能出现具有排他性的文本内容，导致 OOC 的出现。总体来看，本文方法可针对不同语义情况作出较好的 OOC/NOOC 判定，并给出对判定过程的解释。

4 总结与展望

本文提出了一种基于预训练大模型的多模态事实核查方法，并利用该法在公开数据集 COSMOS 上进行了实验，取得了 0.859 的 Accuracy 且能够在每次核查时给出理由，相对基线方法具有更高的准确性和可解释性。同时，本文总结归纳了数据集中不同情况的虚假信息判别场景，证明了本文方法凭借强大的多模态信息语义理解能力，能够在不同复杂情境下完成 OOC 检测任务。

值得注意的是，在两文本构成补充关系的情况下，相关信息是否具有排他性则需要根据实际情况具体判断，但就目前来看，这些细节信息通常难以被全面收集，未来应从证据链构建与参考图文检索方面展开深入研究。此外，当样本被分为 NOOC 时，仅能说明根据当前参考信息无法判断出不同来源信息存在冲突，但两文本均不真实等其他情况仍然需要进一步讨论，因此如何更加周密具体地对多模态事实核查任务进行建模和分析是目前仍待解决的关键应用问题。

参考文献（References）：

[1] Zlatkova D, Nakov P, Koychev I. Fact-checking meets fauxtography: verifying claims about images[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 2099–2108.

[2] Sabir E, AbdAlmageed W, Wu Y, et al. Deep multimodal image-repurposing detection[C]//Proceedings of the 26th ACM International Conference on Multimedia, 2018: 1337-1345.

[3] Luo G, Darrell T, Rohrbach A. Newsclippings: automatic generation of out-of-context multimodal media[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021: 6801–6817.

[4] 杨昱洲, 周杨铭, 应祺超, 等. 基于事实信息核查的虚假新闻检测综述[J]. 中国传媒大学学报(自然科学版), 2023, 30(06): 28-36.

[5] Wang Y, Ma F, Jin Z, et al. Eann: event adversarial neural networks for multi-modal fake news detection[C]// KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 849-857.

[6] Zhang D Y, Shang L, Geng B, et al. Fauxbuster: a content-free fauxtography detector using social media comments[C]// 2018 IEEE International Conference on Big Data (Big Data), 2018: 891-900.

[7] Shang L, Zhang Y, Zhang D, et al. Fauxward: a graph neural network approach to fauxtography detection using social media comments[J]. Social Network Analysis and Mining, 2020, 10: 76.

[8] 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, 2017: 795-816.

[9] Khattar D, Goud J S, Gupta M, et al. Mvae: multimodal variational autoencoder for fake news detection[C]// WWW '19: The World Wide Web Conference, 2019: 2915-2921.

[10] Sun W, Fan Y, Guo J, et al. Visual named entity linking: A new dataset and a baseline[DB/OL]. arXiv:2211.04872, 2022.

[11] Tahmasebzadeh G, Kacupaj E, Müller-Budack E, et al. GeoWine: geolocation based wiki, image, news and event retrieval[C]//Proceedings of the 44th international ACM SIGIR Conference on Research and Development in Information Retrieval, 2021: 2565-2569.

[12] Guo Z, Schlichtkrull M, Vlachos A. A survey on automated fact-checking[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 178-206.

[13] Müller-Budack E, Theiner J, Diering S, et al. Multimodal news analytics using measures of cross-modal entity and context consistency[J]. International Journal of Multimedia Information Retrieval, 2021, 10: 111-125.

[14] Aneja S, Bregler C, Nießner M. Cosmos: catching out-of-context misinformation with self-supervised learning[DB/OL]. arXiv:2101.06278, 2021.

[15] Aneja S, Midoglu C, Dang-Nguyen D T, et al. MMSys' 21 grand challenge on detecting cheapfakes[DB/OL]. arXiv:2107.05297, 2021.

[16] Aneja S, Midoglu C, Dang-Nguyen D T, et al. Acm multimedia grand challenge on detecting cheapfakes[DB/OL]. arXiv:2207.14534, 2022.

[17] Akgul T, Civelek T E, Ugur D, et al. Cosmos on steroids: a cheap detector for cheapfakes[C]// MMSys '21: Proceedings of the 12th ACM Multimedia Systems Conference, 2021: 327-331.

[18] Tran Q T, Tran T P, Dao M S, et al. A textual-visual-entailment-based unsupervised algorithm for cheapfake detection[C]// MM '22: Proceedings of the 30th ACM International Conference on Multimedia, 2022: 7145-7149.

[19] La T V, Dao M S, Le D D, et al. Leverage boosting and transformer on text-image matching for cheap fakes detection[J]. Algorithms, 2022, 15(11): 423.

[20] Zheng P, Chen H, Hu S, et al. Few-shot learning for misinformation detection based on contrastive models[J]. Electronics, 2024, 13(4): 799.

[21] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]// NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, 159: 1877-1901.

[22] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[DB/OL]. arXiv:2303.08774, 2023.

[23] Li J, Li D, Savarese S, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models[C]// ICML'23: Proceedings of the 40th International Conference on Machine Learning, 2023, 814: 19730-19742.

编辑：王谦