

基于立场分析的脱离上下文虚假信息检测方法

袁鑫, 郭捷, 邱卫东, 黄征

(上海交通大学网络空间安全学院, 上海 200240)

摘要: 随着基于人工智能的自动生成技术的发展, 虚假信息以更具欺骗性的形式出现, 对经济和社会秩序造成了巨大危害。脱离上下文虚假信息是最具欺骗性和最容易实现的虚假信息类型之一, 恶意攻击者通过歪曲真实图像中的人物、事件、地点等上下文信息来增强虚假叙事的可信度。针对已有检测算法严重依赖知识库、未能全面考虑待检测信息与互联网证据之间的立场关系的问题, 提出了一种基于立场分析的脱离上下文虚假信息检测方法。对于需要检测的图像-标题对以及在互联网上检索到的文本证据与视觉证据, 该方法根据标题与文本证据中命名实体的共现关系, 为每条文本证据计算一个立场增益分数; 使用互相独立的立场分析网络对图像与视觉证据、标题与文本证据分别进行层次聚类, 并完成基于多个注意力机制与立场分析模块的语义立场表征抽取; 根据语义比较与立场分析的结果对图像-标题对的真实性进行预测。实验结果表明, 得益于立场分析的引入, 所提方法能够显著提高检测效果, 相比其他使用互联网证据进行检测的最佳算法准确率提升了2.3%。

关键词: 脱离上下文; 立场分析; 虚假信息; 互联网证据

中图分类号: TP309

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024017

Out-of-context misinformation detection method based on stance analysis

YUAN Xin, GUO Jie, QIU Weidong, HUANG Zheng

School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract: As artificial intelligence-based automatic generation technology advances, the emergence of misinformation in more sophisticated guises has become a significant threat to the economy and social order. Among its forms, out-of-context misinformation stands out as particularly deceptive and readily executable. This type of misinformation involves malicious actors enhancing the credibility of false narratives by misrepresenting contextual details such as individuals, events, and locations within real images. To address the shortcomings of current detection algorithms,

收稿日期: 2023-05-29; 修回日期: 2023-07-24

通信作者: 郭捷, guojie@sjtu.edu.cn

基金项目: 国家自然科学基金 (No.61972249)

Foundation Item: The National Natural Science Foundation of China (No.61972249)

引用格式: 袁鑫, 郭捷, 邱卫东, 等. 基于立场分析的脱离上下文虚假信息检测方法[J]. 网络与信息安全学报, 2024, 10(02): 133-142.

Citation Format: YUAN X, GUO J, QIU W D, et al. Out-of-context misinformation detection method based on stance analysis [J]. Chinese Journal of Network and Information Security, 2024, 10(02): 133-142.

which heavily depend on knowledge bases and often overlook the stance relationship between the information under scrutiny and online evidence, a stance analysis-based out-of-context misinformation detection method was developed. This method involved several steps for the detection of an image-caption pair along with the corresponding textual and visual evidence retrieved from the Internet. Initially, a stance gain score for each piece of textual evidence was calculated based on the co-occurrence of named entities. Subsequently, independent stance analysis networks were utilized to perform hierarchical clustering on both the image and visual evidence, as well as on the caption and textual evidence. This process involved the extraction of semantic stance representations, facilitated by multiple attention mechanisms and a stance analysis module. The authenticity of the image-caption pair was subsequently predicted based on the outcomes of semantic comparison and stance analysis. Experimental results indicate that the incorporation of stance analysis significantly enhances the method's detection capabilities. Specifically, the accuracy of this method outperforms the state-of-the-art algorithm that employs Internet evidence for detection by 2.3%.

Keywords: out-of-context, stance analysis, misinformation, internet evidence

0 引言

互联网的发展以及智能通信设备的普及加速了信息化时代的到来,新闻媒体与个人用户越来越多地在网络上发布新闻内容或社交信息。互联网以其便捷且近乎实时的传播方式在向受众提供信息内容时,也促进了虚假信息的产生和泛滥,对经济和社会秩序造成了巨大危害。其中脱离上下文虚假信息通过歪曲真实图像中出现的人、发生的事件或其他上下文信息使得真实图像支持虚假叙事,脱离上下文虚假信息示例如图1所示。

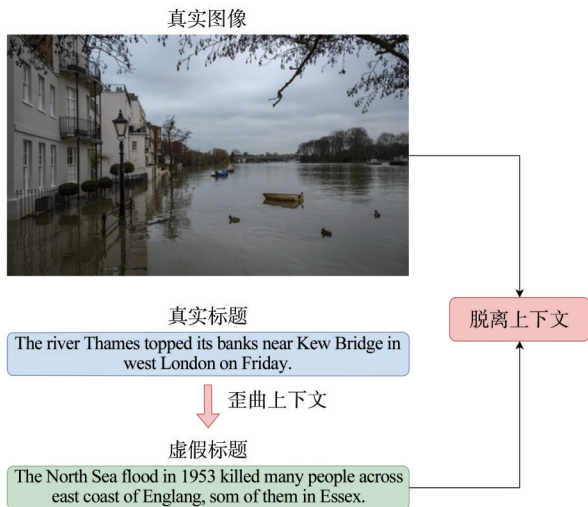


图1 脱离上下文虚假信息示例
Figure 1 An example of out-of-context misinformation

网络上,尤其是新闻中的图像和文本之间的对齐是抽象的,单独观察图像时很难获知其要表达的具体事件,而文本中携有的精确信息则对图像的语义起到了消歧作用^[1]。因此当歪曲文本中

的上下文信息时,即便歪曲是不精细的,不了解具体事件的受众也极易相信其内容的真实性从而受到欺骗。脱离上下文虚假信息具有编造难度低、成本低、迷惑性高等特征,如何对网络上大量存在且持续高速增长的此类虚假信息进行精确检测是一个亟待解决的问题。

目前,脱离上下文虚假信息的检测方法主要分为基于参考数据集的方法、基于信息本身特征的方法以及基于互联网证据的方法三大类。给定需要检测的图像-标题对,基于参考数据集的方法^[2-4]在参考数据集中检索相关信息来验证图像-标题对真实性,但对于现实场景,假设存在一个包含有相关真实信息的参考数据集是不现实的;基于信息本身特征的方法^[5-6]仅依赖图像-标题对本身的语义一致性进行检查,适用范围广,但精妙设计的虚假信息难以被检出;基于互联网证据的方法^[7-8]舍弃了参考数据集,根据图像-标题对中的相关信息在互联网上搜索文本证据或视觉证据来验证图像-标题对真实性,与人工验证的逻辑相符,精确率较高,但现有算法存在的问题在于,未能有效建模证据与待验证信息之间的立场关系,从而忽视了具有不同立场的证据对判断信息是否脱离上下文这一任务的重要性。

为了解决上述问题,本文设计了立场增益分数(stance gain score, SGS)和立场分析网络(stance analysis network, SAN),在脱离上下文虚假信息检测过程中引入立场分析,本文主要贡献如下。

(1) 通过立场增益分数和立场分析网络对脱

离上下文虚假信息和互联网证据之间的立场关系进行全面建模, 有效地保证重要证据的立场表达。

(2) 本文基于公开数据集进行性能评估, 实验结果表明, 相对于目前的检测算法, 本文方法不仅实现了显著的性能提升, 还在多个脱离上下文场景中达到了最高准确度。

1 相关工作

1.1 虚假新闻检测

大多数虚假新闻检测算法仅针对纯文本的情况进行处理。一些研究基于知识库或知识图谱^[9], 通过事实核查的方式对虚假新闻进行检测, 但是对多模态信息的事实核查方法则鲜有研究。近年来, 越来越多的学者对多模态情况下的虚假新闻检测问题进行探索, 这需要梳理多模态的特征进行语义建模^[10-12]。Xue等^[13]将文本特征和视觉语义特征映射到同一语义空间, 以获得跨模态特征表示, 并考虑它们之间的一致性。Sun等^[14]设计了一个双不一致性网络, 同时检测图像与标题之间的跨模态不一致性以及图像一标题对与背景知识之间的内容知识不一致性。

虚假新闻的自动生成对检测提出了更高要求^[15]。Tan等^[16]使用GROVER^[17]和实体感知图像字幕模型^[18], 分别生成虚假文章和虚假图像字幕。InfoSurgeon^[19]通过对知识图谱执行不同类型的编辑操作并使用BART^[20]生成文本来获得知识元素级别的假新闻。人们无法准确分辨出人工写成的文章和机器自动生成的文章, 这使得伪造具有一定欺骗性的虚假信息的成本降低而效率提升, 亟须具有高精度且实时的检测算法对这种自动化的威胁进行防御。

1.2 图像重用与脱离上下文虚假信息

早期与脱离上下文相似的一种虚假信息形式为图像重用。多模态信息篡改数据集^[2]与多模态实体图像重用数据集^[3]分别通过图像文本随机匹配与命名实体替换的方式获取伪造样本。Jaiswal等^[4]使用对抗性的方法同时训练一个伪造元数据进行图像重用的恶意行为者和对图像及元数据进行一致性验证的检测器来进行重用行为检测。然而, 上述3个研究都假设存在一个可供使用且为真的参考数据集, 这对于实际场景是不现实的。

Müller-Budack等^[7]同样采用对文本中命名实体进行操作的方式来获得图像一文本语义不一致的样本, 但细致的替换策略使得不具备背景知识的人很难察觉, 在检测方面其使用文本中提取出的命名实体在网络上搜索视觉证据, 以验证图像与视觉证据之间的语义一致性。为了消除前述几种方法由于操纵文本中的命名实体以获得不一致性而引起的语义偏差, Luo等^[2]通过(contrastive language image pre-training, CLIP)^[21]和其他模型自动检索不相关但令人信服的图像与标题进行匹配。Abdelnabi等^[8]提出了多模态循环一致性核查的概念, 分别根据图像/标题在互联网上搜索文本证据/视觉证据, 对标题/图像进行验证。

1.3 立场检测

立场检测是一个分类问题, 其分类结果为“赞成”“反对”“中立”或其他类似形式^[22]。Zubiaga等^[23]论证了立场有助于谣言检测。一些工作试图将立场检测引入虚假信息检测过程中, 但主要集中在纯文本情况下。在多模态情况下, Yao等^[24]在声明验证阶段生成证据的立场表示, 以预测三分类标签。然而, 至今没有任何研究将立场检测全面引入脱离上下文虚假信息的检测中。

2 基于立场分析的模型设计

2.1 概述

给定一个待检测的图像一标题对 (I^c, C^c, L^c) , 其中 I^c 为图像, C^c 为标题, L^c 为真实性标签, 文本证据集合 $T^c = [T_1^c, T_2^c, \dots, T_M^c]$ 和视觉证据集合 $V^c = [V_1^c, V_2^c, \dots, V_N^c]$, 其中 M 为文本证据数目, N 为视觉证据数目, 本文的目标是根据 I^c 、 C^c 、 T^c 和 V^c 精确预测标签 L^c , 即构建一个满足 $f(I^c, C^c, T^c, V^c) \rightarrow L^c$ 的模型 $f(\cdot)$ 。

基于立场分析的脱离上下文虚假信息检测方法框架如图2所示。视觉立场分析网络与文本立场分析网络分别负责图像与视觉证据、标题与文本证据之间的语义比较以及立场分析, 在文本立场分析网络中, 本文根据命名实体的共现关系为每一条文本证据计算一个立场增益分数, 并将其融合到文本证据的特征中, 最后根据立场分析网络输出预测图像一标题对的真实性标签。

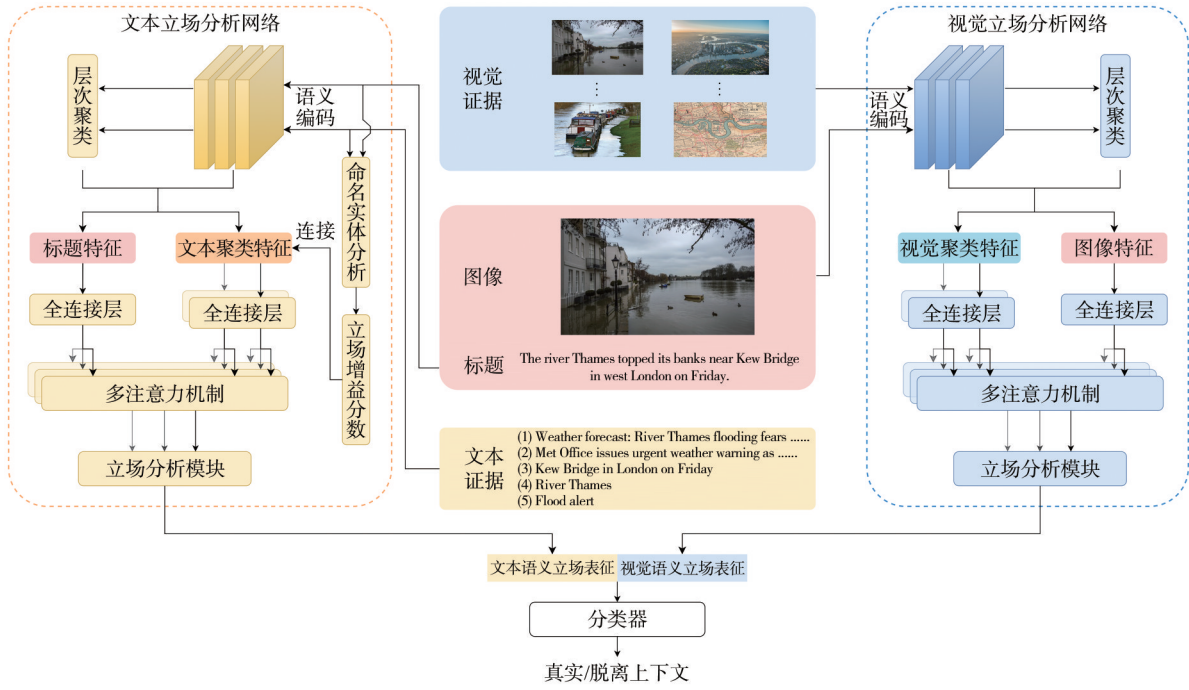


图2 基于立场分析的脱离上下文虚假信息检测方法框架

Figure 2 The framework of out-of-context misinformation detection method based on stance analysis

2.2 立场增益分数

脱离上下文虚假信息相对于其他类型的虚假信息的一个特征为，其上下文中的命名实体很大程度上代表了其语义信息。例如，信息 X 中包含命名实体“苹果、库克”，信息 Y 中包含命名实体“南美洲、美洲杯”，信息 Z 中包含命名实体“苹果公司”，则 X 与 Z 描述的重点都为苹果公司，从而二者在表述同一事件的概率要远大于 X 与 Y 在表述同一事件的概率。有些工作使用简单的二元信息^[8,16]来表示这一特征，未能考虑复杂情形下的特征挖掘，本文设计了立场增益分数来对这种特征进行建模，初步检测脱离上下文的情况。

本文使用 Spacy 命名实体识别对标题 C^e 和文本证据 T_i^e 进行命名实体识别，并得到 C^e 的命名实体集合 E^e 和每条文本证据 T_i^e 的命名实体集合 E_i^e 。对所有 E_i^e 进行统计，得到按频率非递增排序的命名实体集合 E^e ，其中次序可以反映出命名实体的重要程度，即出现次数越多的命名实体是重要信息的概率更大，反之则更有可能是噪声，从而是重要信息的概率更小。按命名实体的对应关系对 E^e 、 E_i^e 和 E^c 进行如式(1)所示的共现关系分析，得到图3中所示的共现关系矩阵 M^{co} ，其中 ent_j 表示集合中第 j 个命名实体。SGS计算

模块如图3所示。

$$M_{ij}^{co} = \begin{cases} 1, & \text{ent}_j \in E_i^e \text{ 且 } \text{ent}_j \in E^c \\ 0, & \text{ent}_j \notin E_i^e \text{ 或 } \text{ent}_j \notin E^c \end{cases} \quad (1)$$

对于第 i 条文本证据 T_i^e ，本文通过式(2)、式(3)计算其立场增益分数。

$$E_i = E_i^e \cap E^c \quad (2)$$

$$\text{SGS}_i = \#E_i + \zeta \cdot \sum_{\text{ent} \in E_i} e^{-\sqrt{\text{idx}_{\text{ent}}}} \quad (3)$$

其中， E_i 表示 $M_{i,j}^{co}$ 中为1的命名实体集合， $\#E_i$ 表示 E_i 中命名实体的数量， idx_{ent} 为命名实体 ent 在 E^e 中的次序。每个同时出现在 T_i^e 和 C^e 中的命名实体，对于SGS除了贡献一个固定的基准增益1外，还按照其次序反映的重要程度贡献一个额外的小幅度次序增益，重要程度与次序增益呈正相关。 ζ 为缩放因子，用来控制次序增益对整体分数的影响幅度。

之后，将 SGS_i 与 T_i^e 在立场分析网络中编码得到的特征在特征维上以连接的方式进行融合，得到 T_i^e 的最终特征表示如图3所示。

SGS是一个有效的对脱离上下文虚假信息与文本证据之间立场关系进行表示的方式，然而本文同样希望检测一种更为复杂的场景，在此场景中，存在对应的命名实体但图像一标题对脱离上

下文。本文设计了立场分析网络来综合考虑图像-标题对与文本、视觉证据之间的语义立场关系。

2.3 立场分析网络

本文对图像声明与视觉证据、标题声明与文本证据使用互相独立但结构相同的立场分析网络进行语义立场表征。以单模态内的待验证声明与证据之间的立场分析为例，本文设计的立场分析网络如图4所示。

对于图像与视觉证据，本文使用在 ImageNet 上预训练的 ResNet152^[25]模型和在 Places365^[26]上预训练的 ResNet50 模型分别提取通用语义特征与场景语义特征。对于标题与文本证据，本文使用预训练的 Sentence Transformer^[27]模型来进行语

义特征编码。

本文对声明和证据进行基于欧氏距离的层次聚类，按照相同的聚类阈值，证据被分为与声明在同一个聚类之内的支持聚类、包含证据数量最多的稠密聚类和剩余证据组成的补充聚类3个分簇。其中，支持聚类与声明语义相近，可以起到支持立场作用，但其有可能为空；稠密聚类为证据中表达相似语义最多的一组聚类，最能代表所有证据的语义；补充聚类中的信息可以对语义与立场分析起到补充作用。

聚类之后本文仿照文献[28]中对多句子处理的方式，使用全连接层处理后的声明特征作为注意力机制中的 Query，同时使用另外2个互相独立的全连接层对证据特征进行处理，分别得到注

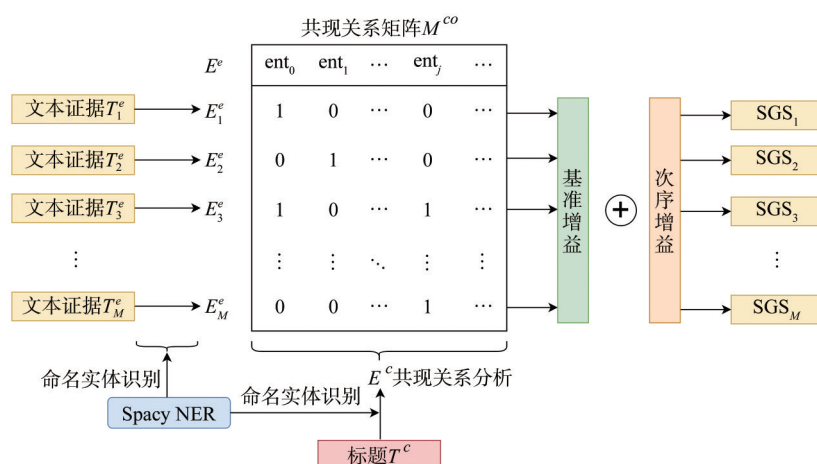


图3 SGS计算模块
Figure 3 SGS calculation module

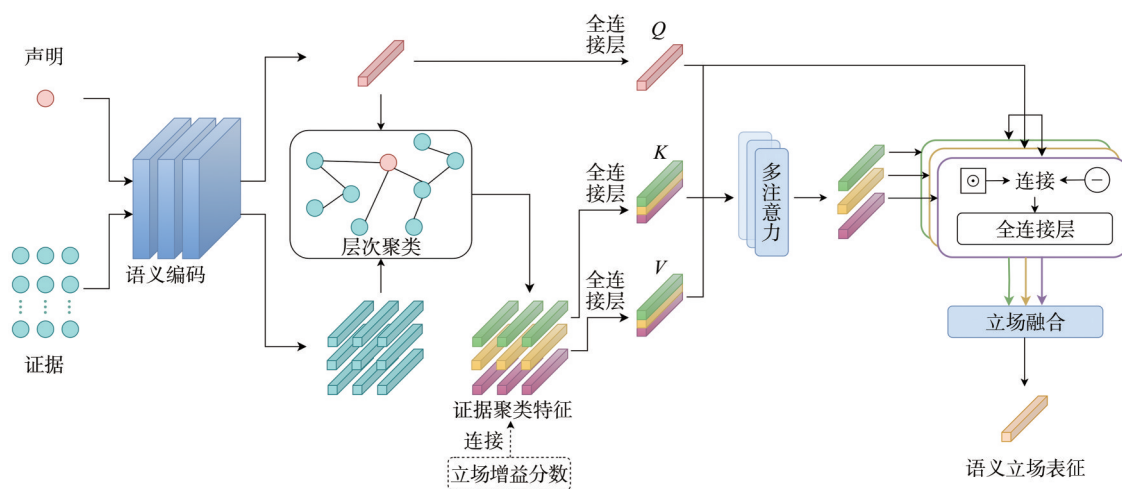


图4 立场分析网络
Figure 4 Stance analysis network

注意力机制中的 Key 和 Value, 表达式中分别用 Q 、 K 、 V 表示。

本文使用互相独立的多个注意力机制处理 3 个聚类, 基于 softmax 与相同的 Q 分别进行权值计算并进行加权求和的求取如式 (4) 所示, 其中 p 表示上述 3 个不同聚类。

$$H_p = \text{softmax}(QK_p^T)V_p, p \in \{1, 2, 3\} \quad (4)$$

然后, 本文对 3 个聚类的立场进行提取、表示和融合, 如式 (5) ~ 式 (7) 所示。

$$\widehat{H}_p = [Q \odot H_p : Q - H_p], p \in \{1, 2, 3\} \quad (5)$$

$$\widetilde{H}_p = W_p \widehat{H}_p + b_p, p \in \{1, 2, 3\} \quad (6)$$

$$H = W[\widetilde{H}_1 : \widetilde{H}_2 : \widetilde{H}_3] + b + Q \quad (7)$$

其中, $\odot$ 为按元素乘, $[\cdot]$ 为连接操作。 W_p 和 b_p 、 W 和 b 在视觉和文本的立场分析网络中具有不同的维度。概括来说, 若 $H_p \in \mathbf{R}^d$, 则 $W_p \in \mathbf{R}^{d \times 2d}$, $b_p \in \mathbf{R}^d$, $W \in \mathbf{R}^{d \times 3d}$, $b \in \mathbf{R}^d$, 在视觉和文本立场分析网络中 d 分别为 1 024 和 768。

得到的 H 为视觉或文本语义立场表征。本文使用 Abdelnabi 等^[8]所采用的分类器的形式, 即对输入进行批标准化后经过一个 2 层的全连接层构成的分类器对结果进行预测, 本文根据所有 SAN 输出的语义立场表征对分类器进行训练, 得到能够同时考虑语义信息与立场信息的分类器。

3 实验

3.1 实验数据集

本文使用数据集 NewsCLIPPings^[5]来衡量本文方法的性能。给定一个原始的图像-标题对 (cap1, img1), 其根据不同策略检索另一个具有不同上下文的图像-标题对 (cap2, img2) 从而匹配出脱离上下文的图像-标题 (cap1, img2)。根据不同的策略, 数据集可以分为 4 个不相交的子集, 分别代表 4 个不同的脱离上下文场景。

(1) 场景 1: Semantics/CLIP Text-Image, 使用 CLIP^[21]检索出与 cap1 具有最高语义相似度的 img2。

(2) 场景 2: Semantics/CLIP Text-Text, 使用 CLIP 检索出与 cap1 具有最高语义相似度的 cap2。

(3) 场景 3: Person/SBERT-WK Text-Text, 使用 SBERT-WK^[29]检索出与 cap1 含有相同人员实体且上下文相似度最低的 cap2。

(4) 场景 4: Scene/ResNet Place, 使用基于 Places365^[26]预训练的 ResNet50^[25]检索出与 img1 场景相似度最高的 img2。

Merged/Balanced 子集将来自前面 4 个子集相等数量的样本进行混合, 以模拟更真实的场景, 其训练集包含 71 072 个图像-标题对, 验证集包含 7 024 个图像-标题对, 测试集包含 7 264 个图像-标题对, 本文使用 Merged/Balanced 子集来进行实验。

对于外部证据, 本文使用 Abdelnabi 等^[8]检索结果。给定一个标题, 其使用谷歌检索最多 10 张图像证据。给定一张图像, 其检索若干命名实体证据和若干句子证据。

3.2 实验设置

在计算立场增益分数的过程中, ζ 设置为 1。在立场分析网络中, 设置图像与视觉证据的聚类阈值 $\tau_i = 0.166$, 标题与文本证据的聚类阈值 $\tau_t = 0.500$ 。

本文使用 PyTorch 来实现上述模型。实验环境为 Ubuntu18.04, Nvidia GeForce RTX 3090。训练模型时设置批大小为 64, 迭代轮次为 60, 优化器使用 Adam, 使用基础学习率 0.000 009、最大学习率 0.000 06 的循环学习率, 损失函数为 Cross-Entropy Loss。

3.3 实验结果

将本文方法与现有其他算法在准确率 (Accuracy, ACC) 指标下进行对比, 模型性能比较结果如表 1 所示。其中为了控制变量以分析本文方法对于检测效果的提升, 一致性检查网络 (consistency-checking network, CCN)^[8]与本文方法都采用 Sentence Transformer 进行文本特征提取, 同时本文方法加入了 CCN 所设置的限制条件与增益, 如证据过滤限制, 以及 CLIP 微调模型增益、图像 Label 增益、Domain 增益。

在 5 种检测方法中, CCN 与本文方法使用了外部证据对脱离上下文虚假信息进行验证, 准确率明显高于 VisualBERT^[30]、自监督蒸馏学习 (self-supervised distilled learning, SSDL)^[6]、CLIP 等不使用外部证据的方法, 说明了相较于依赖提取特征的大模型中隐含的外部知识, 外部证据更加可靠。此外, VisualBERT 和 SSDL 都有过度预测真

实标签的趋势，CLIP则出现了过度预测虚假标签，即脱离上下文标签的现象，这是由于News-CLIPpings使用若干模型来进行细致地匹配，许多虚假的样本都是高度迷惑性的。而在具备外部证据的情况下，CCN与本文方法对两类标签的预测准确度大致相同，表明外部知识能够有效地对真实样本和虚假样本中的上下文进行验证。

表1 模型性能比较结果
Table 1 Model performance comparison results

模型	外部证据	所有数据 准确率	真实数据 准确率	虚假数据 准确率
VisualBERT	×	58.6%	78.4%	38.9%
SSDL	×	65.5%	68.7%	62.2%
CLIP	×	66.1%	64.2%	68.1%
CCN	√	83.9%	82.8%	84.9%
本文方法	√	86.2%	86.1%	86.3%

与CCN相比，本文方法在所有数据上的准确率提升了2.3%，其中在真实数据上的准确率提升了3.3%，高于在虚假数据上提升1.4%。通过全面建模立场关系，本文提出的立场增益分数和立场分析网络更加高效地挖掘了外部证据与待验证信息之间存在的语义关联，从而能够更好地预测真实性标签。

本文将CCN算法与本文方法进行对比，通过比较二者在Merged/Balanced子集中来自4个场景的数据上的表现，来分析二者在不同脱离上下文场景中的性能，不同场景检测准确率对比结果如图5所示。

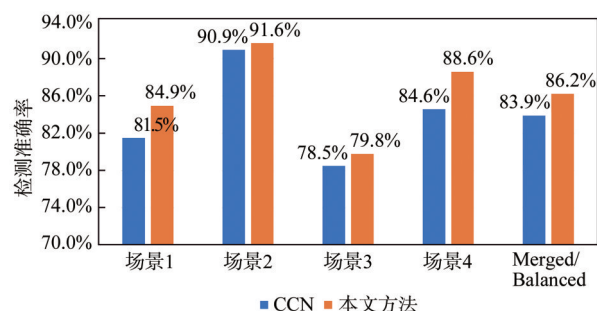


图5 不同场景检测准确率对比

Figure 5 Comparison of detection accuracy in different scenarios

相比于CCN，本文方法在场景1和场景4下准确率提升显著，分别提升了3.4%和4.0%。CCN和本文方法在场景2下均表现优异，都超过

了90.0%，但本文方法取得了准确率0.7%的提升。而在场景3（即来自Person/SBERT-WK Text-Text子集的数据上），准确率提升了1.3%，但2种方法在此场景下的准确率均显著低于其他3个场景，可以看出，场景3最具挑战性，这与Luo等^[5]对数据集的评估结论一致，即Person子集中数据的构建思路为将同一个人描述在另外一个上下文场景中，此场景的问题需要具有更强语义理解能力的模型以及更全面的证据去解决。

3.4 消融实验

为了分析立场增益分数与立场分析网络分别为检测方法带来的性能提升，本文进行了消融实验，实验中各个消融后的网络结构如下。

(1) Base：不使用立场增益分数与立场分析网络，相当于不使用Binary命名实体识别的CCN模型。

(2) SAN：使用视觉和文本立场分析网络做性能提升。

(3) SGS+SAN（视觉）：使用立场增益分数和视觉立场分析网络做性能提升。

(4) SGS+SAN（文本）：使用立场增益分数和文本立场分析网络做性能提升。

(5) 本文方法：使用立场增益分数和视觉及文本立场分析网络做性能提升。

消融实验结果如表2所示。当分别去除立场增益分数、文本立场分析网络、视觉立场分析网络时，在各类数据上的检测性能都出现了下降，其中所有数据准确率会下降0.8%~1.3%，即立场增益分数和视觉及文本的立场分析网络都对检测结果产生了提升效果。而当去除本文提出的所有方法后，所有数据准确率只有80.8%，相当于本文提出的方法在其基础上提升了5.4%，进一步验证了本文方法的有效性。

表2 消融实验结果
Table 2 Results of ablation experiment

模型	所有数据 准确率	真实数据 准确率	虚假数据 准确率
Base	80.8%	81.0%	80.6%
SAN	84.9%	84.0%	85.8%
SGS+SAN(视觉)	85.4%	84.7%	86.1%
SGS+SAN(文本)	85.1%	84.5%	85.6%
本文方法	86.2%	86.1%	86.3%

3.5 定性分析

部分预测样例如图6所示, 包括本文方法3个成功预测例子和1个错误预测的例子, 最后一列给出了本文方法的预测结果及预测是否正确的判断。可以看出, 在前2个例子中, 本文方法从文本与视觉证据中找到了与图像一标题对相对应的元素, 如文本证据中的 Australia、Totoyo、The Stone Roses, 视觉证据中的汽车、标志、聚会等相同或相似的元素, 从而对图像一标题对形成了支持关系。在第3个例子中, 文本证据高频出现的 Germanwings 与视觉证据中高频出现若干橙衣警察对图像一标题对形成了驳斥关系。第4个例子来自前述所说的 Person 子集, 其中文本证据没

有提供更多可以用来判断的信息, 视觉证据与图像都为 Cook 在苹果发布会上的照片, 并没有证据指出这是不同的发布会, 由此造成了预测错误, 进而说明证据的质量是影响预测性能的重要因素。

4 结束语

本文提出一种基于立场分析的脱离上下文虚假信息检测方法, 能够有效地应对此类虚假信息增长的威胁。视觉立场分析网络与文本立场分析网络的使用除了能够分别对图像与视觉证据、标题与文本证据进行语义比较与立场分析外, 其结构也保证了重要证据的立场表达, 文本立场分析

图像-标题对	文本证据	视觉证据	预测结果
 The Toyota Australia plant in Altona Melbourne	['Factory', 'Manufacturing', 'Automotive industry'] (1) End of Australian car manufacturing could cost up to 39,000 jobs Australia news (2) Toyota's assembly line in Altona, Melbourne. Photograph: Joe Castro/AAP (3) Toyota factory (4) Joe Hockey: union 'war with Toyota' is threatening carmaker's future Australia news (5) Toyota and Holden factories to close, end of the line for autoworkers - ABC News	   	真实 ✓
 Ian Brown and Reni of the Stone Roses headline the Media Stage on day 1 of the V Festival at Hylands Park in Chelmsford	['V Festival 2012', 'The Stone Roses', 'Singer-songwriter'] (1) V Festival day 1: Ian Brown and Reni of The Stone Roses headline (2) The V festival in pictures Music (3) Ian Brown Reni V Festival 2012 - Day 1 (4) Ian Brown, Reni - Ian Brown and Reni Photos	   	真实 ✓
 Policemen walk in front of the chalet Milky Way in the Swiss mountain resort of Gstaad where film director Roman Polanski remains under house arrest	['Germanwings', 'First officer', 'French Alps', 'Alps'] (1) Germanwings Flight 4U9525 co-pilot Andreas Lubitz 'hid illness' from employer CBC News (2) Police searched the Dusseldorf apartment of Germanwings co-pilot Andreas Lubitz on Thursday. (3) Prosecutor: Passengers shrieked as jet crashed into Alps - USA Today (4) A police officer stands in front of an apartment building Thursday where the co-pilot of the crashed Germanwings jet may have lived in Dusseldorf, Germany.	   	脱离上下文 ✓
 Apple CEO Tim Cook speaks during an Apple event announcing the iPhone 6 and the iPhone 6 Plus at the Flint Center in Cupertino California	['Projection screen', 'Flat-panel display', 'Wallpaper', 'Computer', 'Multimedia', 'Computer Monitor', 'Projector']	   	真实 ✗

图6 部分预测样例(仅显示部分证据)
Figure 6 Some prediction examples (only showing partial evidence)

网络中还通过加入基于命名实体共现关系计算的立场增益分数来进一步强调立场关系, 最终通过语义比较与立场分析的结果对图像-标题对的真实性进行预测。实验结果表明, 相较于现有的检测方法, 本文方法在检测效果上取得了显著提升。针对此类虚假信息检测的复杂性, 后续还有一些研究方向可以探索, 具体如下。

(1) 在检测方法中引入人员身份识别, 可以有效地解决一些难以判断的样本。

(2) 如何解决虚假信息, 尤其是新闻中图像与文本的抽象关联一直是一个难题, 这需要更加细粒度的语义建模、更加丰富的外部知识以及更加复杂的关联模型来进行探索。

参考文献:

- [1] THOMAS C, KOVASHKA A. Preserving semantic neighborhoods for robust cross-modal retrieval[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2020: 317-335.
- [2] JAISWAL A, SABIR E, ABDALMAGEED W, et al. Multimedia semantic integrity assessment using joint embedding of images and text[C]//Proceedings of the 25th ACM international conference on Multimedia. New York: ACM Press, 2017: 1465-1471.
- [3] SABIR E, ABDALMAGEED W, WU Y, et al. Deep multi-modal image-repurposing detection[C]//Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM, 2018: 1337-1345.
- [4] JAISWAL A, WU Y, ABDALMAGEED W, et al. AIRD: adversarial learning framework for image repurposing detection[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 11322-11331.
- [5] LUO G, DARRELL T, ROHRBACH A. NewsCLIPpings: automatic generation of out-of-context multimodal media[C]//Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 6801-6817.
- [6] MU M, DAS BHATTACHARJEE S, YUAN J S. Self-supervised distilled learning for multi-modal misinformation identification[C]//Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2023: 2818-2827.
- [7] MÜLLER-BUDACK E, THEINER J, DIERING S, et al. Multi-modal analytics for Real-world news using measures of Cross-modal entity consistency[C]//Proceedings of the 2020 International Conference on Multimedia Retrieval. New York: ACM, 2020: 16-25.
- [8] ABDELNABI S, HASAN R, FRITZ M. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 14920-14929.
- [9] THORNE J, VLACHOS A, CHRISTODOULOPOULOS C, et al. FEVER: a large-scale dataset for fact extraction and VERification[C]//Proceedings of the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg: Association for Computational Linguistics, 2018: 809-819.
- [10] IMRAN M, OFLI F, CARAGEA D, et al. Using AI and social media multimodal content for disaster response and management: opportunities, challenges, and future directions[J]. Information Processing & Management, 2020, 57(5).
- [11] WANG Y Q, MA F L, JIN Z W, et al. EANN: event adversarial neural networks for multi-modal fake news detection[C]//Proceedings of the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM press, 2018: 849-857.
- [12] ZHOU X Y, WU J D, ZAFARANI R. $\mathcal{SAFE}$: similarity-aware multi-modal fake news detection[M]. Springer International Publishing, 2020: 354-367.
- [13] XUE J X, WANG Y B, TIAN Y C, et al. Detecting fake news by exploring the consistency of multimodal data[J]. Information Processing & Management, 2021, 58(5).
- [14] SUN M Z, ZHANG X, MA J Q, et al. Inconsistency matters: a knowledge-guided dual-inconsistency network for multi-modal rumor detection[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 1412-1423.
- [15] 李婧文, 李雅文. 深度合成技术应用与风险应对[J]. 网络与信息安全学报, 2023, 9(2): 184-190.
- LI J W, LI Y W. Application and risk response of deep synthesis technology[J]. Chinese Journal of Network and Information Security, 2023, 9(2): 184-190.
- [16] TAN R, PLUMMER B, SAENKO K. Detecting cross-modal inconsistency to defend against neural fake news[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 2081-2106.
- [17] ZELLERS R, HOLTZMAN A, RASHKIN H, et al. Defending against neural fake news[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: 2019: 9054-9065.
- [18] BITEN A F, GOMEZ L, RUSIÑOL M, et al. Good news, everyone! context driven entity-aware captioning for news images [C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 12458-12467.
- [19] FUNG Y, THOMAS C, GANGI REDDY R, et al. InfoSurgeon: cross-media fine-grained information consistency checking for fake news detection[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the

- 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 1683-1698.
- [20] LEWIS M, LIU Y H, GOYAL N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 7871-7880.
- [21] RADFORD A, KIM J W, HALLACY C, et al. Learning transfer-able visual models from natural language supervision [C]//Proceedings of the International Conference on Machine Learning. Virtual Event: PMLR, 2021: 8748-8763.
- [22] ALDAYEL A, MAGDY W. Stance detection on social media: state of the art and trends[J]. Information Processing & Management, 2021, 58(4).
- [23] ZUBIAGA A, AKER A, BONTCHEVA K, et al. Detection and resolution of rumours in social media: a survey[J]. ACM Computing Surveys, 51(2): 32.
- [24] YAO B M, SHAH A, SUN L C, et al. End-to-end multimodal fact-checking and explanation generation: a challenging dataset and models[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM press, 2023: 2733-2743.
- [25] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.
- [26] ZHOU B L, LAPEDRIZA A, KHOSLA A, et al. Places: a 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1452-1464.
- [27] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using siamese BERT-Networks[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). Stroudsburg: Association for Computational Linguistics, 2019: 3982-3992.
- [28] SUKHBAATAR S, SZLAM A, WESTON J, et al. End-to-end memory networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. New York: ACM, 2015: 2440-2448.
- [29] WANG B, KUO C C J. SBERT-WK: a sentence embedding method by dissecting BERT-based word models[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020(28): 2146-2157.
- [30] LI L H, YATSKAR M, YIN D, et al. VisualBERT: a simple and performant baseline for vision and language[EB]. arXiv preprint arXiv: 1908.03557. 2019.

[作者简介]



袁鑫 (1999-), 男, 山东济南人, 上海交通大学硕士生, 主要研究方向为信息安全、多媒体安全。



郭捷 (1976-), 女, 河南南阳人, 上海交通大学副研究员, 主要研究方向为多媒体安全、人工智能安全、大数据隐私保护。



邱卫东 (1973-), 男, 江西修水人, 上海交通大学教授、博士生导师, 主要研究方向为计算机取证、密码分析、人工智能安全、大数据隐私保护。



黄征 (1975-), 男, 四川南充人, 上海交通大学副教授, 主要研究方向为人工智能、隐私保护计算。