基于多模态对比学习的短视频虚假信息早期识别方法

程浩天
广东外语外贸大学，广东 广州 510006
DOI:10.61369/ASDS.2026070009

摘　　要： 短视频虚假信息早期识别面临信息不完整与模态关联模糊的双重困境。多模态对比学习通过拉近同一样本跨模态表示、推远不同样本表示，能够有效捕捉模态间一致性线索。针对短视频发布初期的稀疏模态特征，构建多粒度对比学习框架，分别从实例级、模态级与时序级强化特征可判别性。提出跨模态对齐机制与轻量化快速推理策略，在公开数据集上验证了该方法在早期识别时效性与准确性上的优势。实验结果表明对比增强特征显著提升了虚假信息检测的鲁棒性。

关 键 词： 短视频虚假信息；早期识别；多模态对比学习；跨模态对齐

Early Detection of Misinformation in Short Videos Based on Multimodal Contrastive Learning

Cheng Haotian
Guangdong University of Foreign Studies, Guangzhou, Guangdong 510006

Abstract: Early detection of misinformation in short videos faces the dual challenges of incomplete information and ambiguous modal correlations. Multimodal contrastive learning effectively captures inter-modal consistency cues by pulling together cross-modal representations of the same sample and pushing apart those of different samples. To address the sparse modal features during the early stages of video posting, this paper proposes a multi-granularity contrastive learning framework that enhances feature discriminability at the instance level, modality level, and temporal level. A cross-modal alignment mechanism and a lightweight fast inference strategy are introduced, demonstrating the method's advantages in both timeliness and accuracy for early detection on public datasets. Experimental results show that contrast-enhanced features significantly improve the robustness of misinformation detection.

Keywords: short video misinformation; early detection; multimodal contrastive learning; cross-modal alignment

引言

短视频平台信息传播速度快、覆盖面广，虚假信息在发布后数分钟内即可形成大规模影响[1]。传统虚假检测方法依赖完整视频内容或大量用户反馈，无法满足早期干预需求。早期识别只能获取标题、封面图、前几秒关键帧及片段音频，模态缺失与跨模态矛盾成为主要障碍。对比学习在无监督表示学习中展现了强大的跨模态对齐能力，将其引入虚假信息检测可量化模态一致性。现有研究多关注图文虚假新闻，对短视频多模态时序特性挖掘不足。本文提出基于多粒度对比学习的早期识别方法，设计模态自适应训练策略，解决信息稀疏条件下的特征融合问题，为短视频平台提供可部署的快速检测方案。

一、短视频虚假信息早期识别的多模态对比学习建模

（一）问题形式化与定义

短视频虚假信息早期识别任务定义为在视频发布后短时窗口内，基于部分多模态数据输出真伪标签。设短视频样本包含视觉流、文本流、音频流3个模态。早期识别仅能获取子集，其中视觉子集为封面图与前若干个关键帧，文本子集为标题与光学字符识别文字，音频子集为前若干秒语音片段[2]。检测模型需要从早期子集映射到真实或虚假的二元标签。与完整视频检测不同，早期识别面临信息不完整性、模态组合随机性、时间窗口刚性约束。

作者简介：程浩天，广东外语外贸大学，本科生，研究方向为计算机科学与技术。

信息不完整性指大量语义线索位于后期内容中，模态组合随机性反映平台上传格式差异，时间窗口刚性约束要求模型推理延迟低于短视频传播爆发阈值。虚假信息制造者常利用早期信息的片段性，在开头放置真实内容而后续编造谎言，或在封面与标题中制造误导性关联。这使得模型必须在极有限的信息中捕捉到哪怕细微的跨模态矛盾。形式上定义信息完整度比率不超过0.3时仍有较高检测性能。另外引入时间敏感因子，规定从视频上传到模型输出判定的延迟不超过传播半衰期的十分之一，短视频传播半衰期通常在5至15分钟之间，因此响应时间应小于30秒。

（二）多模态对比学习基础

对比学习通过构造正负样本对学习判别性表示。在跨模态场景中，正样本对为同一短视频的不同模态编码，负样本对为不同短视频的相同模态编码。采用对比损失函数最大化正样本对互信息下界。该机制强制模型捕捉模态间共享语义，过滤模态特有噪声。将对比学习嵌入虚假检测任务，真实短视频的跨模态表示趋于一致，虚假短视频则因人为拼接、音画不同步、图文矛盾等操作呈现较低的对齐得分。进一步分析梯度特性，对比损失对正样本对的梯度方向是缩小其距离，对负样本对的梯度方向是推开其表示，这天然适合检测跨模态不一致性。与传统的三元组损失相比，对比损失利用整个批次构造负样本，避免了难例挖掘的额外开销[3]。

（三）早期识别场景下的挑战与对比学习适配

早期信息稀疏性导致负样本中混入大量假阴性样本，即不同短视频的某模态表示偶然相似却被推远。通过难负样本挖掘策略缓解，对每个批次内相似度最高的负样本施加额外惩罚项。模态不确定性问题表现为部分短视频缺少音频或仅有单帧图像，采用子模态自适应对比学习架构，根据实际可用模态动态构造对比损失，缺失模态不参与对齐计算。时间敏感性要求对比学习模块轻量化，将预训练对比模型中的大型投影头替换为2层线性层，并冻结底层特征提取器，仅微调对比头与分类器[4]。上述适配使模型能够在50毫秒内完成单样本推理。此外早期识别中存在模态时间未对齐的挑战，例如语音内容与关键帧出现时间偏移，对比学习需要引入动态时间规整的软对齐策略。解决方法是计算所有可能的时间偏移下的最小对比损失，用该最小值代替固定时间戳下的对比得分。此操作增加了计算量但可在训练阶段使用，推理时仍采用固定对齐。

二、基于多粒度对比学习的短视频特征提取与跨模态对齐

（一）多模态早期特征提取模块

视觉流抽取封面图与首1至2秒关键帧，间隔0.5秒采样共4帧。使用 EfficientNet-B3提取帧级特征，经平均池化得到1280维视觉嵌入。文本流包含标题与画面光学字符识别文字，标题最长20字符，光学字符识别文本拼接后限长50字符，采用 Chinese RoBERTa-wwm-ext 获得768维文本嵌入。音频流提取前2秒语音，经 wav2vec2.0基础模型转换为512维声学特征，同时将语音转写为文本后与标题文本融合。3个模态特征经线性投影层映射至统一维度256。特征时间戳按帧序号与音频起始点对齐，形成3路并行序列，每路长度不超过5个时间步。对于光学字符识别文字，还额外记录每个文字框在画面中的空间坐标，将坐标信息编码为4维向量与文本嵌入拼接，从而保留文字与图像区域的空间对应关系。此空间对齐线索对虚假检测十分关键，因为篡改视频常出现文字漂浮或位置反常。

（二）跨模态对比对齐机制

构建3组对比学习任务：视觉－文本对齐、视觉－音频对齐、文本－音频对齐。每组任务独立计算对比损失，最终对比损失为3组损失加权和。正样本对构造规则：同一短视频的视觉关键帧序列与标题文本构成多个正对，关键帧与音频片段构成正对，标题与音频转写文本构成正对[5]。负样本对来自同一训练批次内不同短视频的对应模态。引入双向对比损失，同时考虑模态 A 到 B 与 B 到 A 的相似度矩阵，取两者对称化结果。对比对齐后输出一致性得分向量，该向量维度为3，分别表征3组模态对的匹配程度。真实短视频的一致性得分普遍高于0.7，虚假视频常低于0.4。为增强对比对齐的鲁棒性，采用动量更新队列存储历史负样本特征，队列大小设为4096，每个训练步更新队列中当前批次特征。这扩大了负样本池规模，使对比学习更难出现坍塌解。还引入校准偏置项，对每个模态对计算长期平均相似度，用该平均值中心化相似度矩阵，消除模态自身的相似性偏差。

（三）多粒度对比学习增强

实例级对比将整个短视频的融合表示作为锚点，与同批次其他短视频的融合表示进行对比，损失函数采用监督对比学习形式，利用真实标签区分正负实例。模态级对比对每个模态独立实施实例判别任务，强制各模态编码器捕获与虚假检测相关的判别特征。时序级对比针对关键帧序列，从同一视频的连续帧间构造正样本对，从不同视频或打乱顺序的帧间构造负样本对，使模型感知帧间逻辑连贯性。虚假短视频常出现帧间场景突变或重复贴片，其时序对比损失显著高于真实视频。3粒度损失按加权方式融合。多粒度策略迫使模型从局部到全局全面评估内容真实性，尤其适用于早期仅有少量帧的情况。分析梯度流发现，实例级损失主要影响融合层参数，模态级损失直接影响各单模态编码器，时序级损失则调节帧间注意力权重。3部分相互配合避免了单一粒度带来的表示退化。

三、基于对比增强特征的早期虚假信息识别模型

（一）模型整体架构

模型包括输入层、对比学习层、融合层、分类层。输入层接

收多模态早期特征向量。对比学习层执行跨模态对齐与多粒度对比增强，输出3组一致性得分与增强后的模态特征。融合层采用跨模态注意力机制，以文本特征为查询向量，视觉与音频特征为键值对，计算加权融合表示。同时引入门控融合单元，根据一致性得分动态调节各模态贡献权重，一致性低的模态被抑制。分类层由2层全连接网络与 Sigmoid 输出构成，输入为融合表示与3组一致性得分的拼接向量。辅助输出分支直接使用一致性得分均值作为快速预判信号，当均值低于0.3时可直接判定虚假，跳过分类层计算以节省时间。整个模型采用端到端训练，对比学习损失与分类损失共享底层特征提取器，但对比头与分类头参数独立。这种设计使得对比学习任务不会强迫分类特征过度偏向对齐，保留了模态特有信息用于判别。

（二）早期识别专用训练策略

时间感知掩码训练用于模拟真实场景下的信息缺失。每个训练批次中随机丢弃30% 样本的音频模态，随机压缩50% 样本的关键帧数量至2帧，随机屏蔽20% 样本的光学字符识别文本。模型在同一批次内处理完整与缺失混合数据，迫使分类器对不同信息完整度产生鲁棒性。难负样本挖掘针对模态欺骗性强的虚假样本，即那些图文一致但内容虚假的短视频。采用对抗式采样算法，计算每个负样本的当前对比损失，将损失排名前10% 的负样本复制并加入下一轮训练，增加对比难度。联合损失函数包含分类损失、对比损失、一致性正则项3部分。一致性正则项惩罚同一视频不同模态对间得分方差过大的情况。训练过程中还采用渐进式对比学习策略，前10个轮次只使用对比损失预训练对齐模块，之后加入分类损失联合微调。这种预热方式避免早期分类噪声干扰对比学习。

（三）快速推理与阈值自适应

为满足早期识别的时间要求，将对比投影头设计为单层线性映射，推理时仅需一次前向传播即可获得所有模态特征与对比损失。特征缓存机制将已计算的短视频基础特征存入内存池，重复请求相同视频时直接读取。动态阈值调整依据视频发布后已获取的信息量比率计算分类阈值，基础阈值设为0.5，调节系数设为0.2。当信息不完整时提高阈值，减少假阳性误报。推理时间控制在每样本45毫秒以内，模型参数量约12兆，可在移动端部署。另设计早期退出机制，在每个模态特征提取完成后计算当前可用的一致性得分，若得分极端低或极端高则提前终止后续模态提取，直接输出判定。该机制可将部分样本的推理时间压缩至15毫秒。

四、实验设计与结果分析

（一）实验数据集与预处理

选用公开多模态虚假视频数据集 FakeSV 和 TI-CNN。FakeSV 包含15678个短视频，其中虚假视频8234个，涵盖政治、健康、娱乐等领域。TI-CNN 收集自微博平台，总计11234条视频数据。构造早期子集时保留标题、封面图、前3秒关键帧、前2秒音频，丢弃后续所有内容。关键帧采样频率为1帧 / 秒。将数据集按8比1比1划分为训练集、验证集、测试集。评价指标包括准确率、精确率、召回率、F1分数，另设早期识别时效性指标：发布后10秒内完成判定的样本比例。所有实验在单张 NVIDIA A100图形处理器上运行。

（二）对比方法与消融实验

选取6种基线方法：单文本模型 TextCNN、单视觉模型 ResNet50、多模态拼接ConcatMLP、无对比学习的多模态注意力模型 MMA、基于 CLIP 的零样本检测、对比学习检测方法。消融实验设置4个变体：去掉多粒度对比损失、去掉跨模态对齐模块、去掉时序对比、去掉模态掩码训练。下表展示各方法在 FakeSV 测试集上的结果。

表：对比结果

方法	准确率	精确率	召回率	F1分数	10秒内判定比例
TextCNN	0.672	0.654	0.683	0.668	0.98
ResNet50	0.701	0.712	0.689	0.700	0.95
ConcatMLP	0.734	0.745	0.721	0.733	0.93
MMA	0.769	0.778	0.756	0.767	0.89
CLIP 零样本	0.713	0.724	0.698	0.711	0.91
对比学习检测方法	0.801	0.812	0.789	0.800	0.87
本文方法	0.848	0.861	0.832	0.846	0.94
去掉多粒度对比损失	0.812	0.825	0.796	0.810	0.92
去掉跨模态对齐模块	0.778	0.785	0.769	0.777	0.90
去掉时序对比	0.830	0.842	0.816	0.829	0.93
去掉模态掩码训练	0.835	0.848	0.821	0.834	0.91

（三）结果分析与讨论

本文方法在4项分类指标上均优于所有基线，相比对比学习检测方法 F1分数提升4.6个百分点。跨模态对齐模块贡献显著，去除后 F1下降6.9个百分点。多粒度对比损失带来3.6个百分点的提升，其中时序对比贡献约1.7个百分点，说明帧间连贯性对短视频虚假检测尤为重要。模态掩码训练使模型在早期识别场景下保持高召回率，去掉掩码训练的变体10秒内判定比例下降3个百分点但精确率略高，表明掩码训练以微小精确率代价换取响应速度。CLIP 零样本检测效果较差，因预训练任务与虚假检测分布差异大。错误案例分析发现失败样本多为精心制作的深度伪造视频，其跨模态一致性极高，仅依赖模态对齐难以区分。另一类错误为信息极度稀疏的样本，仅有标题无图像，模型表现为假阴性。计算效率方面，本文方法单样本平均推理时间43毫秒，参数量12.3兆，满足移动端实时部署要求。进一步分析混淆矩阵，本文方法的假阳性率为8.2%，假阴性率为10.1%，相比基线分别降低了4.5%和3.8%。在 TI-CNN 数据集上的结果趋势与 FakeSV 一致，F1分数达到0.831，验证了方法的泛化能力。

五、结语

针对短视频虚假信息早期检测中模态缺失与语义碎片化问题，提出了多粒度对比学习框架。该框架融合实例级、模态级、时序级对比损失，强化跨模态对齐与特征判别性。设计时间感知掩码训练策略提升模型对信息不完整场景的鲁棒性，轻量化推理机制实现毫秒级响应。在公开数据集上验证了方法的有效性，F1分数达到0.846，10秒内判定比例94%。未来工作将引入用户评论模态与图神经网络传播结构，进一步挖掘社交上下文线索，同时探索在线持续学习以适应新型虚假生成技术。动态阈值自适应机制在小样本场景下的性能边界也值得深入探讨，可结合强化学习实现阈值的自动调节。

参考文献

[1]Yan F ,Zhang M ,Wei B , et al.FMC: Multimodal fake news detection based on multi-granularity feature fusion and contrastive learning[J].Alexandria Engineering Journal,2024,109376-393.

[2] 张永成，魏小梅，王欢，徐荣康 . 一种不确定模态缺失的多模态对抗虚假新闻检测框架 [J]. 中文信息学报 ,2024,38(06):151-160.

[3] 段钰潇，胡艳丽，郭浩，谭真，肖卫东 . 改进的跨模态关联歧义学习的虚假信息检测方法研究 [J]. 计算机科学 ,2024,51(04):307-313.

[4] 李卓远，李军 . 基于对比学习的多模态注意力网络虚假信息检测方法 [J]. 中国科技论文 ,2023,18(11):1192-1197.

[5]Yang Y ,Ran B ,Weili G , et al.Deep visual-linguistic fusion network considering cross-modal inconsistency for rumor detection[J].Science China Information Sciences,2023,66(12): DOI:10.1007/S11432-021-3530-7.

