Advanced Pattern Recognition

ASSIGNMENT

- SUBMITTED BY:

  TAHNEES SHERAZ

- REGISTRATION NO:

  SP24-RAI-016

- SUBMITTED TO:

  Dr. RASOOL BUKHSH

Department of Computer Science, CUI

TOPIC:

Out-of-Context Misinformation Detection by Multimodal Large Language Model

Paper Title	Author(s)	Methodology	Dataset	Limitations
SNIFFER: Multimodal LLM for Explainable Out-of-Context Misinformation Detection	Qi et al.	Two-stage instruction tuning on InstructBLIP using GPT-4-generated tasks for domain and task adaptation. Combines internal (image-text) and external (retrieved context) reasoning.	NewsCLIPpings for training, News400 and TamperedNews for testing	Relies on GPT-4 and fine-tuned instruction data, which may limit reproducibility. external tools and retrieval systems introduce dependencies.
Open-Domain, Content-Based, Multi-Modal Fact-Checking of Out-of-Context Images	Abdelnabi et al.	Consistency-Checking Network (CCN) using memory networks to assess image-caption pairing by gathering external visual and textual evidence (multi-modal cycle-consistency check).	NewsCLIPpings, Web evidence collected dynamically using Google APIs	Relies on search engine APIs and web scraping which may yield noisy or irrelevant data; external evidence quality is variable; slower due to web access latency.
FLAVA: A Foundational Language and Vision Alignment Model	Singh et al.	Unified multimodal transformers using both unimodal and multimodal pretraining objectives (Global Contrastive, ITM, MMM, MLM, MIM). Combines dual and fusion encoders.	COCO, Visual Genome, Conceptual Captions, SBU Captions, Localized narratives, Wikipedia Image Text etc.	Computationally intensive; may not generalize to novel domains outside the training corpora. Does not explore knowledge-grounded fact-checking.

E2LVLM: Evidence-Enhanced Large Vision-Language Model for Multimodal Out-of-Context Misinformation Detection	Wu et al.	Enhances LVLMs by reranking and rewriting retrieved textual evidence to align with model inputs. Constructs a multimodal instruction-following dataset with judgments and explanations. Employs instruction tuning for improved detection and explanation capabilities.	NewsCLIPpings	Dependent on the quality of external evidence retrieval; rewriting strategies may not always produce accurate alignments; instruction tuning requires careful prompt design and may be sensitive to prompt variations.

Introduction

With the exponential growth of online information, the spread of misinformation has become a critical issue in today's digital society. A particularly subtle yet harmful form of misinformation is out-of-context (OOC) misuse, where real, unaltered images are paired with captions or textual claims that are misleading, incorrect, or refer to unrelated events. This form of misinformation is difficult to detect using traditional methods because the visual content is authentic, and manipulation lies in the contextual mismatch. In response to this challenge, researchers have explored multimodal misinformation detection, leveraging the power of vision-language models to assess the consistency between image and text content.

One of the most notable recent advancements in this area is the SNIFFER model, a multimodal large language model (MLLM) fine-tuned through instruction tuning to detect and explain OOC misinformation. SNIFFER integrates both internal consistency checks (image-text alignment) and external evidence verification (retrieved web sources) to produce highly accurate judgments along with natural language explanations. Despite its strong performance, SNIFFER still has several limitations that hinder its practical adoption and scalability. Specifically, it relies on expensive and restricted-access models like GPT-4 for generating instruction data, making it resource-intensive and difficult to replicate in lower-resource environments. Moreover, the reasoning process that combines internal and external evidence is largely opaque, offering limited insight into how the final decision is made. Lastly, the model

provides only a binary decision—whether a pair is OOC or not—without identifying the type of contextual inconsistency, such as whether the mismatch is about a person, location, or event.

This thesis builds upon SNIFFER's foundation and proposes an enhanced framework that addresses these key limitations. The proposed system introduces a more efficient and scalable instruction tuning pipeline by leveraging smaller open-source language models, self-supervised learning strategies, and active data selection; a transparent composed reasoning module that quantifies and explains the influence of both internal and external evidence sources on the model's decision-making; and an extension to the model's output that allows it to classify the type of OOC misinformation, thereby improving its interpretability and usefulness in real-world fact-checking scenarios. By tackling these limitations, this work aims to advance the field of multimodal misinformation detection and contribute a practical, interpretable, and scalable solution for detecting image-text inconsistencies across digital platforms.

Problem Statement

Despite the significant advancement of the SNIFFER model in detecting and explaining out-of-context (OOC) misinformation using multimodal large language models (MLLMs), several limitations remain unaddressed.

Current state-of-the-art Out-of-Context (OOC) misinformation detectors like SNIFFER rely heavily on large, proprietary language models (e.g., GPT-4) for both generating instruction-tuning data and performing the final 'composed reasoning' step that integrates internal and external evidence. This dependency presents challenges in terms of accessibility, cost, and transparency. While SNIFFER introduces a 'composed reasoning' step to synthesize internal image-text consistency checks and external retrieved evidence for OOC misinformation detection, this final LLM-based integration often lacks transparency in how evidence is weighed and combined. Furthermore, the generation of training data for both the core OOC detection and this reasoning step can be resource-intensive

This research aims to develop an efficient, cost-effective instruction tuning pipeline by exploring open-source and smaller-scale language models LLaMA and leveraging self-supervised and active learning techniques, and to design and implement a more transparent and potentially disentangled composed reasoning module for OOC detection that explicitly models the contribution of different evidence sources, while simultaneously investigating more efficient strategies for generating the necessary instruction data to train such a system, thereby improving both the interpretability and practicality of explainable OOC detectors. These enhancements will enable a more interpretable, scalable, and human-aligned approach to multimodal misinformation detection.

System Model

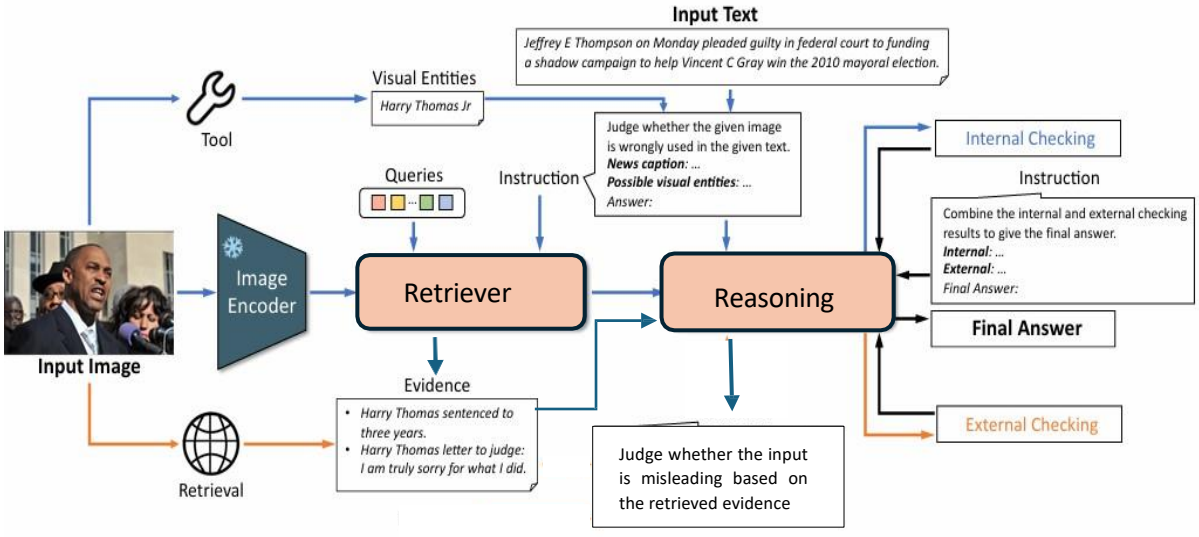
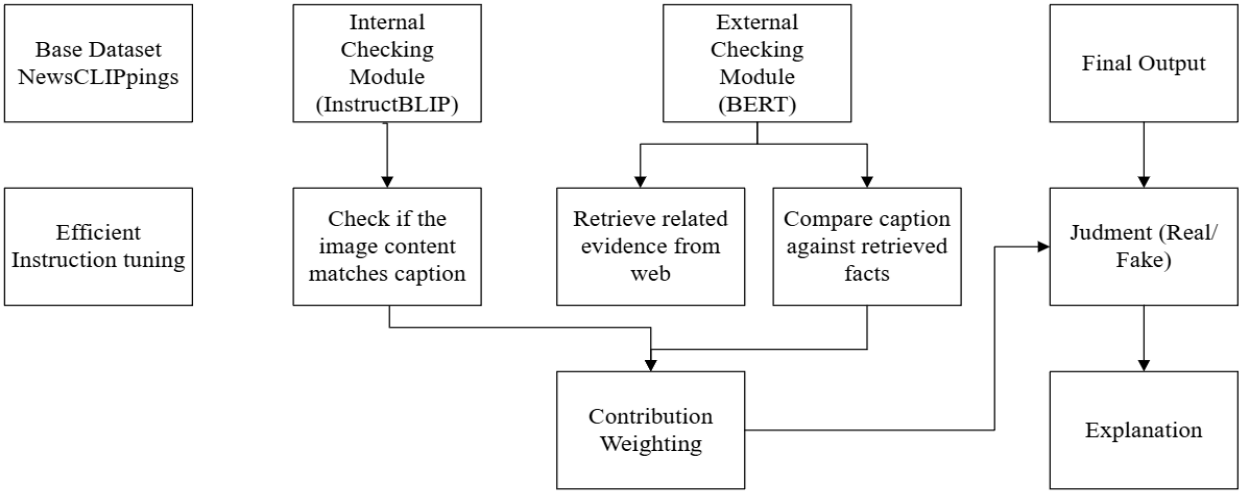


Figure 1: Proposed architecture for multimodal misinformation detection. The system integrates a lightweight image encoder, a noise-mitigated evidence retriever, and a dedicated reasoning module to assess whether an image-text pair is misleading. The framework performs both factual grounding and logical verification and combines internal and external rationales to generate an interpretable final decision. This modular design enhances robustness, accuracy, and explainability compared to prior approaches.

The algorithm takes an image and its accompanying text as input. First, the Image Encoder extracts visual features and key objects from the image. These features, along with the text, are used to create a query that is sent to the Retriever, which searches for relevant evidence from trusted sources. A filtering step removes irrelevant or low-quality information to reduce noise. The clean evidence, along with the original input, is passed to a Reasoning Module that checks if the image and text are consistent with each other and with external facts. Finally, the system combines all the results and generates a clear decision (true or misleading) with a short explanation. This process helps detect misinformation more accurately, efficiently, and transparently

Proposed Model



Pseudocode

The pseudocode for the proposed methodology is given below

Algorithm: OOC_Misinformation_Detector

Input: Image I, Caption C, Retrieved_Evidence E

Output: Label (OOC / Not OOC), Explanation

Self-Supervised Inconsistency Check

if Is_SelfSupervised_Training:

Generate synthetic mismatches (I, C') where C' ≠ original caption

Train model to classify (I, C') as OOC

Internal Reasoning

Internal_Score ← ComputeAlignmentScore(I, C) // via InstructBLIP

External Reasoning

External_Score ← ComputeEvidenceMatch(C, E) // via BERT/RoBERTa

Fusion via Contribution Weighting

Final_Score ← α * Internal_Score + β * External_Score

Label ← Threshold(Final_Score)

Explainability

Explanation ← GenerateExplanation(I, C, E, Internal_Score, External_Score)

return Label, Explanation

Dataset

The primary dataset used in the SNIFFER model is the NewsCLIPpings dataset, which was originally introduced by Grace Luo et al. and is publicly available as part of their work on out-of-context misinformation detection. The dataset builds upon the VisualNews dataset and is designed specifically for evaluating models on detecting image-caption mismatches in real-world news articles.

Name: NewsCLIPpings
Source: *https://huggingface.co/g-luo/news-clippings*
Content Type: Real and synthetically manipulated image-caption pairs from online news articles
Domains: Four major English-language news agencies

- The Guardian
- BBC
- USA Today
- The Washington Post

Size and Structure

Attribute	Description
Total Samples	368,013 real news image-caption pairs (VisualNews)
OOC Pairs (Generated)	71,072 training samples
Validation Samples	7,024
Testing Samples	7,264
Balanced Subset	Includes equal numbers of real and OOC samples

Features per Sample

- Image
- Original caption (real)
- Manipulated caption or mismatched image (OOC)
- Retrieved external evidence (optional)
- Named entities extracted via NER
- Human or model-generated label: OOC or not

Implementation

https://github.com/Tahnees/PRAssignment

References

[1] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2022). *FLAVA: A foundational language and vision alignment model*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 15638–15649). https://arxiv.org/abs/2112.04482

[2] Abdelnabi, S., Hasan, R., & Fritz, M. (2022). *Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 14940–14950). https://arxiv.org/abs/2112.00061

[3] Qi, P., Yan, Z., Hsu, W., & Lee, M. L. (2024). *SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://arxiv.org/abs/2403.03170

[4] Wu, X., Liu, Y., Li, J., Wang, Y., Zhang, Z., & Wang, Y. (2024). *E2LVLM: Evidence-enhanced large vision-language model for multimodal out-of-context misinformation detection*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://arxiv.org/abs/2404.10397