画像を再利用した偽情報の検出に対する情報補完を用いた精度向上手法

渡邉　祐太† 矢崎　孝一† 佐々木佑樹† 北島　信哉†

† 富士通株式会社 データ＆セキュリティ研究所　〒 211–8588 神奈川県川崎市中原区上小田中 4–1–1
E-mail: †{watanabe.yut-14,yasaki.kouichi,sasaki.yuki-01,kitajima.shinya}@fujitsu.com

あらまし　近年，自然災害や政治，公衆衛生など，様々な分野において偽情報が大きな社会問題になっている．偽情報に対抗するために，多くの団体がファクトチェック活動を行っているが，人手によるファクトチェックには膨大な時間がかかるため，その自動化の研究が進められている．中でも，画像を異なる文脈で再利用する文脈外（OOC: Out-of-Context）誤情報は手軽に作成できることから，典型的な偽情報であり，OOC 誤情報を自動的に判定する技術も盛んに研究されている．従来研究として，大規模視覚言語モデルを利用して OOC 誤情報の判定とその根拠説明文を生成する SNIFFER があげられるが，SNS 投稿では背景情報が欠落していることが多いため，十分な精度が得られないという課題があった．そこで本稿では，テキストと画像からなる SNS 投稿を OOC 誤情報の判定対象とし，SNS 投稿の内容や場所，時間の情報を補完することで判定精度を向上する手法を提案する．性能評価の結果，提案手法では 0.80 という高い精度で OOC 誤情報を判定できることを確認した．

キーワード　ファクトチェック，フェイクニュース，偽情報検知，ソーシャルメディア，Out-of-Context misinformation

1 はじめに

近年，Web や SNS（Social Networking Service）の発展にともない，多くの人々が世界中の情報を手軽に共有できるようになった一方で，偽情報の拡散が大きな社会問題になっている．たとえば，令和 6 年能登半島地震では，SNS を介して救助活動や復興活動を妨げるような偽情報が流通し，社会に混乱をもたらした [1]．このような偽情報に対抗するために，報道機関やインターネットメディア，大学関係者などが，言説に含まれる主張が事実かどうかを客観的に検証するファクトチェックという取り組みを行っている．しかし，人手によるファクトチェックは検証作業に膨大な手間と時間がかかるため，自動ファクトチェックの研究が盛んに行われている [1].

偽情報の中でも，文脈外（OOC: Out-of-Context）誤情報とよばれる偽情報が非常に多い [2]．OOC 誤情報とは，事実とは異なる文脈で画像を使用する偽情報である．たとえば，東日本大震災の発生時に撮影された写真を利用して，この写真は能登半島地震の被害状況を表すものであると主張する言説は，OOC 誤情報に該当する．OOC 誤情報は特別な技術を必要とせずに誰でも手軽に発信できるため，典型的な偽情報となっている．

本稿では，テキストと画像で表された言説が与えられたとき，言説の正誤にかかわらず，その画像が事実とは異なる文脈で使用されている場合，その言説を OOC とよぶ．そして，テキストと画像で表された言説が OOC であるか否かを判定するタスクを，OOC 判定とよぶ．また，OOC 誤情報は，OOC のうち言説が誤っているものを指す．ここで，言説の正誤の判定は人間でも基準や判定がわかれる難しいタスクであるため，まずは判断基準を作りやすい OOC 判定を自動化する研究がいくつか行われており，我々も OOC 判定の自動化に取り組んでいる．

OOC 判定のためには，過去の画像を別の文脈で使い回しているかどうかの検証と，テキストと画像の内容が矛盾していないかの検証の，2 つの検証を行う必要がある．これまでに，いくつかの OOC 判定手法が提案されているが，2 つの検証のどちらかのみしか行っていないために一部の OOC しか判定できないという課題や，判定結果に対する根拠説明が十分でないという課題がある [3, 4].

これらの課題を解決するために，SNIFFER という手法が提案されている [5]．SNIFFER は，OOC 判定向けにチューニングされた大規模視覚言語モデル（LVLM: Large Vision Language Model）による検証と，画像のインターネット上の出所情報を用いた検証の，2 つの検証を組み合わせた OOC 判定技術であり，テキストと画像で表された言説に対する OOC 判定と，その根拠説明の生成を行う．SNIFFER は，ニュース記事をもとに構築された OOC 誤情報検出ベンチマークである NewsCLIPpings [6] において，2024 年 3 月時点での最高性能である判定精度 88.4 %を達成している．また，根拠説明の生成についても，テキスト要約の評価指標として知られている ROUGE [7] や人間による評価によって，効果があることを示している．

SNIFFER は，ニュース記事に対しては高い判定精度を達成しているが，SNS 投稿に対する効果は明らかにされていない．しかし，メディアが発信するニュース記事における OOC 誤情報よりも，誰でも発信できる SNS 投稿における OOC 誤情報の方が数が多く，対応が必要不可欠であると考えられる．そこで，SNIFFER がニュース記事だけでなく SNS 投稿に対しても有効かどうかを確認するために，X（旧 Twitter）[2]に投稿されたいくつかの OOC の実例を用いて SNIFFER の判定精度を

1：https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r06/index.html

2：https://x.com/

確認したところ，十分な判定精度が得られなかった．これは，SNS 投稿はニュース記事における画像とそのキャプションとは異なり，主張の背景情報や画像の説明が省略される傾向にあり，OOC 判定に必要な情報を有していないためと考えられる．

そこで本稿では，SNIFFER をベースに，OOC 判定の際に重要な手掛かりとなる SNS 投稿の内容や場所，時間の情報を，SNS 投稿の周辺情報から収集して補完することで，OOC 判定の精度を向上する手法を提案する．

本稿の構成は以下の通りである．まず，2 で OOC 判定の関連研究を説明し，3 で提案手法の詳細を示す．そして，4 で提案手法の評価とその結果について説明したのち，最後に 5 で本稿のまとめと今後の課題を述べる．

2 関 連 研 究

本章では，OOC 判定に関連する既存研究と，その課題について述べる．また，提案手法のベースとなる SNIFFER の概要とその課題について詳しく説明する．

2.1 OOC 判定に関する既存研究

OOC 判定の既存手法として，機械学習ベースの手法や LVLM を用いた手法，外部情報を活用した手法がある．本節では，それぞれの手法について説明する．

Aneja ら [3] は，画像とその画像の内容を説明するキャプションからなるデータセットを用いて自己教師あり学習により作成した機械学習モデルを使うことで，OOC 判定に取り組んでいる．しかし，Aneja らの OOC の定義は，1 つの画像に対して 2 つのキャプションが与えられることを前提に，2 つのキャプションが画像内の同一の対象物を指しているものの意味が異なっていること，となっている．実際に OOC 誤情報が発生するシーンでは，1 つの画像に対して 2 つのキャプションが与えられる事例はまれであり，この手法が実用的であるケースは限られている．

Luo ら [6] は，事前学習済み LVLM である CLIP [8] や visualBERT [9] をニュースドメイン向けにファインチューニングし，これらを用いて OOC 判定を行っている．しかし，LVLM にのみ依存する手法では，判定対象の画像が LVLM の訓練データの中に含まれていない場合，LVLM は判定対象の画像が過去の画像の再利用かどうかを判定できないため，画像を再利用するタイプの OOC の検出は原理的に不可能である．

一方で，Abdelnabi ら [4] は，インターネット上の情報を証拠として活用することで OOC 判定を行う手法を提案した．具体的には，判定対象の画像が使用されている箇所をインターネット上から探し，その箇所のテキスト情報と判定対象のテキストを比較することで OOC 判定を行う．さらに，判定対象のテキストからインターネット上の画像を検索し，その結果の画像と判定対象の画像を比較することでも OOC 判定を行う．しかし，撮影した写真を，写真とは異なる文脈のテキストとともにすぐさま SNS に発信するといった OOC の事例も考えられる．このように，OOC 判定に有効な情報が必ずしもインターネット上で見つかるとは限らないため，判定対象の情報のみからテキストと画像の間の矛盾を検出するといった，インターネット上の情報に依存しない検証も併用する必要がある．

また，これらの研究に共通の課題として，判定結果に対して説得力のある根拠説明を生成できないことが多い．適切な根拠説明が生成できない場合，たとえ判定結果が正しくても説得力に欠けるため，一般のインターネット利用者に対する偽情報への反証効果が減少してしまうおそれがある．またファクトチェックを行う人々にとっても，検証過程や検証結果の根拠をファクトチェック記事に記載できず，ファクトチェック自動化のメリットが損なわれてしまう．

そこで Qi ら [5] は，非特化モデルである事前学習済み LVLM では特定タスクに対する性能に限界があると考え，LVLM を OOC 判定に特化してチューニングした．そして，この OOC 特化型 LVLM とインターネット上の情報を用いたハイブリッドな方式である SNIFFER を提案することで既存研究の課題を解決し，高精度での OOC 判定と同時に適切な根拠説明を可能とした．SNIFFER の詳細については，次節で説明する．

2.2 SNIFFER の概要

本節では，提案手法のベースとして用いた SNIFFER の概要について説明する．SNIFFER は，つぎの機能から構成される．

- 逆画像検索：OOC 判定対象の画像に関するインターネット上の情報の収集
- 内部チェック：OOC 特化型 LVLM を用いた OOC 判定
- 外部チェック：画像のインターネット上の出所情報を用いた OOC 判定
- 合成推論：内部チェックと外部チェックの結果を用いた最終判定

図 1 に，SNIFFER における処理フローを示す．以下，各機能の詳細について説明する．

2.2.1 逆画像検索

逆画像検索では，Google Cloud Vision API[3]を用いて，OOC 判定対象の画像に関する情報をインターネット上から収集する．Google Cloud Vision API の出力は，インターネット上の類似画像をもとに推測された，入力画像に写っているオブジェクトの名称（画像の視覚エンティティ）や，入力画像の出所記事のタイトルと画像のキャプション，出所記事の URL などである．このうち，画像の視覚エンティティは内部チェックにおける判定の補助情報として利用し，画像の出所記事のタイトルと画像のキャプションは外部チェックにおいてテキストと画像の文脈を比較する際に用いる．

2.2.2 内部チェック

内部チェックでは，OOC 判定に特化してチューニングされた LVLM を用いてテキストと画像の整合性を検証することで，OOC 判定とその根拠説明の生成を行う．LVLM は，事前学習済みの汎用 LVLM である InstructBLIP [10] を基盤モデルとし，これにインストラクション・チューニング [11] を施し

3：https://cloud.google.com/vision/docs

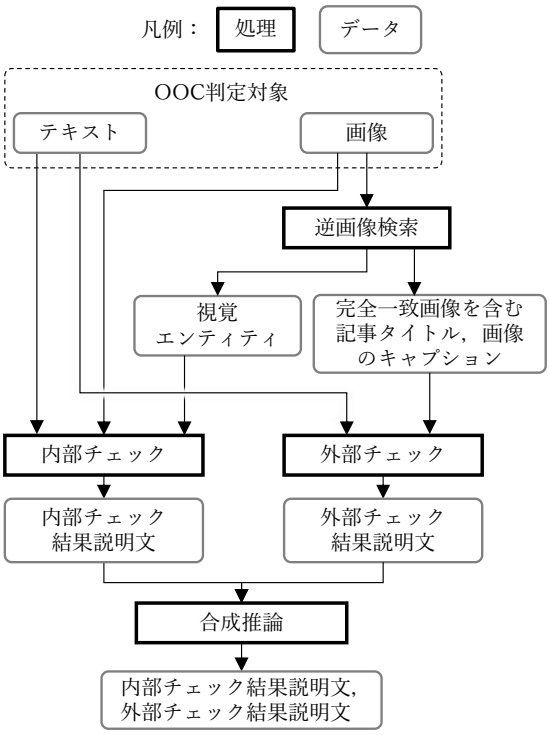

図 1　SNIFFER における処理フロー

た OOC 判定特化型の LVLM を用いる．インストラクション・チューニングの訓練用データセットには，ニュース記事の画像とそのキャプションから構成される NewsCLIPpings を指示応答形式に変換したものを使用する．そして，この LVLM に対して，OOC 判定対象のテキストと画像，および逆画像検索で取得した視覚エンティティとともに，テキストと画像の間に矛盾があるか否かを判定させるプロンプトを入力し，OOC 判定とその根拠説明生成を実行させる．ここで，視覚エンティティの情報をプロンプトに含めるのは，LVLM の学習時には存在しない最新のオブジェクトの名称も加味して推論するためである．

内部チェックでは，たとえば，入力された OOC 判定対象のテキストでは国内の街でのできごとを主張しているにもかかわらず，画像には外国の街並みが写っている場合，矛盾を検出し，OOC であると判定する．

内部チェックでの検証が有効となる対象は，テキストと画像間の矛盾を LVLM の内部知識によって特定できる言説のみである．視覚エンティティはインターネット上から獲得した外部の情報であるが，あくまで判定の補助的な情報として画像内のオブジェクトの名称を与えているだけである．LVLM の内部知識には存在しない過去の別の事象の画像を再利用しているといった，インターネット上の情報を活用しなければ原理的に OOC を検出できないものは，外部チェックにより検出される．

2.2.3　外部チェック

外部チェックでは，LLM を用いてテキストの文脈と画像の出所情報の文脈を比較することで，OOC 判定とその根拠説明の生成を行う．具体的には，OOC 判定対象のテキストと，逆画像検索で取得した画像の出所記事のタイトルおよび画像のキャプションを LLM に入力し，テキストの文脈と画像の文脈が一致しているかどうかを検証する．Qi らは，検証に用いる LLM として Meta 社の Llama2 [12] を利用しているが，他の LLM に置き換えることも可能である [4]．

外部チェックでは，たとえば，入力された OOC 判定対象のテキストが能登半島地震の被害状況を主張しているにもかかわらず，画像の出所情報によって過去に発信された東日本大震災の様子を撮影した画像であると判明した場合，文脈の不一致を検出し，OOC であると判定する．

なお，画像の出所記事は複数見つかる場合もある．その場合は，すべての出所記事のタイトルとすべての画像のキャプションを検証に用いる．一方で，画像の出所記事が 1 つも見つからない場合には，外部チェックは行わない．

2.2.4　合 成 推 論

合成推論では，内部チェックと外部チェックで出力された根拠説明を合成する．これは内部チェックと外部チェックで異なる結論を導く可能性があるため，内部チェックの結果と外部チェックの結果を LLM に与え，両者の出力から最終的な根拠説明文を生成する．なお，逆画像検索において画像の出所記事が見つからなかった場合は，内部チェックの結果を最終判定とする．

2.3　SNIFFER の課題

SNIFFER は，ニュース記事に関しては高い判定性能を有するが，SNS 投稿の判定精度には課題が残る．実際，SNIFFER は，ニュース記事をもとに構築された OOC 誤情報検出ベンチマークである NewsCLIPpings において 2024 年 3 月時点での最高性能を達成した．しかし，ファクトチェック記事をもとに，X に投稿されたいくつかの OOC 誤情報を SNIFFER に OOC 判定させたところ，その多くで正しく判定できなかった．

これは，ニュース記事と SNS 投稿の性質の違いが原因だと考えられる．ファクトチェックの対象となるような SNS 投稿は，テキストと画像が整合性をもつように作成されているため，画像の再利用を検出しないと原理的に判定できないケースが多い．したがって，SNS 投稿に対しては内部チェックは有効性が低く，外部チェックで正しく OOC 判定するべきである．ここで，ニュース記事における画像のキャプションは画像の内容を説明する情報を十分に含んでいるのに対し，SNS 投稿はその主張の背景情報や添付されている画像の説明が欠落し，不明瞭であるという傾向がある．SNIFFER における外部チェックでは，判定対象の発信日以前に作成された異なる文脈の出所記事を特定できたにもかかわらず，テキストの文脈と画像の文脈を比較する際に必要な情報が足りないために誤判定が多くなっていた．

現実世界の偽情報に立ち向かうためには，上記の外部チェックの課題を克服し，SNS 投稿の OOC 判定にも対応することが重要である．そこで，提案手法では，テキストと画像それぞれの詳細な内容やそれに関する場所，時間といった，OOC 判定の際に重要な手がかりとなる情報を SNS 投稿の周辺情報から収集し，元の SNS 投稿には欠落していた判定材料を補った上

4：https://github.com/MischaQI/Sniffer

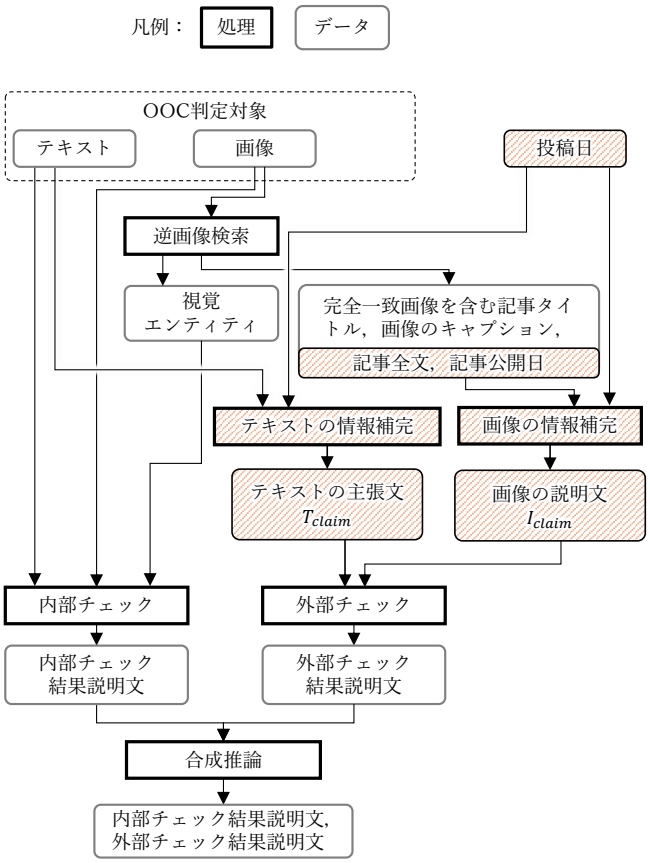

図 2　提案手法の処理フロー

で外部チェックに与えることで，OOC 判定の精度向上を狙う．

3　提 案 手 法

本章では，テキストと画像からなる SNS 投稿に対して Web 検索を用いて投稿の背景情報を収集したのち，背景情報を補完したテキストの主張文と画像の説明文を生成し，それらを比較することで OOC 判定を行う手法を提案する．

3.1　提案手法の概要

図 2 に，提案手法における処理フローを示す．図中の網掛け部分が，提案手法で追加した部分である．SNIFFER を用いて SNS 投稿の OOC 判定を行った際に十分な判定精度が得られない原因は，画像が出所記事と異なる文脈で再利用されているかどうかを判定する外部チェックにあると考え，外部チェックの改善を行う．SNIFFER では，逆画像検索で得られた記事のタイトルや画像のキャプションを画像の文脈とし，SNS 投稿のテキストそのものをテキストの文脈として，両者を外部チェックに与えているのに対し，提案手法ではテキストと画像両方の文脈に情報補完を行ってから外部チェックに与えている．具体的には，SNIFFER では用いていない逆画像検索で見つかった記事全文や公開日から情報を抽出して画像の情報補完とし，SNS 投稿に含まれるメタ情報や投稿時期の時事情報をテキストの情報補完として，両者を外部チェックに与えている．

ここで，外部チェックが出力する判定結果と合成推論の結果で決まる判定結果を区別するために，外部チェックによって OOC と判定された場合は再利用 OOC（ROOC: Reuse OOC）とよび，外部チェックにおけるこの判定を ROOC 判定とよぶ．

SNIFFER では ROOC 判定において，画像の出所情報の文脈がテキストの文脈を支持するかどうかという観点で比較を行っている．ここで，画像の出所情報の文脈がテキストの文脈を支持するとは，画像の出所情報の文脈がテキストの文脈と一致することを意味している．文脈を比較する際には，SNS 投稿のテキストと画像，それぞれが示す内容に加え，場所や時間の情報が重要である．なぜなら，画像とテキストそれぞれが言及している場所，時間に矛盾がないかを確認することが，ROOC 判定の大きな手がかりとなるからである．そこで，提案手法では，SNS 投稿のテキストと画像の両方に対して，それぞれが示す内容のさらなる具体化や，時間，場所の情報を補完する．

以下では，提案手法におけるテキストの情報補完および画像の情報補完，さらに外部チェック処理の改良点について説明する．

3.2　テキストの情報補完

まず，テキストの情報補完で収集する内容，場所，時間を，以下の通り定義する．

- 内容（$T_{content}$）：テキストが主張する内容の主語，述語，目的語に関するものや，主張の背景情報，主張に含まれる代名詞や指示語が示すもの
- 場所（$T_{location}$）：テキストの主張が言及している場所
- 時間（T_{time}）：テキストの主張が言及している，または，その主張内容が発生した日や月，年，期間

テキストの情報補完では，これらの情報を以下の手順にしたがって収集する．本稿では，以下を手作業で行っている．

1. テキストの主張を LLM などを用いて，主語・述語・目的語形式で書き出し，これを $T_{content}$ の初期値とする．また，判定対象の投稿日を，T_{time} の初期値とする．
2. 1. の中で，主語，述語，目的語が欠落していたり，代名詞，指示語が残っている場合は，画像に書かれている内容を用いて，$T_{content}$ の内容を具体化し，$T_{content}$ を更新する．たとえば，主張部分に「このテント」という指示語が残っている場合は，画像から「体育館に置かれたテント」といった情報を読み取り，$T_{content}$ の「このテント」という部分を，「体育館に置かれたテント」に置換して更新する．
3. 判定対象の SNS 投稿の前後ポストから， 内容，場所，時間に関する情報を探索し，新たな情報があれば，$T_{content}$，$T_{location}$，T_{time} をそれぞれ更新する．
4. 内容をより具体化するために，内容に関する背景情報を収集する．具体的には，テキスト本文の内容をもとに LLM を用いて検索キーワードを生成して Web および SNS 内の情報を検索し，T_{time} の周辺日時の時事情報を収集する．ただし，ここでは自治体や公式団体の発表など，信頼できる情報源のみを参照する．収集した時事情報とテキストの主張の間に矛盾がなければ，得られた時事情報を用いて $T_{content}$ をより具体化し，更新する．

以上の手順で生成された $T_{content}$，$T_{location}$，T_{time} を用いて，テキストの情報補完部の出力であるテキストの主張文 T_{claim} を生成する．まず，$T_{claim} = T_{content}$ としたのち，T_{claim} の内容が場所に言及する主張で，かつ場所情報が欠落していれば，$T_{location}$ を用いて T_{claim} を具体化して更新する．さらに，T_{claim} の内容が現在のできごとや特定の日，期間に関して言及している主張で，かつ日付情報が欠落していれば，T_{time} を用いて T_{claim} を具体化して更新する．

この手順により，判定対象のテキスト本文から背景情報や時間，場所の情報を補完した T_{claim} を生成し，これを外部チェックの入力として用いる．

3.3 画像の情報補完

まず，画像の情報補完で収集する内容，場所，時間を，以下の通り定義する．

- 内容（$I_{description}$）：画像の出所記事において，画像について説明されている内容
- 場所（$I_{location}$）：画像に写っている場所や画像が撮影された場所．ただし，画像が会社のロゴマークや世界地図など場所を特定できないものである場合は特定しない
- 時間（I_{time}）：画像が撮影，もしくは作成された日や月，年，期間

画像の情報補完では，これらの情報を以下の手順にしたがって収集する．本稿では，以下の手順を手作業により行っている．

1. 逆画像検索によって，判定対象の画像と完全一致する画像が使われている出所記事 W を収集する．
2. W の中で公開日が最も古い記事 w_{oldest} を可能な範囲で特定する．w_{oldest} が判定対象の SNS 投稿そのものであるならば，$I_{description} = origin$ として，画像の情報補完を終了する．ここで，情報収集対象の記事を $w = w_{oldest}$ とする．
3. 2. で特定した w の中から，画像について説明している箇所を特定する．
4. 3. で特定した画像の説明箇所をもとに，$I_{description}$，$I_{location}$，I_{time} に値を設定する．ここで，I_{time} は画像の撮影日もしくは作成された日を説明箇所から抽出するが，抽出できなければ I_{time} に w の公開日を代入し，w の公開日も不明な場合は $I_{time} = NULL$ とする．また，$I_{description}$，$I_{location}$ について，それぞれ情報が抽出できない場合は，$I_{description} = NULL$，$I_{location} = NULL$ とする．
5. 4. において，$I_{description} = NULL$ または $I_{location} = NULL$ または $I_{time} = NULL$ の場合，W の中で w のつぎに公開日が古いもの（w_{-1}）が存在し，その内容が w と矛盾していないならば，$w = w_{-1}$ として 3. に戻る．W の中で w のつぎに公開日が古いものが存在しない場合，ここで画像の情報補完を終了する．また，$I_{description} \neq NULL$ かつ $I_{location} \neq NULL$ かつ $I_{time} \neq NULL$ の場合，画像の情報補完を終了する．

以上の手順で生成された $I_{description}$，$I_{location}$，I_{time} を用いて，画像の情報補完部の出力である画像の説明文 I_{claim} を生成する．まず，$I_{claim} = I_{description}$ としたのち，I_{claim} の内容と矛盾のない範囲で，$I_{location}$，I_{time} を用いて，どこで，いつ撮影された画像かという内容を I_{claim} に加え，更新する．

この手順により，判定対象の画像の出所記事から，何を説明する画像であって，いつ，どこで撮影もしくは作成された画像かという情報が補完された I_{claim} を生成し，これを外部チェックの入力として用いる．

3.4 外部チェックの処理

提案手法における外部チェックでは，3.2 の出力である T_{claim} と，3.3 の出力である I_{claim} を入力として受け取り，I_{claim} が T_{claim} を支持しているかどうかという ROOC 判定を行う．ただし，$I_{claim} = origin$ の場合，判定対象の画像と一致する過去の画像が見つからなかったということになるため，ROOC でないと判定する．また，$I_{claim} = NULL$ または $T_{claim} = NULL$ の場合は判定不能とする．

提案手法における外部チェックでは，I_{claim} が T_{claim} を支持しているか否かを LLM を用いて判定することで，ROOC 判定を行う．SNIFFER においても，外部チェックは LLM を用いて行っている．SNIFFER の外部チェックで用いられている LLM プロンプトと，提案手法で用いる LLM プロンプトは，どちらも同じ指示内容で，具体例を 1 つ与える One-shot プロンプトであるが，与える具体例が異なっている．以下，SNIFFER の外部チェックのプロンプトと，提案手法の外部チェックのプロンプトの違いを説明する．

プロンプト 1 に，提案手法における外部チェックで用いる LLM プロンプトを示す．プロンプト中の下線部は，SNIFFER が外部チェックで用いているプロンプトからの変更点を表している．1 行目は，LLM に与える指示内容を表す．ここでは，以降のプロンプトにおいて判定対象のテキストの主張文と画像の説明文を与えることを LLM に伝え，画像の説明文がテキストの主張文を支持しているかどうか判定し，判定結果と根拠説明文を生成することを指示している．2 行目から 5 行目は LLM に与える入出力の One-shot 回答例である．具体的には，2 行目は判定対象のテキスの主張文の例，3 行目は判定対象の画像の説明文の例，5 行目は ROOC 判定の回答例を示している．6 行目と 7 行目には，ROOC 判定対象である 3.2 の出力である T_{claim} と，3.3 の出力である I_{claim} を LLM に与えている．

4 評　　価

本章では，提案手法の性能評価を行う．はじめに，性能評価のために SNS 投稿を収集して作成したデータセットについて説明する．つぎに，作成したデータセットを用いて，提案手法の性能を評価する．

4.1 評価データセット

提案手法の性能を評価するために SNS 投稿を用いて作成したデータセットについて説明する．

まず，ファクトチェック団体である日本ファクトチェックセン

プロンプト 1 提案手法における外部チェックの LLM プロンプト（下線部分が SNIFFER からの変更箇所）

1: You will be provided with a claim and some retrieved evidence. You need to determine whether the given claim is supported by these evidence, meaning whether they describe the same news event. It is sufficient if any evidence supports the claim. Please provide your judgment followed by your reasoning.
2: Claim: Due to the uplift caused by the Noto Peninsula earthquake that occurred on January 1, 2024, wave-dissipating blocks were washed up on the beach and became a shoreline obstacle.
3: Evidence: From August 24 to 29, 2020, the Japan Ground Self-Defense Force conducted training at a training ground in Teshio Town, Hokkaido, to construct obstacles on the shoreline to prevent enemy amphibious vehicles from landing.
4: Your answer is:
5: The claim is not supported by the provided evidence. Reasoning: The claim states that on January 1st 2024,the Noto Peninsula earthquake caused the wave-dissipating blocks to become a hazard in the shoreline area. On the other hand, the evidence states that in August 2020, the Japan Ground Self-Defense Force constructed obstacles at a training ground in Hokkaido, so we know that the Ground Self-Defense Force installed the obstacles for training purposes. Therefore, the wave-dissipating blocks in the claim were not affected by the earthquake, and the evidence does not support the claim.
6: Claim: $\{T_{claim}\}$
7: Evidence: $\{I_{claim}\}$
8: Your answer is:

ター（JFC）が作成したファクトチェック記事 [5]において検証された言説のうち，以下の条件をみたすものを収集した．

- SNS 投稿を検証したファクトチェック記事であって，カテゴリが国際または自然災害に分類されており，2025 年 1 月 29 日以前に公開されたもの
- 検証対象の SNS 投稿が，1 文字以上のテキストと 1 枚の画像からなるもの

ここで，JFC によるファクトチェック記事において，記事の数が多いカテゴリは，医療・健康，国際，政治，災害の 4 つのカテゴリである [6]．このうち，医療・健康カテゴリについては，背景情報の収集には医療に関する高度な専門知識が必要となるため，データセットの対象からは除外した．また，政治カテゴリについては，判定対象や背景情報に，政治的な主張や非公開会議の内容に関する憶測，といった裏付けをとることが難しい SNS 投稿が含まれており，ROOC かどうか判断できないものが含まれるため，同じくデータセットの対象からは除外した．

画像を 1 枚に限定したのは，SNIFFER および提案手法はテキストと 1 枚の画像の OOC 判定にしか対応していないためである．また，SNS 投稿に URL が含まれる場合，SNS の機能によって URL 先のページに含まれる画像をプレビューとして表示される場合がある．この機能によって，ユーザから SNS 投稿がテキストと 1 枚の画像に見える投稿も含めた．

上記の条件で SNS 投稿を収集したところ，55 件の SNS 投稿を得た．このうち，2 件は元の SNS 投稿が SNS 上や Web アーカイブから削除されており，再現できないため除外した．

53 件の SNS 投稿のうち，Google Cloud Vision API を用いて投稿内の画像を逆画像検索した結果のうち，完全一致する画像を含む記事の公開日を確認し，対象の SNS 投稿よりも前に公開されている記事が存在する 15 件を，本稿の評価で用いるデータセットの対象とした．SNIFFER では，Google Cloud Vision API による逆画像検索結果に含まれる記事の公開日が，判定対象の投稿日より前かどうかを考慮せずに同一画像が使用されている記事を収集して判定が行なわれている．本稿における評価では，前述した条件を含めて SNS 投稿を収集することで，再利用であることがわかっている投稿のみを用いて評価することとした．

つぎに，この 15 件の SNS 投稿に対して ROOC か NROOC かの正解ラベルを付与し，データセットを作成した．ここで，NROOC は Not ROOC，すなわち ROOC でないことを表すラベルである．正解ラベルはファクトチェック記事を参考にしつつ，ROOC の基準にしたがって人手で与えた．表 1 に，評価用データセットの項目数と，正解ラベルの内訳を示す．また，付録 1 に，作成したデータセットの詳細を示す．

表 1　評価用データセットの正解ラベルの内訳

評価対象	ROOC	NROOC
15	11	4

4.2 評価方法

提案手法の性能評価を行うため，4.1 で作成したデータセットに対して，提案手法と，提案手法のベースとした SNIFFER を用いて ROOC 判定を行う．提案手法と比較するために，SNIFFER における外部チェックを従来手法とした．これは，SNIFFER ではテキストと画像の一貫性を LVLM を用いて検証する内部チェックと，画像の出所情報を用いた外部チェックの 2 つの検証を組み合わせて OOC 判定を行っているが，提案手法では外部チェックの改善を行ったため，外部チェックのみを従来手法として比較することとした．以降，SNIFFER の外部チェック部分を従来手法と呼ぶ．

提案手法における情報補完したテキストの主張，および情報補完した画像の説明については，それぞれ 3.2 と 3.3 の手順にしたがって，あらかじめ手動で作成した．また，提案手法では，SNIFFER の評価で使用されているものと同じ Llama-2-13b-chat-hf[7]という LLM を利用し，この LLM に対してプロンプト 1 を入力して ROOC 判定を行った．

従来手法では，ROOC かどうかの判定結果を「The claim is

5：https://www.factcheckcenter.jp/tag/fact-check/
6：https://www.factcheckcenter.jp/explainer/fact-check/jfc-fact-checking-101/
7：https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

表 2　従来手法と提案手法の正解率

	正解率
従来手法	0.667
提案手法	**0.800**

表 3　従来手法の混同行列

		判定結果のラベル	
		ROOC	NROOC
正解ラベル	ROOC	7	4
	NROOC	1	3

表 4　提案手法の混同行列

		判定結果のラベル	
		ROOC	NROOC
正解ラベル	ROOC	11	0
	NROOC	3	1

not supported by the provided evidence.」または「The claim is supported by the provided evidence.」という形式で出力する．ここで，claim は判定対象のテキスト，evidence は判定対象の画像を逆画像検索して得られる画像の情報である．本評価では，SNIFFER による外部チェックの出力のうち，前者を ROOC，後者を NROOC であるとみなすこととした．

今回の評価では，比較評価のために提案手法の出力形式を SNIFFER の外部チェックに合わせて評価を行った．つまり，プロンプト 1 の 1 行目の文末に「You should answer in the following forms: ‘The claim is not supported by the provided evidence. Reasoning: ...’ or ‘The claim is supported by the provided evidence. Reasoning: ...」という文言を付加して LLM に入力した．

4.3 評価結果

表 2 に，従来手法と提案手法の正解率を示す．ここで，正解率とは，正解ラベルと判定結果のラベルが一致した数を，データセット内の事例の件数で割った値である．表から，提案手法は従来手法よりも正解率が向上していることがわかる．また，表 3 と表 4 に，従来手法と提案手法それぞれの混同行列を示す．表から，提案手法では従来手法に比べて ROOC の正解数は増加しているが，NROOC の誤判定数が増加していることがわかる．

提案手法で誤判定した事例はすべて，正解ラベルが NROOC の 3 件を誤って ROOC と判定したものであった．誤判定の原因は主に次の 2 つである．1 つは，提案手法では背景情報を補完した上で ROOC 判定を行っているため，判定対象と背景情報が一致しているかどうか確認する際に確認すべき観点が増え，重要でない差異を理由として ROOC と判断しているためである．この対策として，LLM に対して補完した情報のどこに着目して比較すべきかを明示的に指示する必要があると考えられる．もう 1 つは，画像の出所記事に画像の内容を説明する記述がほとんどなく，提案手法であっても ROOC 判定に有効な画像の背景情報を収集できていないためである．この対策として，画像の出所記事に画像の内容を説明する十分な記述が存在しない場合，画像自体の内容を解釈することで画像の説明文を補完する必要があると考えられる．

一方で，従来手法で誤判定しているが，提案手法によって正解した事例は 4 件であった．従来手法では，逆画像検索により画像の正しい文脈が含まれる記事を見つけられていたにもかかわらず，文脈を正確に把握するために必要な情報が記事タイトルやキャプションに含まれていなかったため，誤判定していた．提案手法では，逆画像検索により見つけた記事の本文をもとに画像の文脈を補完することで，正しく ROOC であると判定できるようになった．

本稿では評価の対象としていないが，従来手法と提案手法それぞれを適用した際の外部チェックにおける根拠説明文も比較し，考察を行った．根拠説明文は ROOC 判定の根拠を示しているが，判定に用いる情報が不足している場合は妥当な結論を導くことができない．本評価で用いた SNS 投稿は，背景情報が十分に含まれていない事例が多い傾向がある．入力に背景情報が不足している事例では，従来手法によって出力される根拠説明文における理由が正しく生成されず，判定結果の妥当性を判断できないものがみられた．一方で，提案手法では，テキストの情報補完により得られた主張文や，画像の情報補完により得られた画像の説明文を入力に用いて根拠説明文を出力することで，納得性が高い根拠説明文が出力できた．

5 おわりに

本稿では，SNS 投稿に対する OOC 判定の精度向上を目指して，SNIFFER をベースに，SNS 投稿の周辺情報からその内容や場所，時間に関連する情報を収集し，テキストと画像それぞれの背景情報を補完したうえで OOC 判定を行う手法を提案した．

ファクトチェック機関によって背景情報の検証が終わっている X（旧 Twitter）の投稿事例 15 件を用いて，従来手法である SNIFFER における外部チェックと提案手法における外部チェックの OOC 判定精度を比較評価し，提案手法では従来手法より高い判定精度が得られることを確認した．

今後の課題は以下の通りである．

- 情報補完によって情報量が増加することで，重要でない差異をもとに誤判定してしまう事例があったため，収集した情報のどこに着目して比較すべきかを LLM に対して指示する手法を検討する．
- 画像の出所記事に画像の内容を説明する十分な記述が存在せず，OOC 判定に有効な画像の情報を補完できずに誤判定してしまう事例があったため，画像の内容に関する説明が十分に収集できない場合には画像そのものを解釈する手法を検討する．
- 本稿における評価で用いたデータセットは事例数が 15 件と少なく，正解ラベルにも偏りがあるため，より多くの事例を用いて評価を行う．
- 本稿ではテキストの情報補完と画像の情報補完は手動で行

い，情報補完によって OOC 判定の精度が向上することを確認したため，情報補完を高い精度で自動実行する手法を検討する．

謝　　辞

この成果は，NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託業務（JPNP22007）の結果得られたものです．

文　　献

[1] Z. Guo, M. Schlichtkrull, and A. Vlachos, “A survey on automated fact-checking,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, Feb. 2022.

[2] L. Fazio, “Out-of-context photos are a powerful low-tech form of misinformation,” *The Conversation*, vol. 14, p. 1, Feb. 2020.

[3] S. Aneja, C. Bregler, and M. Nießner, “Cosmos: Catching out-of-context misinformation with self-supervised learning,” *arXiv preprint arXiv:2101.06278*, Apr. 2021.

[4] S. Abdelnabi, R. Hasan, and M. Fritz, “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*, pp. 14940–14949, June 2022.

[5] P. Qi, Z. Yan, W. Hsu, and M. L. Lee, “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*, pp. 13052–13062, June 2024.

[6] G. Luo, T. Darrell, and A. Rohrbach, “NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, eds.), (Online and Punta Cana, Dominican Republic), pp. 6801–6817, Association for Computational Linguistics, Nov. 2021.

[7] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, pp. 74–81, July 2004.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*, pp. 8748–8763, PMLR, July 2021.

[9] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “Visualbert: A simple and performant baseline for vision and language,” *arXiv preprint arXiv:1908.03557*, Aug. 2019.

[10] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “InstructBLIP: Towards general-purpose vision-language models with instruction tuning,” in *Thirty-seventh Conference on Neural Information Processing Systems*, Sept. 2023.

[11] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, “Finetuned language models are zero-shot learners,” in *International Conference on Learning Representations*, Jan. 2022.

[12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, July 2023.

付　　録

1　評価のために作成したデータセット

評価で用いた評価データセットを表 5 に示す．表の項目はそれぞれ，SNS 投稿を収集するために参照したファクトチェック記事のタイトルとその記事へのリンク，それぞれの SNS 投稿について情報補完して作成したテキストの主張文と画像の説明文，ROOC かどうかを筆者らが確認して付与した正解ラベルを表している．実際のデータセットには判定対象のテキストや画像なども含むが，ここでは省略している．

表 5　評価用データセットの詳細

No.	ファクトチェック記事タイトル	提案手法によって情報補完された内容		正解ラベル
		テキストの主張文	画像の説明文	
1	東日本大震災での自粛や批判をきっかけに陸上自衛隊で赤飯が廃止された？【ファクトチェック】	東日本大震災後、陸上自衛隊は赤米の提供を中止しました。	2009 年に陸上自衛隊広報チャンネルが YouTube に投稿した人気戦闘食ランキングの一場面。	ROOC
2	台風で駅が浸水しても陽気なスウェーデンの人々の画像？【ファクトチェック】（訂正あり）	台風で駅が冠水したにもかかわらず、スウェーデンの人々は明るさを保っていた。	2018 年、スウェーデン人は洪水に見舞われたものの、スウェーデン第 4 の都市ウプサラ駅ではそれほど心配していなかった。	NROOC
3	スターバックスがアメリカで禁止? パロディアカウントを引用【ファクトチェック】	2025 年 1 月 20 日、スターバックスは米国で禁止されました。	スターバックスのロゴは 2011 年から使用されています。	ROOC
4	オバマ元大統領は「小児性愛者」? 姪と撮影した公式写真【ファクトチェック】	子供を膝の上に乗せて椅子に座るバラク・オバマの姿は、あまりにもひどいものでした。	2015 年 12 月 4 日、バラク・オバマは大統領執務室で姪のサビタ・ンと寄り添った。	ROOC
5	ウクライナに支援してもゼレンスキーの別荘になるだけ? 再拡散【ファクトチェック】	2024 年 5 月 15 日に、ウクライナへの支援がゼレンスキー氏の別荘に送られます。	2016 年にハイグローブ庭園が空から撮影されました。	ROOC
6	イランのヘリ事故でライシ大統領は無事？【ファクトチェック】	2024 年 5 月 19 日、イランのライシ大統領を乗せたヘリコプターがヴァルザカン市近くで墜落しましたが、ライシ大統領は生き残り、タブリーズへ向かっていました。	イラン赤新月社（IRCS）によると、2022 年 7 月 30 日、イランのテヘラン市フィルズクーで、イランのエブラヒム・ライシ大統領が洪水後のフィルズクーを訪問した。	ROOC
7	「（画像）『最も嫌われている国』の地図」は誤り　出典とされた団体が否定【ファクトチェック】	世界中の都市の人々はイスラエルを憎んでいます。	各国で最も嫌いな国 2022 年 8 月 25 日。	NROOC
8	ゼレンスキー大統領がチャールズ 3 世の邸宅を購入したと英国が発表？【ファクトチェック】	2024 年 4 月 4 日、ロンドンで、ウラジミール ゼレンスキーはカール 3 世から邸宅を購入しました。	ロンドン・クライアー紙によると、2024 年 4 月 4 日、ロンドンでゼレンスキー大統領はカール 3 世の邸宅を 2000 万ポンドで購入した。	NROOC
9	オジマンディアス（ファラオ）に発行されたパスポートの画像？【ファクトチェック】	オジマンディアスの遺体は現在エジプトにありますが、カビを除去するために 1970 年代にフランスに移送され、その際に正式なパスポートが発行されました。	2020 年 10 月 20 日、ラムセス 2 世のミイラのパスポートをデジタルで作成した画像がネット上に出回り、1979 年にフランスへの移送に使用されたと偽った。ルーブル美術館の専門家と公法の教授は、この主張の虚偽を確認し、ミイラの移送をめぐる法的背景を説明した。	ROOC
10	「テキサス州が戦争状態に突入」「アメリカで内戦」を伝える動画?【ファクトチェック】	メディア報道が移民をめぐるテキサス州と連邦政府の対立を指摘する中、テキサス州は 2024 年 1 月 27 日に戦争状態に突入しました。	2024 年 1 月 25 日、グレッグ アボットはテキサス州憲法における自衛権に関する声明を発表しました。	ROOC
11	9.11 同時多発テロ　ユダヤ人は犠牲者にひとりもいなかった？【ファクトチェック】	2023 年 9 月 11 日、日本で NHK 記者の長谷川博司氏が「ユダヤ人は全員仕事を休んでおり、被害者には含まれていない」と暴露した後に転落死し、警察は自殺と判断し、真実を語ることが失踪につながる可能性があるとの懸念が高まった。	2001 年 9 月 11 日、ハイジャックされた旅客機がニューヨークの世界貿易センターに墜落しました。	ROOC
12	マウイ島火災発生の直前にレーザー光線？【ファクトチェック】	2023 年 8 月 11 日、レーザーのような光がハワイのマウイ島を襲いました。	2018 年 5 月 22 日、スペース X 社のファルコン 9 がカリフォルニア州セントラルコーストのヴァンデンバーグ空軍基地から打ち上げられた。	ROOC
13	ジョー・バイデン死去？史上最高齢の米大統領、死去情報は繰り返し拡散【ファクトチェック】	2023 年 7 月 1 日、アメリカでジョー バイデンが心臓発作で亡くなりました。	バイデン大統領の公式大統領肖像画が 2021 年 3 月 3 日に公開された。	ROOC
14	（画像）アフリカ 40 カ国の代表がモスクワにいる？会議は開かれているが別の画像【ファクトチェック】	2023 年 3 月 20 日、アフリカ 40 か国の代表者がモスクワを訪れましたが、西側メディアはこれについて報道していませんでした。	2019 年 10 月 24 日にロシア・ソチでロシア・アフリカ首脳会議が開催された。	ROOC
15	各国人が最も嫌っている国の画像は本物？【ファクトチェック】	各国の人々が特定の国を最も嫌っていることが観察されました。	これは捏造です。	NROOC