日本―スペイン、ポーランド　国際共同研究 「レジリエント、安全、セキュアな社会のための ICT」 2023 年度　年次報告書	
研究課題名（和文）	ソーシャルメディアプラットフォームにおけるフェイクニュース検出 (DISSIMILAR)
研究課題名（英文）	Detection of Fake News on Social Media Platforms (DISSIMILAR)
日本側研究代表者氏名	栗林　稔
所属・役職	岡山大学 学術研究院環境生命自然科学学域・准教授 東北大学 データ駆動科学・AI 教育研究センター・教授
研究期間	2021 年 4 月 1 日～ 2024 年 12 月 31 日

１．日本側の研究実施体制

氏名	所属機関・部局・役職	役割
栗林　稔	岡山大学学術研究院環境生命自然科学学域・准教授 東北大学データ駆動科学・AI 研究センター・教授	研究全体の統括
Asad Malik	Aligarh Muslim University・Research Associate Monash University Malaysia・Lecturer	プログラム等の実装

２．日本側研究チームの研究目標及び計画概要

WP②については、 セミフラジール電子透かし技術の実装を進め、圧縮などの正常な範囲の処理では問題なく透かし情報を取り出して検証できるが、意図的な加工や改ざんがなされると、その形跡を確認できる手法の実装を進めて行く。動画像コンテンツをフレームレートに応じて、フレーム単位で切り出し、そのフレームに対して低周波から中周波の成分に対して量子化法に基づく電子透かし法を用いた埋め込みを行う。その際に最新の圧縮コーデックによって埋め込んだ透かし情報が壊れない程度の埋め込み強度を設定できるようにシミュレーション条件を変更しながら確認していく。埋め込む透かし情報は、各フレームおよび音声信号から取り出した特徴成分を秘密鍵を用いて変調させた信号とする。正常な範囲での圧縮においてはロバスト性を担保しつつ、意図的な加工や編集が加えられると変化しやすくするために、この変調操作において誤り訂正符号を活用した手法を採用する。オプションとして、非代替トークンを適用したコンテンツ保護の手法が本システムの実装上で利用できない

か確認する。

WP③では、敵対的攻撃によってフェイクコンテンツを見破る技術が誤検知を引き起こすことを防ぐための、前処理フィルタの実装を進め、既存の複数の識別器(Mesonet, CLRnet など)を基に識別器モデルを設計し、その出力を多数決させる形で最終判定を出力する流れを実装させる。ただし、計算量を考慮して、個々の識別器モデルは軽量化を図る。また、設計した識別器モデルは、複数のフェイクコンテンツのデータセットを用いることで、汎用的に高い精度が得られるように学習させる。また、それらをプラグインとして動作させて全体のシステム上での処理速度の計測を行う。さらに最適化を進めて、計算コストおよび通信コストの削減を図る。

3. 日本側研究チームの実施概要

本プロジェクトでは、フェイクコンテンツを検出するためにハイブリッド型の構造にて処理をするアプリケーションの作成を目指している。一つはコンテンツに含まれる歪みを解析するフォレンジクス技術である。コンテンツの加工・編集などの処理により生じる歪みや、生成系 AI に由来するコンテンツ内の不自然な信号成分を対象としており、受身的な対策である。もう一つは、積極的にコンテンツを守るために原本性保証の情報や改ざん検知信号を電子透かし技術を用いて忍ばせる対策である。

コンテンツの原本性を保証するため、フラジール電子透かしとして埋め込んだ信号パターンを検出できなかった箇所を白画素として、検証用の白黒画像を作成し、視覚的に加工・編集箇所を視認できるようにプログラムを改良した。DISSIMILAR システムの電子透かしプラグインとして実装を進め、加工・編集箇所を白黒二値画像として表示する処理の流れをスペイン側グループと共に開発した。

フォレンジクス技術に関連させて深層学習モデルに対する電子透かし技術の研究においては、重み一定符号を適用することで、モデル軽量化のためのプルーニング処理に耐性を持たせる手法を提案した。コンテンツの正当な加工・編集は履歴を誰でも検証できるような枠組みについては、構想のアイディアを具体的に実現する方法をいくつか検討しており、ブロックチェーンに紐づける方法を構想したところである。具体的な実装評価は今後の課題として残っている。

DISSIMILAR システムのフォレンジクスプラグインにおける敵対的攻撃への対策として、フィルタの強度を変更することで、深層学習モデルの出力がどのように変化するのかを観測し、出力成分の変化量を特徴成分として分類する手法をこれまでに提案してきた。フィルタ操作により取り除かれる信号成分は、敵対的ノイズと高い相関があることから、フィルタ前後の画像の差分を特徴成分として解析するアプローチを新しく考案し、実装評価を行った。その結果、計算量を従来手法と比べて半分以下に抑えつつ、ハイパーパラメータが異なる既知の敵対的攻撃や未知の敵対的攻撃に対しても高い分類精度が得られることを確認した。

フェイクコンテンツの識別においては、学習済みモデルをファインチューニングしてフェイクか否かを二値判別できる識別器を設計し、公開されているフェイクコンテンツのデータセットを用いて訓練させることで、核となる二値判別器を作成した。プラグインとして動作させるために、入力には個々に指定した画像もしくは動画ファイルだけでなく、対象となるコンテンツ一式をまとめたフォルダを指定できる仕様とした。また、出力においては、二値判別器が出力する確信度に応じて、加工・編集された箇所において重要となる箇所のヒートマップ画像を生成し、可視化された結果を出力させる仕様とした。

フェイクコンテンツの作成手段を推定には、既知の作成方法をクラス分けする推定器を CNN モデルにて設計した。その出力における確信度の値に応じてオンオフを切り替える活性化関数を提案し、そのクラスに応じた二値判定器に繋げるアーキテクチャを設計した。最終的に重み付き多数決判定にてフェイクであるか否かを判定する。生成 AI の登場に伴って、

既存のフェイクコンテンツのデータセットだけでなく、利用可能な画像生成 AI を用いて画像を生成し、提案手法の性能を検証した。その結果、推定器においては 90%以上の高い精度でクラス分類できることが確認できた。活性化関数や二値判別器におけるハイパーパラメータの設定については、今後の課題として残っている。