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」 2022 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　学術研究院自然科学学域　准教授
研究期間	2021 年 4 月 1 日～2024 年 3 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・学術研究院自然科学学域・准教授	研究全体の統括
Asad Malik	Aligarh Muslim University・Research Associate	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

中心的な役割を担っている WP③に対して、敵対的生成ネットワーク(GAN)によって作成されたフェイクコンテンツに含まれる不自然な信号を検出してフェイクを見破る技術を実現するシステムの提案と評価実験を行う。コンテンツに対して、局所的な特性に応じた領域分割を行い、その領域内における歪みを解析するだけでなく、各局所領域における特徴を空間領域と時間軸領域共に関連付けて最終的な判別が行えるような手法を実装していく。

また、WP②においてコンテンツの原本性を確保するための枠組み作りにおいても、漏洩元を追跡する電子指紋情報を扱った情報埋め込み手法および抽出法の検証と、暗号プロトコルも含めたシステムの設計を目指す。

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。

フォレンジクス技術では、対象となるコンテンツに対して、人の顔部分を検出した後に、人工的に作成・加工・編集された形跡の有無を判定する。生成系 AI より人工的に創造された人の顔写真と、正常な写真を判定するために、大量の画像データを公開されたデータセットから収集するだけでなく、生成系 AI を用いて作成した。このデータセットを用いて、識別器を学習させることでシミュレーションを進めてきた。フェイクコンテンツの識別器の設計では、深層学習技術によって設計された既存の画像識別器を、フェイクと正常な画像を判別する二値分類器に修正することで実現する手法を選択した。日本側研究チームで作成した画像データセットを用いて、この二値分類器を学習させた結果、同じ生成系 AI によって創造される画像においては 95%以上の高い精度でフェイク画像を識別できることが確認できた。一方で、異なる生成系 AI によって創造された画像に対してはその識別精度が下がる傾向も見られた。識別に有効な色空間に変換する前処理を施すことで、提案した二値判別器の性能を向上させることにも成功している。畳み込みニューラルネットワークでは、画像の局所領域における特徴成分を収集することに対して、本前処理では画像の色空間の特性において識別に有益な成分だけを抽出することが可能であり、その結果として高い識別性能を得ることができたものと理解している。

フォレンジクス技術は AI 技術など機械的に判別する手法であることから、意図的にノイズを加えて識別器が誤動作を引き起こすようにフェイクコンテンツを加工する可能性も考えられる。特に、敵対的攻撃と呼ばれる AI システムを誤動作させることに特化した攻撃がなされると、フォレンジクス技術が正常に動作しなくなる。その対策として、提案した二値分類器に入力する前に、敵対的攻撃がなされていかを検証するための処理を行う防御システムの提案と検証を行った。シミュレーションの結果、90%以上の精度で敵対的攻撃からフォレンジクス技術のシステムを防御できる手法であることが確認された。