信頼される AI システムを支える基盤技術
2020 年度採択研究代表者

2021 年度
年次報告書

越前　功

情報・システム研究機構　国立情報学研究所
情報社会相関研究系
教授

インフォデミックを克服するソーシャル情報基盤技術

§1. 研究成果の概要

本研究課題は，AI により生成されたフェイクメディア（FM）がもたらす潜在的な脅威に適切に対処すると同時に，多様なコミュニケーションと意思決定を支援するソーシャル情報基盤技術を確立することを目的とする．具体的には，AI により生成されたフェイク映像，フェイク音声，フェイク文書などの多様なモダリティによる FM を用いた高度な攻撃を検出・防御する一方で，信頼性の高い多様なメディアを積極的に取り込むことで人間の意思決定や合意形成を促し，サイバー空間における人間の免疫力を高めるソーシャル情報基盤技術を確立する．具体的には，(1)多様なモダリティによる高度な FM 生成技術，およびそれらの FM に対する(2)検出・防御技術，(3)FM 無毒化技術，(4)インフォデミックを緩和し多様な意思決定を支援する情報技術の確立を目標としている．

2021 年度の主だった成果は以下の通りである．(1)多様なモダリティによる高度な FM 生成技術では，メディアクローン型の手書き文書画像・動作映像・顔画像 FM の生成，やトピックに応じたテキスト生成による個人的嗜好を考慮したなりすまし投稿生成など，多様なモダリティを対象とした高度な FM 生成技術を検討するとともに，プロパガンダ型 FM 生成に向けたデータセット構築の検討を行った．(2) FM 検出・防御技術では，フェイク顔映像検出のための WebAPI からなるプログラム群を開発し，判定対象となる映像のアップロードから，判定結果を示した映像をダウンロードするまでの全てのプロセスを Web API として利用可能なアプリケーションを実装した．また，自然言語を対象とした自動ファクトチェックの検討，映像における印象操作の対象人物の検出と印象推定，フェイクメディア検知に有用なアテンションマップを利用した説明可能な AI 技術，頑健な FM 検出に有用な秘密鍵を用いた DNN モデルのアクセス制御などを検討した．(3)FM 無毒化技術では，ノイズ除去による敵対的サンプル（AE）型 FM の検出に加えて，企業の財務諸表における欠損及び不正確な数値の補修といった，新たな視座からも検討を行った．(4)インフォデミックを緩和し多様な意思決定を支援する情報技術の確立では，インフォデミックにおいて Bot が偽情報の拡散に及ぼす影響の分析や，インフォデミック下における反ワクチン運動の実態調査など，計算社会科学の手法を用いて，現実の脅威を科学的に解明するための検討を行った．

§2. 研究実施体制

(1)越前グループ(研究機関別)

① 研究代表者:越前 功 (国立情報学研究所　情報社会相関研究系　教授)

② 研究項目

・FM 検出・防御技術

・FM 無毒化技術

(2)馬場口グループ(研究機関別)

① 主たる共同研究者:馬場口 登 (大阪大学　工学研究科　教授)

② 研究項目

・多様なモダリティによる高度な FM 生成技術

・FM 無毒化技術

(3)笹原グループ(研究機関別)

① 主たる共同研究者:笹原 和俊 (東京工業大学　環境・社会理工学院　准教授)

② 研究項目

・インフォデミックを緩和し多様な意思決定を支援する情報技術

【代表的な原著論文情報】

1. Huy H. Nguyen, Sébastien Marcel, Junichi Yamagishi. Isao Echizen, "Master Face Attacks on Face Recognition Systems", IEEE Transactions on Biometrics, Behavior, and Identity Science (IEEE TBIOM), 14 pages, Early Access, April 2022
2. Mahdi Khosravy, Kazuaki Nakamura, Yuki Hirose, Naoko Nitta, and Noboru Babaguchi, "Model Inversion Attack by Integration of Deep Generative Models: Privacy-Sensitive Face Generation From a Face Recognition System", IEEE Transactions on Information Forensics and Security, Vol.17, pp.357-372, January 2022.
3. Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, "OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild", in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10097–10107, November 2021.
4. Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara, "SCOUTER: Slot Attention-based Classifier for Explainable Image Recognition," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5422–5432, November 2021.
5. Xu Wentao and Kazutoshi Sasahara, "Characterizing the roles of bots on Twitter during the COVID-19 infodemic", Journal of Computational Social Science, 5(1), pp.

591-609, August 2021

以上