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

2020 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，(2)FM 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．

2020 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，顔認識器の出力（人物名）から入力画像を推定する Model Inversion Attack に着目し，顔認識モデルのホワイトボックスおよびブラックボックス条件下で，入力画像を生成する基本手法を確立した．また，メディア処理に耐性を持つ敵対的サンプル生成の基本手法を確立した．(2) FM 検出・防御技術では，FM 検出性能向上を目指した高品質な大規模データセットを構築するとともに，顔映像を対象とした FM 検出手法の改良，敵対的サンプル攻撃に対するモデル防御法を確立した．(3)FM 無毒化技術では，データセットのバイアス低減を無毒化と捉え，バイアス低減手法の基礎検討を行った．(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，COVID-19 に関する大規模なソーシャルメディアデータを収集し，誤情報拡散の分析を行った．このほか，FM 検出手法の技術移転を見据えた取り組みを開始した．

§2. 研究実施体制

(1)越前グループ

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

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

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. M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, N. Babaguchi: ″Model Inversion Attack: Analysis under Gray-box Scenario on Deep Learning based Face Recognition System″, KSII Transactions on Internet and Information Systems, Vol. 15, No. 3, pp. 1100-1118 (2021-03). DOI: 10.3837/tiis.2021.03.015
2. 吉村駿佑，中村和晃，新田直子，馬場口登: ″構造未知の画像認識器に対する Model Inversion Attack の検討″，電子情報通信学会 2021 年総合大会, D-12-13, p. 54 (2021-03).
3. 川上蒼太，岡田渓，新田直子，中村和晃，馬場口登: ″半教師あり学習による非視覚センサ値を用いた時間軸をもつ画像列生成″，電子情報通信学会パターン認識・メディア理解(PRMU)研究会，電子情報通信学会技術研究報告, vol. 120, no. 409, PRMU2020-72, pp. 19 - 24 (2021-03).
4. 馬場口登: ″REAL それとも FAKE″，電子情報通信学会マルチメディア情報ハイディング・エンリッチメント(EMM)研究会，電子情報通信学会技術研究報告, Vol. 120, No. 418, EMM2020-74, pp. 40-45 (2021-03-04).
5. 森勇登，中村和晃，新田直子，馬場口登: ″画像認識器に対する認識器クローン作成攻撃とその検知″，第 24 回画像の認識・理解シンポジウム (2021-07). [発表予定]
6. 内田祐生，新田直子，中村和晃，馬場口登: ″画像の印象操作のためのオブジェクトの外観変換″，電子情報通信学会 2021 年総合大会, D-12-30, p. 71 (2021-03).
7. Marc Treu, Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, Isao Echizen, “Fashion-Guided Adversarial Attack on Person Segmentation”, Workshop on Media Forensics, Computer Vision and Pattern Recognition 2021, 10 pages, accepted, June 2021 (arXiv: https://arxiv.org/abs/2104.08422)
8. Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen, “OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild”, the 2021 International Conference on Computer Vision (ICCV 2021), 8 pages, submitted
9. Canasai Kruengkrai, Xin Wang, Junichi Yamagishi, “A Multi-Level Attention Model for Evidence-Based Fact Checking”, Findings of ACL2021, accepted, August 2021
10. Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen, “Preeminence of the Capsule-Forensics Network in Deepfake Detection”, Handbook of Digital Face Manipulation and Detection (Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez and Christoph Busch, eds.), Chapter 13, 25 pages, Springer, in press, 2021
11. April Pyone MAUNG MAUNG, Hitoshi KIYA, ″Block-wise Image Transformation with Secret Key for Adversarially Robust Defense″ IEEE Trans. on Information Forensics and Security, pp.

2709-2723, March 2021.
12. 濱崎直紀，中村和晃，新田直子，馬場口登: "GAN 識別器のアンサンブル学習による真正画像とクローン画像の識別"，電子情報通信学会パターン認識・メディア理解（PRMU）研究会，電子情報通信学会技術研究報告，vol. 120, no. 409, PRMU2020-70, pp. 7-12 (2021-03).
13. 大迫勇太郎，河野和宏，馬場口登: "敵対的生成ネットワークによる映像改ざん検出法の改良"，電子情報通信学会マルチメディア情報ハイディング・エンリッチメント（EMM）研究会，電子情報通信学会技術研究報告，vol. 120, no. 418, pp. 28-33 (2021-03). << EMM 研究会優秀学生発表賞受賞 >>
14. 栗林稔，舩曳信生，Huy Hong Nguyen，越前功: "複数のフィルタ強度による CNN 画像分類器の応答特性を用いた敵対的事例の検出法"，電子情報通信学会マルチメディア情報ハイディング・エンリッチメント（EMM）研究会，電子情報通信学会技術研究報告，Vol. 120, No. 418, EMM2020-70 (2021-03-04).
15. Wentao Xu, Kazutoshi Sasahara, "Characterizing the roles of bots during the COVID-19 infodemic on Twitter", Journal of Computational Social Science, 17 pages, submitted (arXiv: https://arxiv.org/abs/2011.06249)