ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名：
AIセキュリティの研究

実施時期：2025年度
所属機関名：情報セキュリティ大学院大学
代表者氏名：大塚　玲

成果概要：
敵対的サンプル,バイナリ解析,脆弱性の自動修復のそれぞれで成果があった．敵対的サンプル（Adversarial Example; AE）の，ガウス過程理論に基づて敵対的サンプルの攻撃成功確率の上界を理論的に示し，ABCIを用いた実験により確認した．プログラムの挙動解明を目的とするバイナリ解析において，解析戦略を策定するManagerと，解析ツールを実行するWorkerに役割を分割したマルチエージェント構成が正答率の向上と，トークンコストを抑えられることをABCIを用いた実験により示した．さらに，LLMエージェントによる自動修復において，モンテカルロ木探索と自己反省に基づき，脆弱性の自動修復で高い性能を示した．

成果のポイント：
【成果1: 敵対的サンプル】
本研究では，あるカーネル関数を用いたガウス過程(Gaussian Process; GP)分類において，特定のデータセットに対する敵対的サンプル（Adversarial Example; AE）攻撃の成功確率が，異なるラベルをもつ最近傍点間の距離の関数によって上界づけられることを示した．GP分類におけるさまざまなカーネルパラメータのもとでImageNetを用いた実験を行い，本研究の理論結果を確認した．実験結果は理論結果とよく整合していた．カーネル関数のパラメータを変更すると理論的上界も変化することを示した．Shamir らによりDimpled Manifold Model は，入力データは入力空間において多様体上に分布しており，敵対的サンプルは多様体外に作られるというモデルである．入力データに対する多様体への射影は有効な防御手法となりうるが，識別器の性能劣化を引き起こす可能性があった．本研究においては，Isometric Autoencoder による多様体への射影が本来の推論結果に対する劣化を及ぼさないこと，頑健性が向上することをガウス過程での推論に対して理論的および実験的に示した．
【成果2：LLMエージェントによるバイナリ解析】
プログラムの挙動解明を目的とするバイナリ解析は，LLMエージェントの導入により，計画からツール実行までを自律的に遂行可能な新たな段階を迎えている．しかし，複雑なバイナリを対象とする場合，解析ステップが増大し，LLMのコンテキスト長制約に起因する情報の忘却や解析の中断が課題となる．そこで本研究では，解析戦略を策定するManagerと，解析ツールを実行するWorkerに役割を分割したマルチエージェント構成を提案する．Managerは関数コールグラフ（FCG）を解析全体を俯瞰する「地図」として活用し，解析状態を構造的に管理する．そして，Workerを無記憶で運用することでコンテキスト枯渇の抑制を図る．CTFの課題を用いた評価実験の結果，提案手法はFCGを用いないベースラインと比較して，正答率を向上させるとともに，トークンコストを抑えた効率的な解析を実現した．
【成果3: LLMエージェントによる自動修復 】モンテカルロ木探索（MCTS）と自己反省（Self-Reflection）を統合した LATS（Language Agent Tree Search）に基づくエージェントを開発し，ExtractFix および AI×CC Nginx Challenge を用いて評価を行い，修復タスクに対して高い性能を示すことができた．

成果についてより詳細な情報を提供しているWebページ，発表論文などの情報：
Hiroaki Maeshima, Akira Otsuka. “Robustness bounds on the successful adversarial examples in probabilistic models: Implications from gaussian processes.” New frontiers in artificial intelligence: LNAI 15692, pp. 340–354, Springer Nature Singapore, 2025.
Minami Someya, Akira Otsuka. “Empowering LLM-based Malware Analysis with Synthetic Code”, Ps-02, AICompS2025.
前嶋, 大塚, “確率的分類モデルに対する敵対的サンプルへの等長オートエンコーダを用いた防御手法”, 暗号と情報セキュリティシンポジウム SCISム2026.
染谷, 大塚, “関数コールグラフによる解析状態管理に基づくバイナリ解析エージェント”, 暗号と情報セキュリティシンポジウム SCISム2026.
Tomonori Yoneda, Ryotaro Nakata, Akira Otsuka, “Automated Vulnerability Repair based on Language Agent Tree Search,” Pr-31, AICompS 2025.