＜年次報告書＞
第 1 年度（2025 年度）
日 ASEAN 科学技術・イノベーション協働連携事業（NEXUS）
日本－シンガポール「AI」研究領域

研究課題名（和文）	合成データ生成によるロバストな連合基盤モデル
研究課題名（英文）	Robust Federated Foundation Model with Synthetic Data Generation
研究期間	2025 年 4 月 1 日～2028 年 3 月 31 日

氏名	所属 部署　役職
佐久間 淳	東京科学大学 情報理工学院　教授
村上 隆夫	統計数理研究所 学際統計数理研究系　准教授

1. 研究進捗状況と成果

1 連合学習向けの DP（差分プライバシー）
25 年度では、DP を満たす連合学習（DPFL）の最新技術、特にポイズニング攻撃に耐性を持つ Byzantine-robust DPFL を調査した。また、頻度分布や key-value statistics などの統計解析を対象に、従来の shuffle model よりもポイズニング攻撃に強い効率的な DP プロトコルを提案した。提案手法は、shuffler がサンプリングやダミー追加を行う augmented shuffle model に基づき、高い耐性を実現する。本成果は NDSS 2026 に採択された。
2 連合学習実行中に取得されるモデルや勾配の悪用・盗用リスクの評価
25 年度では、複数のファインチューニング済みモデルを追加学習なしに統合する model merging により、第三者が専門能力を低コストで継承できてしまう問題に対し、不要なモデル統合を事前に抑止する手法を提案した。従来の透かしやフィンガープリンティングが不正利用の事後検出を目的としていたのに対し、本研究は、単体性能を維持しつつ、他モデルと統合された場合に性能が劣化するようモデルパラメータを変更する proactive defense を実現した。本成果は ICCV 2025 で発表した。
「国際研究交流」活動
25 年度は、シンガポールチームとの共同研究のための共有環境を整備し、定期的にオンラインミーティングを実施した。また、2025 年 9 月 17 日の RIKEN AIP AI Security and Privacy Team: Achievement Workshop において本プロジェクトを紹介する講演を行った。さらに、日本側から PI 二名（佐久間、村上）および研究者 1 名（Yu）が A*STAR を 2025 年 9 月 28 日から 10 月 2 日に訪問し、対面での研究打合せを行った。

2. 代表的な成果事例

1. T. Murakami, Y. Sei, R. Eriguchi, Augmented Shuffle Differential Privacy Protocols for Large-Domain Categorical and Key-Value Data, Proceedings of the 33rd Network and Distributed System Security Symposium (NDSS 2026), pp.1-20, 2026.
   概要：統計解析において、データポイズニングに対して従来の shuffle model よりも高い耐性を持ち、かつ効率的な augmented shuffle model の DP プロトコルを提案した。提案プロトコルの DP、有用性、データポイズニングへの耐性、効率性などを理論解析および評価実験を通して示した。

2. Junhao Wei, Yu Zhe, Jun Sakuma: Disrupting Model Merging: A Parameter-Level Defense without Sacrificing Accuracy. ICCV (IEEE/CVF International Conference on Computer Vision) 2025: 17698-17707
   概要：第三者が複数のファインチューニング済みモデルを追加学習なしに統合し、専門能力を低コストで継承できてしまう model merging の問題に対し、初の proactive defense を提案した。具体的には、MLP パラメータの再配置と attention head のスケーリングにより、単体性能を維持しつつ、他モデルと統合された場合に性能が大きく劣化するよう保護対象モデルを変換する。画像分類、画像生成、テキスト分類の実験により、通常性能を保ちながら model merging を大きく阻害できることを示した。