AI セキュリティ：誤認識と情報漏洩を防ぐ防御技術

情報科学研究科
知能工学専攻

教授　**田村　慶一**　TAMURA，Keiichi

教員情報

キーワード　AI セキュリティ、敵対的攻撃、連合学習、データセキュリティ、機械学習、深層学習

研究シーズの概要

深層学習（ディープラーニング）が社会に普及する一方で、そのセキュリティ上の脆弱性が大きな課題となっています。データに細工を施して深層学習モデルに誤った判断をさせる「敵対的攻撃」や、公開されたモデルの挙動から学習に使用した元データを推定され、機密情報が漏洩するリスクなどが指摘されています。そこで、これらのセキュリティ課題を克服し、安心・安全に深層学習を運用するための高度な防御・学習技術の研究開発を行っています。

研究シーズの詳細

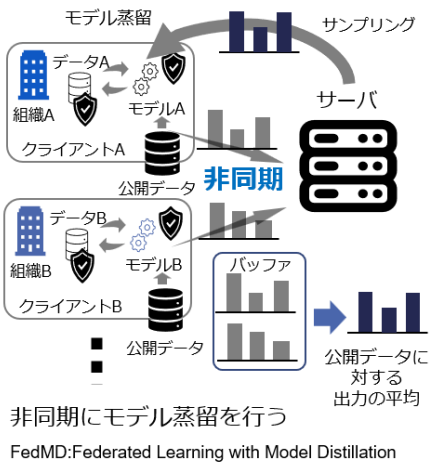
◆研究例◆

【①敵対的攻撃の防御手法に関する研究】

AI を騙そうと意図的に細工された不正な入力（敵対的サンプル）を精度高く検知し、モデルの誤判断を未然に防ぐ、堅牢な防御システムの構築に取り組んでいます。

【②機密性を保つモデル蒸留ベースの連合学習】

連合学習では、データの機密性を保ったまま機械学習モデルを構築できます。データのみならずモデルも秘匿化することで機密性をさらに高めるモデル蒸留に基づく分散型の連合学習手法の開発を行っています。



非同期にモデル蒸留を行う

FedMD:Federated Learning with Model Distillation

【応用先】

提携工場や顧客先へエッジ実装として配布した AI モデルに対し、挙動から製造レシピや設計パラメータ（企業秘密）を逆算・推定しようとする「モデル反転攻撃」を無効化など

応用例・セールスポイント

本技術の導入により、元データやモデル構造を外部に一切開示することなく、安全に組織間での AI 共同学習（連合学習）が進められます。「自社の持つ社外秘データや顧客データを使って安全に AI を学習させたい」「運用中の AI システムが外部からのサイバー攻撃やデータ推定（リバースエンジニアリング）に耐えられるか不安がある」といった課題をお持ちの企業様との共同研究・技術相談を歓迎します。

【問い合わせ先】 広島市立大学　地域共創センター　〒731-3194　広島市安佐南区大塚東 3-4-1（情報科学部棟別館 1F）
TEL:082-830-1764　／　FAX:082-830-1555　／　E-mail:ken-san@m.hiroshima-cu.ac.jp