特集論文

AIの安心・安全を守るセキュリティー技術とプライバシー技術

小関義博*
Yoshihiro Koseki
中井綱人†
Tsunato Nakai

*情報技術総合研究所
†同研究所(博士(工学))

Security and Privacy Technologies for Safe and Secure Artificial Intelligence

要　旨

深層学習の発展によって，AIをビジネスに活用する取組みは，近年ますます活発になってきており，三菱電機でも様々な分野での活用検討が行われている。一方で，AIが組み込まれたシステムの活用が広がるにつれて，悪意を持った攻撃者がAIの持つ脆弱(ぜいじゃく)性を突いて攻撃してくる機会も増加していく。当社では，AIの性能向上や活用検討といった取組みだけではなく，たとえAIが悪意のある攻撃にさらされる環境にあったとしても，安心・安全に活用するためのセキュリティー技術・プライバシー技術の研究開発を行っている。その取組みの中で，物体検知のAIが入力に細工をされても正しい結果を出力する技術と，AIのモデルから学習データに関する情報の漏洩(ろうえい)を防ぐ技術を開発した。

1. ま え が き

2010年代前半に再注目されたニューラルネットワークでの深層学習は，その圧倒的な精度によって画像・映像，言語，音声や時系列データなど様々な分野でのAIに革命をもたらした。近年では，大規模言語モデルを用いた対話型AIや拡散モデルを用いた画像生成AIなどの生成AIが単なる研究対象としてだけでなく，ビジネスへの適用まで含めた大きなブームになっており，AIの実社会での活用は今後も加速度的に進んでいくことが予想される。一方，AIの活用範囲が拡大していくにつれて，AIを用いたシステムへの攻撃機会もまた増加していく。特にAIを用いたシステムでは，ITシステムに対するOSやアプリケーションの脆弱性などを突いたこれまでのサイバー攻撃手法だけでなく，AI固有の脆弱性を突いた攻撃手法についても考慮しなければならない。

本稿では，AIの安心・安全な活用を実現するためのセキュリティー技術やプライバシー技術の研究開発に関する当社の取組みについて述べる。

2. AIでのセキュリティーとプライバシー

近年活発に研究が行われている深層学習は，機械学習というAI技術の一種である。機械学習の利用に当たっては**図1**に示すとおり，学習データを入力してモデルを生成する学習フェーズと，生成されたモデルを用いて入力された画像やテキストなどに対して物体検知や翻訳などのタスクを実行する推論フェーズの二つのフェーズが存在する。この学習フェーズや推論フェーズで，モデルに入力されるデータに細工をすることで意図しない結果を出力させたり，モデルが出力するデータを観察することで学習データに含まれる個人情報などの意図しない漏洩を引き起こしたりする手法などがAIを対象とした攻撃である。AIを対象とした攻撃については，近年精力的な研究が行われており，例えば攻撃手法の一種である敵対的サンプル攻撃に限定しても，2024年３月の時点で8,000本以上の論文が公表されている(1)。

総務省によるAIセキュリティ 情報発信ポータル(2)で，AIを対象とした攻撃の手法がまとめられたAIセキュリティ・マトリックスが提供されている。AIセキュリティ・マトリックスでは攻撃の分類として，データ汚染，モデル汚染，敵対的サンプル，データ窃取，モデル窃取の５分類が定義されている。各分類のイメージと概要を**図1**と**表1**に示す。

データ汚染，モデル汚染，敵対的サンプルの３種類の攻撃は，主にモデルの正しい動作に関連するものであり，データ窃取とモデル窃取は，学習データやモデルのプライバシーに関するものである。本稿では，これらのAIを対象とした攻撃手法に対抗するためのセキュリティー技術・プライバシー技術の当社研究開発での取組みの中から，**３章**では敵対的サンプル攻撃の一種である敵対的パッチ攻撃への対策技術について述べて，**４章**ではデータ窃取攻撃の一種であるメンバーシップ推論攻撃への対策技術について述べる。

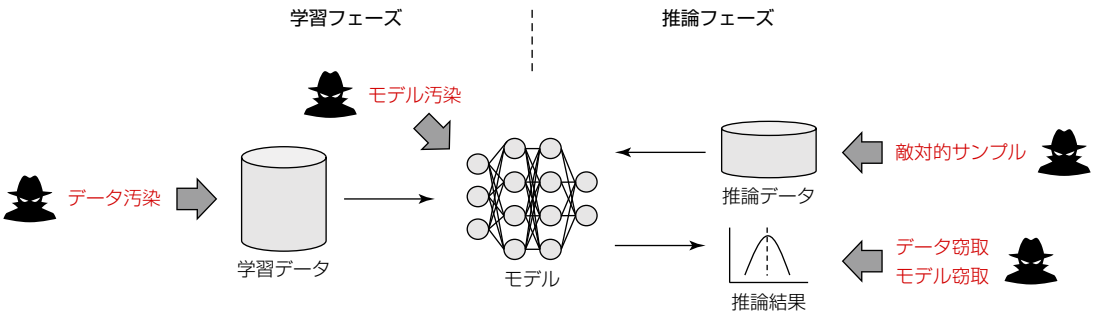

図１−AIを対象とした攻撃のイメージ

表１−AIを対象とした攻撃の概要（AIセキュリティ・マトリックス[2]から引用）

分類	概要
データ汚染	汚染データと呼ばれる特殊なデータを学習データに注入することで，特定の入力データに対して攻撃者が意図した動作をさせるバックドアの設置やモデルの精度低下を引き起こす
モデル汚染	学習後のモデルに細工を行うことで，事前学習済みモデルとして利用者に配布するモデルに，特定の入力データに対して攻撃者が意図した動作をさせるバックドアを設置する
敵対的サンプル	モデルへの入力データに摂動と呼ばれる改変を加えることで，画像の分類や音声認識の結果などを間違ったものに変更する
データ窃取	モデルに複数のデータを入力し，その出力を観測することによって学習に使われたデータの情報を盗み出す
モデル窃取	モデルに複数のデータを入力し，その出力を観測することによってモデルの構造やパラメーターなどの内部情報を盗み出す

3．敵対的パッチ攻撃への対策技術

物体検知とは，入力された画像に写っているそれぞれの物体に対して，物体の位置を示す囲みの座標と，その物体の種類を示すラベルを出力する技術であり，自律走行車での歩行者の検知や監視カメラでの不審人物の検知などに活用されている。物体検知を対象とした敵対的パッチ攻撃は，特殊な方法によって作成されたパッチ画像を入力画像に写った物体の近くに配置することで，その物体の検知を阻害する攻撃であり，物体検知を活用したシステムへの脅威になる。**図２**は既存研究[3]での敵対的パッチ攻撃を実施した例である。左側の人物は囲みと人物であることを示すpersonというラベルが表示されており，正しい検知が行われている。その一方で，右側の人物については囲みとラベルが表示されておらず，配置されたパッチ画像によって正しい検知ができていないことが分かる。またこのパッチ画像は，画像上にデジタル的に配置されているのではなく，印刷して実際の人物の近くに配置した状態でカメラに入力されても効果を発揮している。このような敵対的パッチ攻撃は，物理的敵対的パッチ攻撃と呼ばれ，特に重要な脅威として認識されている。

図２−敵対的パッチ攻撃の例（既存研究[3]から引用）

当社は，敵対的パッチ攻撃が行われた場合であっても，正しく物体を検知するための対策技術を開発した[4]。この技術では，入力された画像に対して，まず通常の物体検知を行う。このとき物体検知の出力には，入力された画像での各物体の位置を示す囲みの座標とともに，そこに物体が存在する確率が物体度スコアとして出力される。物体検知では，この物体度スコアがしきい値以上の値を持つときに，その場所に物体があるとして検知される。一方で，敵対的パッチは周囲の物体の物体度スコアを低下させる効果を持っており，パッチを配置された物体の物体度スコアをしきい値以下にすることで，物体の検知を阻害する。今回の対策技術では，物体度スコアがしきい値未満まで低下している囲みに着目する。**図３**左の画像で点線で示した囲みが，しきい値を下回る物体度スコアを持った囲みであり，この技術では，それぞれの囲みを黒く塗りつぶして複数の塗りつぶし画像を生成する。そして生成された塗りつぶし画像のそれぞれに対して，再度物体検知を行う。敵対的パッチの一部又は全部を塗りつぶした画像では，物体度スコアを低下させる効果が失われ，配置された物体の物体度スコアがしきい値以上になり，正しく物体として検知することが可能になる。複数の塗りつぶし画像それぞれに対して出力された検知結果と，塗りつぶし前の画像に対して行った検知結果に対して，最後に統合処理を行うことで物体検知としての最終的な出力が得られる。

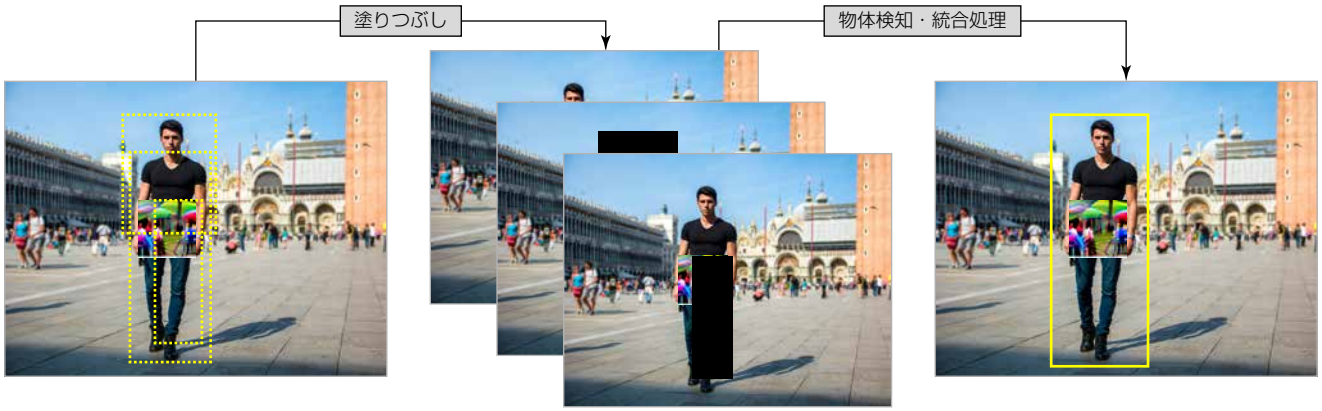

図３-敵対的パッチ攻撃の対策技術での塗りつぶし処理

4. メンバーシップ推論攻撃への対策技術

メンバーシップ推論攻撃は，サンプルデータがモデルの学習に使用されたかどうかを識別する攻撃であり，攻撃者は，サンプルデータに対するAIの推論結果から，データの持ち主が学習データの提供者かどうか特定できる。**図４**にメンバーシップ推論攻撃の概要を示す。メンバーシップ推論攻撃そのものは，学習データに関する情報漏洩として比較的軽微な内容ではあるが，メンバーシップ推論攻撃に耐性を持つ学習モデルは，学習データの情報漏洩に関する他の攻撃に対しても耐性を持つ。そのため，メンバーシップ推論攻撃に対抗する技術は，モデルのプライバシーを実現するために重要になる。

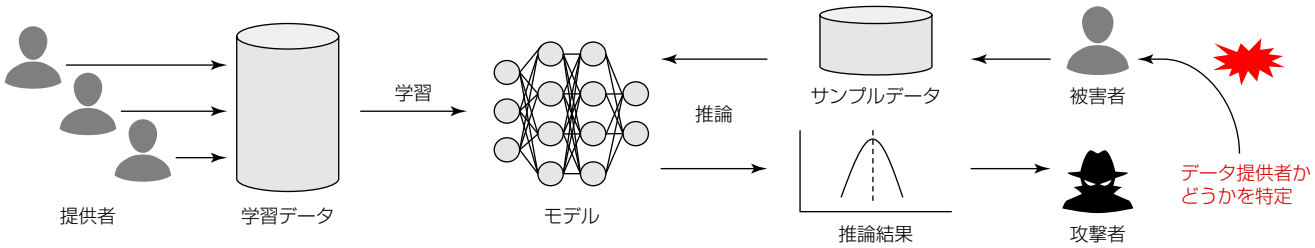

図４-メンバーシップ推論攻撃の概要

メンバーシップ推論攻撃が成功する要因は，モデルが学習データに含まれるデータに対して，他のデータよりも高いスコアを出力するような，強い反応を示してしまうオーバーフィッティング(過剰適合)だと考えられている。当社は，メンバーシップ推論攻撃に対抗する技術として，過剰適合を抑制する学習方式SEDMA(SElf-Distillation with Model Aggregation

for Membership Privacy)を開発した[5]。**図5**にSEDMAの概要を示す。SEDMAは，学習データのラベルを，その学習データを学習に用いていないモデルの推論結果(ソフトラベル)に置き換えて，ソフトラベル付き学習データを新たに生成する。ソフトラベル付きの学習データで生成した対策済みモデルは，元の学習データに対する過剰適合が発生しにくいため，メンバーシップ推論攻撃に強い。ソフトラベルの生成に必要なモデルは，学習データを分割し，それぞれの分割学習データで学習したモデルの集約(モデル間でのパラメーターの加重平均)によって生成した集約モデルを用いる。各集約モデルは，集約された各モデルの学習データに含まれていない学習データにソフトラベルを付与する。このモデルの集約にSEDMAの特長があり，既存の対策方式に比べて，対策によるモデルの精度劣化と対策強度の優れたトレードオフ性能を少ない計算コストで実現している。

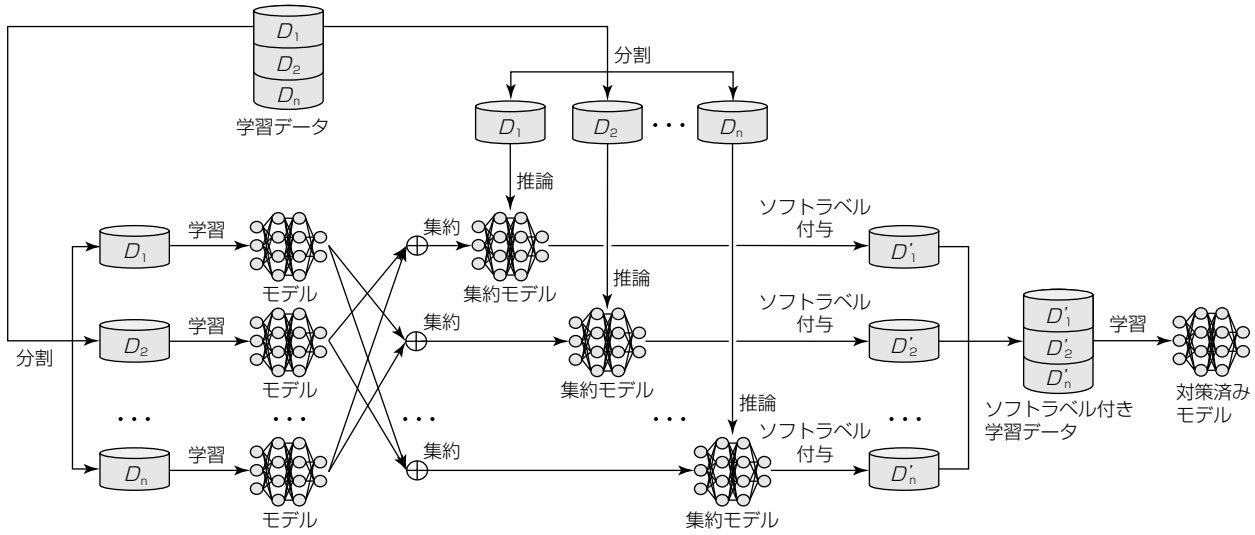

図5-SEDMAの概要

最近では，膨大な学習データを必要とする大規模言語モデルの学習に対して，学習データに関する情報漏洩のリスクが懸念されている。当社は，大規模言語モデルに対しても，メンバーシップ推論攻撃による学習モデルからの情報漏洩リスク評価，既存対策(差分プライバシーなど)の効果検証や新たな対策方式の検討を進めている[6][7]。

5. むすび

AIを対象とした攻撃に対して，物体検知での敵対的パッチ攻撃への対策技術や，メンバーシップ推論攻撃への対策技術についての当社の取組みを述べた。今後は，生成AIとして注目されている大規模言語モデルや，画像や音声なども同時に扱うマルチモーダル大規模言語モデルなども対象として，セキュリティー技術やプライバシー技術の研究開発を更に進めていくことで，AIの安心・安全な活用とその促進に貢献していく。

参考文献

(1) Carlini, N.：A Complete List of All (arXiv) Adversarial Example Papers (2019)
https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html
(2) 総務省，ほか：AIセキュリティ 情報発信ポータル
https://www.mbsd.jp/aisec_portal/
(3) Thys, S., et al.：Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 49～55 (2019)
(4) 小関義博：物体検知における敵対的サンプルパッチ攻撃の無効化技術について，第37回人工知能学会全国大会論文集，2A6-GS-2-01 (2023)
(5) Nakai, T., et al.：SEDMA: Self-Distillation with Model Aggregation for Membership Privacy, Proceedings on Privacy Enhancing Technologies, 494～508 (2024)
(6) 中井綱人，ほか：プロンプト・チューニングは大規模言語モデルの安全性を高めるか？，情報処理学会コンピュータセキュリティシンポジウム，4F2-4 (2023)
(7) 東　拓矢，ほか：大規模言語モデルのファインチューニング手法LoRAにおいて差分プライバシーは有効か？，電子情報通信学会暗号と情報セキュリティシンポジウム，2F2-3 (2024)