機械学習セキュリティ研究のフロンティア

A Guide to Research Frontier of Machine Learning Security

森川郁也 Ikuya MORIKAWA

アブストラクト 機械学習技術は、適切な処理方法を人間が明示するのではなくデータから学ばせるという新しいソフトウェアパラダイムであり、幅広い応用分野に大きな進展をもたらしつつある。しかし未解明の部分も多く、それゆえに機械学習特有のセキュリティ上の弱点が見つかっており、現在盛んに研究されている。本稿では機械学習のセキュリティに関する研究の動向を紹介し、セキュアな人工知能 (AI) 応用システム実現への展望を述べる。

キーワード 機械学習, セキュリティ

Abstract Machine learning is an assembly of technologies that can learn an appropriate way of processing data from data themselves, rather than by explicitly specified by humans. It is a new software paradigm that brings about much progress in many application domains, but it also has many uncharted areas. Recently, it is found that there are types of security flaw specific to machine learning, which are now being actively studied. In this article, we illustrate the research trends of machine learning security and also show a perspective to realize secure applications of artificial intelligence (AI).

Key words machine learning, security

1. はじめに

1.1 AI はだまされる

皆さんは図 1 の画像をご存知だろうか？ これは機械学習に対する攻撃の一つ、敵対的サンプル (adversarial examples) の著名な画像である⁽¹⁾。左は正常な画像であり、機械学習を用いた画像認識器はこれを「パンダ」と判定する。しかし微小な摂動を意図的に加えた右の画像は、人間の目では元の画像と変わらないにもかかわらず、画像認識器に「テナガザル (gibbon)」と誤判定されてしまう。

近年、ニューラルネットワークなど機械学習技術の著しい発展

図 1 敵対的サンプルの例 (Goodfellow ら⁽¹⁾より)

により、画像認識などの応用において人間並みの判断能力をもつソフトウェアが生まれている。一方で、意図的な攻撃にさらされる環境では、前述した敵対的サンプルのように思わぬ脆弱さを露呈することも分かってきた。機械学習を安心して利用するためには、このような攻撃がどの程度可能なのか、なぜ可能なのか、攻撃を防ぐにはどうしたらよいかを明らかにする必要がある。このような試みは機械学習セキュリティ (machine learning security, AI security) や敵対的機械学習 (adversarial machine learning) と呼ばれ、ここ数年、人工知能 (AI) とセキュリティの二つの研究領域において大きな注目を集めている。

本稿では機械学習セキュリティの研究動向を紹介し、セキュアな AI 応用システムの実現に向けた現時点での展望を述べる。AI が社会へ普及していくためにセキュリティの知識は不可欠なので、具体的な事例を挙げながら、初学者やエンジニア一般の読者にも分かりやすい平易な説明を目指す。より深く知りたい方には Papernot らの SoK 論文⁽²⁾をおすすめる。なお、本稿は機械学習特有の脆弱性に注目し、一般的な AI の品質問題や秘匿機械学習 (データを暗号化した状態で行う機械学習) は扱わないが、ご容赦頂きたい。

1.2 準備

最初に機械学習のコンセプトについて簡単に整理しておく。

図 2 に機械学習の一般的な処理フローを示す。本稿では、実施したいデータ処理 (図 2 の下側) のことをタスクと呼び、原則として入力データサンプルを c 個のクラスのいずれかに分類する分類タスクを扱う。分類タスクは各クラスに属する確率ま

森川郁也 富士通株式会社

E-mail morikawa.ikuya@fujitsu.com

Ikuya Morikawa, Nonmember (Fujitsu Limited, Kawasaki, 211-8588 Japan).

電子情報通信学会 基礎・境界サイエンス

Fundamentals Review Vol.15 No.1 pp.37-46 2021 年 7 月

©電子情報通信学会 2021

表 1 機械学習に対する攻撃の分類

攻撃の種類	被害の種類	攻撃フェーズ	Black-box	White-box
敵対的サンプル	推論結果の改ざん	推論	可能	可能
ボイズニング	推論結果またはモデルの改ざん	訓練	—	—
モデル抽出	訓練済みモデルの情報漏洩	推論	可能	容易
訓練データ推定	訓練データの情報漏洩	推論	可能	容易

図 2 機械学習の処理フロー

たは確信度を出力する。分類タスクを実行するプロセスを分類器と呼ぶ。分類器は、 d 次元の実数ベクトル $\mathbf{x}$ を入力とし、 c 次元の実数ベクトル $\mathbf{y}$ を出力する関数 $f(\cdot)$ と表現できる。

一般のソフトウェア開発では、分類タスクの処理 $f(\cdot)$ は人間が仕様を決め、プログラミングによって実現される。これに対し機械学習では、事前に用意した入出力の組 $(\mathbf{x}, \mathbf{y})$ の集合である訓練データセット D_T を学習器が見て、適切な関数 $f(\cdot)$ を学習取ろうとする。このフェーズを訓練と呼び、実際にタスクを実行するフェーズである推論と区別する^(注1)。訓練フェーズでは、関数 f は自由に選べるわけではなく、実施したいデータ処理にあうように人間が選んだ関数の「かたち」の範囲で選ぶ。この関数のかたちをモデルと呼ぶ。これに対し、訓練によって選ばれた f を訓練済みモデルと呼ぶが、単にモデルと呼ぶこともある。

そのほかの機械学習処理フロー^(注2)に対しても攻撃が考えられているが、本稿では分かりやすさを優先し、上記のフローを主な題材として扱う。

2. 機械学習に対する攻撃

まず機械学習に対する攻撃を四つの種類に分類し、各種別の処理フローと既知の事例について見ていこう。表 1 に攻撃を分類した表を示す。

ここで攻撃者のモデルへのアクセス条件であるブラックボックスとホワイトボックスについて説明しておく。ホワイトボックス (white-box) 条件では、攻撃者は訓練済みモデルそのものを入手して、内部のパラメータや動作状態も含めて自由に情報を入手できる。ソフトウェアやハードウェアとしてモデルが手に入る場合はおおむねホワイトボックスに該当し、攻撃者にとって有利な条件である。一方、ブラックボックス (black-box) 条件では、攻撃者はモデルに直接アクセスできず、ユーザとして問合せ (クエリ) を行って結果を得るという形で入出力の組だけを入手できる。クラウドなどリモートサービスとしてモデル

の機能が提供される場合が典型的で、攻撃者にとって不利である。表 1 にあるように、攻撃の種類によって考慮すべきアクセス条件が異なる。

このあと本節では、表 1 の各攻撃種類について基本的な考え方、典型例、具体的な攻撃手法と防御手法などを紹介していく。なお、ほとんどの攻撃事例はセキュリティ研究者らが発見し報告しているものであり、現実被害が起きた事例ではないことをあらかじめ申し添えておく。現実の攻撃や対策に関する考察は 4. 節で述べる。

2.1 敵対的サンプル

敵対的サンプル (adversarial examples) は、攻撃対象の訓練済みモデルが誤った推論をするように、攻撃者が意図的に加工した入力の意味する。推論結果がある種の入力の検知に使われる場合、検知を回避するという意味で回避攻撃 (evasion) と呼ばれることもある。画像認識モデルに対して 2013 年に報告⁽³⁾されてから大きな注目を集めており、図 1 で紹介したパンダの画像⁽¹⁾が有名である。

敵対的サンプルを使った攻撃の処理フローを図 3 に示す。攻撃者は攻撃対象の分類器に合わせて加工した入力 $\mathbf{x}_{adv}$ を用意する。それが分類器 f にわたると、 f は誤った分類結果 $\mathbf{y}_{err} = f(\mathbf{x}_{adv})$ を出力する。 $\mathbf{y}_{err}$ の使われ方によっては被害が生じてしまう。

図 3 敵対的サンプル攻撃のフロー

図 1 のパンダの例は画像認識に対する攻撃であり、現実的な有効性については次の二つの疑問がある。一つめの疑問は、敵対的サンプルにより誤認識が起きても、動物の種類を間違えるという被害に留まるという点である。意図的に誤認識を誘発できると、どんな被害が考えられるのだろうか？ 二つめの疑問は、攻撃者は加工した画像を直接入力しているという点である。図 3 の処理フローにもあるように、実際のシステムではカメラによる撮影 (データ取得) や前処理が入ることが多く、意図的に加工した画像データそのものを入力することは難しい^(注3)。

(注3)：一方で、攻撃者が画像データを直接入力できる場合も考えられる。例えば分類器の API にアクセスできる、またはカメラや中継路を掌握して画像データを改ざんできる、といった場合である。ただし、攻撃者がそこまで深くシステムに関与できる場合は、敵対的サンプルは脅威の一つにすぎず、機械学習に限らず様々な攻撃の可能性を考慮すべきだろう。

(注1)：訓練を学習または適合 (fit)、推論を予測またはテストと呼ぶこともある。
(注2)：回帰や生成モデルなど分類以外のタスクや、訓練データセットが正解ラベルをもたない教師なし学習など。

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5' 0°					
5' 15°					
10' 0°					
10' 30°					
40' 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

図4 交通標識に対する敵対的サンプル (Eykholt ら⁽⁵⁾)

しかしこれらの疑問に対し、より現実的なシナリオでも有効な敵対的サンプルも報告されているので紹介しよう。

Eykholt ら⁽⁵⁾は、画像データを直接操作するのではなく、現実の交通標識に細工することで意図的な誤認識を誘発できることを報告している。図4にその結果の一部を示す。この研究では、人間には「曲がれ」や「止まれ」に見えるが画像認識器には「40 マイル制限」と誤認識されやすい敵対的サンプル画像を作成し、紙への印刷や標識へのステッカー貼付によって現実世界に用意した。それらを様々な角度や距離からカメラで撮影した後に画像認識させても、66.67%から100%という高い確率で誤認識される。自動運転において道路標識の認識を画像認識に頼る場合、こうした誤認識は交通事故につながりかねない。

ほかにも、誤認識される物体を3Dプリンタで作った事例⁽⁶⁾、眼鏡フレームで顔認識器に誤認識させる事例⁽⁷⁾、部分的な模様(パッチ)で誤認識させる事例⁽⁸⁾など数多くの報告があり、現実的なシナリオでも敵対的サンプルは有効と考えるべきである。

ここまで画像データに対する事例を見てきたが、敵対的サンプルは画像以外のデータにも適用可能である。画像に対する敵対的サンプル以前に、マルウェア分類器やスパムメールフィルタに対する回避攻撃が報告されており、10年以上の歴史がある⁽⁴⁾。このような機械学習のセキュリティ応用は敵対的サンプルが現実的となりうる利用形態といえる。なぜなら、そもそも第三者から受け取ったデータを検査するのが目的なので攻撃しやすく、また攻撃者に検知を回避するインセンティブがあるためである。ほかにも、音声認識⁽⁹⁾、自然言語処理⁽¹⁰⁾、金融や医療などでの表形式データの分析⁽¹¹⁾など、画像以外のデータについても敵対的サンプルの研究が行われており、それぞれに制約や難しさはあるものの、原理的にはどんなデータに対しても誤認識は起こりうる。

2.1.1 攻撃の原理と手法

攻撃者はどのように敵対的サンプル $\mathbf{x}_{adv}$ を見つけるのだろうか？ここでは画像認識の分類モデルを例に説明する。

図5は敵対的サンプル探索の概念図である。ある正常なパンダの画像 $\mathbf{x}$ をモデル f は正しくパンダに分類したとする。 $\mathbf{y} = f(\mathbf{x})$ は各クラスに分類すべき確率(確信度)のベクトルな

図5 最適化による敵対的サンプルの探索 (l_2 ノルムの場合)

ので、 $\mathbf{y}$ のうちパンダに対応する第 k 成分が最大となっている。つまり $k = \operatorname{argmax} f(\mathbf{x})$ である。

このとき敵対的サンプル $\mathbf{x}_{adv}$ は「人間にはパンダに見えるが、モデルはほかの動物に分類する」入力である。このうち「パンダに見える」という条件は人間の判断に依存するが、元画像 $\mathbf{x}$ の各画素の輝度を少しだけ変えても人間の判断はほぼ変わらないので、 $\mathbf{x}$ のごく近傍の画像を $\mathbf{x}_{adv}$ の候補とする。この近傍範囲で f の分類結果が変化しているものを見つければ、それが敵対的サンプルである。すなわち近傍の上限を ϵ として

$$k \neq \operatorname{argmax} f(\mathbf{x}_{adv}) \text{ かつ } \|\mathbf{x}_{adv} - \mathbf{x}\| \leq \epsilon$$

を満たす $\mathbf{x}_{adv}$ を探す問題といえる。 $\|\cdot\|$ は距離関数で、画像の場合 l_2 ノルム(ユークリッド距離)や l_∞ ノルムがよく用いられる。このような $\mathbf{x}$ から $\mathbf{x}_{adv}$ への変動($\mathbf{x}_{adv} - \mathbf{x}$)を敵対的摂動と呼ぶ。

ここで $f(\mathbf{x})$ の各成分は対応するクラスの確率なので、もし上記の条件を満たす $\mathbf{x}_{adv}$ が存在するならば、近傍範囲内でパンダの確率を最小化すればそのとき $\mathbf{x}$ は条件を満たす。つまり、

$$\min_{\mathbf{x}_{adv}} f(\mathbf{x}_{adv})_k \text{ s.t. } \|\mathbf{x}_{adv} - \mathbf{x}\| \leq \epsilon$$

という最適化問題に帰着できる^(注4)。求解には $f(\mathbf{x})_k$ に関する損失関数 $L(f, \mathbf{x}, k)$ の傾き(勾配)を偏微分または近似で求めて利用するのが一般的である。

最も単純な探索方法は Fast Gradient Sign Method (FGSM)⁽¹⁾である。この手法では損失関数の勾配の各成分の符号だけを見て、 $\mathbf{x}$ の各成分を ϵ だけ摂動させる。

$$\mathbf{x}_{adv} = \mathbf{x} + \epsilon \operatorname{sign}(\nabla_{\mathbf{x}} L(f, \mathbf{x}, k))$$

より強力な探索方法として、FGSMをより小さな摂動で複数回繰り返す Projected Gradient Descent (PGD) がある⁽¹²⁾。

上記は元画像の近傍での探索だが、人間が画像データを似ているとみなす条件は必ずしもこの範囲に限らない。変化させる画素数(l_0 ノルム)や変化させる画素の位置を制限して探索することで敵対的サンプルを見つけることも可能である。またす

(注4)：この定式化は本来のクラス k 以外の任意のクラスへの誤分類をさせる攻撃(generic attack, non-targeted attack)である。特定のクラス t に誤分類させる攻撃(specific attack, targeted attack)の場合は $f(\mathbf{x}_{adv})_t$ を最大化する問題となる。

でに紹介した撮影距離や角度が変わっても通用する敵対的サンプルは、縮尺や回転などの変換を加えても正しいクラスへの分類確率が下がるような摂動を探索することで実現されている。

なお、攻撃対象モデルの損失関数や勾配を求めるにはモデル内部の情報が必要なので、ここまで述べてきた攻撃はホワイトボックスアクセスを前提としている。しかし、ブラックボックスであっても複数回の問合せにより勾配を近似的に得ることは可能である。また一般に類似したモデル間では同じ敵対的サンプルが通用しやすい性質（転移性, transferability）があり、これを利用するとブラックボックス条件の攻撃者でも攻撃対象のモデルの代理モデルを用意してホワイトボックスと同様に攻撃可能と指摘されている⁽¹³⁾ので、注意を要する。

2.1.2 防御手法と適応的攻撃

敵対的サンプルの動向を紹介してきたが、これらの攻撃を防ぐことはできるのだろうか？ 本節では敵対的サンプルに対する防御の動向を紹介する。防御側の目的は敵対的な入力の影響を受けにくい分類器を実現することであり、この性質は敵対的ロバスト性 (adversarial robustness) と呼ばれる。

敵対的サンプルに対する防御手法は、攻撃の発見以来数多く提案されており、攻撃検知とモデル自体のロバスト化に大きく分けられる。検知手法の分類としては、複数のモデルを使って結果を比べる方法、主成分分析などで特徴量の貢献度を見る方法、自然なデータとの分布の違いを見る方法、ランダムな摂動への耐性を見る方法、などがある⁽¹⁴⁾。一方、ロバスト化手法の分類としては、入力値を「ぼかす」などして情報を欠落させる方法、微分不可能な関数やランダム性を導入して勾配を得にくくする方法、訓練データに敵対的サンプルも含めて学習させる方法、複数のモデルの合議（アンサンブル）を導入する方法などが挙げられる。しかし、後の評価によれば攻撃を十分に防げないものも多く、効果があっても特定のデータやタスクに限定され、あるいは副作用としてモデル本来の推論性能が劣化することがしばしばあり、決定的な防御手法はまだ確立されていない。

Carlini らは、2017 年に 10 種類の攻撃検知手法に対して評価を行い、それらは既知の攻撃に対しては一定の効果があるものの、検知手法を知った上で適応的に行う攻撃には十分な検知能力を発揮しないことを示した⁽¹⁴⁾。また彼らは 2018 年にも、3 か月前の会議で発表された九つのロバスト化手法に適応的な攻撃を行い、うち七つに対して攻撃に成功したと報告した⁽¹⁵⁾。特に勾配を得にくくする方法は初歩的な敵対的サンプルを防ぐことはできるが、そのロバスト化手法に対して適応的に作られた敵対的サンプルを防げないことが報告されている。この種のロバスト化手法は勾配の導出を難しくしているだけで、モデルの決定境界は変わっていないので、勾配を近似するなどの方法で攻撃可能とされる。こうした適応的攻撃は防御手法の評価に有効であるため、ベストプラクティスの共有など方法論としての整備も始まっている^{(16), (17)}。

訓練データに敵対的サンプルも含めて学習させる防御手法は敵対的訓練 (adversarial training)⁽¹²⁾と呼ばれ、上記の適応的な攻撃に対しても一定の効果を保つとされる。しかし敵対的サ

図 6 ポイズニング攻撃のフロー

ンプルの生成とモデルの訓練に多くの計算リソースが必要なことで、防御性能は訓練に入れる敵対的サンプルやデータの種類の依存することなどの課題もあり、現在も活発に研究されている。

比較的新しい防御手法として certified robustness が挙げられる^{(18), (19)}。Certified robustness の基本的な考え方は、所与の入力値 x に対して l_p 近傍内の摂動では分類結果が変わらないことを保証するというものである。理論的にはこの保証は可能で、防御手法として期待できるが、未知の入力値に対しての保証はできないこと、実装上はモンテカルロ法などによる近似で実現するので敵対的な摂動を見逃しうることなど課題もある。

2.2 ポイズニングとバックドア

2016 年、Microsoft は Tay という名のチャットボットを Twitter 上にデプロイし、若者と会話で交流する実験を始めた。しかし Twitter ユーザの発言を学習して自身の発話能力を向上する機能が裏目に出て、Tay は人種差別的な発言をするようになり、Microsoft は実験を中止した⁽²⁰⁾。この事件は AI の倫理問題とポイズニング攻撃の可能性を世に知らしめることとなった。

ポイズニング (poisoning, 汚染) は、訓練データを汚染することにより訓練後のモデルに誤りを加え、推論フェーズにおいてある種の入力に対して誤った推論結果を出力するようにしむける攻撃である (図 6)。ナイーブな攻撃方法として、前述の Tay に対するポイズニングのように大量の誤ったデータを訓練データセットに加えることが考えられるが、訓練フェーズで推論性能が上がらず人間に気付かれやすい。そのため実環境でのみ性能が下がる、または特定の入力に対してだけ誤る、といった巧妙な汚染方法が主な研究対象となっている。

これまでに、SVM、ロジスティック回帰、深層ニューラルネットワークなど様々な分類モデルに対してポイズニングが可能なが報告されている^{(21), (22)}。図 7 は Muñoz-González ら⁽²²⁾による簡単な実験の例である。左は汚染前、右は汚染後の分類の様子を示す。青いひし形の点が汚染により加えられた点である。左右を比べると分かるとおり、一つの汚染点の影響で分類の境界が大きく変わっていることが見てとれる。

ポイズニング攻撃は表 1 の分類では唯一訓練フェーズに働きかける攻撃であり、攻撃者が訓練データにアクセス可能という条件をもつのが特徴である。このため推論フェーズに働きかけるほかの攻撃より難しいが、訓練データを外部から得る場合、たとえば他社に提供してもらい、インターネットからダウンロードする、不特定のユーザから集める、などの場合に攻撃されやすい。Tay の事例のように、追加のデータで再訓練する場合 (オ

図 7 ポイズニングによる決定境界の変化 (Muñoz-González ら⁽²²⁾に基づく)

図 8 バックドアのトリガーを含む画像の例 (Bagdasaryan ら⁽²⁶⁾より)

ンライン学習)にも注意を要する。

2.2.1 ポイズニングに対する防御

既に述べたように、ポイズニングは単なる推論性能の低下ではなく、ごく一部のデータまたは訓練フェーズに見たことのないデータに対して誤りを起こそうとする攻撃であり、テストデータセットに対する精度だけを見て防ぐのは難しい。一方で、汚染された訓練データセットには攻撃者の意図を反映した偏りがあるはずなので、汚染されたデータ点は少数であるという仮定のもとで外れ値検知で検知・除外する防御手法⁽²³⁾や、各データ点から訓練結果への負の影響が特に大きいものを検知・除外する防御手法⁽²⁴⁾、これら二つの方針を組み合わせた手法⁽²⁵⁾などが提案されている。ただし、そうした偏りは攻撃者の意図ではなくデータ傾向の自然な変化である可能性もある。また Tay の事例で見られるように、「悪い」訓練データが大量に現れた場合にどうすべきか、またそもそも差別的か否かを誰がどのように決めるのか、といった課題もあり、防御手法は自動化ではなく人間の判断の支援や効率化を目的に利用されるべきだろう。

2.2.2 バックドア

ポイズニングに似た攻撃としてバックドア攻撃が挙げられる。バックドアは訓練済みモデルが特定のトリガーを含む入力に対してのみ誤った推論結果を出力する性質であり、その特定の入力を知っている攻撃者は文字どおり裏口として利用できる。図 8 はバックドアへのトリガーを含む画像の例である⁽²⁶⁾。左は微小な画素パターンのトリガー、右はより自然な実体物の映り込みで実現できる配色パターンのトリガーである。また画像以外の

テキストデータなどでもバックドアが実現できることが報告されている。バックドアはモデルを汚染して特定の誤りを起こすという点でポイズニングと似ているが、その実現手段は訓練データの汚染に限らず、攻撃者自身がモデルを訓練してバックドアを直接仕込むことができるので、攻撃はより容易である。

バックドアが脅威となるのは他人が用意した訓練済みモデルを利用する場合である。訓練フェーズにはしばしば大規模なデータセットと計算リソース、および複雑なノウハウを要するので、クラウドを含む他者の計算機やサービスを利用して訓練する場合や既に訓練済みのモデルをダウンロードしたり購入したりして利用する場合も多い。また訓練済みの汎用的なモデルを個別の用途に合わせて調整 (fine tuning) して使う転移学習も普及し始めており、バックドア攻撃の現実性は高まっている。

バックドア攻撃に対する防御として、訓練済みモデルが含むバックドアを発見して取り除く方法や、訓練済みモデルを簡略化してバックドアの影響を受けない別のモデルを作る方法 (蒸留) などが提案されている。ただしバックドア攻撃はポイズニングよりも防御の実施が難しい傾向にある。なぜなら前述の通り訓練データセットや訓練のための計算リソース・ノウハウを用意できない場合が攻撃対象なので、防御の実施にも制約が大きいためである。したがって、信頼できるベンダーから購入する、またはオープンソースソフトウェア (OSS) 配布のようなコミュニティ的なアプローチをとるなど、防御手法を適用済みのモデルを効率良く再利用するための社会的・経済的な仕組みも求められるだろう。

2.3 モデル抽出

今日では、画像認識や機械翻訳など機械学習を使った便利なサービスがクラウドで提供されている。このとき訓練済みモデルはクラウド上で稼働しており、利用にはユーザ登録や料金の支払い、利用回数制限などの制約が生じる。このとき、もしもクラウド上のモデルを手元に複製し、制約なく自由に使えるとしたらどうだろう？

モデル抽出 (model extraction) はこのような状況でモデルの複製を作る攻撃である。サービス提供者側から見ると、訓練済みモデルには訓練データの収集や訓練フェーズの実施に時間や金銭、ノウハウといったコストがかかっているため、複製により価値が奪われることは脅威である。また第三者が訓練済みモデルを入手できるという想定外のことが起きると、ほかにもセキュリティ上のリスクにつながる。このことは 3.1 節で述べる。

図 9 にモデル抽出の攻撃フローを示す。モデル抽出はブラックボックスアクセスを前提とした攻撃であり、ホワイトボックスは考えない。ホワイトボックスなら訓練済みモデルは既に攻撃者の手元にあるので、複製するまでもなく自由に使い、内部の情報を知ること可能だからである。

モデル抽出の基本的なアイデアは、攻撃者の手元で新たなモデルを訓練して作ることである。訓練には訓練データセットが必要だが、それを攻撃対象のモデルに問い合わせで準備する。すなわち、攻撃者は複数の入力値 $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}$ を用意して対象

図 9 モデル抽出攻撃のフロー

モデル f に入力し、推論結果として $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ を得る。これらを組にした $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})$ を訓練データセットとして新たなモデル $\hat{f}$ を訓練すれば、 f と類似した訓練済みモデルを得ることが期待できる。うまく複製できたかどうかは、テストデータセットに対する精度の近さや推論結果の一致度（忠実度 fidelity）などの指標で評価される。

このとき、入力値 x の取りうる値全てを $\{x^{(i)}\}$ として問い合わせれば f の完全な複製を得られるのは自明だが、現実には入力回数に応じてコスト（時間、料金、攻撃行為の発見されやすさなど）がかかるので、できるだけ少ない数の問合せで f に類似した $\hat{f}$ を復元する工夫が求められる。「入力は動物の画像である」などの仮定をおける場合は外部知識などを活用し、例えばインターネットで動物の画像を集めるなどして $x^{(i)}$ を選ぶ。また手元のモデル $\hat{f}$ の訓練を段階的にを行い、初期の $\hat{f}$ がうまく複製できていない領域から重点的に次の $x^{(i)}$ を選んで改善していく。ほかにも訓練時の設定（ハイパーパラメータ）や訓練データセットの部分知識など、よい複製をするのに有用な情報は多々あり、攻撃者がそうした外部知識をもつことをどの程度仮定するかによっても様々な研究が行われている。

初期の研究事例として、2016 年に Tramèr らは API のみが公開されたクラウド上の決定木とニューラルネットについてモデルを複製する手法を示した⁽²⁷⁾。その後の研究としては、能動学習 (active learning) を活用して訓練データセットに関する知識を仮定せずにモデル抽出を行う Knockoff Nets⁽²⁸⁾や、敵対的サンプルにより決定境界付近の入力値を生成して問合せを効率化する Juuti らの手法⁽²⁹⁾が挙げられる。これらの攻撃は小規模な画像認識モデルを対象とし、おおむね 10 万回前後の問合せで実現されている。

モデル抽出に対する素朴な防御手法として問合せ回数を制限する方法が挙げられるが、現実的な上限を設定できるかはユースケースに依存する。正常な問合せとモデル抽出攻撃の問合せの分布の違いを検知する手法も提案されている (PRADA⁽²⁹⁾など) が、攻撃者は正常な振る舞いを混ぜて攻撃することもできるので、攻撃コストは増すものの決定的な対策とはいえない。

2.4 訓練データ推定

訓練済みモデルは訓練データセットに基づいて推論結果を導出するので、訓練データに関する何らかの情報を「覚えて」い

図 10 訓練データ推定攻撃のフロー

ると考えるのが自然である。この情報を取り出すことはできないのだろうか？ この疑問に答えるのが訓練データ推定と呼ばれる攻撃である。

本稿では、訓練データに関する秘匿されるべき情報を推定する攻撃全般を訓練データ推定と総称する。攻撃方法や推定する情報によってモデル反転 (model inversion)、訓練データ再構成 (training data reconstruction)、メンバーシップ推定 (membership inference) など様々な呼称の攻撃手法がある。また守るべき情報の性質として、企業秘密などの一般的な情報秘匿と個人の情報を守るプライバシーの二つの方向性がある。しかし、ここではそれらを明確に区別せずにまとめて扱う。

訓練データ推定攻撃のフローを図 10 に示す。攻撃者は基本的にブラックボックスアクセスで、自身が用意したデータを訓練済みモデルに入力し、推論結果を得る。これらの入出力の組を見て、訓練データセットにどのようなデータが含まれていたか推定する。大抵の場合一度の推定では不確実なので、入力を調整し再度出力を得て推定することを繰り返す。

分類タスクの場合、推論結果 y は入力 x の各クラスへの所属確率（または確信度）を意味する。分類モデルは訓練データに基づいて結果を導き出すので、特に汎化が不十分なモデルでは x が訓練データ点に近づくほど確率が高くなりがちで、確率が高くなるように x を動かして探索することでいずれは訓練データ点に到達することが期待できる。

このように素朴な訓練データ推定攻撃はブラックボックス関数 f に対する極大値探索問題とみなすことができ、いかに少ない f への問合せで効率良く最適解にたどりつけるかが鍵になる。画像のように入力 x の次元が大きい場合には問合せ回数が爆発してしまうので、事前知識（「人間の顔画像である」など）も活用して効率化を図る。またよく汎化されたモデルでは訓練データ点で確率が極大になるとは限らないので、モデルの特性や周辺での関数の振る舞いも考慮する。

図 11 に顔認識モデルからの訓練データ推定攻撃⁽³¹⁾の処理構成を示す。左端の顔画像は「正解」の訓練データであり、右端の顔画像は攻撃により推定したデータである。高次元の顔画像を効率よく探索するため、確率ベクトル y から入力 x を復元する生成モデル（右側の赤いニューラルネットワーク）を活用する。このように事前知識に基づく生成モデルを利用することで、

図 11 顔認識モデルからの訓練データ推定の例 (Yang ら⁽³¹⁾より)

もっともらしい訓練データの候補のみを探索する方法が主流になっている。ただし筆者らの研究⁽³²⁾によれば、図に示したような顔画像の場合、探索を工夫しても数万回規模の問合せを要しており、現実には攻撃可能かはシステム形態に依存する。一方で、攻撃者がモデルを入手できる場合、つまりホワイトボックスアクセスを仮定する場合は、計算リソースと時間の許す限りいくらかでも問合せが可能であり、またモデル内部の構造や状態を直接見ることもできるので、訓練データ推定攻撃はずっと容易であり、大きな脅威になりうる。

顔画像は個人のプライバシーと直結するため脅威として分かりやすいが、原理的には画像に限らずどんな種類のデータに対しても訓練データ推定攻撃を適用できる。本稿執筆時点ではまだプレプリントだが、2020 年末には大規模な自然言語処理モデルである GPT-2 から訓練データとして使われたテキストの情報を抽出できたという報告⁽³³⁾がされている。一般に訓練データは企業秘密やプライバシー情報を含むことが多いので、今後も動向を注視する必要がある。

訓練データ推定に対する防御として、問合せ回数の制限や訓練データの加工といった素朴な方法のほか、モデル自体にランダム性をもたせる方法が複数提案されている (De Cristofaro の概説⁽³⁴⁾を参照) が、精度劣化や実用性の低下といった副作用があるので、ユースケースに応じた検討が必要である。

3. 機械学習セキュリティの広範さ

3.1 攻撃同士の関係性

ここまで攻撃を四つに分類してそれぞれ紹介してきたが、これらは独立ではなく、相互に関係している。

特に注意すべきなのは、敵対的サンプルの節でも触れた**転移性 (transferability)** だろう。転移性とは、同じようなタスクを行う訓練済みモデル同士が、たとえ異なるモデル構造や訓練過程によるものであってもある程度似た性質を共有し、そのために一方に通用する攻撃がもう一方にも通用してしまう性質である。このためブラックボックスアクセスしかできない攻撃者であっても、攻撃対象と類似した代理モデルを手元に用意できればホワイトボックスアクセスで敵対的サンプルを見つけることができる⁽¹³⁾。モデル抽出攻撃はまさにこの代理モデルを作る手段なので、モデル抽出に弱いシステムは敵対的サンプル攻撃も受けやすいといえる。ポイズニングや訓練データ推定にも転移

性の影響が議論されており、モデルの秘匿以外の観点でもモデル抽出への対策を考慮すべきである。

ほかにも秘密情報の推定を容易にするためのポイズニング攻撃⁽³⁵⁾や、敵対的サンプルへの対策が訓練データ推定の一環であるメンバーシップ推定を容易にするとの報告⁽³⁶⁾もあり、機械学習セキュリティには総合的な視点が必要といえる。

また機械学習の利用を助けるものとして注目されている説明機能 (説明可能性, explainability) についても触れておきたい。説明機能とは、モデルが推論結果以外の情報を出力して、その結果に至った原因や結果を変えるためのヒントを人間に示唆するものである。しかしセキュリティの観点からは、攻撃者が推論結果以外の情報を得られるので、ブラックボックスでの攻撃を容易にする場合がある。実際に敵対的サンプルの探索や効率のよいモデル抽出、訓練データの推定に説明機能を利用する研究が報告されている⁽³⁷⁾。また説明機能による説明を誤らせる敵対的サンプルも指摘されている⁽³⁸⁾。こうしたことから、説明機能による追加情報の開示にあたっては悪影響も考慮し、利用者を限定するなど慎重な設計が求められる。

3.2 近年の傾向

本節では、やや主観的になるが、機械学習セキュリティの研究の 2020 年の傾向を述べる。

一つの切り口として、2020 年のいわゆる三大セキュリティ国際会議 (S&P, USENIX, CCS) における機械学習セキュリティ関連の論文の傾向を見てみる^(注5)。筆者の調べでは、三つの会議で機械学習セキュリティを主題とする論文は 23 本あり、これは論文数の合計 392 本の 5.9%にあたる。

23 本の論文で扱われている攻撃種別は、敵対的サンプル 8、ポイズニング 7、訓練データ推定 6、モデル抽出 2 であった (重複あり)。このことから、セキュリティ系の会議では敵対的サンプルだけに偏らずに攻撃・防御が検討されていることが見てとれる。

対象データについては画像が多いが、テキストを対象とした研究も増えている。表形式データもしばしば扱われているが、医療や金融などセキュリティやプライバシーが重要な用途には表形式も多いので自然であろう。

機械学習モデルの種類については、基本的な分類や回帰モデル以外への広がりが見られる。2019 年頃から連合学習 (federated learning) を扱う研究が増えており、2020 年にも研究対象とされている。新しい傾向としては、生成モデルやモデル更新を対象とした攻撃が見られた。また機械学習の前処理に対する攻撃、具体的には自然言語の埋め込み (embedding) や画像のスクーリングを対象とした攻撃が現れたのは興味深い。

このように機械学習セキュリティの研究は、機械学習の用途やテクニックの広範さに合わせて広がっていく必要があり、国際会議でもその傾向が表れている。

(注5)：併設ワークショップやポスター発表は除く。また特に敵対的サンプルについては NeurIPS, ICML, ICLR など機械学習系の国際会議で扱う大数の論文があるが、ここでは含めない。

4. AI システムを守るには？

ここまで見てきたように、機械学習技術に対して様々な攻撃の可能性が指摘されており、敵対的環境下での利用には十分な注意が必要といえる。現状では攻撃側が有利な状況であり、決定的な防御手段は確立されていないが、攻撃の研究自体がまだ発展途上である。防御手段を評価するためにもまず攻撃手段の解明が必要で、どんな攻撃が、どこまで、なぜ可能かを明らかにすることが研究としては重要である。

とはいえ、機械学習を応用した AI システムは様々な用途が期待されており、実用化も始まっている。自動運転に代表されるサイバーフィジカルシステム、医療・健康、金融・商取引、社会インフラ、あるいは個人データを扱う用途などリスクの高いアプリケーションについては、研究としての興味や価値とは別に、機械学習の外側のシステムや要件・運用など人間的側面も含めたトータルシステムとして今すぐ対策を考えるべきだろう。

本節ではトータルシステムとしての対策を強調するため、対象を取って「AI システム」と呼び、AI システムに対するセキュリティに関する動向と今後の展望について述べる。

4.1 AI システムセキュリティの動向

まず現実世界で AI システムがどれだけ攻撃を受けているか見てみよう。Black Hat USA 2020 での講演において Herbert-Voss は、実用化された機械学習モデルに対する攻撃事例として次の七つを挙げた⁽³⁹⁾。

- (1) 自動運転車の車線認識を誘発
- (2) 多量のスマホで偽の渋滞を作成
- (3) 暗号通貨取引ボットを誘導
- (4) 推薦エンジンをだまして自社を有利に
- (5) マルウェア判定をすり抜ける
- (6) 電子メールフィルタをすり抜ける
- (7) オンライン詐欺検知をすり抜ける

筆者の知る限り、このうち実際に悪意をもった攻撃者が実行したのは (3)、(4)、(7) の三つのみで、ほかは善意のセキュリティ研究者が報告したものである。前者の三つの事例に共通するのは、推論結果の利用が自動的に行われており、なおかつ攻撃者の金銭的なメリットに直結している点である。この観測から次の三つのことがいえるだろう。

- 現時点では悪意をもって攻撃された事例はまだ少ない。
- 攻撃者にとって金銭的なメリットを得やすいものが早く攻撃される。
- 推論結果が人間に提示されるのではなく、自動的に利用される事例が被害につながりやすい。

ただし悪意をもった攻撃でなくても脆弱性対応の負担や設計・開発の手戻りが生じるので、無害とはいえない。

次に AI システムを開発・運用する産業界の現状に目を向けてみよう。Kumar らは 28 社の企業にインタビューを行い、機械学習セキュリティに対する取り組みや現状認識を聞きとって、

2020 年に報告している⁽⁴⁰⁾。これによれば、機械学習セキュリティに対して具体的な取り組みを始めている企業は 28 社中 3 社のみであった。また学術的な研究では敵対的サンプルが最も盛んに扱われているのに対し、企業が最も心配している脅威は上位からボイズニング、モデル抽出（盗難）、訓練データ推定（モデル反転）、次いで敵対的サンプルであった。これは一つのインタビュー結果にすぎず、このまま敷衍できる知見とはいえないが、製品レベルの AI システムをセキュアにするという意味では研究と実践に大きなギャップがある可能性を示唆している。またこの報告では、ソフトウェアの脆弱性対策で培った Secure Development Lifecycle (SDL) を参考に AI システムのセキュリティ実現に取り組む提言もされている。

セキュリティ実現に向けた活動を広めるためには標準化が担う役割も大きい。AI システムのセキュリティに関する標準化の動きはまだ始まったばかりである。米国の NIST は 2019 年に“Taxonomy and Terminology”と題する文書のドラフトを公開し、文字どおり概念と用語の整理を試みている^(注6)。また 2020 年には、MITRE と Microsoft が中心になってまとめた“Adversarial ML Threat Matrix”^(注7)が公開された。Threat Matrix は既知の攻撃手法を分解・整理したもので、AI システムのセキュリティに携わる人々の間での意思疎通の円滑化を目的としている。サイバーセキュリティ分野で活用されている ATT&CK^(注8)の枠組みを踏襲し、偵察 (reconnaissance) や初期アクセス (initial access) など攻撃の進行段階に分けて多くのテクニックをリストアップしている。また本稿の執筆時点では 13 のケーススタディを掲載しており、各ケースで使われた攻撃テクニックを Matrix と対応づける。これら二つの例に見られるように、情報共有と相互理解を進めるための概念整理が進められており、やがてベストプラクティスやガイドラインがまとまってくると予想されるが、具体的な作業や基準に落とし込むのは容易ではないだろう。

なお、日本においては QA4AI コンソーシアムや産業技術総合研究所が機械学習や AI システムの品質の観点からガイドラインを発行しており、セキュリティについても概略が記載されている^{(注9) (注10)}。

実装面では、本稿で解説した攻撃や一部の防御方法を実施するためのツールボックスと称した OSS が出現している。Adversarial Robustness Toolbox (ART)^(注11)、Foolbox^(注12)、CleverHans^(注13)が著名な例である。ただしこれらは文字どおり「道具箱」であり、その道の専門家が役立てるツール群である。適応的攻撃の方法論でも自動化はまだ難しいとされており、AI システムのセキュリティの実装やテストには当面は専門家の関

(注6) : Draft NISTIR 8269. <https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf>

(注7) : <https://github.com/mitre/advmthreatmatrix>

(注8) : MITRE ATT&CK. <https://attack.mitre.org>

(注9) : <http://www.qa4ai.jp/QA4AI.Guideline.202008.pdf>

(注10) : <https://www.cpsec.aist.go.jp/achievements/aiqm/>

(注11) : ART. <https://github.com/Trusted-AI/adversarial-robustness-toolbox>

(注12) : Foolbox. <https://github.com/bethgelab/foolbox>

(注13) : CleverHans. <https://github.com/cleverhans-lab/cleverhans>

与が必要そうだ。

以上をまとめると、研究分野の盛り上がりとは対照的に、実践的な活動はまだ着手されたばかりといえる。現実的な被害はまだ稀だが、AI システムの普及にともなって増加が予想され、自動化システムや金銭的価値を扱うアプリケーションに特に注意が必要である。しかし研究レベルでも攻撃・防御ともに未解明で、対策は不確実性ははらんでおり、実施には専門家の関与と時間的・金銭的なコストを要する。

4.2 今後の展望

これらを踏まえ、セキュアな AI システムを実現するにはソフトウェアの脆弱性対策を参考に取り組むべきと筆者は考える。機械学習に限らず一般にソフトウェアからバグや脆弱性を完全になくすることはできないが、これまで我々は失敗を重ねながらも、セキュリティを取り入れた設計・開発・運用のプロセスを徐々に構築してきた。機械学習には人間が仕様を明示しないことに起因する特有の難しさがあるが、ソフトウェアのセキュリティにおけるこれまでの知見を基盤にして取り組むべきだろう。

Kumar らの提言⁽⁴⁰⁾と重なるところも多いが、次の五つのレイヤーで並行して取り組むことを推奨したい。

- **機械学習モデル:** できるだけセキュアなモデルを実現する。本稿で紹介してきた攻撃検知やロバスト化といった防御手法を利用し、できるだけセキュアなモデルを作る。決定的な防御手法はまだ見つかっておらず、また完全にセキュアなモデルを作るのは非現実的だが、何が防げるか、なぜ防げるのかを明らかにすることで、以下に述べる他レイヤーの取組みへの有益な示唆が得られるだろう。

- **データと知識:** より多くを学習させる。機械学習に対して攻撃が可能なのは、訓練データが足りていないからでもある。機械学習においてはデータ処理の仕様は訓練データが決めるので、訓練データ不足は仕様不足を意味する。ただし「ゴミから得られるのはゴミだけ」(garbage-in, garbage-out) という警句が示すようにデータ量さえ多ければよいわけでもないで、品質の担保も鍵になる。機械学習セキュリティの実現にはドメイン知識が必須という見解⁽⁴¹⁾もあるので、人間だけがもつ知識や決められる条件を効率良く取り入れる、いわゆる human-in-the-loop のアプローチも求められる。

- **実装と運用:** 安全な実装やノウハウを積み上げる。セキュリティの観点で実装を精査し、ベストプラクティスを積み上げていくというこの方法は、従来のソフトウェアやシステムでの実績があり、ユースケースごとの違いが大きいという特有の難しさもあるとはいえ、AI システムでも有効と考えられる。その上で、AI システムにおいてはデータに関する施策の強化が求められるだろう。著名な OSS のように、出自を確認可能な同一のライブラリやモデル、データセットを多くのユーザが検証しながら使うことによるコミュニティ保証も有望である。各ステークホルダの行動や判断にも依存するため、適切な知識の普及や教育も重要だろう。

- **セキュリティ要件:** 重要なところを優先して対策する。AI

システムに限らず、セキュリティの対策にかけられるリソースは限られるので、重要なところに絞って対策に取り組むべきである。そのため個々の AI システムにおいて何を守るべきか、どのレベルの攻撃まで想定するか（例えば内部犯も想定するのか）、モデルの推論結果はどう使われるのか、といったセキュリティ要件の明確化は非常に重要である。このため筆者らは AI システムの脅威分析にも力を入れている⁽⁴²⁾・⁽⁴³⁾。要件が明確なら、想定被害はごく小さいので注意喚起に留める、推論結果の利用方法を変えてフェイルセーフを導入する、などモデル自体をセキュアにするより確実に低コストな対策も可能になる。また機械学習への攻撃が可能な条件下（例えば内部犯あり、ネットワーク上でデータを改ざん可能、など）では機械学習によらない攻撃のほうが容易かつ深刻な場合も多く、システム全体として対策を考慮すべきである。

- **有事への備え: 事後対処・早期警戒の仕組みを構築する。** 全ての攻撃を未然に防ぐことは困難なので、AI システムは少なくとも効率良くログを記録する機能をもち、攻撃や被害が発覚した後でも何が起きたのか把握可能にすべきである。Adversarial ML Threat Matrix が示唆するように、攻撃者が事前に周辺情報や弱点を探る動きも考えられるので、将来的にはシステムや組織の枠を超えた早期警戒や情報共有の仕組みが実現できると望ましい。このレイヤーではマルウェアやサイバー攻撃に対する成功事例が参考になるだろう。

5. おわりに

本稿では、近年大きな盛り上がりを見せている機械学習セキュリティの研究動向について解説した。機械学習技術の活用は学術的な興味だけでなく実践的・社会的な需要も大きく、そのセキュリティは喫緊の課題であるが、セキュリティは生来的に人間や社会と不可分であるため、多くの方の理解や議論・活動への参加が求められている。その一助として、本稿が読者にとってよき水先案内となれば幸いである。

謝辞 執筆にあたり、早稲田大学の吉岡信和上級研究員から有益な示唆を頂きました。ここに感謝の意を表します。

文 献

- (1) I. Goodfellow, J. Shlrens, and C. Szegedy, “Explaining and harnessing adversarial examples,” ICLR, 2015.
- (2) N. Papernot, P. McDaniel, A. Sinha, and M.P. Wellman, “SoK: security and privacy in machine learning,” IEEE EuroS&P, 2018.
- (3) C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint, 2013. <https://arxiv.org/abs/1312.6199>
- (4) B. Biggio and F. Roli, “Wild patterns: ten years after the rise of adversarial machine learning,” Pattern Recognition, vol. 84, pp. 317–331, 2018.
- (5) K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. prakash, T. Kohno, and D. Song, “Robust physical-world attacks on deep learning visual classification,” CVPR, 2018.
- (6) A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, “Synthesizing robust adversarial examples,” ICML, 2018.

- (7) M. Sharif, S. Bhagavatula, L. Bauer, and M.K. Reiter, “Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition,” ACM CCS, 2016.
- (8) T.B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial patch,” NIPS, 2017.
- (9) N. Carlini and D. Wagner, “Audio adversarial examples: targeted attacks on speech-to-text,” DLS, 2018.
- (10) J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, “HotFlip: white-box adversarial examples for text classification,” ACL, 2018.
- (11) F. Cartella, O. Anunciacao, Y. Funabiki, D. Yamaguchi, T. Akishita, and O. Elshocht, “Adversarial attacks for tabular data: application to Fraud detection and imbalanced data,” SafeAI, 2021.
- (12) A. Mądry, A. Makalov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” ICLR, 2018.
- (13) N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z.B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” ASIA CCS, 2017.
- (14) N. Carlini and D. Wagner, “Adversarial examples are not easily detected: bypassing ten detection methods,” arXiv preprint, 2017. <https://arxiv.org/abs/1705.07263>
- (15) N. Carlini and D. Wagner, “Obfuscated gradients give a false sense of security,” ICML, 2018.
- (16) N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Mądry, and A. Kurakin, “On evaluating adversarial robustness,” online document, 18 Feb. 2019. <https://github.com/evaluating-adversarial-robustness/adv-eval-paper>
- (17) F. Tramèr, N. Carlini, W. Brendel, and A. Mądry, “On adaptive attacks to adversarial example defenses,” NeurIPS, 2020.
- (18) A. Raghuathan, J. Steinhardt, and P. Liang, “Certified defenses against adversarial examples,” ICLR, 2018.
- (19) J.M. Cohen, E. Rosenfeld, and J.Z. Kolter, “Certified adversarial robustness via randomized smoothing,” ICML, 2019.
- (20) O. Schwartz, “In 2016, Microsoft’s racist chatbot revealed the dangers of online conversation,” IEEE Spectrum, Nov. 2019. <https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>
- (21) B. Biggio, B. Nelson, and P. Laskov, “Poisoning attacks against support vector machines,” ICML, 2012.
- (22) L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E.C. Lupu, and R. Fabio, “Towards poisoning of deep learning algorithms with back-gradient optimization,” AISEC, 2017.
- (23) J. Steinhardt, P.W. Koh, and P. Liang, “Certified defenses for data poisoning attacks,” NIPS, 2017.
- (24) B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, “Bagging classifiers for fighting poisoning attacks in adversarial classification tasks,” Multiple Classifier Systems (MCS), 2011.
- (25) 清水, 森川, “機械学習モデルに対するポイズニング攻撃意図の検知について,” 暗号と情報セキュリティシンポジウム SCIS, 2020.
- (26) E. Bagdasaryan and V. Shmatikov, “Blind backdoors in deep learning models,” arXiv preprint, 2020. <https://arxiv.org/abs/2005.03823>
- (27) F. Tramèr, F. Zhang, A. Juels, M.K. Reiter, and T. Ristenpart, “Stealing machine learning models via prediction APIs,” USENIX Security Symposium, 2016.
- (28) T. Orekondy, B. Schiele, and M. Fritz, “Knockoff nets: stealing functionality of black-box models,” CVPR, 2019.
- (29) M. Juuti, S. Szyller, S. Marchal, N. Asokan, “PRADA: protecting against DNN model stealing attacks,” EuroS&P, 2019.
- (30) M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” ACM CCS, 2015.
- (31) Z. Yang, J. Zhang, E.-C. Chang, and Z. Liang, “Neural network inversion in adversarial setting via background knowledge alignment,” ACM CCS, 2019.
- (32) 樋口, 森川, 清水, “分類モデルに対する VAE を用いた教師データ推定攻撃,” 暗号と情報セキュリティシンポジウム SCIS, 2020.
- (33) N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel, “Extracting training data from large language models,” arXiv preprint, 2020. <https://arxiv.org/abs/2012.07805>
- (34) E. De Cristofaro, “An overview of privacy in machine learning,” arXiv preprint, 2020. <https://arxiv.org/abs/2005.08679>
- (35) S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, and G. Hanaoka, “Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes,” IEEE PST, 2017.
- (36) L. Song, R. Shokri, and P. Mittal, “Privacy risks of securing machine learning models against adversarial examples,” ACM CCS, 2019.
- (37) R. Shokri, M. Strobel, and Y. Zick, “On the privacy risks of model explanations,” arXiv preprint, 2021. <https://arxiv.org/abs/1907.00164>
- (38) A. Ghorbani, A. Abid, and J. Zou, “Interpretation of neural networks is fragile,” AAAI, 2019.
- (39) A. Herbert-Voss, “Practical defenses against adversarial machine learning,” briefing at Black Hat USA, 2020.
- (40) R.S.S. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissioner, M. Swann, and S. Xia, “Adversarial machine learning – industry perspectives,” arXiv preprint, 2020. <https://arxiv.org/abs/2002.05646v2>
- (41) A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Mądry, “Adversarial examples are not bugs, they are features,” NeurIPS, 2019.
- (42) 森川, 清水, 樋口, 前田, 矢嶋, “敵対的機械学習のための脅威分析手法の提案,” コンピュータセキュリティシンポジウム CSS, 2019.
- (43) 矢嶋, 清水, 森川, 大久保, “機械学習システムに潜む AI セキュリティ脆弱性の分析手法に関する一考察,” 暗号と情報セキュリティシンポジウム SCIS, 2021.

(ISEC 研究会提案, 2021 年 3 月 5 日受付,
2021 年 4 月 9 日再受付)

森川 郁也

1996 東京大学電子工学科修士課程卒。1999 同博士課程中退, 株式会社富士通研究所入社。セキュリティ技術の研究開発に従事。Web セキュリティ, ソフトウェアの脆弱性対策, システムの脅威分析, 暗号応用, 機械学習セキュリティを中心に活動。