AI Securityはなぜセキュリティ分野か

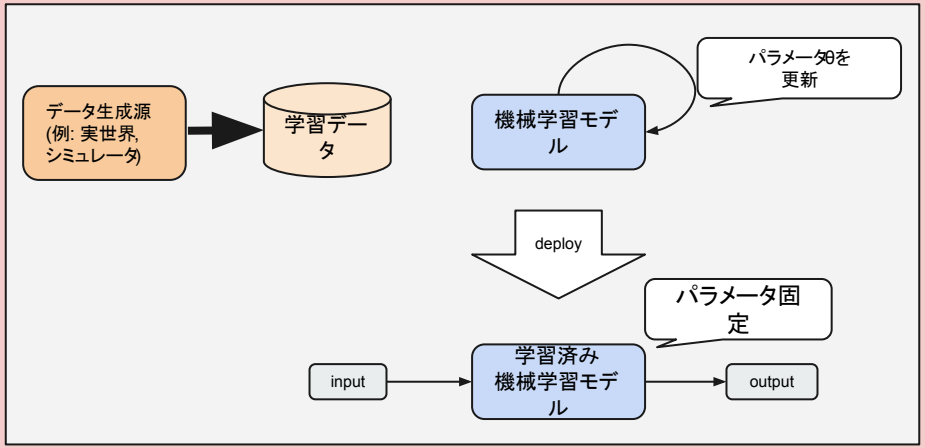
情報科学若手の会 2023 in 軽井沢

Issa Sugiura
B4, Osaka University, Japan

機械学習(ML)とセキュリティ

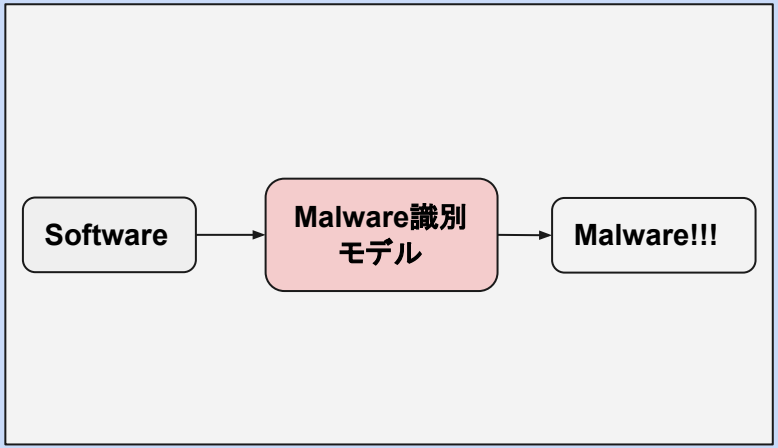
Security for ML(AIセキュリティ)(←今回のテーマ)

- MLシステム自体の脆弱性を研究



ML for Security

- 機械学習をセキュリティ分野に応用(マルウェア検知, 異常検知, etc)



目標

- AIセキュリティは一見機械学習の性質を調べてるだけっぽい.
- AIセキュリティと従来のセキュリティの類似点, 相違点を知り,
  AIセキュリティをセキュリティの観点から見つめ直す.
  →セキュリティの知見をAIセキュリティに取り入れたい

目次

- 機械学習の基礎
- AIセキュリティとは
  - 敵対的サンプル攻撃
  - データポイズニング攻撃
  - AIセキュリティにおける重要な概念
- AIセキュリティとセキュリティの関係
  - AIセキュリティとセキュリティの類似点
  - AIセキュリティとセキュリティの相違点
- AIセキュリティの今後

機械学習の基礎

機械学習 = 関数を自動的にプログラミングする手法

….?

古典的なプログラミングによるFizzBuzz

プログラミングは
関数を記述する行為

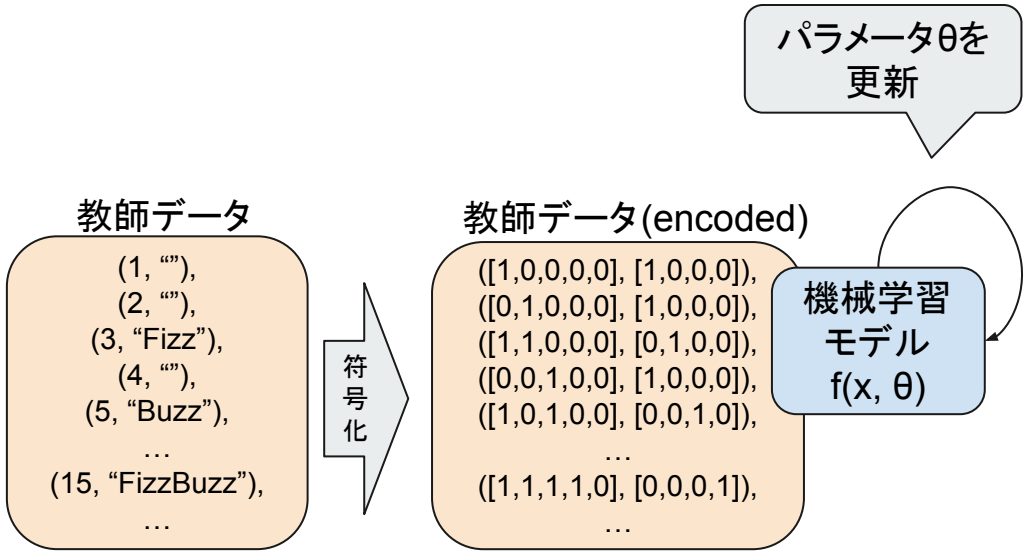
```
def fizz_buzz(x: int)->str:
    if x % 15 == 0:
        return 'FizzBuzz'
    elif x % 5 == 0:
        return 'Buzz'
    elif x % 3 == 0:
        return 'Fizz'
    else:
        return ""
```

機械学習によるFizzBuzz

Goal: fizz_buzz()と同じ挙動

方法:

1. 複雑な関数を表現できる機械学習モデルf(x, θ)を用意.
2. 大量の教師データを用いてfizz_buzz()と挙動が同じになるようにパラメータθを更新
3. 欲しい関数f(x, θ)が得られる

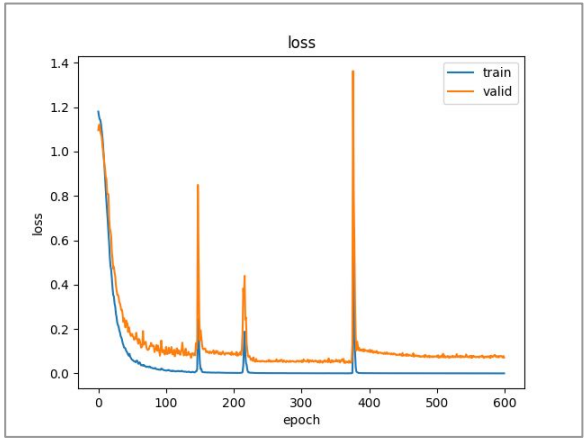


u++の備忘録, PyTorchでFizzBuzz: https://upura.hatenablog.com/entry/2019/01/21/202653

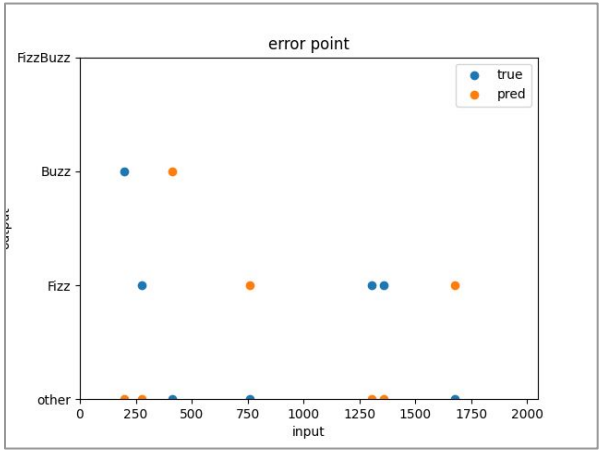
機械学習によるFizzBuzz

- 単純な深層学習モデルで正答率: 0.965

test data 200個	training data 1~2048からrandomに(2048-200)個



学習過程での損失



間違えた入力 7 / 200

古典的なプログラミングと機械学習の違い

機械学習

- 処理を明確に記述できない場合に有効

欠点

- データ処理方法が**Black Box**
- (一般的に)学習データが**大量に必要**
- 性能や挙動が**学習データに大きく依存**

深層学習

- 大雑把に言えば線形関数と非線形関数を繰り返し合成した関数f(x, θ)
  - 単純ゆえ大規模な並列計算が可能
    - GPT3のパラメータ数は175Billion*
- 自動微分により効率的に微分値を求められる
- 普遍性定理により任意の関数が近似できる**

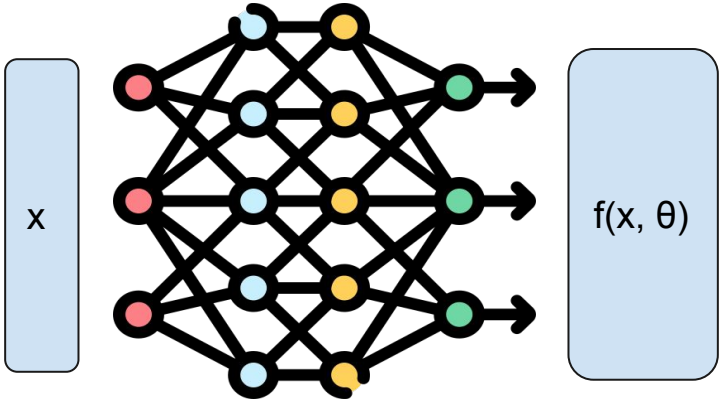


image: Flaticon.com

*Brown+, Language Models are Few-Shot Learners, 2020

** 厳密なstatementはこちら: 深層学習の統計理論 , 鈴木　大慈, 日本統計学会誌 第 50 巻, 第 2 号, 2021 年 3 月 p.229~ p.256

深層学習の広がり

深層学習は多様な問題に応用され，最高性能を叩き出している

- 大規模言語モデル (ChatGPT, etc)
- 画像生成モデル (DALL･E, etc)
- 画像分類モデル
- 音声認識(Siri, etc)
- ゲームAI (AlphaGo, etc)
- etc

https://openai.com/research/dall-e

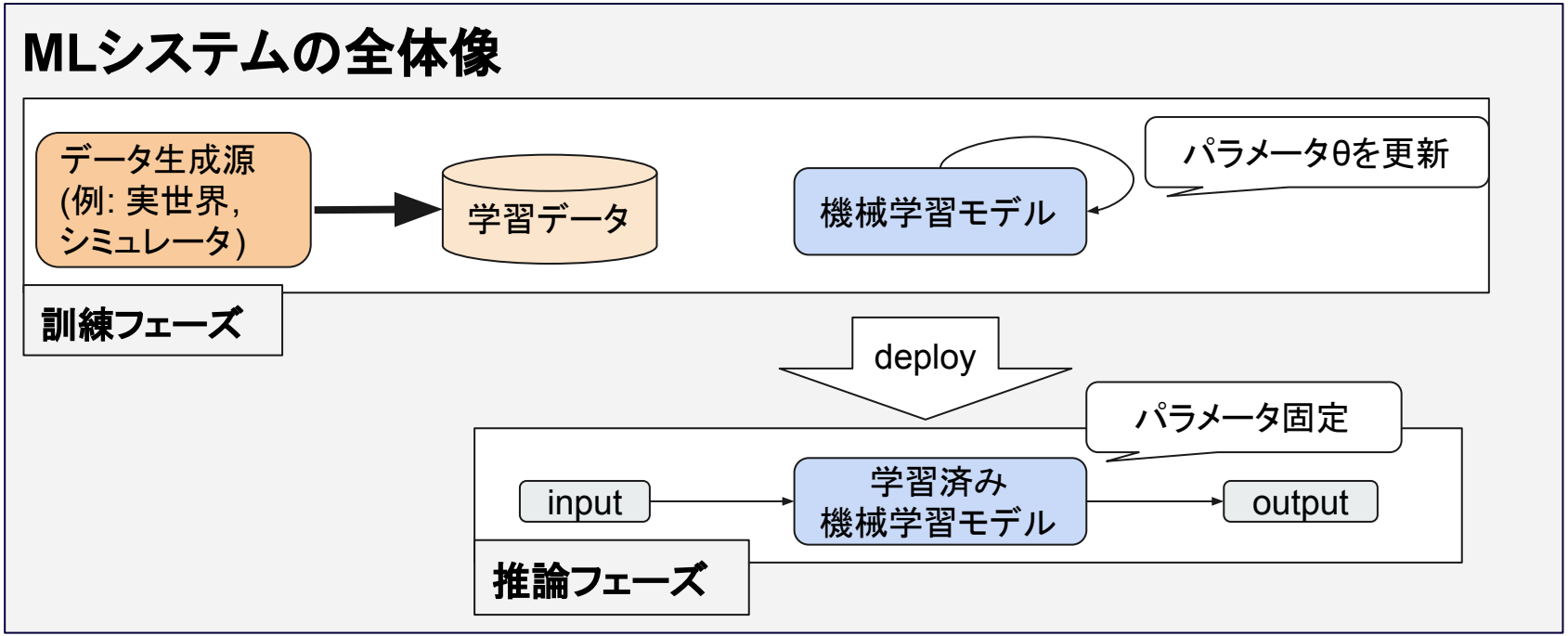
機械学習の基礎: まとめ

- 機械学習は関数を自動的にプログラミングする手法
- 処理がブラックボックス
- モデルの挙動がデータに大きく依存

AIセキュリティとは

AIセキュリティとは

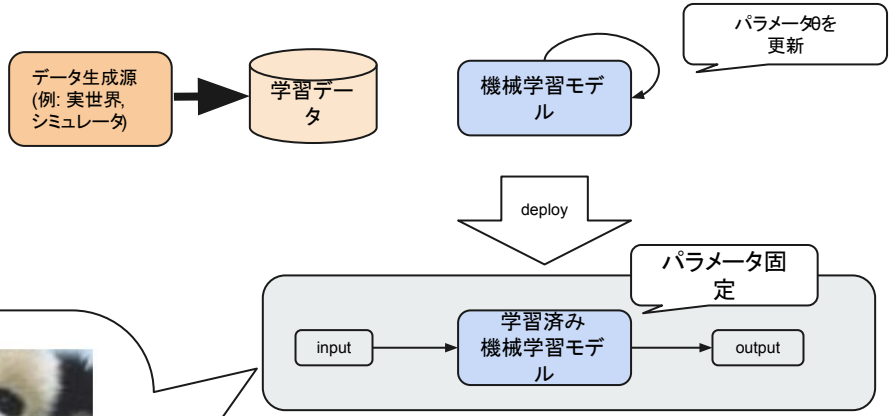
MLシステムの脆弱性についての研究分野

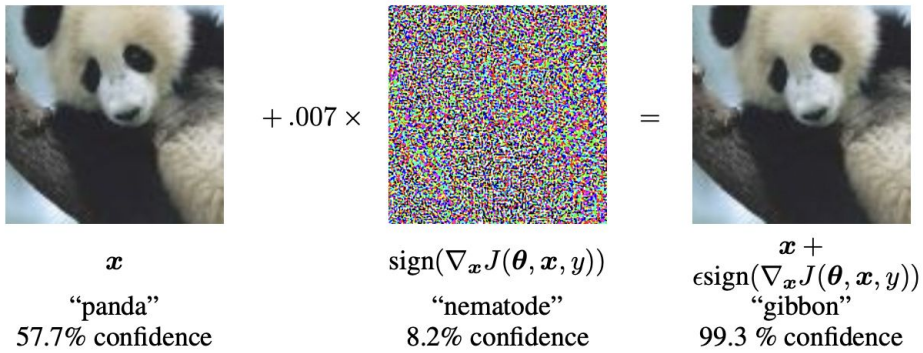


敵対的サンプル攻撃

敵対的サンプル(Adversarial Example)攻撃

- 学習済みモデルへの攻撃
- **人間にはわからないほど**小さな摂動を入力に加え, モデルに誤識別させる攻撃





Goodfellow+, "EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES", Figure 1

防御手法その1: 敵対的トレーニング

idea: 入力に摂動が加わっても同じラベルを出力させる

method: 敵対的サンプルを学習データに加えて学習する

欠点:
- 対症療法的
- 未知の攻撃に弱い

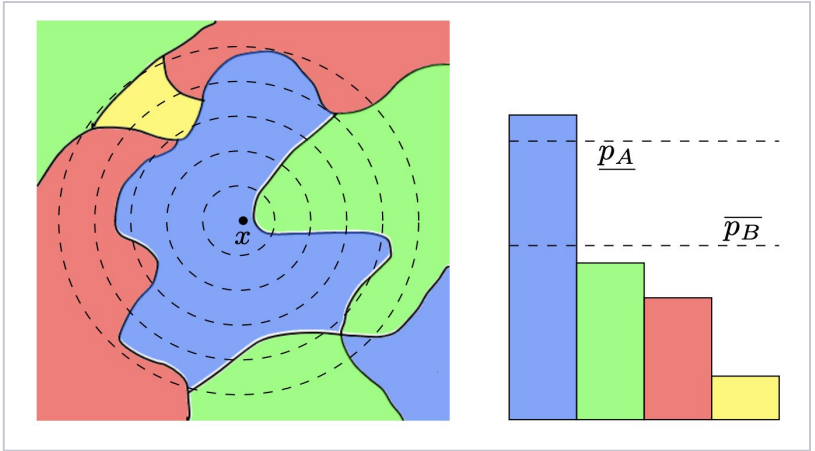
防御手法その2: Randomized Smoothing

idea: 識別モデルf(x)から, 一定の範囲の摂動に対して出力が変化しないよう設計されたモデルg(x)を構成する

method:
入力xに対して, x周辺で予測ラベルの多数決をとる
- xに近いほど一票の重みが大きい
- 僅差なら予測しない

選ばれたラベルをg(x)とする.

効果:
- 一定の範囲の摂動に対して同じ出力となることを理論的に保証
- 識別精度とRobustnessにトレードオフ

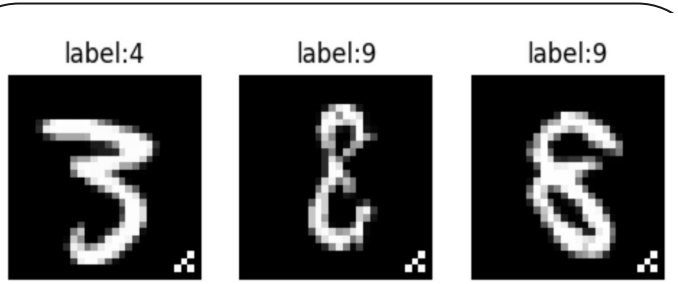


Cohen+, Certified Adversarial Robustness via Randomized Smoothing, 2019, Figure 1

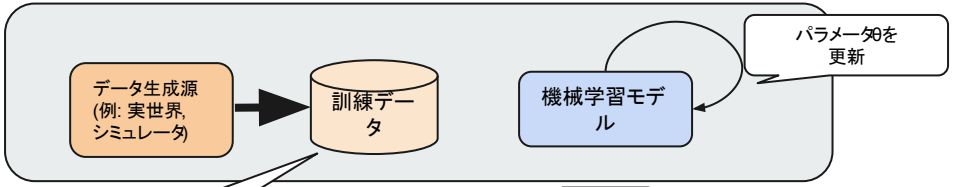
データポイズニング攻撃

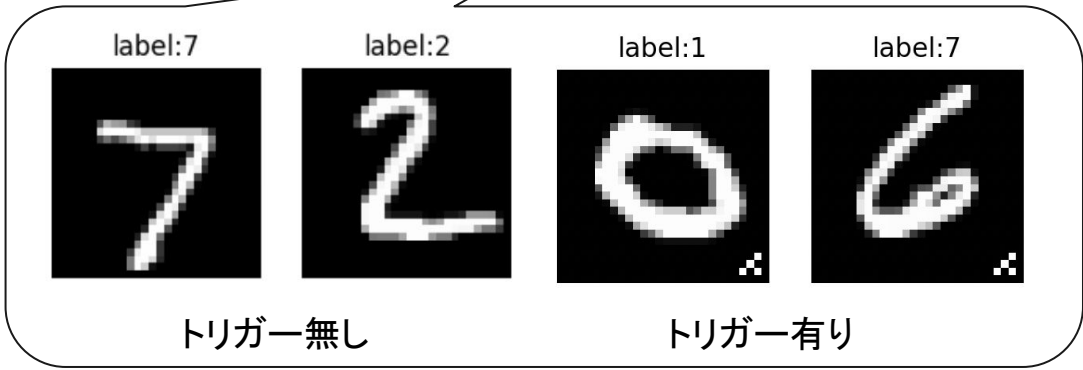
データポイズニング攻撃

訓練データを汚染し，モデルの挙動を意図的に操作する攻撃



訓練データにトリガーを加えたデータのラベルを+1したものを混入させる



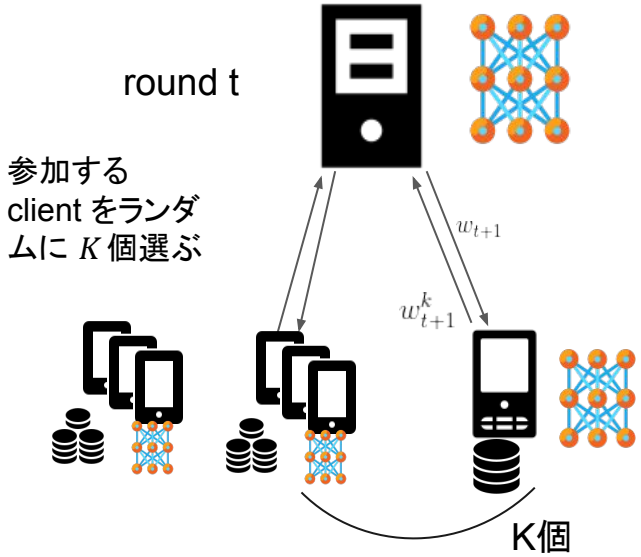


Saha+, Hidden Trigger Backdoor Attacks, 2019

データポイズニング攻撃は実現可能? 連合学習の観点

連合学習: 各自が持つデータセットを外部に送らず, 全体が協調してMLモデルを学習する学習手法

- 悪意あるクライアントが汚染したデータを使ってモデルを学習する可能性



Server:

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$$

w_{t+1}

client k: 各時のデータセットを用いてパラメータ更新

$$w \leftarrow w - \eta \nabla l(w; b)$$

send w to the Server

McMahan+, Communication-Efficient Learning of Deep Networks from Decentralized Data, 2016

データポイズニング攻撃は実現可能? LLMの観点

大規模言語モデル(LLM)は事前学習, fine-tuningの2ステップで学習
事前学習ステップでは, データセットとして

- wikipedia
- web上に存在するコンテンツ
- 論文
- etc

といった大量のデータを用いる.

人の目によるチェックはとても難しく, NLP技術を用いてフィルタリング
→ 悪意のあるデータ混入チャンス!

防御手法: 検知によるアプローチ

異常・分布外検知, クラスタリング等を用いてトリガーを検知

例:

- activation clustering: 活性化関数の値でクラスタリングし, トリガーを含むデータを検知

Chen+, Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering, 2018

他にも様々な攻撃が存在

- Model Inversion Attack
  - MLモデルから訓練データを推定する攻撃
- Membership Inference Attack
  - あるデータが訓練データに含まれているかを推定する攻撃
- Model Theft(Model Extraction Attack)
  - 対象のモデルの入出力等から模倣モデルを作成する
- etc

OWASP AI Security and Privacy Guide, https://owasp.org/www-project-ai-security-and-privacy-guide/

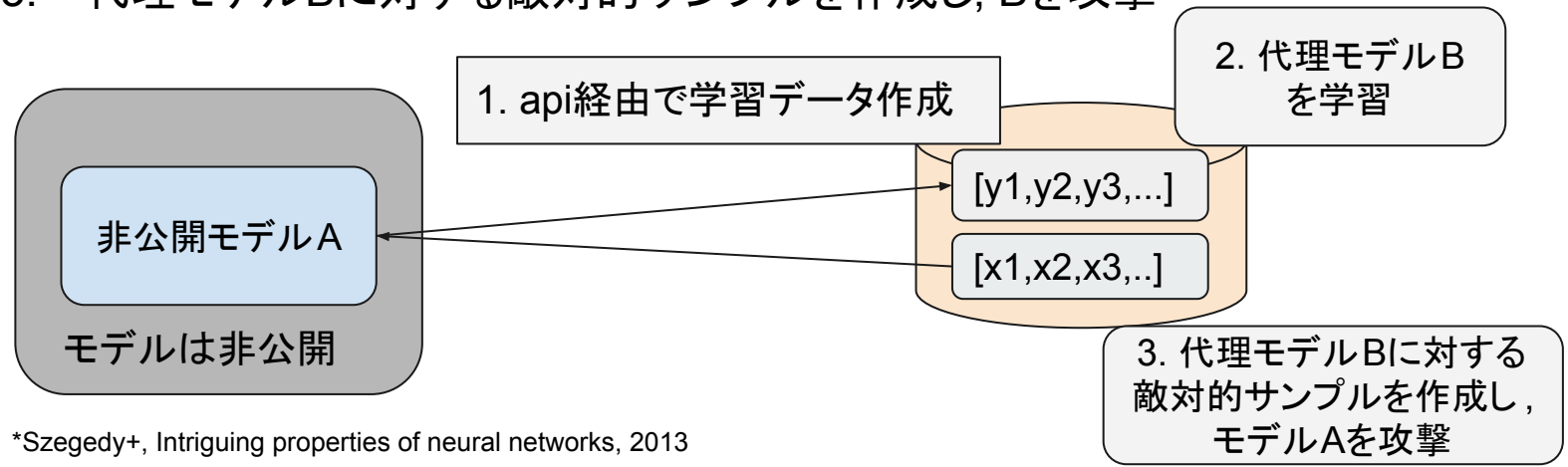
AIセキュリティにおける重要な概念

Transferability(転移性)*

Transferability: あるモデルで有効な敵対的サンプル入力値が, しばしば他のモデルに対しても有効となるという実験的事実

転移性の利用例:

1. 非公開モデルAの入出力から学習データを作成
2. 代理モデルBを学習
3. 代理モデルBに対する敵対的サンプルを作成し, Bを攻撃



*Szegedy+, Intriguing properties of neural networks, 2013

AIセキュリティとセキュリティの関係

AIセキュリティとセキュリティの類似点

セキュリティと性能のトレードオフ

サイバーセキュリティの世界
- アクセス可能な人を制限
- API制限
- ログイン機能
- 速度低下

AIセキュリティの世界
- モデルの非公開
- モデルのロバスト化に伴う性能低下

攻撃者のヒントになる情報を難読化(obfuscation)

サイバーセキュリティの世界
サイドチャネル攻撃やタイミング攻撃に対する防御としての難読化

- 例: 暗号処理で不要な計算を差し込む

AIセキュリティの世界
敵対的サンプル攻撃で利用される勾配を求められにくいよう難読化(有効性に対する指摘も)*

https://openai.com/research/attacking-machine-learning-with-adversarial-examples
*Athalye+, Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, 2018

AIセキュリティとセキュリティの相違点

パッチはコード vs データ

サイバーセキュリティの世界

- 従来のソフトウェアにおけるバグはコードに存在
- 静的解析や動的解析等で問題を特定できる(できないことも…)

AIセキュリティの世界

- MLシステムにおけるバグはモデル内部やデータに存在
- 問題の特定が難しい

AIセキュリティの今後

AIセキュリティと大規模言語モデル(LLM)

- LLMはパラメータ数, データ量ともに大規模
  →モデルだけでなく, データもブラックボックス化が進む

- 創発(*An ability is **emergent** if it is not present in smaller models but is present in larger models*)の指摘*(諸説あり**) →AIセキュリティの知見は適用できる?
- LLMに意思決定が任される可能性
  - ex. Open AIのFunction Calling***
    - 関数を呼び出すタイミング, 関数の引数をLLMがjsonで出力
    - クラウドリソースを無限に消費する攻撃などありえる(?)

*Wei+, Emergent Abilities of Large Language Models, 2022
**Schaeffer+, Are Emergent Abilities of Large Language Models a Mirage?, 2023
***https://platform.openai.com/docs/guides/gpt/function-calling

AIセキュリティの重要さ

- 機械学習が応用されるほど重要になる
  - 自動運転, セキュリティ製品, 医療現場, etc
- AIセキュリティの知見で安全性を担保し, 機械学習の適用可能な領域を増やす

*Rob van der Veer , How Artificial Intelligence attacked my family and other AI security lessons,
https://www.softwareimprovementgroup.com/how-artificial-intelligence-attacked-my-family-and-other-ai-security-lessons/

まとめ

機械学習の基礎
- 機械学習はデータから関数を自動的にプログラミングする

AIセキュリティとは
- MLシステムの脆弱性についての研究分野
- 多様な攻撃手法と防御手法がある

AIセキュリティとセキュリティの関係
- AIセキュリティとセキュリティには似た概念がいくつも見つかる

AIセキュリティの今後
- 急速な機械学習の発展とともに重要性が高まる

References

- 瀧川一学, 機械学習と自動微分: https://speakerdeck.com/itakigawa/ji-jie-xue-xi-tozi-dong-wei-fen-2023
- 岡野原 大輔, 大規模言語モデルは新たな知能か, 岩波書店
- https://course.mlsafety.org
- https://github.com/Trusted-AI/adversarial-robustness-toolbox
- セキュリティエンジニアのための機械学習 Chebbi 著, 新井 他 訳, オライリー・ジャパン
- 深層学習 改訂第二版, 岡谷 貴之著, 講談社