Orchestrating a brighter world NEC

資料1

AIベースシステムの事業化における課題

2019年3月5日

日本電気株式会社 セキュリティ研究所
谷　幹也

\Orchestrating a brighter world

未来に向かい、人が生きる、豊かに生きるために欠かせないもの。
それは「安全」「安心」「効率」「公平」という価値が実現された社会です。

NECは、ネットワーク技術とコンピューティング技術をあわせ持つ
類のないインテグレーターとしてリーダーシップを発揮し、
卓越した技術とさまざまな知見やアイデアを融合することで、
世界の国々や地域の人々と協奏しながら、
明るく希望に満ちた暮らしと社会を実現し、未来につなげていきます。

本日の説明

1. **AI(＝機械学習型）を活用する場合のビジネス課題**

2. **AIシステムのセキュリティ課題**

1. AI(＝機械学習型）を活用する場合のビジネス課題

1-1. 変革しつつあるビジネス形態における課題

1-2. 将来的に出てくるビジネス課題

1-1. 変革しつつあるビジネス形態における課題(1)

従来型のシステム開発・SIビジネスから、データ分析ビジネスへ移行する際に発生する課題と対策・取り組み

値付け/納品/検収

- 費用積上げ型のSIビジネスから顧客の受取価値ベース、成功報酬型ビジネスへ

SOAと品質保証 / 認証システム

- 提供サービスの精度についての保証は困難
- システムとしての品質保証

データの確保

- 必要データの確保とデータの品質

所有権の問題

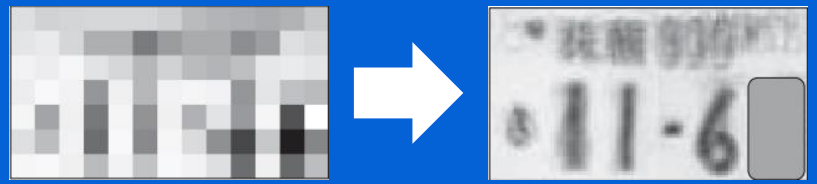
1-1. 変革しつつあるビジネス形態における課題:データがAIの競争力を決める

データの有効活用に向けては、多様なデータソースの確保、データのクオリティの確保、およびデータ活用基盤の整備が必要

多様なデータソースの確保

- ドメインを超えてデータを確保・流通
- オープンデータも活用し量を確保

データの良質化



- 雑音やバイアスを除去
- 通常では使えないデータも有効活用
- リアルタイムなデータの確保

研究者
データサイエンティスト

- 測定バラつきへ配慮
- 倫理観に基づいた選定
- ドメイン知識とセットでの活用

1-1. 変革しつつあるビジネス形態における課題(2)

センサーやログ情報が突き合わされることによるプライバシー問題と
オープンな環境下で発信された情報の信頼性が課題

収集した情報の利活用におけるプライバシー問題

- データ集約することにより、個々の情報では問題にならない個人の特定が可能になる場合が存在

オープンな環境下での信頼性の課題

- 工場や企業が発信した情報を元にダイナミックにサービスの連携などがなされる際、どのように発信された情報を信用するのか？
- 現在の企業間システムの連携は、契約ベース。今後、ＡＩが最適なサービス連携等を考える場合、どうサービスを信頼していくか？が課題。

1-1. 変革しつつある技術的課題(⇒2章にて詳述)

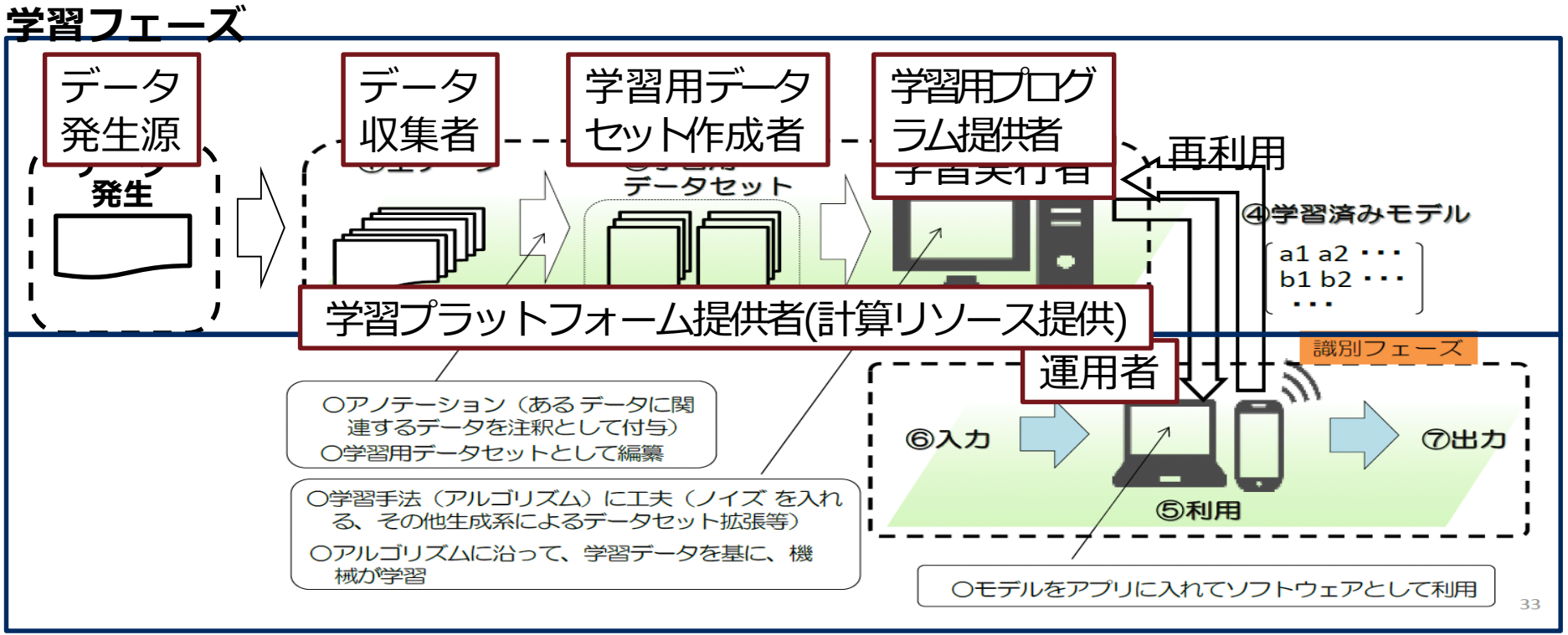
機械学習で作り出されたモデルに対する攻撃により、データの漏えいや間違った振る舞いを誘導する技術的課題が存在する

- 学習済モデルに対して悪意ある問い合わせを繰り返すことにより個人情報の漏えい

- 学習データに悪意のあるデータを混ぜることにより、間違った学習を行ってしまう課題

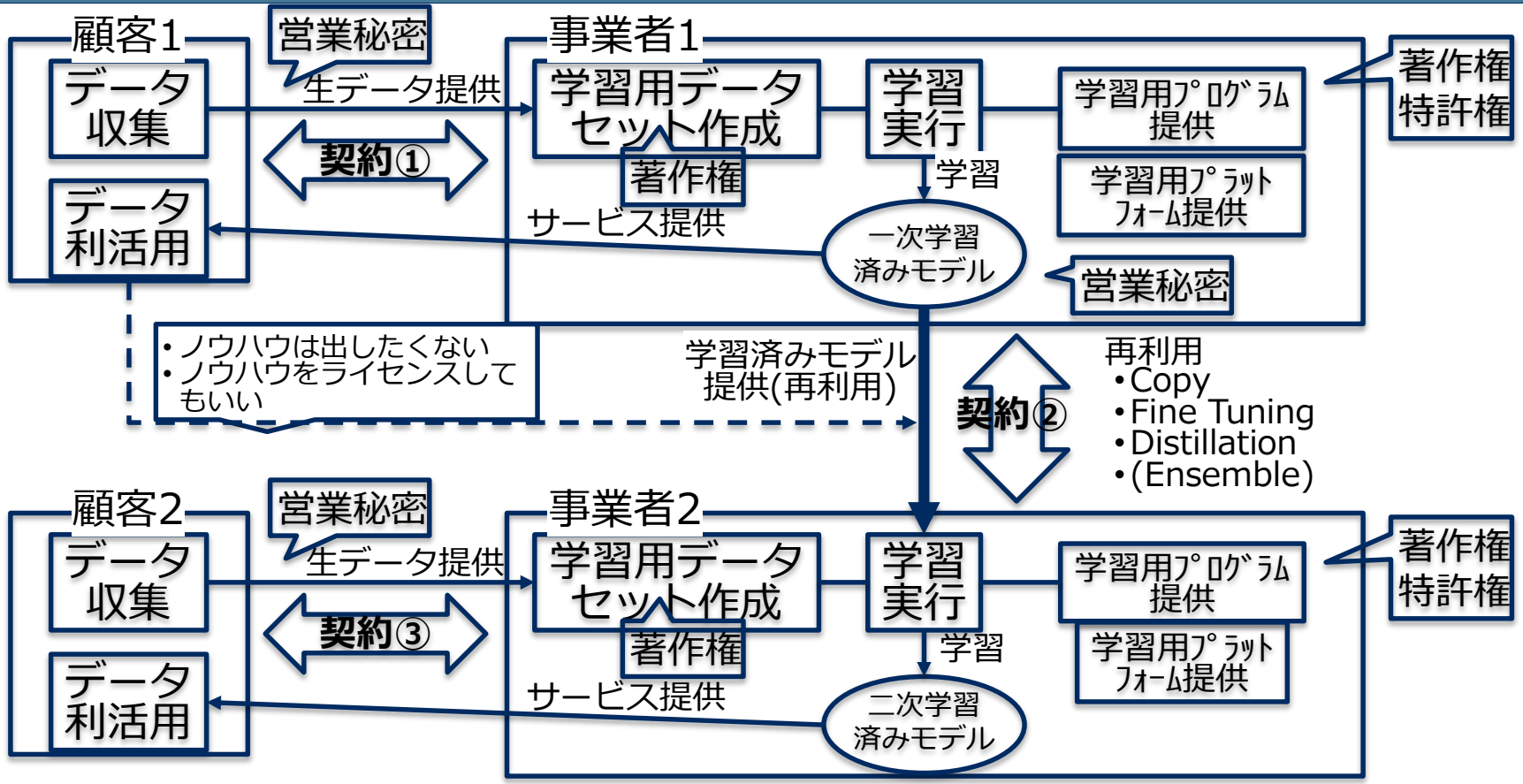
- 学習済モデルの特性を利用して誤った結果を導き出す攻撃の存在
  - 例)Adversarial example攻撃

1-2.将来的に出てくるビジネス課題：ステークスホルダーの複雑化



出典：産業構造審議会 商務流通情報分科会 情報経済小委員会
分散戦略ワーキンググループ 第7回 資料2 2016年8月29日 p.75

1-2. 将来的に出てくるビジネス課題：仮想事例



江村克己(2016), 人工知能の活用と共有経済の進展から考察するこれからの知的財産, 知財研フォーラムVol.107.pp.30-37 より

1-2. 将来的に出てくるビジネス課題(1)

契約だけでは解決が困難な学習済モデルの模倣

- 不正なCopy, Fine Tuningに関する契約違反、Distillationによる模倣問題
- 資本主義経済の立場とシェアリングエコノミーの立場におけるせめぎ合い
- フリーライドの規制

契約当事者間の力の不均衡を背景とする不公正な契約

- プラットフォーム提供者による囲い込みとこれを背景とする不公正契約の可能性

演繹的アルゴリズムにおけるサービスの品質保証

- サービス・製品の責任に関する保証問題
- 納品、検収等の商習慣における基準の変革

データ利用に関するELSI問題

1-2. 将来的に出てくるビジネス課題：**ELSI（Ethical, Legal and Social Issues）の重要性**

- IoTにより集められる情報には、沢山のパーソナル情報が含まれているが、プライバシーを守りながら利便性を損なわないためには？
- AIが人に代わって高度な制御を行うようになった場合、その安全を担保するには？

- IoT・AIにより実現される新たな経済社会のもたらす多様な影響や課題について多角的に検討を行うことが不可欠
- **ELSI「倫理的、法的、社会的課題」の検討**が重要

1-2. 将来的に出てくるビジネス課題(2)

社会システムにおける全体調整

- 単体システムの幸福と全体システムの幸福

AI間での調整(競争・競合・調停・協調・戦争,…)

- 個別AI間での調整をどのように行うのか？

2. AIシステムのセキュリティ課題

ＡＩシステムのセキュリティ課題：概観

現状 機械学習がコモディティ化
様々な場面でのAI活用に期待高まる

AI＝
- 特定用途にデータを凝縮したDB
- 特定業務を代替するロボット
- 能動的に変化を捉えるセンサー

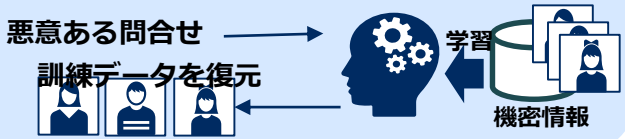
AIを運用する際に様々なリスクあり

Adversarial Example

悪意のあるインプットで誤作動を誘引
- 自動運転の最大の障害

Memory Leakage

学習モデルから訓練データを推定・復元
- AIプラットフォーム事業の障害



Data Poisoning

訓練データに悪意のあるデータを混入
- オンライン学習、モデル更新の障害

ＡＩシステムのセキュリティ課題: モデルの安全性(1)

“ML Models is FRAGILE”

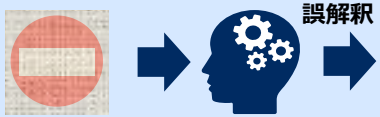
AIに対する“セキュリティ”

**自動運転や常時監視などの
ロボットやセンサーが自律的に
業務を遂行する際に問題となる
重大な社会課題の一つ**

Adversarial Example

悪意のあるインプットで誤作動を誘引

- 自動運転の最大の障害

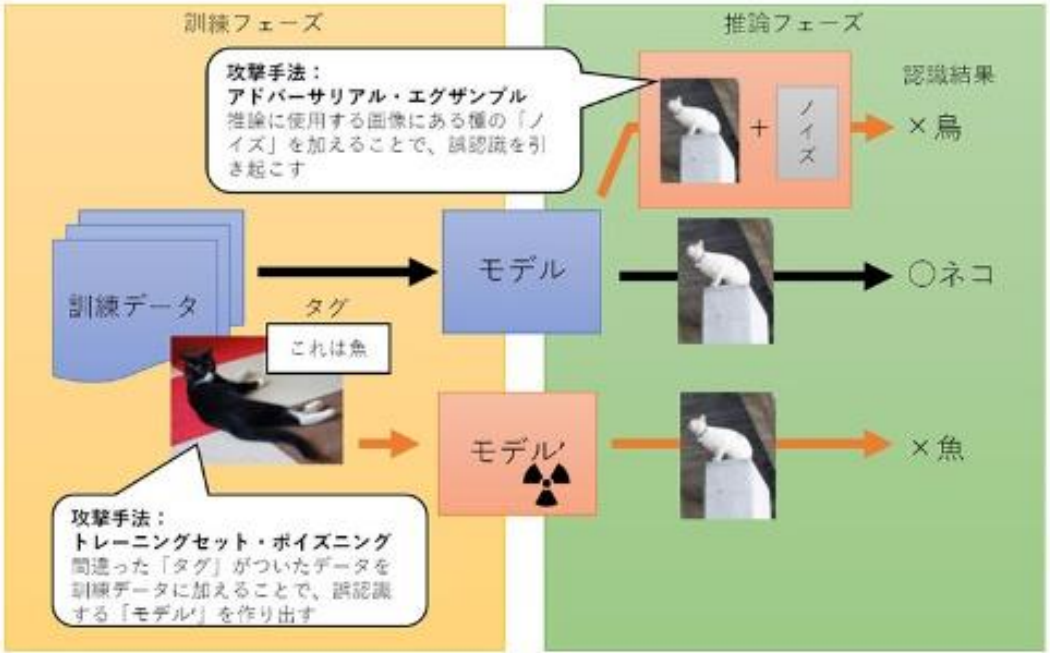


Data Poisoning

訓練データに悪意のあるデータを混入

- オンライン学習、モデル更新の障害

ＡＩシステムのセキュリティ課題: モデルの安全性(2)



ディープラーニングにもセキュリティ問題、AIをだます手口に注意
中田 敦＝シリコンバレー支局, *2017.11.15*
http://tech.nikkeibp.co.jp/it/atcl/column/15/061500148/111400137/

Adversarial Example

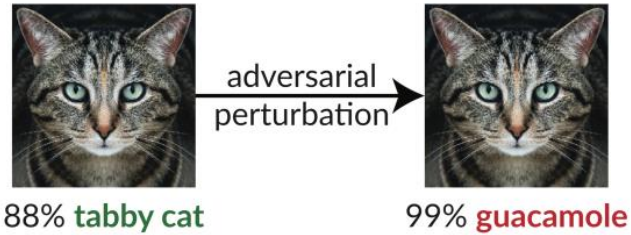
ＡＩの安全性に関するクリティカルな問題 “**Adversarial Example**”

- Adversarial Example (AX)：ＡＩが誤判断するよう意図的に生成した入力データ
- 任意の機械学習モデルで生じ得る問題
- 特に人が介在しないリアルタイムなアプリで多大な損害が生じる恐れ

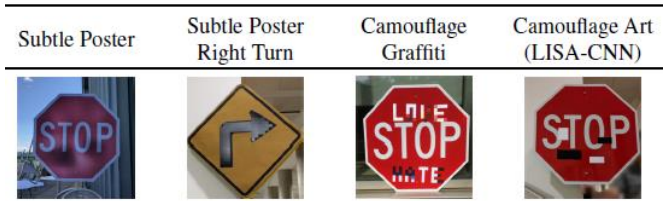
攻撃者優位な状況。ヒューリスティックな防御はすべて破られている

- ヒューリスティックな防御は、理論的な保証がなく、簡単に“穴”を発見されている
- 理論保証のある防御は、莫大な計算時間もしくはメモリを必要とする

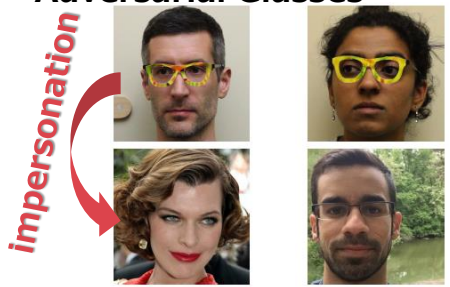
Adversarial Example (AX)



Crafted AXs (result in “speed limit 45”)



Adversarial Glasses



なぜAdversarial Exampleが生じるか

学習が不十分

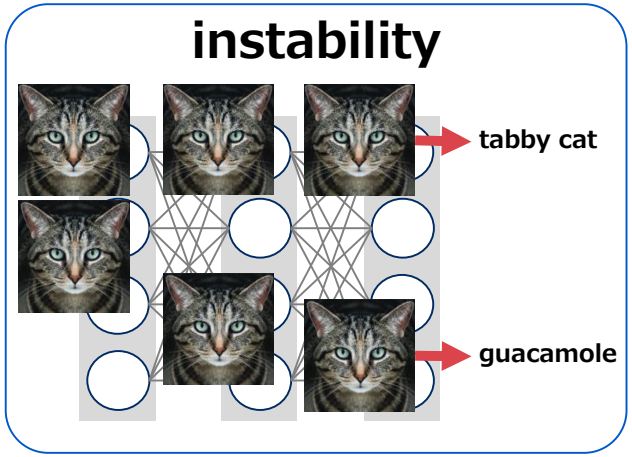
- 訓練データからの微小な変動をモデルが捉えられていない。解決には莫大な訓練データが必要

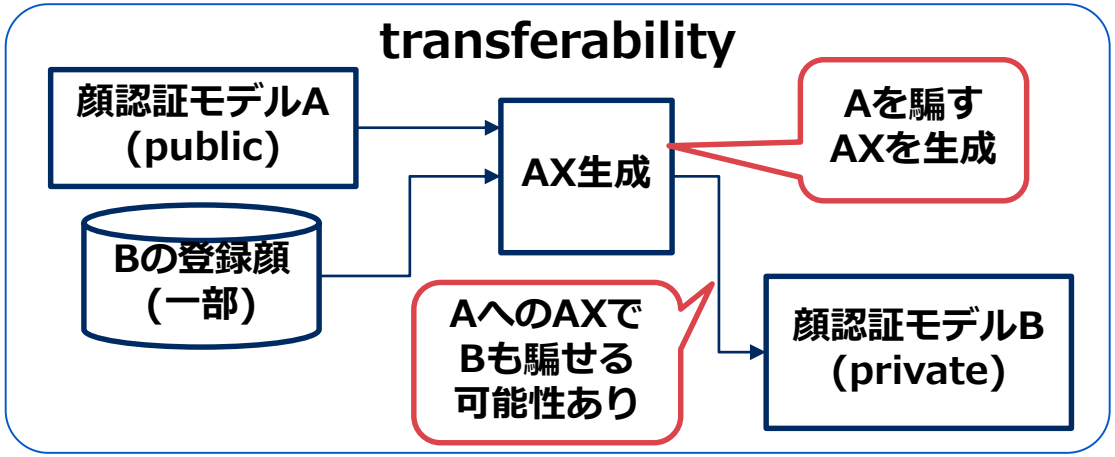
モデル内の挙動が不安定 (Instability)

- 入力時は近いデータであっても、モデルの内部・出力時に距離が遠くなってしまう現象

攻撃の転移 (Transferability)

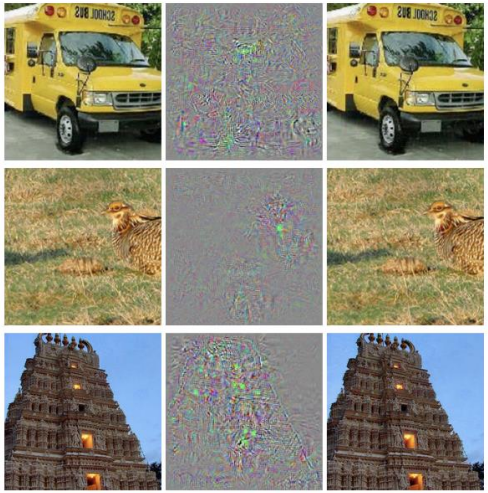
- あるタスクのモデルから学習した攻撃は、(black-boxな) 同じタスクのモデルに共有に有効



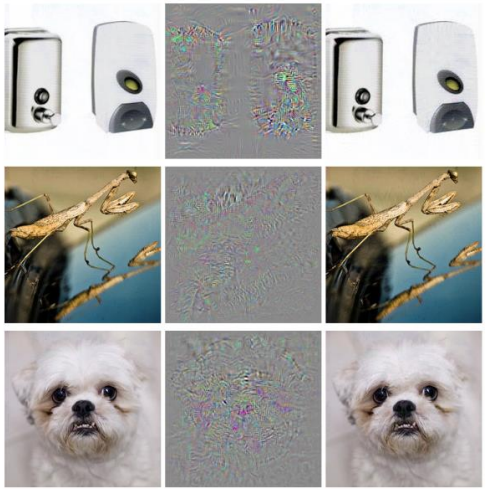


Adversarial Example(1)

元画像に対して、人間の目には違いがわからない摂動画像を重畳することで
他の画像だとご認識させることができる



元画像　摂動　ダチョウ



元画像　摂動　ダチョウ

Intriguing properties of neural networks
Christian Szegedy, et.al
(Submitted on 21 Dec 2013 (v1), last revised 19 Feb 2014 (this version, v4))
https://arxiv.org/abs/1312.6199

Adversarial Example(2)

TABLE IV: Sample experimental images for the attacks detailed in Table V and Table VI at a selection of distances and angles.

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%

Robust Physical-World Attacks on Deep Learning Models
Ivan Evtimov, et.al,
(Submitted on 27 Jul 2017 (v1), last revised 13 Sep 2017 (this version, v4))
https://arxiv.org/abs/1707.08945

Adversarial Example(3)

Attack on a Stop Sign using Black/White Art Stickers

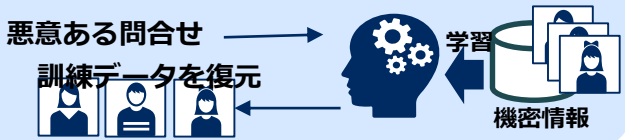
https://youtu.be/1mJMPqi2bSQ

ＡＩのセキュリティ課題: モデルからのリーク

Memory Leakage

学習モデルから訓練データを推定・復元

- AIプラットフォーム事業の障害



“ML Models Remember Too Much”

AI時代における新しいタイプの“データセキュリティ”

お客様のデータやオープンデータを活用し課題を解決していくＡＩ事業における考えるべきリスクの一つ


Orchestrating a brighter world
NEC