資料５－４

画像識別AI（CNN）に対する脅威と対策

2025年11月21日

三井物産セキュアディレクション株式会社

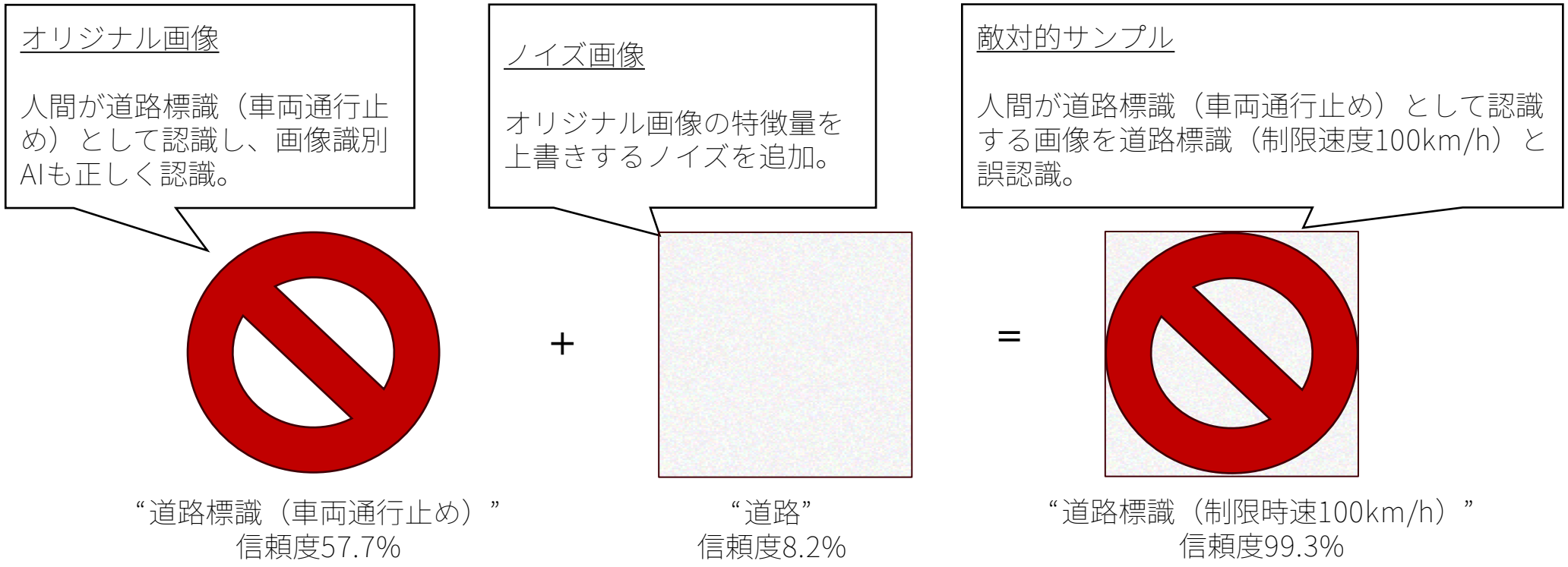
アジェンダ

- 画像識別AIの代表的な脅威
- 代表的な脅威における対策
- その他の脅威と対策

画像識別AIの代表的な脅威

画像識別AIの代表的な脅威（敵対的サンプル）

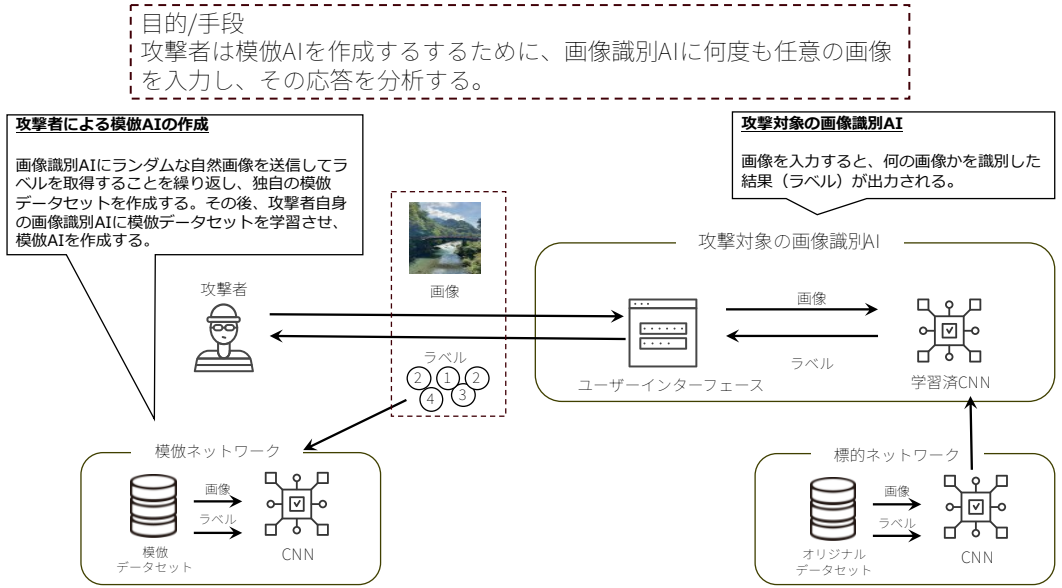
- **脅威の概要**
  - 入力画像に微小なノイズを加え、AIが捉える特徴を別の物体の特徴へと上書きすることで誤識別を誘発する（**敵対的サンプル**）。
  - 以下の図は、オリジナル画像に特徴量を上書きするノイズを加えることで、画像識別AIが、道路標識（車両通行止め）を道路標識（制限速度100km/h）と誤認識する例を示したもの。
  - 画像識別AIを対象とした敵対的サンプルについては、多くの研究事例が知られている。



出典： https://arxiv.org/abs/1412.6572v3
※上図は、出典の論文を基に新規作成した内容

画像識別AIの代表的な脅威（モデル抽出攻撃）

- **脅威の概要**
  - 画像識別AIの挙動を観察して、同等の性能を持つ「模倣AI」を作成する（**モデル抽出攻撃**）。
  - 以下の図は、攻撃者が画像識別AIの挙動を観察することで、同等の性能を持つAIを複製するイメージを表している。
    - 攻撃者は、標的となる画像識別AIに多数の画像を入力し、それらに対する応答(分類結果のラベル)を収集する。
    - 入力画像とその分類結果を紐付けて、独自の「模倣データセット」を作成する。
    - その後、攻撃者は自身のAIに模倣データセットを学習させ、標的となる画像識別AIとほぼ同等の性能を持つ「模倣AI」を作成する。
  - 画像識別AIを対象としたモデル抽出攻撃については、研究事例が知られており、実際の商用サービスにおけるものが報告されているほか、視覚言語モデル（VLM）においても同様な手法でモデル抽出攻撃の報告がある。

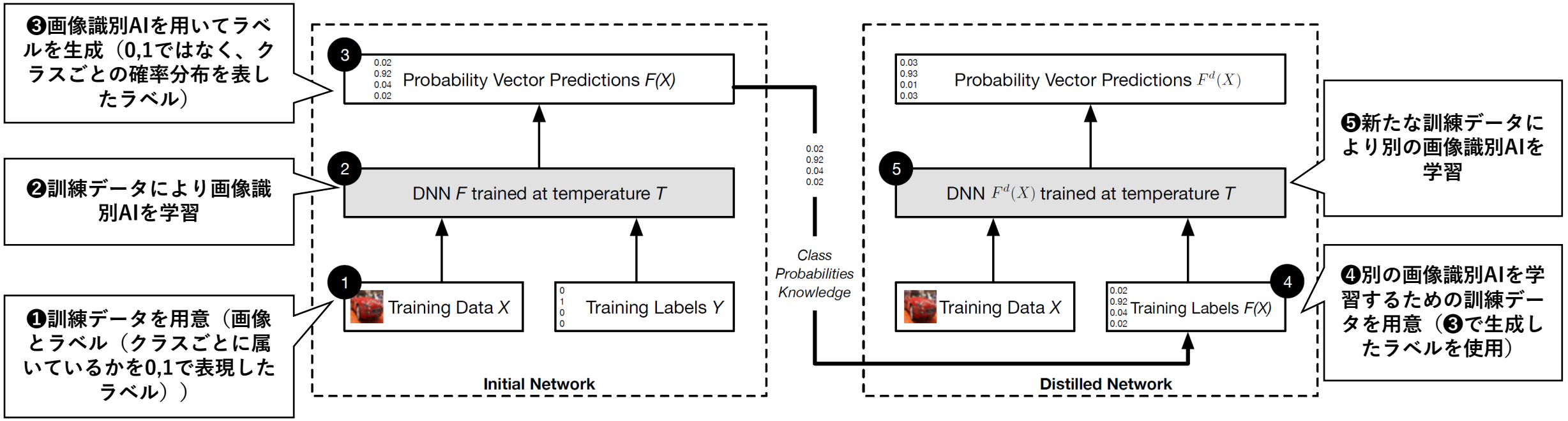


出典：https://arxiv.org/abs/1806.05476
※上図は、出典の論文を基に新規作成した内容

代表的な脅威における対策

代表的な脅威における対策（敵対的サンプル）

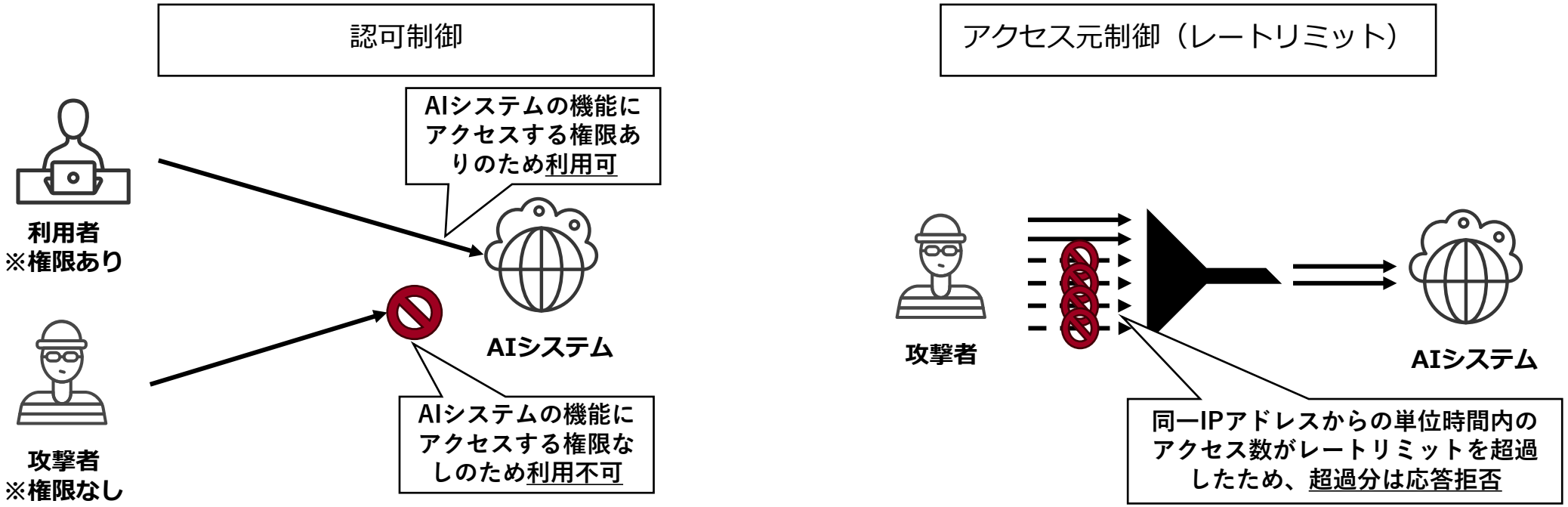
- AIの学習時に、通常の学習データに敵対的サンプルを加え、敵対的サンプルの特徴を学習する**「敵対的学習」(Adversarial Training)**により、敵対的サンプルによる誤分類を抑制することができる。
  - 出典：https://arxiv.org/abs/1412.6572
- また、敵対的サンプルの対策として、**「防御的蒸留」(Defensive Distillation)**という手法が提案されている。
  - 「元となる画像識別AIを用いて、蒸留用の訓練データ（ラベル）を生成」「蒸留用の訓練データにより新たな画像識別AIを学習させる」という手順により、防御的蒸留を行う。
  - 蒸留用の訓練データ（ラベル）により、モデルの決定境界が滑らかになり、入力のわずかな摂動に対する感度が低下するため、敵対的サンプルが生成されにくくなる。



出典： https://arxiv.org/abs/1511.04508v2

代表的な脅威における対策（モデル抽出攻撃）

- モデル抽出攻撃の対策には、AIシステムへの**「認可制御」**や、同一IPアドレスからの単位時間内のアクセス数にレートリミットを設定して制限する**「アクセス元制御」**が有効である。
  - 「認可制御」では、AIシステムの機能へのアクセスについて、利用者が適切な権限を持っているかを確認し、権限が無ければ応答を拒否する。模倣データセットの収集を防ぐことで、攻撃を抑制することができる。
  - 「アクセス元制御」では、同一IPアドレスからの単位時間内のアクセス数に回数制限（レートリミット）を設けて、レートリミットを超過した場合は応答を拒否する。模倣データセットの作成には多くの試行回数を要するが、多数の試行に対して制限を掛けることにより、攻撃を抑制することができる。



その他の脅威と対策

その他の脅威と対策

- その他の脅威と対策としては、文献調査の結果、以下のようなものが知られている。

〇スポンジ攻撃（DoS攻撃）：画像識別AIに対して、処理負荷が高まるように細工をした画像を入力することで、想定以上の計算負荷を生じさせ、画像識別AIの応答の遅延や停止を引き起こす攻撃（出典：https://arxiv.org/abs/2006.03463）

- 対策としては、通常の入力の処理に必要な時間を基に閾値を設定しフィルタリングすること、平均ケースだけでなく、最大遅延・最大消費を設計に織り込むことなどが提唱されている。

〇メンバーシップ推論攻撃：画像識別AIへの画像入力に対する出力を分析することで学習に使われたデータセットが推測され、情報漏洩※につながる攻撃（出典：https://arxiv.org/abs/1610.05820）

- 対策としては、モデルの過学習を抑えてメンバー／非メンバーでのモデル振る舞いの差を小さくする、出力される確信度のスコア丸めることなどが提唱されている。

※具体的な脅威として、例えば、ある患者の臨床記録画像がその病気を対象とした画像識別AIに含まれていたという事実が判明すれば、その患者がその病気である可能性が高いという情報漏洩になる可能性が指摘されている。

〇モデル反転攻撃：画像識別AIの出力（確信度等）を利用して、学習に使われた画像データを逆算し、元データに近い画像を復元してしまう攻撃（出典：https://dl.acm.org/doi/10.1145/2810103.2813677）

- 対策としては、出力される確信度のスコア丸めること、出力をモデルラベルのみに制限することなどが提唱されている。