大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

トップエスイー ソフトウェア開発実践演習

自動運転システムにおける機械学習システムを守るための対策の検討

株式会社日立製作所　石野 正典　　株式会社NTTデータ・アイ　伊藤 匡哉　　NTTテクノクロス株式会社　大澤 亮孝
キヤノン株式会社　斎藤 健太郎　　株式会社NTTデータグループ　清宮 聡史

AIセキュリティの現状と課題

- AIシステムの普及とともにAIに対する攻撃の存在を踏まえたセキュリティ対策が求められている。
- その一方、産業界ではAIシステムの開発、運用現場に適用可能なAIセキュリティ対策の具体的手法が十分に検討されていない。

手法の検討と実機検証

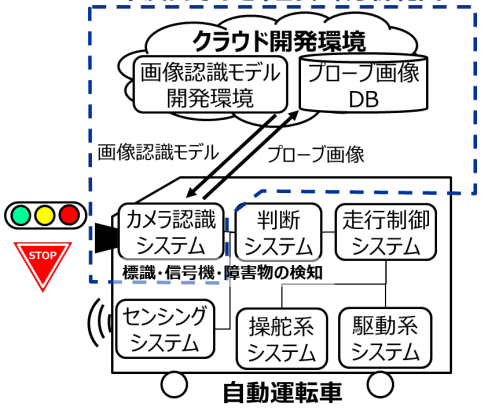
- リスク分析として、AI 固有の脅威をCRSS とSTAMP/STPAを用いて抽出し、対策優先度の高い脅威を抽出した。
- 机上での分析結果の妥当性検証のため、AI への攻撃・防御検証フレームワークを用いた実機検証を実施した。

本演習での検討アプローチ

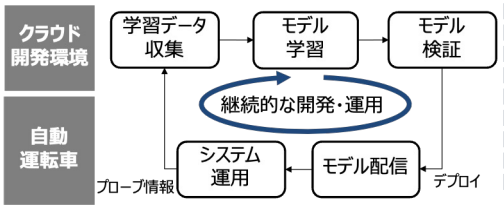
モデル化

■ 分析対象システム
自動運転車の走行・走行制御のための画像認識システム

本演習でのセキュリティ分析範囲



■ 画像認識システムの
AI 開発・運用ライフサイクル



リスク分析（机上検討）

CRSS分析(定量的・ルールベース)

保護資産	攻撃分類	場所	CRSS値
モデル	モデル汚染	クラウド	9.67
センサー情報	データ汚染		8.77
モデル	敵対的サンプル		7.05

⋮

・MLシステム特有の資産、クラウドへの攻撃がCRSS値が高い。

STAMP/STPA(定性的・シナリオベース)

ハザード誘発要因	シチュエーション
外部環境	窓ガラスに映った標識を実物と誤認識する。
内部システム/クラウド関連	入力画像に対してノイズが乗る。クラウドで学習データやモデルが改ざんされる。

⋮

・自動運転車の画像認識システムにおいて、主な原因は外的要因、物体検知モデルの誤作動による事故が考えられる。

実機検証

- ML開発工程、CRSS値、STAMP/STPAから選定した８種類の攻撃を検証
- 攻撃の容易性や効果に机上検討との違いに気づきあり
- 企業では、各業界のユースケースを想定し、ML開発の各工程で対策が必要

本演習でのポイントと成果

◆ **開発・運用ライフサイクルのモデル化**
ML特有の攻撃を考慮した継続的な評価を行うためのモデルを考案した。

◆ **机上検討（脅威分析と安全性解析）**
2つの異なる分析手法の組合せにより、クラウドサービス上の攻撃リスクを抽出、対策優先度づけに成功した。

◆ **実機検証（AIセキュリティOSSの活用）**
AIセキュリティOSSを活用することで自動運転車の画像認識システム特有の脅威と対策を把握できた。

今後の課題

✓ **実践的なノウハウ・ガイダンスの普及**
各業界/各産業分野横断で活用可能な机上検討および実機検証を踏まえたガイダンス

✓ **日々高度化する脅威への追従**
AIセキュリティに関する学術的研究および産学官民での連携の推進

✓ **AIセキュリティへの意識向上**
企業や研究機関での教育プログラム、トレーニングの推進