氏名: 程　ヨウ，試験区分: 情報科学区分，希望研究室: 情報基盤システム学研究室，現在の専門: 都市システム工学

小論文

1　これまでの研究内容

1.1　研究の背景・課題

現在, 世界中にある自動車の合計数はすでに 10 億台を超えており、2035 年までにこの数字は 20 億台に達すると予測されている [1]。自動車の台数の増加に伴って、交通渋滞や交通事故の発生率が増加している [2]。一方、情報技術の急速な発展に伴い、人々の移動手段はより複雑化・多様化している。このような背景の中、自動車のインターネット (Internet of Vehicles, IoV) が生まれた。自動車用無線通信技術 (Vehicle to Everything, V2X) を基盤した IoV は自動車、路側機器 (Road-side Unit, RSU)、サービスプロバイダーを 1 つのネットワークシステムとして接続し、それら間の全面的な通信を実現する [3]。IoV を通じて、サービスプロバイダーはユーザと道路環境に関するデータを取得し、これらのデータに基づいた機械学習や深層学習を通じて、自動運転、経路計画、衝突警告、車載インフォテインメントなどの多様なサービスを提供できるようになった。これらのサービスにより都市の交通問題を効率的に緩和し、運転の安全性と快適性を向上させることで、ユーザー体験 (Quality of Experience, QoE) の向上にも寄与している。

しかし、これらのデータはユーザー行動の予測やプライバシーの侵害のため利用される可能性がある。例えば、攻撃者がユーザーの車両の位置情報を取得し、ユーザーの移動パターンを分析することで、ユーザーの生活習慣や行動パターンを把握することができる。その他、現在の車は多くのカメラを搭載しており、これらのカメラが撮影した画像は歩行者や他の車両の識別のため利用されることにより、プライバシー侵害となる可能性がある。そのため、IoV のプライバシー保護手法は、IoV の発展において重要な課題となる。

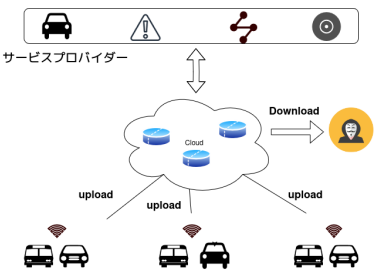

図 1　IoV のシステム構成

1.2　提案手法

そこで本稿では、敵対的サンプル攻撃に基づくプライバシー保護手法 (PIoV,Privacy Internet of Vehicle) を提案する。従来、敵対的サンプルは敵対的機械学習研究 [4, 6] において分類器への攻撃手法と見なされるが、強いセキュリティ手法として利用することもできる [7]。PIoV は敵対的サンプルを利用してデータにわずかな摂動を加えることでユーザーの機密情報を識別する機械学習分類器を欺き、プライバシー侵害を防ぐ手法である。PIoV の中核であるのは「プライバシーメディエーター」。このプライバシーメディエーターは収集されたデータに基づいて摂動の生成とフィルタリングする機能を有する。

1.2.1　敵対的サンプル

敵対的サンプルとは、機械学習モデルに誤った出力を生成させるため特別に設計された入力データである [4]。敵対的サンプルの生成手法としては Fast Gradient Sign Method(FGSM)[5] や Jacobian-based Saliency Map Attack(JSMA)[6] が代表的である。

1.2.2　FGSM

Ian ら [5] はニューラルネットワークの線形性に着目して FGSM という敵対的サンプルを高速に計算する手法を提案した。

$$\tilde{x} \approx x + \epsilon \cdot \mathrm{sign}(\nabla_x \mathcal{J}(\theta, x, y)) \tag{1}$$

(1) 式は損失を最大化する方向に入力 x を変化させるという意味の式である。$\tilde{x}$ は敵対的画像、x は元の画像、y は x に対する入力ラベル、ϵ は摂動の大きさを表示する乗数、θ はモデルのパラメータ、J はモデルの学習に使用した損失関数、sign は勾配の方向を表す。画像内の各ピクセルがどれくらい損失値に貢献しているかを求め、それに応じて摂動を追加することで、損失の最大化を達成する。FGSM はモデルのパラメータ θ と入力データ x に対する勾配情報を取得できれば、1 ステップで敵対的サンプルを生成できる高速な手法として知られている。現実の場合、車は頻繁に移動して異なるサーバーへ接続するため、高速的な摂動生成手法が重要である。JSMA は再帰的な手法であり、摂動の計算に時間がかかるため IoV 環境で適用するのが難しい。そのため本稿では FGSM を PIoV の敵対的サンプル生成手法として利用する。

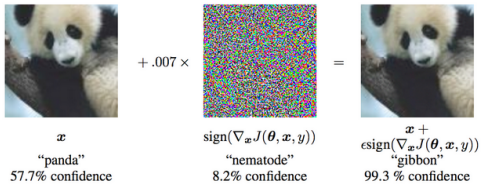

図 2　テナガザルと認識されたパンダの例

1.3 PIoV のプロセス

PIoV の全体プロセスは以下の 4 段階で構成される：

1. 車が収集したデータをメディエータへ伝送
2. メディエータが摂動を生成して, 摂動があるデータセットをサーバーへ送信
3. サーバーがデータを分析し、データより生成したアクション命令コマンドをメディエータへ送信
4. メディエーターがアクション命令コマンドをフィルタリングして, 適切なコマンドだけを車へ転送

IoV に敵対的サンプル攻撃を適用する際、メディエーターが生成した摂動が車に影響を与えることを避けなければならない。そのため、メディエーターは元のデータを変更せず、摂動を加えた新しいデータセットを送信する。しかし、サーバーは摂動を加えた新しいデータセットに基づいて分析すると、誤ったアクション命令コマンドを生成する。このままアクション命令コマンドを実行すると、車は誤った動作を実行し、極めて危険である。そのため、プライバシーメディエーターにはフィルタリング機能が装備されており、受信した異常的なアクションを検出・除去し、適切なアクションだけを通過させる。

1.4 まとめと今後のアプローチ

本研究では IoV のユーザーのプライバシーを保護するため、敵対的サンプル攻撃に基づくプライバシー保護手法 PIoV を提案した。具体的には、センサーデータに摂動を加えることで、第三者の機械学習分類器の精度を低下させ、IoV 設備の機能を損なわずにプライバシー保護を実現する。これからは、PIoV の有効性を検証するための実験と評価を行い、研究の実用化に向け研究を進める予定である。

2 NAIST で取り組みたい内容

2.1 研究の背景

近年、機械学習と深層学習の研究が急速に進展し、ディープフェイクという技術が登場した。ディープフェイク (deepfake) は本来、深層学習を使用して 2 つの画像や動画の一部を結合させ元とは異なる動画を作成する技術である。しかし、ディープフェイクは悪意のある目的で利用され、経済や社会に深刻な影響を与える可能性がある。Zhang らの研究 [8] ではディープフェイクで偽造した音声や画像を用いて、自動認証システムを搭載した車のドアの解錠やエンジン始動などの機能を制御できることを示し、自動認証システムの脆弱性とり堅牢なセキュリティ対策の必要性を掲示した。

2.2 課題

研究背景で述べた通り、より信頼性が高い自動認証システムの実現が求められている。具体的には、ディープフェイクが生成した画像や音声の検出手法とユーザーのプライバシー保護手法の開発が必要となる。しかし現在ディープフェイクの検出手法には 2 つの課題がある。

1. カスタム生成 AI モデル技術の進歩により、攻撃者が大量のカスタム AI を作成するのは容易である
2. 視覚基盤モデルを利用し、検知回避型のディープフェイクが生成される可能性がある

これらの課題は大学院入学後に取り組みたいと考えている。

2.3 NAIST を志望する動機

私が NAIST を志望する理由は、貴学の自由な研究環境に感銘を受け、研究に専念できる環境が整っていると感じたからである。貴学のオープンキャンパスに参加した際、学生の研究の熱意を感じて、自分もその一員として研究に取り組みしたいと思った。そして、自分の研究は大量なデータが必要であり、現場での実験は不可欠である。そのため、現場重視の情報基盤システム研究室の研究環境は私にとって理想的であると考える。希望通り御研究室に配属されれば，社会に貢献できる研究を行い、実用性のある成果を上げたいと考えている。

以上の理由より、私は貴学の情報基盤システム研究室を志望する。

参考文献

[1] Sharma S, et al. "A survey on internet of vehicles: applications, security issues & solutions", Vehicular Communications, 2019, 20:100182

[2] Vegni A M, et al. "A survey on vehicular social networks",IEEE Commun Surv Tut, 2015, 17(4): pp.2397-2419

[3] Boban M, et al. "Connected roads of the future:use cases,requirements, and design considerations for vehicle-to-everything communications",IEEE Veh Technol Mag, 2018, 13(3):pp.110-123

[4] Christian Szegedy, et al. "Intriguing properties of neural networks.", arXiv preprint, 2013. arXiv:1312.6199.

[5] Ian J Goodfellow, et al. "Explaining and harnessing adversarial ex- amples. arXiv preprint", 2014. arXiv:1412.6572.

[6] Nicolas Papernot, et al. "Practical black-box attacks against machine learning." ACM ASIACCS,pp.506–519, 2017

[7] Yunna Lv, et al. "Three-in-One: Robust Enhanced Universal Transferable Anti-Facial Retrieval in Online Social Networks", IEEE TIFS,pp.4408-4421,2025

[8] Xingli Zhang, et al. "From Virtual Touch to Tesla Command: Unlocking Unauthenticated Control Chains From Smart Glasses for Vehicle Takeover",IEEE S&P,2024.