Mobility21
A USDOT NATIONAL
UNIVERSITY TRANSPORTATION CENTER

Carnegie Mellon University

Penn
UNIVERSITY of PENNSYLVANIA

THE OHIO STATE UNIVERSITY

CCAC

Towards Privacy-Preserving Networked Autonomous Mobility: Analysis, Tools Development, and Real-World Evaluation

Ding Zhao (0000-0002-9400-8446),
Zhiwei Steven Wu (0000-0002-8125-8227)

FINAL RESEARCH REPORT

Contract # 69A3551747111

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Problem:

The development of public autonomous mobility, including self-driving vehicles and delivery robots, which benefits from data-driven machine learning and deep learning algorithms, requires urgent investigation into the privacy issue. To protect the sensitive information in the collected data, the learning process should be studied to preserve privacy and also to measure privacy leakage.

This project aims to investigate:

1. existing works about differential privacy theory and privacy-revealing techniques,
2. how to effectively apply privacy-preserving methods to autonomous mobility tasks in order to achieve privacy-preserving under the differential privacy framework, and
3. the privacy risk of using private datasets in the training process.

To this end, we aim to design a privacy-preserving method that can be used in normal autonomous mobility settings and design an attack method to verify how much sensitive information can empirically be revealed during the training process.

Methodology and Findings:

The first stage is to provide a summary of privacy theories and privacy issues in networked autonomous mobility, especially the works about differential privacy. We conducted a literature review on the theory and applications of differential privacy, which aim to preserve privacy. Among individual-level privacy, population-level privacy, and proprietary-level privacy, we focused on individual-level privacy and investigated the privacy-preserving algorithms for individual-level privacy. Specifically, we selected the Gaussian mechanism as the potential method to preserve sensitive information. Moreover, we investigated the possibility of revealing the privacy risk in reinforcement learning algorithms by employing attack methods, namely gradient inversion and model inversion, which aim to reveal private information from gradients and trained models.

The second stage is to study the possibility of using the Gaussian Mechanism to protect the private information in the training dataset. We focused on reinforcement learning (RL) algorithms, which are widely used in self-driving and robotic tasks. We observed that reinforcement learning algorithms usually require large Gaussian noise compared to supervised learning methods used in other tasks, for example, face recognition, due to the smaller dataset and larger epoch. To reduce the demand for large noises, larger datasets and fewer training iterations are critical, and thus the primary conclusion is to choose data-efficient algorithms. We also found that the large gradient variance causes RL algorithms to be more sensitive to large noises compared to supervised learning algorithms. To reduce the sensitivity towards noises and improve the performance of specific reinforcement learning algorithms, we proposed weight-sharing methods. Specifically, we used a recurrent structure to replace the stack of blocks, achieving weight-sharing outside layers. We also proposed to use a low-rank matrix to replace the normal weight matrix, achieving weight-sharing inside layers. The first modification reduced the number of parameters, and the second one alleviated the large gradient variance. Both can increase the signal-to-noise ratio of the sanitized summed gradient given the same noise level. Therefore, our method can achieve higher accuracy under the same differential privacy budget. This method is evaluated in D4RL and the active perception task, an indoor navigation task that allows the robot to utilize mobility to achieve high-quality perception. The network structures are both Decision Transformer. For D4RL, our method significantly improved the performance compared to naively applying the Gaussian Mechanism to the training process of the Decision Transformer. For active perception, our method also achieved better results when the noise level is in reasonable regions.

The third stage is to study privacy leakage in the training process. Two factors can be attacked by adversaries to reveal private information, namely the shared gradient and the publicly released model. We first investigated gradient inversion aiming at the gradient and then studied model inversion. For gradient inversion, we investigated the attack methods for reinforcement learning algorithms and developed a package that defines a general framework. As the assessment of privacy leakage risk in a practical manner has been rarely explored for reinforcement learning algorithms and self-driving and navigation tasks, we studied the possibility of gradient inversion and model inversion in reinforcement learning. For gradient inversion, we proposed a pipeline that can reconstruct the state, action, reward, and estimated Q-value used in the training process from the gradient of the policy network, based on the observation that the gradient inversion of linear layers is much easier and more accurate than the convolution layers and norm layers. Our method requires at least one linear layer at the end of the network, which is widely satisfied by self-driving algorithms. We tested both multi-layer perceptrons and deep convolutional networks, and the results showed that gradients of either structure can be used to reveal the training data. Our method is compatible with both the on-policy algorithm, for example, REINFORCE, and the off-policy algorithm, for example, DQN, and SAC. Our method can successfully reconstruct both the single-modal state and the multi-modal state. To the best of our knowledge, we are the first to study the gradient inversion of multi-modal input data. We evaluated the performance of gradient inversion in the active perception task and the self-driving task. For active perception, the multi-modal inputs include an RGB image (resolution 150×150), a depth image (resolution 150×150), and a coordinate vector. For the self-driving task, the multi-modal inputs include a bird-eye-view image (resolution 256×256), a LiDAR image (resolution 256×256), an RGB image (resolution 256×256), and a vector indicating the reference trajectory. Although thorough experiments have only been done for the setting with batch size 1, this work reveals the privacy risk of federated learning, collaborative learning, and decentralized learning frameworks for self-driving and autonomous navigation tasks. While these frameworks aim to preserve private information by holding the dataset on the local server and only sharing the gradient used to update the model, our work shows that the private information is still at risk due to the private information embedded in the gradient. Furthermore, instead of a theoretical analysis of the privacy budget, this work provides a possibility to empirically evaluate the privacy-preserving ability of the reinforcement learning algorithm. The experiments in the simulator, Carla, are finished. The experiments on the real-world dataset are ongoing.

For the third stage of this project, we summarized the results of gradient inversion in the active perception task as a conference paper and submitted it to the Conference on Robotic Learning (CoRL) [1]. The results of gradient inversion in the self-driving task for both online methods run in the simulator, Carla, and the offline methods conducted on an offline dataset, nuPlan, will be summarized and submitted to IEEE Transactions on Intelligent Transportation Systems (ITS). The code of our gradient inversion framework will be released soon.

Conclusions:

Existing self-driving algorithms to the privacy-preserving framework is still a challenging and unsolved problem. The experiments showed that the performance under privacy constraints can be improved by designing an appropriate network structure. By investigating the gradient inversion, we revealed the privacy risk of federated learning, which is supposed to protect private information. We also experimentally evaluated the privacy leakage risk of learning algorithms.

References:

[1] https://arxiv.org/abs/2306.09273