PrivAttack: A Membership Inference Attack Framework Against Deep Reinforcement Learning Agents

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recently, the substantially improved performance of deep reinforcement learning models at the research level has motivated the employment of these models in real-world domains such as health care, self-driving cars, robotics, and recommender systems. However, due to the concerns stemmed from the sensitive nature of some of these domains to the privacy leakage and lack of enough research in this field, their application has been limited. In particular, while several studies have assessed the privacy of supervised models, the semi-supervised sequential decision making algorithms have not been studied much in this regard. Here, we propose a generic attack framework to test the vulnerabilities of two established deep reinforcement learning algorithms to membership inference attacks. We perform the attack in three high-dimensional continuous locomotion tasks and show that our proposed attack model can predict the vulnerability of the reinforcement learning models with high precision and accuracy.

1 Introduction

Despite the recent advances in the performance of deep reinforcement learning (RL) algorithms in complex domains, these models still struggle to generalize when they move to a new complex environment [9, 5, 17]. There exists a rich body of literature in machine learning that addresses how lack of generalizability leads to potential privacy breaches [15, 6]. However, the focus on RL algorithms in this regard has been minimal. A recent study on the privacy of deep RL models by Pan *et al.* [8] shows that deep RL models potentially breach privacy. In particular, their attack system can infer the floor plans in grid world navigation tasks as well as the transition dynamics of continuous control environments. However, to the best of our knowledge, there has been no empirical study on the potential leakage of collected data employed in training RL agents. In this paper, we introduce the first demonstration of a white-box membership inference attack framework against deep RL agents. In particular, we show that our proposed framework can recognize the membership of a particular data-point (in the form of a trajectory) in a private training set used to train the target deep RL model. To show the effectiveness of our proposed attack framework, we run our proposed attack against two state-of-the-art deep RL models in three high-dimensional continuous control tasks for different trajectory lengths. Our attack framework infers the membership of the training trajectories with considerably high accuracy ranging between 85% to 90%, while the baseline random guess accuracy varies between 44% to 55%. Our results show that the two deep RL models breach the privacy of the training trajectories even in very high-dimensional domains with high variance in the model predictions.

Problem statement: The off-policy algorithms we have used in this study are *Deep Deterministic Policy Gradients* (DDPG) [7] and *Soft Actor Critic* (SAC) [4]. The deep RL agent has no prior knowledge of the underlying environment dynamics, and through interaction with the environment, uses exploration policy π_b to collect training samples from the environment and learns the target policy π_a. We assume that the attacker has the same level of access to the environment as that of the target model. The attacker does not know the private seed number used to train the target model and has only query access to the trained policy π_a. The input to the attack model is composed of: 1) a trajectory from the target model private training set, and 2) a test trajectory generated by the trained policy π_a. The attacker must subsequently determine if the training and the test trajectories belong to the same deep RL agent. The length of input trajectories may vary. A trajectory is a sequence of temporally correlated tuples. Each tuple within a trajectory is in the form of $\langle state, action, reward\rangle$, and the dimensionality of $state$ and $action$ depend on the environment with which the agent interacts.

2 Related Work

There exists an extensive body of literature on membership inference attacks against supervised machine learning models [12, 11, 16]. For the first time, Shokri *et al.* [12] introduced *shadow model* training technique and performed membership inference attack against a deep classifier. Shadow model training is an intuitive approach to designing membership inference attacks by replicating the behavior of the target model through training shadow models on data sets drawn from the same distribution as that of the private data set used to train the target model. The use of shadow model training was subsequently adopted by other follow-up studies [16] [11] [10]. Salem et al. [11] proposed and performed successful attack strategies based on shadow-model training. Yoem et al. [16] showed that overfitting is sufficient for the adversary to perform membership inference attacks against several machine learning models, such as regression and deep convolutional neural networks (CNNs).

In the field of reinforcement learning, Pan *et al.* [8] proposed a black-box attack framework against deep RL algorithms to infer the transition model used to train the target policy. The proposed attacks study the effect of over-fitting on revealing information regarding the agent's training environment as well as the model parameters. However, there is no prior work in the context of deep RL that addresses the problem of membership inference at a microscopic level, where the attacker infers the membership of a particular data point in the training set of a trained policy. Our work is the first implementation of a membership inference attack in a semi-supervised setting where the target model is trained on the environment accessible to the adversary with the same query access level as that of the target model.

3 Methods

Figure 1 depicts the general architecture of our proposed membership inference attack framework against an off-policy deep RL agent. The main components include: 1) *Private Target Trainer*- It uses private seed number, takes as input the number of training time-steps, and privately trains the target model in interaction with the shared environment. The adversary subsequently employ the trained target model to produce test trajectories. 2) *Non-Private Shadow Trainer*- It takes as input the number of shadow models n and the number of training time-steps, and subsequently generates n independent random seeds to train n independent shadow models as well as n independent training and test data sets. The Shadow Trainer has the same access level to the environment as the Target Trainer. 3) *Data Formatter*- It pairs the train and test trajectories uniformly at random and labels them as 'matched' or 'mismatched' if the trajectories belong to the same model or not, respectively. 4) *Attack Trainer*- It trains a classifier that takes as input pairs of trajectories generated by the shadow models and assigns to that the probability that trajectories belong to the same trained model.

The shadow and target trainers use independent random initialization seeds to ensure independent training sets. During the training phase, the training trajectories are collected by each model. The collected trajectories are passed into memory for policy training. In the context of reinforcement learning, each trajectory represents a data-point, and thus the input type for the attack model is in the form of trajectory pairs. We use a probabilistic classifier [3] for the attack model, which output the matching probability between the trajectories. Employing a probabilistic classifier complicates the inference task from the attacker's point of view since the adversary requires to map the probabilistic quantity to a binary outcome (*i.e.* match/mismatch). We resolve this challenge by defining a set of threshold $0 < \beta$<1, above which the probability is mapped to 1, and 0 otherwise. In order to

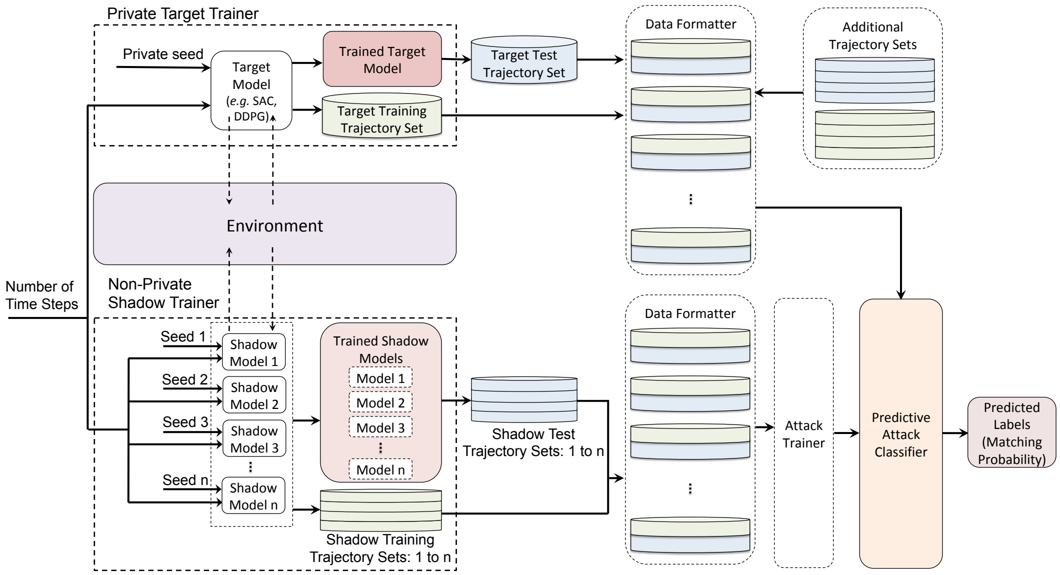

Figure 1: PrivAttack architecture

tune the threshold, we subsequently use the method of *Geometric Mean Relative Absolute Error* (GMRAE) with respect to a given choice of threshold $0 < \beta$<1. Finally, we adopt the following standard performance metrics used in the classification literature [13] to evaluate the performance of our proposed attack against deep RL agents: 1) *prediction accuracy* or *attack accuracy*, which captures the overall performance of the attack classifier; 2) *precision*, which captures the level of agreement between the true labels and the members inferred by the attack classifier. In other words, it shows the fraction of the input pairs classified as matching pairs that are indeed coming from the same model; 3) *recall* or sensitivity, which captures the performance of attack classifier in identifying the true members, or in other words, the fraction of training pairs that the attack classifier can correctly infer as matching pairs.

4 Experiments

We assess the privacy of the two established deep RL models Deep Deterministic Policy Gradients (DDPG) [7] and Soft Actor-Critic (SAC) [4], as well as the performance of the PrivAttack framework. We train the deep RL agents on three high-dimensional continuous control MuJoCo tasks [14] from OpenAI Gym OpenAI GYM [2] *Hopper-v2*, *Half Cheetah-v2* and *Humanoid-v2*. SAC and DDPG implementation used for the experiments are forked from OpenAI spinning-up project [1] (Refer to figure 2 for the benchmark results in these three environments.) We design experimental scenarios to observe the impact of the epoch length on the vulnerabilities of the deep RL models to the membership inference attack. We further study the performance of our proposed attack model using the three standard metrics *accuracy*, *precision*, and *recall*.

To capture the impact of trajectory length on the vulnerability of the deep RL models to membership inference attacks, we train multiple sets of shadow/target models with three different trajectory lengths 50, 500, and 1000 time steps. We train the classifier using 2 to 20 shadow models at a time, with the acceptance threshold ranging from 0.1 to 0.9, and the attack classifier training set size up to 10000 labelled trajectory pairs. The obtained results (Table 1) show that in general, longer trajectory lengths lead to less private algorithms. Note that a longer trajectory carries more information about an individual participating in the data set, which consequently leads to more vulnerability of the individual to membership inference attacks. However, despite the general increasing trend in attack accuracy upon increasing the trajectory length, there are still some exceptions. For instance, while increasing the trajectory length for the SAC agent in Hopper leads to up to 8% increase in the membership inference accuracy, in Half-Cheetah, we see a 5% decrease. These exceptions show that

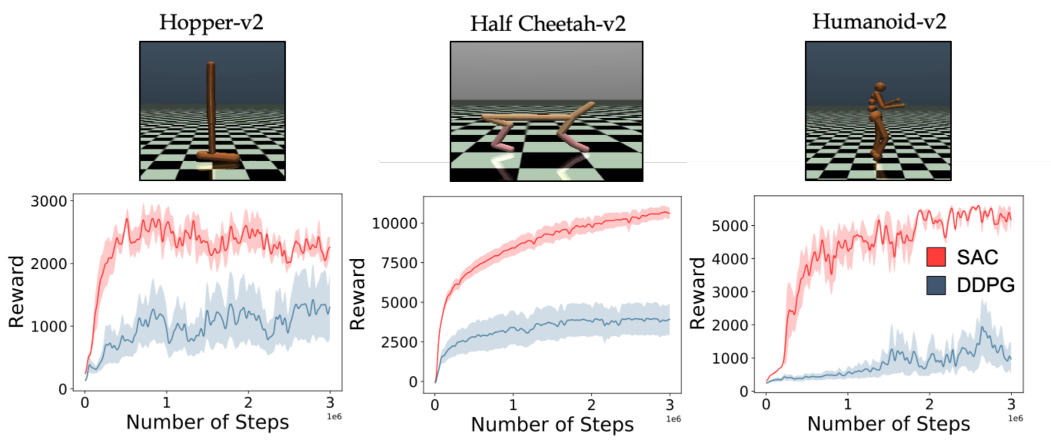

Figure 2: Benchmark results on three high-dimensional locomotion tasks from OpenAI Gym environment. The results are averaged over 5 independent runs with 5 random seeds.

there may be other factors apart from the trajectory length that affect the privacy level of these deep RL algorithms, which can be an interesting direction for the future studies.

Hopper-v2 — Target Model Training Size

Model	100K (Trajectory size)			500k (Trajectory size)			1m (Trajectory size)		
SAC	50	500	1000	50	500	1000	50	500	1000
Attack Accuracy	.792	0.83	.866	.818	.836	.854	0.824	.826	.900
Random-Guesser Accuracy	.552	.502	.484	.506	.508	.48	.484	.502	.536
Attack Precision	.997	.985	.997	.995	1.0	.997	.997	1.0	.995
Attack Recall	.793	.840	.867	.821	.84	.855	.825	0.826	.903
DDPG	50	500	1000	50	500	1000	50	500	1000
Attack Accuracy	.824	0.85	.872	.810	.820	.848	.790	.818	.886
Random-Guesser Accuracy	.526	.494	.512	.496	.516	.54	.492	.500	.470
Attack Precision	.997	.992	1.0	.985	.976	.982	.992	.992	1.0
Attack Recall	.825	.855	.872	.824	.821	.852	.794	.822	.886

Half Cheetah-v2 — Target Model Training Size

Model	100K (Trajectory size)			1m (Trajectory size)			2m (Trajectory size)		
SAC	50	500	1000	50	500	1000	50	500	1000
Attack Accuracy	.798	.81	.846	.828	.812	0.87	.808	.840	.889
Random-Guesser Accuracy	.514	.484	.498	.524	.474	.52	.49	.504	.51
Attack Precision	.982	.997	.998	.992	.997	.998	.987	.992	1.0
Attack Recall	.801	.812	.847	.831	.814	.872	.839	.843	.886
DDPG	50	500	1000	50	500	1000	50	500	1000
Attack Accuracy	.808	.824	.820	.822	.830	.857	.802	.824	.887
Random-Guesser Accuracy	.498	.555	.442	.482	.496	.492	.544	.510	.53
Attack Precision	.982	.997	1.0	.987	.983	1.0	.990	.978	.998
Attack Recall	.819	.826	0.82	.830	.841	.857	.810	.842	.888

Humanoid-v2 — Target Model Training Size

Model	100K (Trajectory size)			1m (Trajectory size)			2m (Trajectory size)		
SAC	50	500	1000	50	500	1000	50	500	1000
Attack Accuracy	.78	.824	.886	.820	.831	.902	.800	.826	.884
Random-Guesser Accuracy	.548	.488	.478	.496	.522	.466	.524	.470	.478
Attack Precision	.997	.997	1.0	.997	.982	.995	.982	.998	1.0
Attack Recall	.781	.825	.886	0.82	.842	.905	.817	.828	.884
DDPG	50	500	1000	50	500	1000	50	500	1000
Attack Accuracy	.816	.844	.866	.846	.870	.890	.808	.818	.891
Random-Guesser Accuracy	.489	.516	.498	.506	.487	.528	.49	.472	.478
Attack Precision	1.0	1.0	.986	1.0	.981	.992	.990	.994	.996
Attack Recall	.816	.844	.876	.846	.884	.885	.801	.819	.881

Table 1: Tabular representation of membership inference attack performance as a function of trajectory length and total number of steps. The experiments are conducted with 5 shadow models and an acceptance threshold of 0.9.

5 Conclusion

The lack of studies that examine the vulnerability of deep RL models against potential membership inference attacks has turned to a real obstacle to such models' industrial application. To address this challenge, in this paper we propose a generic membership inference attack framework. We demonstrate the performance of our attack framework in different epoch-length regimes. Moreover, our attack framework reveals the substantial vulnerability of two established deep reinforcement learning models to the white-box membership inference attack. Finally, our study demonstrates the impact of trajectory size on the vulnerability of the deep RL models to membership inference attacks. This pivotal factor should be considered in the design of privacy-preserving deep RL models. Investigating the impact of other variables such as task dimensionality and algorithmic stability on the privacy of the deep RL models as well as the attack performance is an interesting future direction.

References

[1] J. Achiam. Spinning Up in Deep Reinforcement Learning. 2018.

[2] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[3] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[4] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870, 2018.

[5] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] B. Jayaraman and D. Evans. Evaluating differentially private machine learning in practice. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1895–1912, 2019.

[7] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[8] X. Pan, W. Wang, X. Zhang, B. Li, J. Yi, and D. Song. How you act tells a lot: Privacy-leakage attack on deep reinforcement learning. *arXiv preprint arXiv:1904.11082*, 2019.

[9] A. Rajeswaran, K. Lowrey, E. V. Todorov, and S. M. Kakade. Towards generalization and simplicity in continuous control. In *Advances in Neural Information Processing Systems*, pages 6550–6561, 2017.

[10] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jegou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567, 2019.

[11] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *NCSS*, 2019.

[12] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[13] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[14] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 5026–5033. IEEE, 2012.

[15] Y.-X. Wang, J. Lei, and S. E. Fienberg. Learning with differential privacy: stability, learnability and the sufficiency and necessity of erm principle. *The Journal of Machine Learning Research*, 17(1):6353–6392, 2016.

[16] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

[17] A. Zhang, N. Ballas, and J. Pineau. A dissection of overfitting and generalization in continuous reinforcement learning. *arXiv preprint arXiv:1806.07937*, 2018.