A Study on Protecting Privacy of Machine Learning Models

Younghan Lee*, Woorim Han*, Yungi Cho*, Hyunjun Kim*, Yunheung Paek*
*Dept. of Electrical and Computer Engineering and Inter-University Semiconductor Research Center (ISRC), Seoul National University

Abstract

Machine learning model gained the popularity in recent years as multi-national companies have incorporated machine learning in their services. Such service is called machine learning as a service (MLaSS). Such services are provided to users based on charge-per-query which triggers the motivations for adversaries to steal the trained victim model to reduce the cost of using the service. Therefore, it is important for companies that provide MLaSS to protect their intellectual property (IP) against adversaries. It has been arms race between the attack and defence in a context of the privacy of machine learning models. In this paper, we provide a comprehensive study of recent development in protecting privacy of machine learning models.

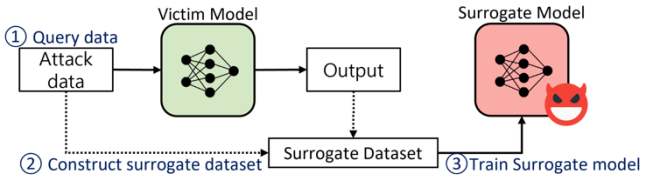
1. Introduction

In accordance with a recent development in big data, the performance of machine learning model has benefitted hugely. As all machine learning models are trained with dataset of various types, collecting a huge database as a training dataset for machine learning models have a tremendous impact in the performance of the trained model. However, it is not always affordable for single user to collect and retrieve a huge database compared to a multi-million companies where they inherit substantial resources. Therefore, in most cases, users must settle with only using the high-performance model from MLaSS and be charged on the number of queries sent to the server. Therefore, in order to minimize the cost of using MLaSS, adversarial users attempt to steal or extraction such models trained with a huge database and have high accuracy. Such attack methods include model extraction attack which aims to extract the functionality of the victim model by training the surrogate model with numerus queries and their output from the model. Another popular attack method is membership inference attack which threatens the privacy of the training dataset by determining if a certain data is included in the training set of the victim model. On the contrary, the companies proving MLaSS seek to protect the intellectual property of their models with various mechanisms and methods. Most common method is to add noise or perturbation to the output of the model to mislead or confuse the adversaries training phase. In this paper, we delve into the arms race of attacks and defences against the privacy of machine learning models.

2. Model Extraction Attack (MEA)

The main purpose of model extraction attack is to mimic the functionality of the victim model with high performance. The adversary is a limited access to the victim model with only access to its output result of the query sent to the victim model. Knockoffnets [1] suggested that by capitalizing the output result of the intended query data, the adversary can train the surrogate model which can be used for free of charge with unlimited queries. Figure 1 surmises the general flowchart of the model extraction attack. Firstly, the adversary must query the data to the victim model to retrieve the output from the victim. Afterwards, by paring the output and the attack query dataset, the adversaries construct the surrogate dataset. Such phase is necessary as the adversaries do not have the access to the training dataset of the victim model. Otherwise, the adversaries can simply train the surrogate model with such training set for very similar functionality as the victim model. Finally, the surrogate model is trained with the surrogate dataset. While such attack method is promising in generating the surrogate model, one serious pitfall in such method was that the attacker needed to query the victim model so many times that the actual cost of extracting the victim model was high.

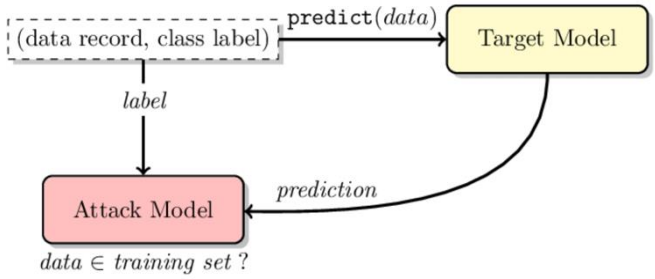
As a result, CloudLeak [2] proposed a new method that tremendously reduce the number of queries needed to successfully steal the victim model. The main idea was to generate adversarial examples of the trained surrogate model in each iteration of the attack to query the victim model with more useful data. Such idea is rooted from the fact that adversarial examples lie on a plain that is very close to the decision boundaries of the machine learning model. Since the functionality and performance of the model rely on the shape of the decision boundaries, if the adversaries can extract more information regarding the boundaries, the attack will be more effective, hence requiring much smaller number of attack queries. By sending queries that will extract the most useful information from the victim, the adversaries can maximize the efficiency of the attack. Such efficient method managed to reduce the number of queries required to complete the attack by a factor of 100 compared to the previous attack. Most attack method is assumed as a black-box attack meaning that the adversaries have no access to the internal information about the victim model and its training dataset.



(Figure 1) The flowchart of model extraction attack

3. Membership Inference Attack

Membership inference attack is a threat to the privacy of the training dataset of the victim model. As the service providers must also protect the training dataset which is a crucial key in generating a high-performance machine learning model, it is adversaries' interest to determine if a certain data is included in the training dataset of the victim. Figure 2 illustrates the general flowchart of such attack as described in a primitive paper [3]. The figure is directly extracted from [3]. The adversaries send the prediction query to the victim model to retract the prediction result. By training the attack model with prediction result and the label, the adversaries attempt to prediction if such data is included in the training set. A typical threat model of such attack is based on the black-box attack which is very similar to model extraction attack.



(Figure 2) The flowchart of membership inference attack [3]

4. Defences against MEA and MIA

The initial defence of model extraction attack rely on hindering the adversaries attempt to construct the surrogate dataset by adding noise to the output of the victim model. Prediction poisoning [5] proposed to *poison* the prediction result to confuse the adversaries in training the surrogate model. If the result of the victim model is *poisoned*, the surrogate model must struggle to predict correctly also. Such idea managed to tremendously decrease the effectiveness of the model extraction attack. However, since the prediction of the model is added with the noise or perturbation, the accuracy of the model was also decreased inevitably.

In terms of protecting the membership inference attack, a very similar attempt was proposed by Memguard [5]. Membguard suggested to add noise to the prediction result of the model to confuse the attack model prediction. One very interesting trick in adding the noise, in order to achieve the best defence effectiveness, the authors used the adversarial examples as a guide. Since adversarial examples result in providing wrong prediction result, the defender can use such samples to add perturbation to the prediction result of the model.

5. Experiments and results

We carried out experiments to check if the architecture of victim model and the surrogate model have effect in the effectiveness of model extraction attack. We used Knockoffnets attack method with random samples strategy which selects the attack query by random sampling.

The metrics used to check the effectiveness of the attack are accuracy and fidelity. Accuracy measures how well the surrogate model predict the data correctly according to the ground truth. Fidelity measures how well the surrogate model's prediction resembles the prediction of the victim model. In other words, to achieve a high fidelity, the surrogate model must falsely predict the data if the victim also falsely predicted the data.

The architecture of the victim model for the first experiment is Wresnet-16 which has 16 layers in the design. The training dataset for the victim was cifar-10 and the victim model achieved both high accuracy and fidelity. The surrogate model is trained with various model architectures and the attack dataset was cifar-100. The result is shown in the brackets of Table 1. For example, 0.90x represent that 90% of the functionality of the victim model was stolen by the surrogate model. We can observe that Wresnet-16 and Wresnet-22 achieved the best attack result with 0.90x and 0.87x in accuracy and fidelity respectively. The worst case was 0.55x and 0.59x with Restnet-18.

The second experiment was carried out with VGG-16 model as the victim. The accuracy of the victim model was slightly lower compared to the first experiment. However, the result of model extraction attack showed a similar effectiveness. Wresnet-22 achieved the best result with 0.88x and 0.82x in accuracy and fidelity respectively. In conclusion we can conclude that that the architecture of the models do not have much impact on the result of model extraction attack.

	Architecture	Dataset	Accuracy	Fidelity
Victim	wres16	cifar-10	95.1 (1x)	100 (1x)
Attacker	wres16	cifar-100	85.63 (0.90x)	87.31 (0.87x)
	wres22	cifar-100	85.78 (0.90x)	87.09 (0.87x)
	res32	cifar-100	79.97 (0.84x)	83.32 (0.83x)
	res18	cifar-100	51.95 (0.55x)	59.05 (0.59x)
	vgg16	cifar-100	68.52 (0.72x)	70.34 (0.70x)

(Table 1) The result of model extraction attack

	Architecture	Dataset	Accuracy	Fidelity
Victim	vgg16	cifar-10	90.6 (1x)	100 (1x)
Attacker	wres16	cifar-100	78.82 (0.87x)	81.06 (0.81x)
	wres22	cifar-100	79.50 (0.88x)	81.69 (0.82x)
	res32	cifar-100	78.13 (0.86x)	80.53 (0.81x)
	res18	cifar-100	65.10 (0.72x)	66.82 (0.67x)
	vgg16	cifar-100	76.47 (0.84x)	79.57 (0.80x)

(Table 2) The result of model extraction attack

6. Discussion

In arms race between the attack and defence of the privacy of the machine learning models, many papers have been published in pursuit of conquering their goals. In future studies the adversaries can utilize the method of generation tools to generate their own attack dataset without having to use any pre-existing dataset for better efficiency. Also, the defence mechanism must strive to maintain the original accuracy of the model as the current defence mechanism only fucuses mainly on protecting the privacy by adding perturbations to the prediction of the model.

7. Conclusion

With emerging MLaSS by many companies, many new types of attack were introduced in recent years. Therefore, protecting the privacy the high-performance model is one of the main interests of the service providers. If the adversaries can extract the privacy of the model by the model extraction attack or the membership inference attack, MLaSS will no longer exist as the surrogate model can substitute the service entirely. Therefore, it is important to develop effective defence mechanisms against such attacks to protect the intellectual property of the service providers. The future defence mechanism should focus on maintaining the original accuracy of the model while being robust to the adversaries attempt to steal the model.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020R1A2B5B03095204) and **by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00230, Development on Autonomous Trust Enhancement Technology of IoT Device and Study on Adaptive IoT Security Open Architecture based on Global Standardization [TrusThingz Project])**. **This work was also supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2021 and by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2020-0-00325, Traceability Assuarance Technology Development for Full Lifecycle Data Safety of Cloud Edge)**

References

[1] OREKONDY, Tribhuvanesh; SCHIELE, Bernt; FRITZ, Mario. Knockoff nets: Stealing functionality of black-box models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. p. 4954-4963.

[2] YU, Honggang, et al. CloudLeak: Large-Scale Deep Learning Models Stealing Through Adversarial Examples. In: *NDSS*. 2020.

[3] SHOKRI, Reza, et al. Membership inference attacks against machine learning models. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017. p. 3-18.

[4] OREKONDY, Tribhuvanesh; SCHIELE, Bernt; FRITZ, Mario. Prediction poisoning: Towards defenses against dnn model stealing attacks. *arXiv preprint arXiv:1906.10908*, 2019.

[5] JIA, Jinyuan, et al. Memguard: Defending against black-box membership inference attacks via adversarial examples. In: *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 2019. p. 259-274.