SCIS 2022 2022 Symposium on
Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics,
Information and Communication Engineers

モデル抽出攻撃の定式化を通じた体系的な整理
Systematization of Model Extraction Attacks via formalizations

小松みさき * 花谷嘉一 *
Misaki Komatsu Yoshikazu Hanatani

キーワード AI セキュリティ、モデル抽出攻撃、攻撃モデル、定式化

あらまし

機械学習の発展に伴い、AI をターゲットとした攻撃についての研究が増加している。その代表的な脅威の一つであるモデル抽出攻撃は、学習済みの機械学習モデルを盗む攻撃である。攻撃者は、攻撃対象のモデルに対し問合せを行い入出力ペアの情報を取得する。取得した情報を用いて、そのモデルと似た振る舞いをする別のモデルを構築することを目的とする。

モデル抽出攻撃が生じると、モデル情報つまり知的財産の流出だけでなく、技術的ノウハウも流出する恐れがある。またモデルに対する問い合わせを API として提供する有償サービスの場合、ビジネスモデルの崩壊も生じる。さらに、盗み出したモデルを用いて別攻撃への転用[1]が可能であることが明らかにされており、さらなる被害が引き起こされる可能性がある。これらの被害を未然に防ぐため、未知の攻撃手法に対しても効果的と言える対策手法を検討していくことが重要と考えられる。

有効な対策手法を検討していくため、既存攻撃手法を体系的に整理し、定式化を図ることは有効な手段の一つと言える。

本稿では、モデル抽出攻撃における攻撃者の能力について注目する。これまでに提案されたモデル抽出攻撃では、攻撃者が攻撃対象のモデルに問い合わせを行う前に与えられる事前情報などの前提条件や攻撃対象のモデルへの問い合わせの作成方法について違いが見られる。提案された攻撃手法の前提条件等は、攻撃間の関係性を整理するために重要と考えられる。

モデル抽出攻撃について体系的な整理や定式化の検を行った先行研究はいくつか存在する[2][3][4]。ところが、我々の知る限り、モデル抽出攻撃の前提条件の１つである、攻撃者に与えられる事前情報を精査して、明示的に定式化を行ったものは存在していない。

本稿では、攻撃者の事前情報に焦点を当てた既存手法のモデル化について検討を行う。具体的には、検討対象とした複数の既存モデル抽出攻撃を、攻撃者と学習済モデルを所有する挑戦者との二者間におけるゲームとして抽象化する。そして、攻撃者・挑戦者それぞれが事前に取得する情報、二者間でやり取りされる情報、攻撃の評価、について表を用いて整理する。この表をベースに、モデル抽出攻撃の攻撃モデルの定式化を検討する。

これにより、今まで十分に検討されてこなかったモデル抽出攻撃を行う攻撃者の事前知識に関する差を表現可能な攻撃モデルの提案をする。

参考文献

[1] N. Papernot, P. McDaniel, I. Goodfellow , S. Jha , Z.B. Celik , A. Swami , “Practical black-box attacks against machine learning", Proceedings of the ASIA CCS, ACM, 2017, pp. 506-519

[2]B.Liu,M.Ding,S.Shaham, .Rahayu,F.Farokhi,Z.Lin. “When machine learning meets privacy: A survey and outlook.”,ACM Comput. Surv., 54(2), March 2021

[3]V.Chandrasekaran,K.Chaudhuri,I.Giacomelli,S.Jha,S.Yan, “Model extraction and active learning.” CoRR, abs/1811.02054, 2018

[4] M.Jagielski, N.Carlini, D.Berth-elot, Alex Kurakin, N.Papernot, “High accuracy and high fidelity extraction of neural networks.”, In29th USNIX, 2020, pp.1345-1362

* （株）東芝 研究開発センター サイバーセキュリティー技術センター,〒212-8582 川崎市幸区小向東芝町１番地, Toshiba Corporation Corporate Research & Development Center,1, Komukai Toshiba-cho, Saiwai-ku, Kawasaki-shi, 212-8582, Japan.