

電子指紋は機械学習の二段階攻撃に使えるか？ Evaluation with Fingerprints on Two-Stage Attacks for Machine Learning Models

岩花一輝 * 三浦堯之 †* 奥田哲矢 † 矢内直人 *
Kazuki Iwahana Takayuki Miura Tetsuya Okuda Naoto Yanai

キーワード 電子指紋, モデル抽出攻撃, 二段階攻撃, 敵対的サンプル, メンバーシップ推定

あらまし

機械学習モデルを攻撃者が抽出し得る脅威として、モデル抽出攻撃 (Model Extraction / Model Stealing) が研究されている。モデル抽出攻撃は、機械学習モデルという秘匿に守るべき知的財産の流出に繋がることに加えて、さらなる攻撃の踏み台として使われる可能性を持つ。具体的な脅威としては、敵対的サンプル (Adversarial Examples) やメンバーシップ推定 (Membership Inference) が挙げられる。敵対的サンプルは、機械学習モデルに対して、決定境界付近のクエリを摂動を加えつつ送り込むことで誤った判定を誘発する攻撃である。メンバーシップ推定は、あるデータサンプルが機械学習モデルの訓練に用いられていたか否かを推定する攻撃で、プライバシー上の脅威となる。これら、モデル抽出攻撃をさらなる攻撃の踏み台として、敵対的サンプルやメンバーシップ推定などの攻撃を行うことを、機械学習モデルに対する二段階攻撃と呼ぶ [1]。

先行研究 [1] では、これら二段階攻撃に有効なモデル抽出攻撃を評価する指標として、忠実度 (Fidelity) を定義している。忠実度は、機械学習モデルへの入力クエリと出力ラベルが、モデル抽出のターゲットモデルとクローンモデルでどれだけ一致するかの割合を示す指標である。先行研究 [1] では、忠実度の高いモデル抽出攻撃が、二段階攻撃に有効であること、特に敵対的サンプルの生成に有効であることを実験的に示している一方で、メンバーシップ推定への応用については言及に留まり、実験的な

結果が示されているとは言えない。そこで本研究では、モデル抽出攻撃からの二段階攻撃への応用として、敵対的サンプルに加え、メンバーシップ推定について、実験的に評価を行った。

さらに本研究では、これら二段階攻撃に有効なモデル抽出攻撃を評価する指標として、新たに電子指紋 (Fingerprint) を導入する。機械学習モデルの電子指紋は、学習済みモデルを特定 (identify) するために提案された指標である [2]。先述の忠実度 [1] との差異として、忠実度ではテストデータという分布に従った (In distribution) 入力に対するモデル出力の top 1 ラベルの一致する割合を計算する一方、電子指紋では分布外 (Out of distribution) の入力、特にモデル決定境界付近の入力に対するモデル出力の top 1 ラベルの一致する割合を計算する。二段階攻撃は、特にモデルの決定境界の情報を狙うことが想定されるため、電子指紋の方が、忠実度に比べて、二段階攻撃の成功確率を評価する指標として、より有効であると考えられる。

本研究では、電子指紋として Cao らによる方式 [2] などを用いて、これらの指標の有効性、具体的には、これらの指標と二段階攻撃の関連性を実験的に評価した。

参考文献

- [1] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1345–1362, 2020.
- [2] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Ipguard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 14–25, 2021.

* 大阪大学大学院情報科学研究科, 大阪府吹田市山田丘 1-5, Graduate School of Information Science and Technology Osaka University, 1-5 Yamadaoka, Suita, Osaka, Japan

† NTT 社会情報研究所, 東京都武蔵野市緑町 3-9-11, NTT Social Informatics Laboratories, 3-9-11, Midori-cho Musashino-shi, Tokyo, Japan