

勾配系の説明付きモデルに対するデータフリーモデル抽出攻撃 Data-Free Model Extraction Attacks against Gradient-based Explainable AI

三浦堯之 *†
Takayuki Miura

芝原俊樹 *
Toshiki Shibahara

矢内直人 †
Naoto Yanai

キーワード データフリーモデル抽出攻撃, 説明可能な AI, Vanilla Gradient, SmoothGrad

あらまし

深層学習をはじめとする機械学習の発展にともなって、訓練済みモデルをその入出力から復元してしまうモデル抽出攻撃 (Model Extraction Attack) の脅威に注目が集まっている。特に、攻撃者がクエリに用いる入力データを収集する代わりに、生成モデルを適応的に学習させて生成するデータフリー方式のモデル抽出攻撃が提案されている [1, 2]。この攻撃は、ラベルなしデータをすら収集することが難しい希少なデータで学習されたモデルに対しても行うことができる従来より汎用性がある攻撃である。

さらに近年、説明可能な AI の出力する勾配情報を用いてデータフリー方式のモデル抽出攻撃を効率的にする手法、MEGEX (Model Extraction attack against Gradient-based EXplainable AI) が提案されている [3]。この攻撃は、Vanilla Gradient という勾配を用いた説明可能な AI に対して、説明として出力される情報を使い、生成モデルを従来よりも効率的に学習させる攻撃である。説明可能な AI は、深層学習モデルなどの機械学習モデルの社会実装に重要であるため、こうした脅威の詳細の解明は重要な課題である。

既存研究 MEGEX[3] では、Vanilla Gradient による説明可能な AI にしか攻撃を検討していなかったが、この手法は、近似的には他の勾配系の説明可能な AI に対して攻撃できると考えられる。本稿では Vanilla Gradient をロバストにした SmoothGrad[4] という勾配系の説明

可能な AI に MEGEX 同様の攻撃を行った際、どの程度の攻撃が成功できるかを検証した。加えて、損失関数をより一般的な形にしたもので攻撃を定式化し直した。

実験の結果、SmoothGrad に対する MEGEX 攻撃は、正確な勾配を用いる Vanilla Gradient に対するものよりは、効率精度は落ちるものの、説明を用いない従来手法 [1] よりは強力な攻撃になっていることが確かめられた。

参考文献

- [1] Jean-Baptiste Truong, Pratyush Maini, Robert J. Walls, and Nicolas Papernot. Data-free model extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.
- [2] Sanjay Kariyappa, Atul Prakash, and Moinuddin K Qureshi. Maze: Data-free model stealing attack using zeroth-order gradient estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13814–13823, June 2021.
- [3] Takayuki Miura, Satoshi Hasegawa, and Toshiki Shibahara. Megex: Data-free model extraction attack against gradient-based explainable ai. *arXiv preprint arXiv:2107.08909*, 2021.
- [4] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

* 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

† 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