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

ブラックボックス型モデル反転攻撃における ユーザ類似性を考慮した生成モデルの検討

Generative Model with User Similarity in Black-Box Model Inversion

井田 天星 * 竹内 廉 * ヴォ ゴック コイ グエン * 西垣 正勝 *
Tensei Ida Ren Takeuchi Vo Ngoc Khoi Nguyen Masakatsu Nishigaki
大木 哲史 *
Tetsushi Ohki

キーワード AI セキュリティ，モデル反転攻撃

あらまし

機械学習 (ML) システムの普及に伴い，顔識別や音声認識など個人の識別タスクにも ML アルゴリズムが利用されるようになった．個人の識別タスクに利用される ML モデルはユーザの顔画像や音声といったセンシティブなデータを用いて訓練される．このことから，訓練済みモデルはパラメータとしてユーザのプライバシーに関わる機密情報を間接的に所有し，プライバシー面での懸念が存在する．近年では，訓練済みモデルに対する攻撃がいくつか存在する．このような ML モデルに対する攻撃がどのような制約で実行可能であり，どの程度有効なのか知ることは，堅牢な ML システムの設計や開発，インシデント発生時の影響調査などにおいて重要である．

訓練済みモデルに対する既知の攻撃として，訓練データに用いられたユーザの機密情報を推測するモデル反転（Model Inversion，MI）攻撃が存在する．MI 攻撃は，攻撃対象モデルが深層ニューラルネットワーク（DNN）のように複雑である場合や，モデルや攻撃対象ユーザの知識が無い場合には，訓練データの推測が難しくなることが経験的に知られている．MI 攻撃の制約に関しては，ホワイトボックス型の攻撃とブラックボックス型の攻撃が検討されている．ホワイトボックスの場合，攻撃対象モデル内部の詳細な情報を利用して高いなりすまし精度の攻撃が可能になる．ブラックボックスの場合，攻撃対象モデルの入出力のみ利用可能で高いなりすまし精度の攻撃は困難になる．高いなりすまし精度の攻撃が可能となるブラックボックス型の MI 攻撃が存在する場合，大きな脅威となる．DNN に対してなりすまし精度が高くなる既存手法として，インターネットから誰でも入手可能な顔画像の公開データと敵対的生成ネットワーク（GAN）を用いる Zhang らの手法が存在する．これは，公開データさえあれば DNN に攻撃できるため，現実的な脅威となるが，攻撃対象モデルの特徴抽出層にアクセスが必要なホワイトボックス型の攻撃であり，顔画像探索時には攻撃対象ユーザとの類似性を考慮しているが，GAN の事前学習時には類似性を考慮していない．これらの指摘事項は，どちらも事前学習プロセスに起因するものである．

そこで本研究では，攻撃対象モデルが DNN で構築された顔画像識別器であるという前提で，公開データを用いて推定データの探索空間を構築するブラックボックス型 MI 攻撃の事前学習プロセスに着目する．探索空間の構築方法が攻撃精度に与える影響を調べるために，Zhang らの攻撃手法における事前学習プロセスに攻撃対象ユーザとの類似度の概念を導入することで，攻撃対象ユーザに特化した探索空間を構築可能な手法を提案する．類似度の概念を導入して構築される探索空間は，類似度を考慮せずに構築した探索空間に比べ，攻撃対象ユーザとの類似度が高いサンプルが多くなる．これにより，広い高次元の探索空間からランダムにサンプルを取り出した際の，類似度の期待値が既存手法より高くなることが期待される．実験では復元した顔画像によるなりすまし精度を示すとともに，提案手法の有効性について考察する．また，なりすまし精度をもとに探索空間を可視化し，既存手法と提案手法の探索空間の差を視覚的に確認する．

* 静岡大学, 静岡県浜松市中区城北 3 丁目 5-1, Shizuoka University, 3-5-1 Jo-hoku, Naka-ku, Hamamatsu City, Shizuoka, Japan