知っておきたいキーワード

Model Inversion Attack

中村和晃†

† 大阪大学

"Model Inversion Attack" by Kazuaki Nakamura (Osaka University, Osaka)
キーワード：深層学習，顔認識，機密情報，ホワイトボックス，ブラックボックス，勾配降下法

まえがき

深層学習を活用してさまざまなAIシステムが作成されるようになり，Web上で公開される例も増えてきた．同時に，AIシステムが有するモデル（深層学習などにより訓練されたモデル）に対する攻撃の危険性を指摘し，その手法を提案する研究が活発化している．Model Inversion Attack（MIA）[1]とは，そのような攻撃法の一つである．

MIAの概要
～顔認識器を例として～

本稿では，AIシステムの具体例として顔画像認識器を取り上げ，その認識モデルに対する攻撃法という観点からMIAの解説を試みる．

顔認識モデルは，抽象化して考えれば，入力顔画像xに対し認識結果$y = f(x)$を返す写像fと捉えられる．出力yは種々の形で定義し得るが，最も単純な形として，例えば人物名を想像して頂ければよい．前提として，fを有する顔認識システムは公開されているものとする．一方，fの訓練に際し訓練データとして用いられた顔画像は，プライバシー上の問題を伴うことから非公開とする．以上の前提のもと，MIAを目論む攻撃者の目標は，任意に定めた人物名$\hat{y}$について，その人物の顔として妥当な画像を得ること，すなわち，攻撃対象モデルfの出力が$\hat{y} = f(\hat{x})$となるような入力$\hat{x}$を推定することである．この攻撃が成功すれば，人物名のみからその人物の顔画像が生成されたことになり，非公開の訓練データが流出した場合と同様の事態を招く（図1）．具体的には，プライバシー漏洩等の問題を生ずるほか，昨今の世情を鑑みれば，DeepFake等のツールと組み合わせてのフェイク画像生成という悪用法も想定し得る．

上記の$\hat{x}$は，人物$\hat{y}$の訓練データと

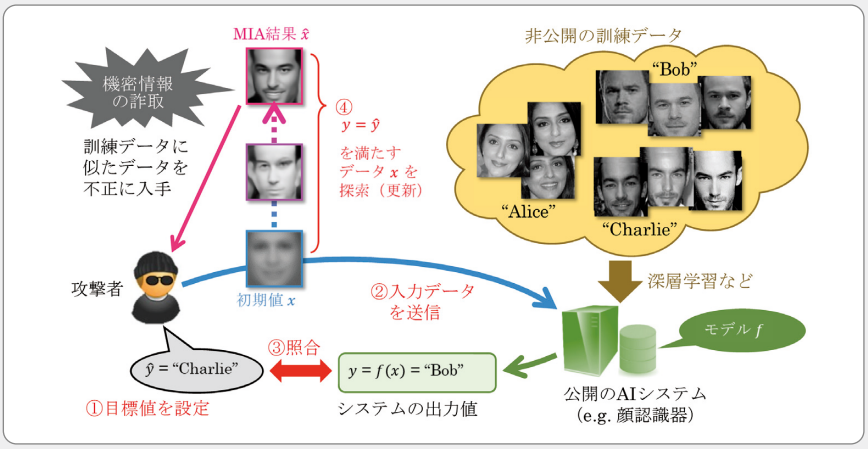

図1　Model Inversion Attackの概要

映像情報メディア学会誌 Vol. 75, No. 3, pp. 384～386（2021）

なった顔画像に類似していることが推測される．このことから，MIAは「攻撃対象モデルからその訓練データを復元する攻撃」と定義されることもある．しかし，実際の訓練データに類似してはいても，それと完全に一致する$\hat{x}$が得られることは稀である．したがって，MIAは現実的には「訓練データに含まれる機密情報を詐取する攻撃」として位置づけた方が妥当と言える．

より一般的には，xを機密（sensitive）部分sと非機密（non-sensitive）部分aに分け，$x=(s,a)$と捉える．顔画像で言えば，aは画像上における顔の位置や向き，サイズがわかる程度の大まかな成分，sは目・眉・口といった細部のテクスチャを決定する詳細な成分，といったイメージである．このとき，認識結果を$y=f(s,a)$として与えるモデルfに対し，$\hat{y}$および$\hat{a}$を所与として$\hat{y}=f(\hat{s},\hat{a})$を満たすような$\hat{s}$を推定し，$\hat{x}=(\hat{s},\hat{a})$として$\hat{y}$に対応する$\hat{x}$を得る．これが，より一般的なMIAの定式化である．ただし，解説の高難度化を避けるため，以降では非機密部分は考慮せず，$x=s$として議論する．

MIAの前提条件 ～攻撃者の知識～

MIAでは，攻撃者の知識に関して以下の前提を置く．

- モデルfの訓練データは未知．これは，機密情報が最初から公開されていれば攻撃自体が不要となるため，当然の前提である．
- 攻撃者はモデルfに対し任意のデータxを任意回数送付でき，対応する認識結果$f(x)$は毎回必ず得られる．
- モデルfの種類や構造，パラメータ等について，これらを既知とする条件をホワイトボックス条件，未知とする条件をブラックボックス条件と呼ぶ（図2）．ホワイトボックス条件下では，認識結果$f(x)$に加え，勾配$\partial f(x)/\partial x$などの追加情報が利用可能である．ブラックボックス条件では追加情報は一切得られないため，より現実的かつ挑戦的な条件と言える．

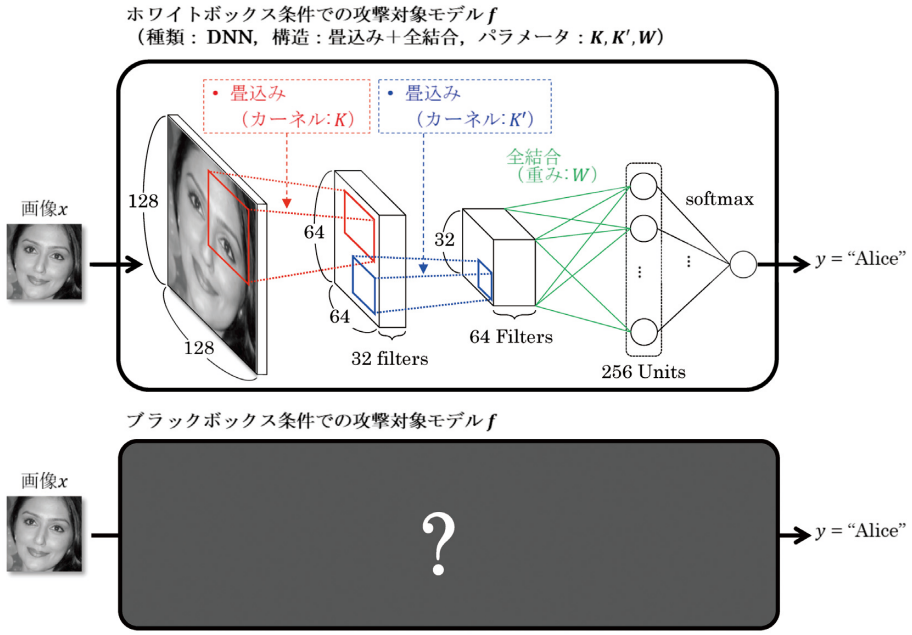

図2　ホワイトボックス条件とブラックボックス条件

MIAの実行手法 ～概論～

モデルfは一般に多対一の写像である．実際，顔認識器は，顔の位置や向き，表情等によらず，同一人物の顔画像はすべて同じクラスに分類する．このことを踏まえ，クラス$\hat{y}$に属する顔画像の分布を$p(x|\hat{y})$とすると，MIAは$p(x|\hat{y})$を最大化するx，すなわち

$$\hat{x}=\arg\max_x p(x|\hat{y})$$

を求める問題に帰着される．その具体的な実行手法は，攻撃対象モデルfの種類や構造，あるいはホワイトボックス条件かブラックボックス条件か，などに応じてさまざまである．

MIAの実行手法
～具体的な事例～

具体的な事例として，fがDeep Neural Network（DNN）による顔認識モデルである場合のMIA実行手法を紹介する．ただし，条件としてホワイトボックス条件を仮定する．

DNNによる画像認識モデルでは一般に，入力画像xが各クラスに属する尤度をsoftmax関数等により擬似的に計算する．そこで以降では，xのi番目の人物らしさを$y_i \in [0,1]$とし，出力yを$y = (y_1\ y_2\ \cdots)^{\mathrm{T}} = f(x)$と定義し直す．また，$\hat{y} = (\hat{y}_1\ \hat{y}_2\ \cdots)^{\mathrm{T}}$は特定人物を表すone-hotベクトルとして与える．例えば，k番目の人物の顔画像生成を目論む場合，$\hat{y}_k = 1$とし，他の$(i \neq k$なる任意の$)\ i$に対し$\hat{y}_i = 0$と定める．以上の前提の下，まずベイズの定理に従って$\mathrm{p}(x\,|\,\hat{y}) = \mathrm{p}(x)\mathrm{p}(\hat{y}\,|\,x)/\mathrm{p}(\hat{y})$とおく．$\hat{y}$は所与であるため$\mathrm{p}(\hat{y})$は定数である．また，事前分布$\mathrm{p}(x)$も，これを一様と仮定して定数とみなす．すると，$\mathrm{p}(x\,|\,\hat{y})$の最大化は$\mathrm{p}(\hat{y}\,|\,x)$の最大化に等しくなる．ここで，$\mathrm{p}(\hat{y}\,|\,x)$は，画像$x$の認識結果が$k$番目のクラスとなる尤度，すなわち$y_k$に他ならない．したがって，$\mathrm{p}(x\,|\,\hat{y})$の最大化は$y_k$の最大化と等価であり，さらに言えば，対数関数の単調性から$-\log y_k$の最小化と等価である．一方で，上記の$\hat{y}$の下では

$$-\log y_k = -\sum_i \hat{y}_i \log y_i$$

であり，これはyと$\hat{y}$の間の交差エントロピーに一致する．したがって，深層学習において一般的に用いられる交差エントロピー損失関数をLとおくと，$\hat{x}$は

$$\begin{aligned}\hat{x} &= \arg\max_x \mathrm{p}(x\,|\,\hat{y}) \\ &= \arg\min_x L(y\,|\,\hat{y}) \\ &= \arg\min_x L(f(x),\hat{y})\end{aligned}$$

として求められる．この最小化問題は勾配降下法により容易に解ける（図3）．

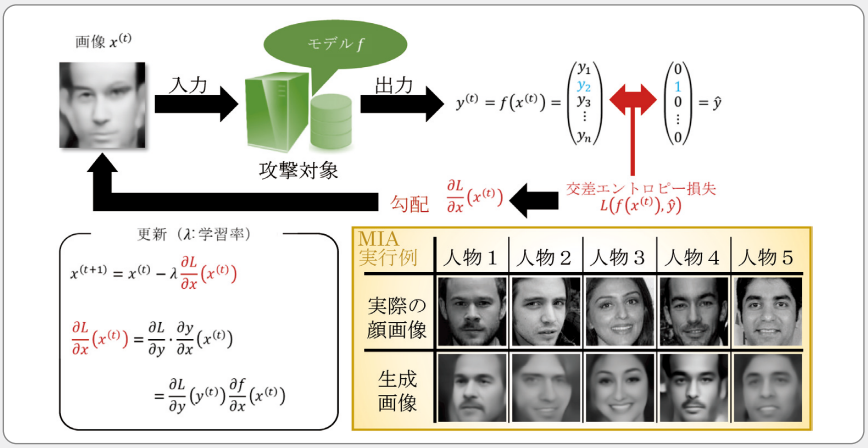

図3　顔認識モデルに対するMIAの具体的な実行手法

他種のモデルを対象としたMIA

MIAの研究では，そのわかりやすさから顔認識モデルが対象となることが多いが（文献[2]など），最後に一つ，別の事例として，推薦システムを標的としたMIA[3]を紹介する．ユーザの過去の動画視聴履歴に基づいて当該ユーザが好みそうな未視聴動画を推薦する動画推薦システムを想像されたい．紙面の都合上，単純化して述べるが，これまでの動画視聴履歴をx，推薦モデルをf，i番目の被推薦動画に対するユーザの評価値の推定値をy_iとすると，推薦は$y = (y_1\ y_2\ \cdots)^{\mathrm{T}} = f(x)$を計算する処理（評価値の高い動画を推薦する）として定式化できる．このとき，ある動画が推薦された事実を何らかの理由で攻撃者に知られれば，それに基づいて適当に$\hat{y}$を定めることで，攻撃者はMIAにより当該ユーザの視聴履歴を特定できる可能性が生ずる．このように，顔認識モデルのみならず$y = f(x)$で表現される任意のモデルがMIAの対象となり得る．

むすび

MIAの考え方が最初に提唱されたのは2014年のことであり，それからの数年で攻撃法の研究は一定の進歩を見せている．この結果，MIAの危険性が必ずしも絵空事ではなくなりつつある一方で，防御法に関する研究はまだあまり見られず，今後の研究の進展に期待したい．また，攻撃法については，「攻撃」という枠を離れ，何らかの別タスクへの応用といった展開が起きるかに注目したい．

（2021年1月28日受付）

参　考　文　献

1) M. Fredrikson, S. Jha and T. Ristenpart: "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures", Proc. 22nd ACM SIGSAC Conf. On Computer and Communications Security, pp.1322-1333 (2015)
2) Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li and D. Song: "The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks", Proc. 2020 IEEE Conf. On Computer Vision and Pattern Recognition, pp.253-261 (2020)
3) S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto and G. Hanaoka: "Model Inversion Attacks for Prediction Systems: Without Knowledge of Non-Sensitive Attributes", Proc. 15th Annual Conf. On Privacy, Security and Trust, pp.115-124 (2017)

中村（なかむら）　和晃（かずあき）
2005年，京都大学工学部情報学科卒業．2010年，同大学大学院情報学研究科博士後期課程研究指導認定退学．2010年，同大学大学院法学研究科助手．現在，大阪大学大学院工学研究科助教．画像・映像の認識・理解・生成，動作映像解析，機械学習モデルに対する攻撃とその防御法に関する研究に従事．博士（情報学）．