AIセキュリティ
―学習データを暴く攻撃とその対策―

東京理科大学 工学部 情報工学科 准教授 中村（なかむら） 和晃（かずあき）

1. はじめに

2010年代の半ばごろから，テレビやネットニュース等で「人工知能（artificial intelligence; AI）」という用語を目にする機会が増えていることと思います．AIが社会に与えるインパクトは大きく，医療診断の自動化による医療従事者の負担軽減や，マーケティングデータの解析による企業経営の効率化など，様々なメリットが期待できます．一方で，AIの登場と普及に伴って発生した社会問題もまた存在し，AIを悪用したフェイクニュースの流布などは，その典型例と言えます．このような「AIを悪用した攻撃」は今や一般にも広く知られているところですが，実は，AIには，「攻撃をする側」だけでなく「攻撃を受ける側」となる可能性も存在します．AIを標的としたサイバー攻撃として具体的にどのようなリスクが考えられるか，どのようにすればそれを防ぐことが可能か，といったテーマに関する研究が活発化しています．こうした研究は，AIサービスを開発する側の個人や組織にとって，サービスの安心・安全な利用を実現する上で欠かせない視点と言えます．

本稿では，AIを標的としたサイバー攻撃の代表例として「モデル反転攻撃（model inversion attack; MIA）」を取り上げ，その内容を筆者らのグループの研究成果も交えて紹介致します．なお，以降では，攻撃の対象として主に顔認識AIを取り上げます．

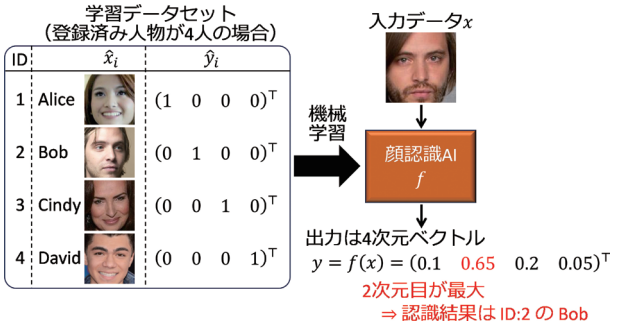
2. 顔認識AIの仕組みとその学習

MIAを紹介する前に，まず「AIとは何か？」について論じたいと思います．数学的な観点では，AIは一種の関数と言えます．すなわち，入力データ x から出力データ $y=f(x)$ を導く関数 f がAIです．例えば，株価予測AIであれば，x は過去数日分の株価，y は明日の株価を表します．将棋AIであれば，x は現在の盤面の状態，y は次の一手として最適な手，となります．そして，顔認識AIでは，x は顔画像，y は人物名です（正確にはもう少し詳細な情報を含みますが，この点については後述致します）．

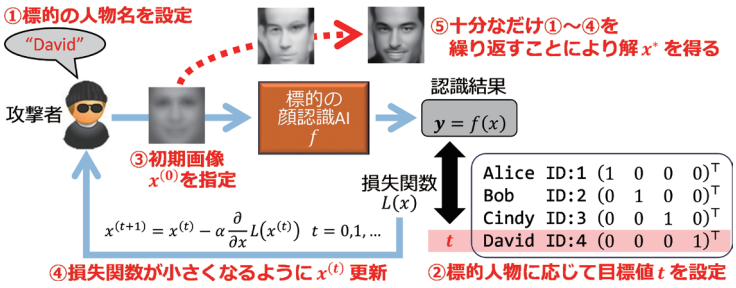
近年のAI開発では，関数 f の「構造」は人間が事前に設計しておき，その「パラメータ」を機械学習により定める，という手順を取ることが一般的です．これは次のようなイメージで捉えると分かりやすいと思います．まず，「構造を設計する」とは，一次関数や二次関数などの種別を指定することを意味します．ここで，例えば二次関数を指定したとすると，f は $y=f(x)=ax^2+bx+c$ の形で表されることになります．この式における定数 a,b,c がパラメータであり，これらの値として最適なものを機械学習により自動決定します．

以上の枠組みにおいて，機械学習を実行するためには，入力が $\hat{x}_i$ のとき出力は $\hat{y}_i$ になる，といった対応関係が既知であるような入出力データの組が必要であり，このような組の集合 $D=\{(\hat{x}_i,\hat{y}_i)\,|\,i=1,\ \ldots\ ,N\}$ を「学習データセット」と呼びます．顔認識AIで言えば，学習データセットとは，人物名 $\hat{y}_i$ が既に分かっている顔画像 $\hat{x}_i$ の集合のことになります．機械学習は，個々の学習データ $(\hat{x}_i,\hat{y}_i)$ が再現されるように f のパラメータを定めるプロセスに相当します．これは，上記の二次関数の例で言えば，関数 $y=f(x)$ の描く曲線が点 $(\hat{x}_1,\hat{y}_1),(\hat{x}_2,\hat{y}_2),\ \ldots\ ,(\hat{x}_N,\hat{y}_N)$ を可能な限り正確に通るようにパラメータ a,b,c を定める問題と言えます．

ここで，後の議論を円滑化するために，顔認識AIにおける出力データ y の扱いについて，もう少し詳しく述べておきます．通常，顔認識AIでは，事前に登録済みの人物のみが認識の対象となります．それら登録済みの人物が n 人存在するものとすると，各人物には1から n までの整数値がそれぞれ割り当てられます．その上で，顔認識AIの出力データ y を n 次元ベクトル $y=(y_1\ \cdots\ y_n)^\top$ として定義します．ここで，y における k 次元目の値 y_k は，入力顔画像 x の「k 番目の人物らしさ」を表す0以上1以下の実数値（これを「尤度」と言います）となり，その値が最も高い番号の人物が認識結果として出力されます．なお，学習データ $(\hat{x}_i,\hat{y}_i)$ における $\hat{y}_i$ は（この添字 i は「i 番目の人物」ではなく「i 番目の学習データ」を意味することにご注意下さい），$\hat{x}_i$ が j 番目の人物の顔画像であるとき，j 次元



【図 1】顔認識 AI の仕組み



【図 2】モデル反転攻撃（MIA）の概要

目のみが 1 でそれ以外の次元は全て 0 の n 次元ベクトルで与えます．以上の内容をまとめたものを【図 1】に示します．

3. モデル反転攻撃：学習データを暴く攻撃

3. 1. 理論的定式化

モデル反転攻撃（MIA）とは，一言で言えば「AI の学習データを暴く攻撃」です．一般に AI では，クラウドサービスなどの形で AI そのものを公開することはあっても，学習データを公開することは基本的にありません．これは，学習データにはプライバシーにかかわる情報が含まれ得るためです．例えば，顔認識 AI の学習データは個人の顔画像そのものであり，個人情報の典型例と言って良いでしょう．このため，通常は非公開とされますが，これを暴く攻撃が MIA です．以下，その具体的な手順を紹介します．

前節で述べたように，j 番目の人物に関する学習データ（顔画像）に対しては，j 次元目のみが 1 でそれ以外の次元は全て 0 のベクトルが割り当てられています．従って，そのようなベクトルを t_j とおくことにすると，$f(x)=t_j$ を満たすような x を求めることができれば，j 番目の人物の学習データを暴けたことになります．これは，前述の二次関数の例で言えば，方程式 $y=f(x)=t_j$ すなわち $ax^2+bx+c=t_j$ を解くことと等価になります．ただし，実際の顔認識 AI では，f は二次関数とは比較にならないほど複雑な関数となるため，厳密に $f(x)=t_j$ を満たす解が存在しないこともよくあります．一方で，$f(x)$ と t_j がある程度近くなるような近似解程度であれば，今度は無数に存在し得ます．特に，画像 x はピクセル数と同数の次元数を持つ高次元データ（128×128 画素の画像でも 16384 次元）であるため，顔に見えないような不自然な見た目の画像でも $f(x)=t_j$ を近似的に満たすことは珍しくありません．このため，実際の MIA は，

$$L(x)=\|f(x)-t_j\|^2+R(x) \qquad (1)$$

のような損失関数 L を最小化する x を求める処理として定式化されます．ここで，$R(x)$ は，x の見た目が顔として自然であることを保証するための正則化項です（x の見た目が不自然なほど $R(x)$ の値が大きくなるように設計する）．正則化項 $R(x)$ の与え方に応じて様々な MIA 手法が提案されていますが[1-3]，ここでは，その詳細に深入りすることは避けたいと思います．

上述の最小化問題（1）は，次の解法により解くことができます．まず，初期値として $x=x^{(0)}$ を適当に定め，そのときの損失関数の値 $L(x^{(0)})$ を求めます．次に，この値が小さくなるように $x^{(0)}$ を少しだけ変化させ，$x^{(1)}$ とします．このときの変化量を $\Delta x^{(0)}$ とすると，$x^{(1)}$ は $x^{(1)}=x^{(0)}+\alpha\Delta x^{(0)}$ として計算されます．ここで，α は学習率と呼ばれる正の定数です．この処理を十分なだけ反復することにより，すなわち，$t=0,1,2,\ldots$ に対して順次

$$x^{(t+1)}=x^{(t)}+\alpha\Delta x^{(t)} \qquad (2)$$

として $x^{(t)}$ を更新することにより，最終的に解 $x^*=\underset{x}{\operatorname{argmin}}\ L(x)$ を得ます．以上の処理において，変化量 $\Delta x^{(t)}$ は，$x=x^{(t)}$ における L の偏微分係数（勾配ベクトル）に基づき $\Delta x^{(t)}=-\frac{\partial}{\partial x}L(x^{(t)})$ で与えます．ここまでの内容を【図 2】にまとめます．

3. 2. 顔認識 AI を標的とした攻撃結果の例

MIA の脅威度を具体的に確かめた例として，筆者らのグループが開発した MIA 手法[3] を顔認識 AI に対して実際に適用した結果を紹介致します．

【図 3】は，9000 人以上の人物に関する顔画像を収録した画像データセットである VGGFace2 からランダムに 2141 人分を選出して顔認識 AI を学習し（$n=2141$），それを標的として実験的に MIA を実行した結果を示したものです．MIA により求めた顔画像は，ややピントがぼけたような見た目となってはいるものの，実際の個人の顔画像（顔認識 AI の学習データ）に極めて近い特徴を有していることが分かります．MIA 結果の顔画像が実際の個人の特徴をどの程度反映しているのかを定量的に評価するために，標的の顔認識

AIとは別の顔認識AIを構築し，その別AIにMIA結果の画像を入力した場合の顔認識精度を検証する実験も行いましたが，その結果は約79%となりました．

以上の実験結果は，顔認識AIに対するMIAにより個人の顔が暴かれるリスクが相応に高いことを示しています．ここでは顔認識AIを標的として実験を行いましたが，同様の攻撃は，顔認識AIに限らず，同じ仕組みで動作する画像認識AI一般に対して成立するものです．従って，AIによる画像認識サービスの開発・運用に際しては，何らかのMIA対策を組み込むことが求められると言わざるを得ません．

4. モデル反転攻撃の防御

4. 1. 防御が難しい理由

前節の例をはじめとするMIA研究の結果を受けて，MIAに対処するための防御技術の研究も進みつつありますが[4]，攻撃法に関する研究ほど活発であるとは言い難いようです．これは，MIAを防ぐことが本質的に容易ではないためと推測されます．

前述のように，MIAは式（1）の損失関数を最小化する問題として定式化されます．従って，j番目の人物の顔が暴かれることを防ぐためには，その人物の実際の顔画像がxとしてfに入力された際にt_jとは大きく異なる出力が得られるように（$||f(x)-t_j||^2$が大きくなるように）顔認識AIを予め設計しておけば良い，ということになります．しかし，実際にこのような策を講じると，出力の尤度ベクトル$f(x)$のj次元目の値（j番目の人物に関する尤度）が極端に小さくなり，j番目の人物の顔画像を正しく認識できる確率が著しく低下します．つまり，認識性能が犠牲になってしまいます．このような理由から，認識性能を維持したままMIAを防ぐことは本質的に困難な課題となります．

4. 2. ダミー認識器を活用した防御手法

上述の課題に対し，筆者らのグループでは，ダミー認識器を用いることにより，認識性能を極力維持したままMIAを防御する手法[5]を検討しています．

先述の通り，顔認識AIは，入力の顔画像xに対しn次元の尤度ベクトル$y=f(x)$を出力する関数です．このfをMIAから守るために，別の関数f'を用意します．このf'がダミー認識器であり，fと同様，画像xを入力としてn次元尤度ベクトル$f'(x)$を出力する関数ですが，どのような画像xに対して出力ベクトルの各次元が大きな値を取るかは，fとf'で全く異なるように設計します．その上で，fとf'を並列に連結した複合認識AIを作成し，これをfの代わりに公開します**【図4】**．この複合認識AIをgとおくと，gの出力は

$$g(x)=\lambda f(x)+(1-\lambda)f'(x) \qquad (3)$$

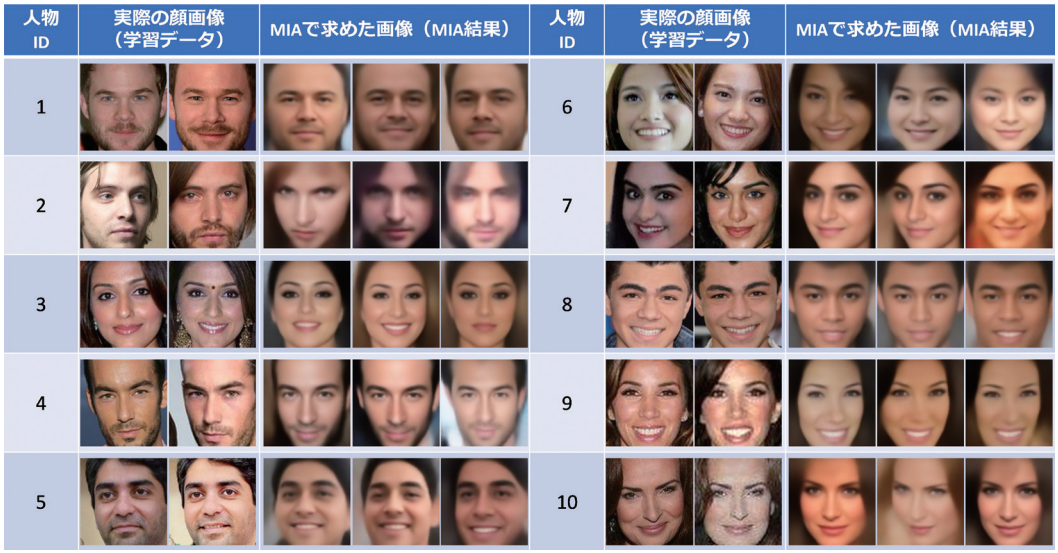
で与えられるようにします．ここで，λは，fとf'の複合比率を定める正の定数であり，$\lambda<1-\lambda$となるよう，0以上0.5未満の値を設定します．重要な点として，一般のAI利用者にはgの内部構造は確認できません．これは攻撃者も同様ですので，この場合のMIAは，式（1）の代わりに

$$L(x)=||g(x)-t_j||^2+R(x) \qquad (4)$$

を最小化する問題に帰着されることになります．

ここで，式（4）の第1項は

$$\begin{aligned}||g(x)-t_j||^2&=||\lambda f(x)+(1-\lambda)f'(x)-t_j||^2\\&=||\lambda(f(x)-t_j)+(1-\lambda)(f'(x)-t_j)||^2\end{aligned}$$



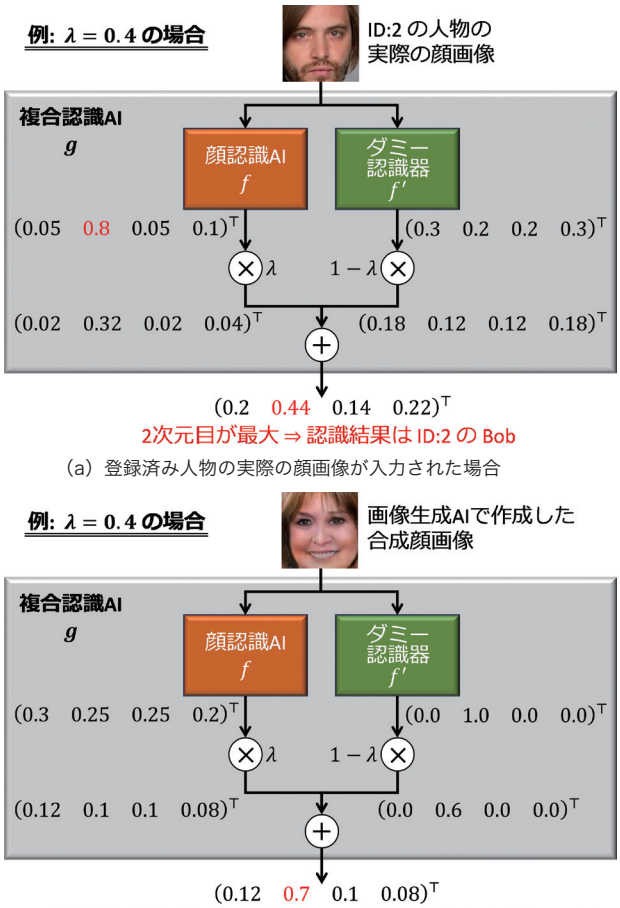
【図3】顔認識AIに対するMIA結果の例

のように変形できますが，定数 λ を $\lambda < 1-\lambda$ となるように設定していることから，式（4）を最小化する x は，方程式 $f'(x)=t_j$ の近似解として与えられることになります．このことを見越して，ダミー認識器 f' の学習に際しては，画像生成 AI を用いて作成した合成顔画像を学習データとして利用します．これにより，$f'(x)=t_j$ の近似解 x^* は合成顔画像に近い画像となり，結果として，MIA の結果を合成顔画像へと誘導することが可能となります．一方，登録済み人物（j 番目の人物）の実際の顔画像が x として g に入力された場合には，$f'(x)$ の値は全体的に小さくなりますが，$f(x)$ の方では j 次元目の値が 1 に近い値を取りますので，$g(x)$ の j 次元目は λ 前後の値を維持します．この結果，g の認識 AI としての性能はさほど低下しないことが期待できます．

【図 5】は，3．2．節で紹介した実験内容と同じ条件の下で上記の防御手法を実際に適用し，その場合における複合認識 AI の認識精度とそれに対する MIA の成功率を調査したものです．なお，この調査では，参考までに，0 以上 1 以下の全ての λ について認識精度および MIA 成功率を評価しました．【図 5】を見ると，$\lambda \geq 0.25$ の範囲であれば，複合認識 AI の顔認識精度はほとんど低下しないことが確認できます．一方で，MIA 成功率は $\lambda = 0.7$ 付近から減少傾向にあり，$0.25 \leq \lambda \leq 0.3$ の範囲では 30% 未満に抑えられています．元の顔認識 AI に対する MIA 成功率が約 79% であったことを考慮すると，認識性能を維持したまま MIA 成功率を大幅に減少させることに成功しており，ダミー認識器を活用した防御手法の有効性が示唆されます．

5. おわりに

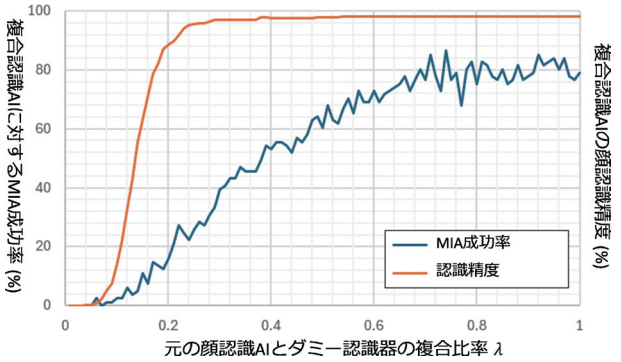
本稿では，AI を標的とするサイバー攻撃の一例として，AI の学習データを暴く攻撃である「モデル反転攻撃（MIA）」を取り上げ，その具体的な攻撃技術と防御法を紹介致しました．冒頭でも述べたように，AI が広く社会に普及しつつある一方で，AI が攻撃を受ける側になるケースもあることはまだあまり知られていないように思われます．こうした問題を AI 開発者のみならず一般の方々にも広く知って頂くための一助として本稿が役立てば，筆者としては幸甚です．



（a）登録済み人物の実際の顔画像が入力された場合

（b）合成顔画像が入力された場合

【図 4】ダミー認識器の併用による MIA 防御



【図 5】提案する防御手法の有効性評価

［参考文献］

1）M. Fredrikson *et al.*, “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures”, Proc. ACM SIGSAC Conf. on Computer and Communications Security, pp. 1322–1333, 2015.

2）Y. Zhang *et al.*, “The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks”, Proc. 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 253–261, 2020.

3）M. Khosravy and K. Nakamura *et al.*, “Model Inversion Attack by Integration of Deep Generative Models: Privacy-Sensitive Face Generation from a Face Recognition System”, IEEE Trans. on Information Forensics and Security, Vol. 17, pp. 357–372, 2022.

4）T. Wang *et al.*, “Improving Robustness to Model Inversion Attacks via Mutual Information Regularization”, Proc. AAAI Conf. on Artificial Intelligence, pp. 11666–11673, 2021.

5）小辻，中村，“ダミー認識器の併用によるモデル反転攻撃の性能低減”，電子情報通信学会 2024 年総合大会講演論文集，D–21–01，2024.