2ZA-08

BlackBox において生体認証を回避する Adversarial Example の検討

Vo Ngoc Khoi Nguyen† 寺田 崇倫 †† 西垣 正勝 †† 大木 哲史 ††
† 静岡大学情報学部 †† 静岡大学大学院総合科学技術研究科

1 はじめに

近年では機械学習アルゴリズムに対する脆弱性を利用する Adversarial Example（以下，A.E.）を用いた攻撃手法の検討が数多く行われている．顔認証においても A.E. を用いた攻撃の危険性が存在するが，既存研究では，攻撃対象の認証モデルが既知（以下，ホワイトボックス）という前提で，A.E. による攻撃を行う限定的な検討に留まっている．本研究では，2 枚の顔画像間の類似度を出力する顔認証において，攻撃対象の認証モデルが未知（以下，ブラックボックス），つまり入力に対して認証の成否のみが得られるという現実的な仮定のもとで，生体認証を回避する A.E. を作成する手法のコンセプトを示す．

2 関連研究

2.1 Adversarial Examples

A.E. とは，入力に対して、人間の目に判別できない程度の細微な摂動を与えることで機械学習識別器を誤認識させる手法である．識別器の識別関数を f，入力を $\boldsymbol{x}$，$\boldsymbol{x}$ に対応するラベルを y とする時に $f(\boldsymbol{x}) = y$ が得られる．A.E. を利用して，入力 $\boldsymbol{x}$ に微細なノイズ $\boldsymbol{\eta}$ を付与することで，識別器に入力 $\boldsymbol{x}+\boldsymbol{\eta}$ を別のラベル $y' \neq y$ として誤認識させる．A.E. は次式で表される．

$$f(\boldsymbol{x}+\boldsymbol{\eta}) = y' \tag{1}$$

A.E. の生成手法の一つである Fast Gradient Sign Method(以下，FGSM) におけるノイズ $\boldsymbol{\eta}$ は，モデルのパラメータ をθ，固定値を ϵ として次式で表される．

$$\boldsymbol{\eta} = \epsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right) \tag{2}$$

本研究では，式 (2) で表される FGSM を用いて，A.E. を生成する．

2.2 深層学習を用いた顔画像認証

深層学習を用いた顔認証は，学習データに含まれる人物であるかを多クラス分類により判定する手法（以下，多クラス分類方式）と複数人の顔画像から学習した深層学習モデルを特徴抽出器として使用し，モデルから抽出された特徴ベクトルに基づいて，入力された 2 枚の画像が同一人物の画像であるかを判定する手法（特徴抽出方式）に分類される．深層学習を用いた顔認証モデルも A.E. による攻撃に対して脆弱であると考えられるが，より現実的な環境下においては顔認証機器から A.E. 作成に必要なニューラルネットワーク（以下，NN）の内部情報を知ることは困難であることが予想される．そこで本研究では，特徴抽出方式の顔認証モデルにおいて，2 枚の顔画像特徴ベクトル間の距離のみを観測可能，という条件の下で A.E. を生成する手法を提案する．

2.3 Jacobian-based Data Augmentation Method

Papernot らはブラックボックスモデルに対して有効な A.E. 作成手法を提案した [2]．対象の NN の内部情報が未知であるという課題に対し，対象モデルの出力ラベルのみをサブモデルの正解ラベルとして利用し，さらにサブモデルをトレーニングするためのデータセットを水増ししつつ再トレーニングを繰り返すことで，対象モデルとほぼ一致する判定境界を持つサブモデルを作成可能としている．このサブモデルをもとに作成された A.E. は対象モデルに効果があると報告されている．

3 サブモデルによる A.E. 作成手法の提案

本研究では，特徴抽出型の顔認証を回避可能な A.E. を作成することを目的とする．特徴抽出型の顔認証においては，2.3 節と異なり，入力顔画像の正解のラベルではなく，入力された 2 枚の顔画像間の距離のみを利用する点を考慮しなくてはならない．ここでの目的は，対象顔認証モデル T とほぼ一致する判定境界を持つサブモデル S を作成し，作成したサブモデル S により A.E. を作成することである．提案手法の概要を以下の 3 つのステップにまとめる．

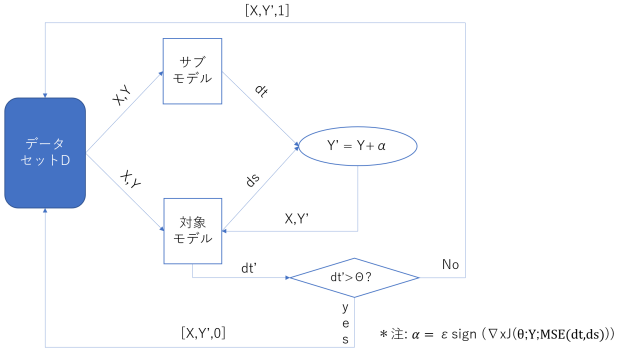

図 1　サブモデルのトレーニングプロセス

1．判定しきい値 θ の推定：　サブモデルの学習にあたって，対象モデルの T が本人画像同士を正しく判定可能な判定しきい値 θ を推定しておく必要がある．ここでは，対象モデルに対して本人間照合を一定回数繰り返して得られるスコア分布から，しきい値 θ を決定する．本検討においては，$FRR < 0.05$ となる点をしきい値 θ と定めることとした．

2．サブモデルの学習：　サブモデルの学習は図 1 のように行われる．S と T に 2 枚の画像を入力と，それらの出力の間隔を縮小させるために平均二乗誤差を求める．求まった誤差を用いて FGSM で新たに画像を作成し，データセットに追加する．またこのデータセットを学習データセットとして利用する．サブモデル作成のアルゴリズムの概要を **Algorithm 1** に示す．

3．A.E. の作成：　S を用いて A.E. を作成する．A.E. の転送可能性を利用して，S によって誤認証される A.E. も T によって誤認証されると予想する．

Algorithm 1: サブモデル作成

Input: max ρ ：繰り返し回数
fs, ft : S モデル，T モデルの出力
D：データセット
λ：ノイズの変化率;
θ：同一人物の最大判定値;

```
for i in max ρ do
    train(S,D);
    for data in D do
        ds=fs(data[0],data[1]);
        dt=ft(data[0],data[1]);
        data[1]=data[1]+λsign(Δ
         J(data[1],dt));
        dt=ft(data[0],data[1]);
        if dt ≤ θ then
            D.append([data[0],data[1],1]);
        else
            D.append([data[0],data[1],0]);
        end
    end
end
```

Result: サブモデル S

4　まとめと今後の課題

本稿では，2 枚の顔画像間の類似度を出力するブラックボックス型顔認証に対する A.E. 作成手法のコンセプトを示した．今後は既存のブラックボックス型生体認証方式を対象に本手法の有効性を検証していく予定である．

参考文献

[1] Alexey Kurakin, Ian Goodfellow, Samy Bengio, ”Adversarial Machine Learning at Scale”, arXiv:1611.01236, 2016.

[2] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, A.Berkay Celik, Ananthram Swami, ”Practical Black-Box Attacks against Machine Learning”, arXiv:1602.02697, 2016.