A study of the possibility of impersonation Black-box Adversarial Attack on face recognition

一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
ISEC2021-22,SITE2021-16,BioX2021-23,
HWS2021-22,ICSS2021-27,EMM2021-27(2021-07)

Black-box Adversarial Attackによる顔認証へのなりすまし可能性に関する検討

ヴォゴックコイ　グエン†　寺田　崇倫†　西垣　正勝†　大木　哲史†

† 静岡大学　〒432–8011 浜松市中区城北 3 丁目 5–1
E-mail: †{nguyen,terada}@sec.inf.shizuoka.ac.jp, ††{nisigaki,ohki}@inf.shizuoka.ac.jp

あらまし　本稿では，顔認証に有効な Black-box Adversarial Examples（以下 A.E.）攻撃を提案する．これまでの顔認証に対する Black-box A.E. では，攻撃成功確率が低い，攻撃対象が限定される，計算量が大きいといった，実用的なシナリオ下での攻撃に関する複数の問題があった．そこで，本研究では，Huang らの低次元埋め込み空間での Black-boxA.E. 作成手法に対して，顔認証に適した攻撃用サブモデルを作成することで，顔認証システムに有効な Black-box 攻撃を行う手法を提案する．評価実験では，本手法および公開データセットを用いて，顔認証システムに登録された任意および特定の人物に対して攻撃を行い，Black-box Adversarial Attack による攻撃可能性を示す．

キーワード　敵対的サンプル，顔認証，Black-box 攻撃

A study of the possibility of impersonation Black-box Adversarial Attack on face recognition

Vo N. K. NGUYEN†, Takamichi TERADA†, Masakatsu NISHIGAKI†, and Tetsushi OHKI†

† Shizuoka University　3-5-1 Johoku, Naka-ku, Hamamatsu City, Shizuoka, 432–8011 Japan
E-mail: †{nguyen,terada}@sec.inf.shizuoka.ac.jp, ††{nisigaki,ohki}@inf.shizuoka.ac.jp

Abstract　In this paper, we propose a Black-box Adversarial Examples (A.E.) attack that is effective for face recognition. In the past, Black-box A.E. for face recognition had multiple problems such as low probability of successful attack, un-targeted attack only, or large computational complexity which lead to impracticality in many real word scenarios. Therefore, in this research, by creating an attack substitute model suitable for face recognition based on the A.E. creation method of Huang et al., we propose a more effective method of attacking face recognition system using Black-box A.E.. For evaluation, this method and the public dataset are used to attack arbitrary and specific people registered in the face recognition system which points out the possibility of a Black-box Adversarial Attack against face recognition system.

Key words　Adversarial Examples , Face Recognition, Black-box attack

1. はじめに

生体認証技術は学術研究のみならず，携帯端末での利用をはじめとして日常生活においても多く活用されている．特に，顔認証システムはここ数年の機械学習技術の発展に伴い飛躍的な進歩を遂げ，人間と遜色ない段階まで識別精度が向上するといった成果をあげている [1]．一方，近年では機械学習アルゴリズムに対する攻撃として，Adversarial Examples（以下 A.E.）を用いた手法が多く提案されている [2]．A.E. とは，入力に対して，人間の目に知覚できない程度の微小な摂動を加えることで機械学習識別器を誤認識させる手法である．

A.E. による誤認識の脅威は，機械学習アルゴリズムに基づく顔認証システムにおいても同様に存在すると考えられる．そこで，より安全な顔認証システムの実現を目的に，顔認証システムを対象とした A.E. による攻撃手法やその対策を検討することが必要である．ここで，顔認証システムに対する攻撃は，攻撃者が入手可能な情報の種類によって White-box 攻撃と Black-box 攻撃に分類される．ネットワークの重みパラメータをはじめとするモデルに関する知識が攻撃者に既知である場合を White-box 攻撃，これらが未知である場合を Black-box 攻撃と分類することが多く，本研究もこの定義に従う．定義をふまえれば，Black-box 攻撃は，MLaaS として生体認証が利用される場合など，より現実的な状況を仮定したモデルと考えることができる．本研究では生体認証システムに対する Black-box 攻

撃によるなりすましおよびその対策の検討を目的とする．

機械学習アルゴリズムに対し，Black-box 攻撃を仮定した A.E. を作成する手法は多様である．一例として，転送ベース攻撃手法は，ローカルモデルを用意し，それに対して White-box 攻撃により A.E. を作成，作成された A.E. を用いて対象モデルを攻撃する [3]．また，スコアベースやラベルベースの攻撃手法は，対象モデルの出力である損失を利用し，クエリ問い合わせを繰り返すことで対象モデルの勾配を予測する [4] [5]．しかし，転送ベース攻撃手法には，攻撃成功確率が高くない，任意ラベルを対象とした攻撃に効果が見られないといった課題があり，一方スコアやラベルベース手法には，攻撃成功確率は高いが，大量のクエリ数が求められるという課題があった．これに対し，Huang らが提案した TREMBA(TRansferable EMbedding based Black-box Attack) は，転送ベースの攻撃手法とスコアベースの攻撃手法，ラベルベースの攻撃手法の 3 つを組み合わせ，事前に学習されたローカルモデルを利用して，より効率的な A.E. の探索を実現した [6]．これにより，Huang らは，攻撃成功確率を高水準で保ちながら，少クエリ数で多くのモデルに適用できる A.E. 作成手法を実現した．

TREMBA をはじめとする多くの A.E. の研究は顔認証を対象とせず，検証は一般物体認識にとどまっている．一般物体認識の場合，その目的は物体形状からその物体の属するクラスを識別することが主な目的であり，クラス内の部位や，個体の識別を対象としない．一方，顔認証システムは，個人の識別を主な対象とするため，顔形状だけでなく，鼻，目，口など全ての顔形状に共通の部位に着目し，これらの個人間の違いを考慮する点が一般物体認識と異なると推定できる．このため，既存の A.E. 作成手法を顔認証システムに対して適用した場合には十分な攻撃性能が得られない可能性がある．そこで，本研究では，TREMBA が利用する Autoencoder による埋め込みネットワークの学習過程に着目し，顔のみを学習する顔認証システムの特徴を利用して，深度が異なるローカルモデルを利用した攻撃手法を提案する．また，深度および埋め込みベクトルの次元が異なる複数のネットワークによる攻撃実験を実施し，提案手法により攻撃成功確率とクエリ数の改善が可能であることを示す．

2. 関 連 研 究

2.1 Black-box 攻撃

A.E. に関する研究は活発に行われているが，本節ではその中でも，本研究が対象とする Black-box 攻撃の代表的な研究について述べる．

2.1.1 転送ベース攻撃

Papernot らは，A.E. においても通常の機械学習モデルと同様の Transferability（転送性）が存在し，あるモデルで生成した A.E. を用いて，異なるモデルへの攻撃が可能であることを示した [3]．転送ベース攻撃手法とは，この A.E. の転送性を利用した Black-box 攻撃である．攻撃者は自身がアクセス可能なローカルモデルに対して攻撃を行い A.E. を作成し，作成された A.E. を攻撃対象である未知のネットワークに転送することで，攻撃を行う．この手法は攻撃対象モデルに対して問い合わせを行わなくても良い利点があるが，現在のところ高い攻撃成功確率を達成する手法は見つかっていない．特に，顔認証システムのような特定のクラス（人物）への攻撃を目的とした攻撃に関しては更に困難性が高まることが指摘されている [3]．

2.1.2 スコアベース攻撃

スコアベース攻撃は，攻撃者が攻撃対象認証システムの出力であるスコアにアクセス可能である仮定で実行される攻撃である．なお，スコアは多くの場合は信頼度か信頼確率の形で表される値である．多くのスコアベース攻撃は，サンプリング法を用いて得られたスコアから正しい勾配への近似を行う．Chen らはスコアベース攻撃の代表的な手法の一つである AutoZoom を提案した [5]．AutoZoom では，Autoencoder とバイリニア変換を用いることでサンプリング空間を縮退させ，攻撃に必要なネットワークへのクエリ数を減少させることに成功した．また，Ilyas らは現在の時間軸方向のデータに関する事前知識を利用することで，ネットワークへのクエリ数と攻撃失敗率をさらに減少させることに成功した [7]．そして，Moon らは勾配近似法を全く使用せず，組合せ最適化を利用することでより少ないクエリ数で効果がある A.E. を作成できることを示した [8]．

2.1.3 ラベルベース攻撃

ラベルベース攻撃は，攻撃対象認証システムの出力であるラベルにのみアクセス可能な仮定で実行される攻撃であり，Black-box 攻撃では，最も厳しい仮定である．一方で，MLaaS などで提供される顔認証・識別サービスにおいては入力に対する判定結果のみが利用者に必要な情報であることは自明なため，これらのサービスにおいては最もよく使われる設定であるとも言える．Brendel らはローカルモデルの勾配とデータバイアス，2 つの事前知識を利用して A.E. を作成することで Google Cloud Vision API に対して自由に出力ラベルを操作することに成功した [9]．TREMBA はラベルベース攻撃をもとに，転送ベース攻撃手法とスコアベース攻撃手法を組み合わせることで，攻撃対象ネットワークへの効率的な Black-box 攻撃を実現した Huang らにより提案された手法である [6]．TREMBA では，Autoencoder を用いることで，攻撃対象の特徴を考慮した摂動を作成する．これらを用いて攻撃対象ネットワークへ Black-box 攻撃を行うことで，より少ないクエリ数かつ高い成功確率での攻撃を実現した．

2.2 既存研究の問題点と本研究との関連

本研究では，多くの市販されている認証モデルを想定したラベルベース攻撃を行うことを目的とする．Brendel ら [9] の手法は顔認証モデルに対して有効な A.E. を作成することに成功したが，膨大なクエリ数が発生することに加え，攻撃成功確率が低いという問題が存在する．また，転送ベース攻撃とスコアベース攻撃を組み合わせた TREMBA [6] は攻撃成功確率で Brendel らの手法を上回り，Autoencoder をローカルモデルとして使用することで意味的摂動を作成できる．ここで，意味的摂動とは，人間から見ても，一定程度，元画像の特徴が推定可能な摂動のことであり，意味的摂動には高い転送性が存在することが Huang らによって示されている [6]．

本稿では，TREMBA を基にした機械学習モデルに対する

Black-box 攻撃を行うが，特に顔認証システムを対象にした場合に有効な埋め込みネットワークについて検討を行い，回避攻撃（Concealer Attack）となりすまし攻撃（Spoofing Attack）の実験を通してその有効性を検証する点が既存研究と異なる．

3. 提 案 手 法

3.1 前 提 条 件

3.1.1 ニューラルネットワーク

本論文は，n 次元顔画像 $\boldsymbol{x} \in \mathbb{R}^n$ を入力とし，顔画像に対応する人物ラベル $y \in [1, m]$ を出力とする m クラス識別のニューラルネットワークで構成される顔認証システム $F(\boldsymbol{x}) = y$ を対象とする（以降，これを DNN 顔識別器と呼ぶ）．ここで，F は softmax 関数を含む完全なニューラルネットワークとして定義する．ここで，$Z(\boldsymbol{x})$ を softmax 関数を除いたネットワークの出力として定義すれば，DNN 顔識別器の定義は式 (1) とすることができる．

$$F(\boldsymbol{x}) = \arg\max\left(softmax(Z(\boldsymbol{x}))\right) = y. \tag{1}$$

3.1.2 攻撃者の目的

生体認証システムに対する攻撃として，(1) 攻撃者自身と特定されることを回避する回避攻撃（Concealer Attack），および (2) 自分とは異なる特定人物へのなりすまし攻撃（Spoofing Attack），の 2 つの目的が考えられる．

回避攻撃とは，入力 $\boldsymbol{x}$ に微小な摂動 δ を付与することで，DNN 顔識別器に入力 $\boldsymbol{x} + \delta$ を，y とは異なるラベル $y' \neq y$ として誤認識させる攻撃のことである．ここで，y' は y 以外のラベルであればどのようなラベルであっても構わない．

一方，なりすまし攻撃とは，DNN 顔識別器に入力 $\boldsymbol{x} + \delta$ を，y とは異なる特定のラベル t として誤認識させることを目的とした攻撃のことである．つまり，なりすまし攻撃は，$F(\boldsymbol{x} + \delta) = t$ となるような δ を見つけることを目的とする．ここで，摂動 δ は L_p 空間上において $\|\delta\|_p \leq \epsilon$ に制限される．ϵ は摂動を制御するパラメータであり，$\epsilon > 0$ とする．

3.2 TREMBA に基づくラベルベース攻撃

本研究では TREMBA [6] を基に，ラベルベース攻撃を実施する．そのため，ここではまず TREMBA における A.E. 作成手法について述べる．TREMBA における A.E. 作成は以下の二つのステップに分けられる．

- ステップ 1：回避攻撃、あるいはなりすまし攻撃のために使用される摂動を作成する Autoencoder を学習する．
- ステップ 2：Autoencoder の低次元埋め込み空間上において，任意のネットワークに対して A.E. の探索を行う．

3.2.1 Autoencoder を用いた摂動作成

エンコーダ E とデコーダ D から作られる生成ネットワークを G とする．そして，A.E. は生成ネットワーク G によって作成することが可能である．エンコーダ E は $\boldsymbol{x}$ を入力として，低次元埋め込みベクトル（潜在ベクトル）$\boldsymbol{z} = E(\boldsymbol{x})$ を出力する．そして，デコーダ D は $\boldsymbol{z}$ を入力として，$\boldsymbol{x}$ と同じ次元数を持つ摂動 $\delta = \epsilon \tanh D(\boldsymbol{z})$ を作成する．

なお，本稿では，DNN 顔識別器 F_s を誤認証させられるように生成ネットワーク G を学習させる．学習に用いる学習データセットを $\{(x_1, y_1), ..., (x_n, y_n)\}$ としている．

回避攻撃の場合，Carlini らによる手法 [10] で使用されるヒンジ損失関数を最小にすることで，生成ネットワーク G の学習を行う．このとき，回避攻撃におけるヒンジ損失関数は式 2 で表される．

$$\mathcal{L}_{untarget}(x_i, y_i) = \max\left(Z(\epsilon\tanh(G(x_i)) + x_i)_{y_i} - \max Z(\epsilon\tanh(G(x_i)) + x_i)_j, -\kappa\right) \tag{2}$$

また，なりすまし攻撃におけるヒンジ損失関数は式 (3) で表される．

$$\mathcal{L}_{target}(x_i, t) = \max\left(\max Z(\epsilon\tanh(G(x_i)) + x_i)_j - Z(\epsilon\tanh(G(x_i)) + x_i)_t, -\kappa\right) \tag{3}$$

式 (3) において，t は攻撃対象クラスである．TREMBA では，L_∞ 空間上でベクトル間の距離を算出する．なお，$\delta = \epsilon\tanh(D(\boldsymbol{z}))$ で設定したため，$\|\delta\|_\infty \leq \epsilon$ が既に満たされる．

3.2.2 低次元埋め込み空間上での探索

TREMBA における攻撃対象は，出力であるラベルにしかアクセスできない識別器 $F_t(x)$，本稿においては DNN 顔識別器である．Ilyas ら [11] の手法から，NES [12]（Natural Evolution Strategy）を用いることで，有効な A.E. を作成するために，別の損失関数（以下ローカル損失関数 $\mathcal{L}$ と呼ぶ）の勾配を概算することができる．NES は A.E. 追加後のロス勾配である $\nabla_\delta \mathcal{L}(x + \delta, y)$ を直接計算せず，更新アルゴリズム $\nabla_\delta \mathbb{E}_{\omega\sim\mathcal{N}(\delta,\sigma^2)}[L(x + \omega, y)]$ を用いて摂動 δ を更新する．更新アルゴリズムの展開式を式 (4) に示す．

$$\begin{aligned} &\nabla_\delta \mathbb{E}_{\omega\sim\mathcal{N}(\delta,\sigma^2)}[L(x + \omega, y)] \\ = \quad &\mathbb{E}_{\omega\sim\mathcal{N}(\delta,\sigma^2)}\left[L(x + \omega, y)\nabla_\omega \log \mathcal{N}(\omega|\delta, \sigma^2)\right] \end{aligned} \tag{4}$$

式 (4) を期待する勾配に近似するまで更新をし続けることで，攻撃可能な摂動 δ を作成する．ここで，η を学習率，b をバッチサイズ，ω_k をガウス分布のサンプル，$\|\delta\|_p \leq \epsilon$ を抑えるクリッピング操作を $\prod_{[-\epsilon,\epsilon]}$ とすると，摂動 δ の更新式は以下の式 (5) に示される．

$$\delta_{j+1} = \prod_{[-\epsilon,\epsilon]}\left(\delta_j - \eta \cdot sign\left(\frac{1}{b}\sum_{k=1}^{b}\mathcal{L}(x + \omega_k, y)\nabla \log \mathcal{N}(\omega_k|\delta_j, \sigma^2)\right)\right) \tag{5}$$

ただし，TREMBA では式 (5) における符号関数 $sign$ は使用しない．これは，Ilyas らの手法を実行する際，勾配の近似に符号関数 $sign$ を使用することは適切ではないことが Li ら [13] によって明らかにされているためである．

次に，TREMBA は入力空間で探索を行わず，埋め込み空間 $\boldsymbol{z}$ 上で探索を行う．入力を x，出力を y とすると，初期埋め込み

アルゴリズム 1 - 低次元空間探索による A.E. 作成

（文献 [6] Algorithm 1 を一部変更して引用）

Input:対象認証システム F_t; 入力 $\boldsymbol{x}$; 出力 y; エンコーダ E; デコーダ D; 標準偏差 σ; 学習率 η; バッチサイズ b; 繰り返し回数 T; 摂動の制限 ϵ

Output: A.E. 摂動

$z_0 = E(\boldsymbol{x})$

for $j = 1$ to T do **do**

　ガウス摂動 $\nu_1, \nu_2, .., \nu_b \sim \mathcal{N}(z_{t-1}, \sigma^2)$ を生成

　$\mathcal{L}_i \leftarrow \mathcal{L}_{untarget}(x, y)$ もしくは $\mathcal{L}_{target}(x, t)$

　$z_j \leftarrow z_{j-1} - \frac{\eta}{b} \sum_{k=1}^{b} \mathcal{L}_i \nabla_{z_{j-1}} \log \mathcal{N}(\nu_i | z_{j-1}, \sigma^2)$

end for

return $\delta = \epsilon \tanh(D(z_T))$

ベクトル $z_0 = E(x)$ となる．そこで，j 回の繰り返し後に NES によって予測される埋め込みベクトル z_j の勾配は次のように求められる．

$$\begin{aligned} & \nabla_{z_j} \mathcal{L}\Big(x + \epsilon \tanh(D(z_j)), y\Big) \\ \approx\ & \nabla_{z_j} \mathbb{E}_{\nu \sim \mathcal{N}(z_j, \sigma^2)}[\mathcal{L}(x + \epsilon \tanh(D(\nu)), y)] \\ \approx\ & \frac{1}{b} \sum_{k=1}^{b} \mathcal{L}(x + \epsilon \tanh(D(\nu)), y) \nabla_{z_j} \log \mathcal{N}(\nu_k | z_j, \sigma^2). \end{aligned} \tag{6}$$

式 (6) において，ν_k はガウス分布 $\mathcal{N}(\nu_k | z_j, \sigma^2)$ のサンプルである．本節の詳細な実行手順はアルゴリズム 1 に示す．

3.3 顔認証に適したネットワーク構成の検討

2.2 節では，既存の Black-box 型 A.E. の研究は一般物体認識モデルを攻撃対象としており，これを顔認証に適用する場合にはいくつかの課題が存在することを述べた．加えて，2.1.3 節で述べた TREMBA を用いた攻撃手法は，少ないクエリ数で高い攻撃成功確率を持つ A.E. 作成法でありながらも，顔認証におけるなりすまし攻撃に適さない可能性がある．

顔認証は入力として顔画像を想定している．顔画像は目・鼻・口など人物を問わず共通の構造を持つため，顔画像の特徴抽出に適した Autoencoder を用いることで攻撃を効率化できる可能性がある．本研究では，Autoencoder の深度を変化させることで，A.E. 作成に有用な顔特徴の抽出に影響を与えられると仮定し，Autoencoder の深度を変化させた複数の攻撃ネットワークを作成して攻撃を行う．

本研究では 3 つの Autoencoder を S1，S2，S3 として作成する．特に顔認証システムを対象にした場合に有効な埋め込みネットワークについて検討を行うために，S1，S2，S3 は深さや埋め込みベクトルの次元が異なるネットワークとして作成する．TREMBA で用いられるネットワークおよび作成した S1, S2, S3 をそれぞれ，図 1 (a)~(d) として示す．

4. 実験・検証

3.3 節で示した手法によって作成した複数の攻撃ネットワークによって作成した A.E. による攻撃を評価する．評価にあたり，攻撃ネットワークで作成した A.E. を有効性および効率の 2 つの指標で評価する．有効性は作成された A.E. がシステムの認証を回避できるか，また任意のターゲットになりすますことに成功できているかを示す指標である．また効率はかかる時間・計算量を示す指標である．本実験において攻撃成功までには攻撃対象の顔認証システムに A.E. を入力し，得られた結果に基づき A.E. を更新するプロセスを複数回繰り返す必要がある．攻撃対象顔認証システムに顔画像を 1 回入力することを 1 クエリとし，攻撃成功までに要したクエリ数で攻撃の効率を評価する．

なお，本研究では TREMBA をベースライン手法とし，提案した 3 つの Autoencoder ネットワークを用いた攻撃に関して，これらの 2 つの指標を用いて TREMBA との比較を行う．

表 1: 実行環境

使用言語	Python 3.7.1
GPU	Geforce GTX TITAN X (11GB) × 2
CUDA コア数	3584 基
CPU	Intel(R) Core(TM) i7-5930K CPU
搭載メモリ	64GB
OS/Kernel	Ubuntu 16.04 / Linux 4.15.0-74-generic

4.1 実験環境およびデータセット

本実験を実施した実験環境を表 1 に示す．また，本実験では，データセットとして CASIA-Webface データセット [14] を用いた．CASIA-Webface データセットは，10,575 人の 453,453 枚の顔画像で構成されるデータセットである．本研究の実験では CASIA-Webface データセット内から，ランダムに 1,021 人を選択し，選択した人物の 81,680 枚の画像（一人当たり約 80 枚）を使用した．使用する全ての画像に対して，顔画像の前処理として，全画像は MTCNN(注1)で顔検出を行い，顔領域を含む縦 160 ピクセル × 横 160 ピクセルに切り出しおよびリサイズを行った．

実験は攻撃対象認証モデルの転移学習，および攻撃対象認証モデルへの攻撃実施の 2 つのプロセスから構成されるため，一人当たり約 80 枚の画像を，学習用および攻撃用に 1:1 の割合で分割して使用した．

学習プロセスでは，攻撃対象モデルとして VGG2 データセット [15] によって事前学習された facenet [16]（InceptionResNetv1 [17]）を選択し，学習用画像を用いて攻撃対象モデルの出力を 1,021 クラスとした転移学習を行った．また，攻撃プロセスでは，攻撃用画像を用いて Autoencoder の学習を行い，その後，NES による攻撃対象モデルに対する A.E. の探索を行った．なお，本実験において，学習用画像を用いて転移学習を行った攻撃対象認証モデルの正確度（Accuracy）は 97.06% であった．

4.2 回避攻撃評価の手順

回避攻撃評価にあたっては，攻撃用画像を攻撃対象モデルに入力し，損失関数として式 2 を用いて攻撃用画像の元のラベル y とは異なるラベル y' となるような摂動 δ を探索した．条件を満たす摂動 δ が見つかった場合に攻撃成功とし，50,000 回の探索を行っても攻撃が成功しない場合には攻撃失敗とした．全ての攻撃用画像を用いた攻撃を実施し，攻撃成功確率を（攻撃成

(注1)：https://github.com/timesler/facenet-pytorch

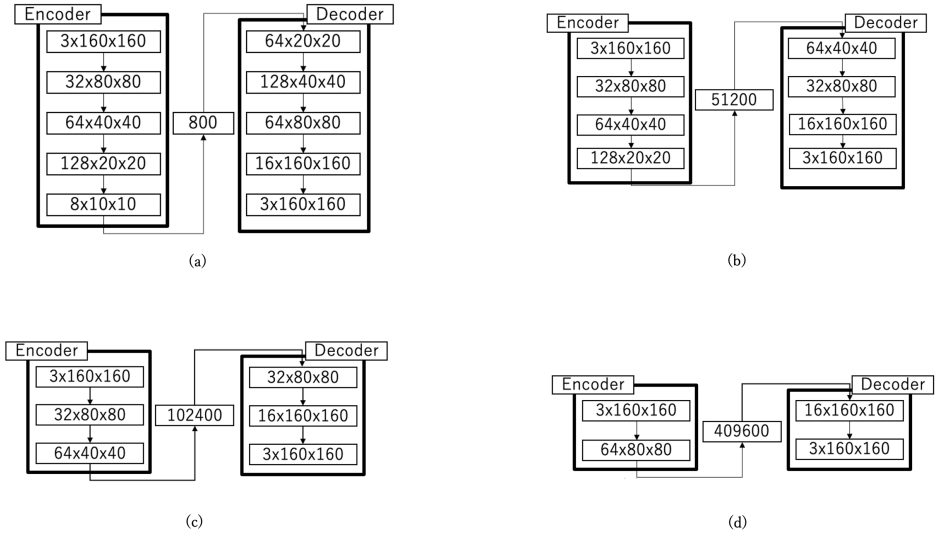


図 1: Autoencoder の構造
(a)Huang らの提案手法におけるネットワーク (b)S1　(c)S2　(d)S3

表 2: なりすまし攻撃実験の結果

攻撃手法	攻撃成功確率	平均クエリ回数
TREMBA	85.71%	4718.22
S1	93.85%	**1920.09**
S2	**96.15**%	3078.76
S3	93.85%	4399.48

功回数／全画像数）×100% として算出した．また，平均クエリ数を（攻撃成功時のクエリ数／攻撃成功回数）として算出した．

4.3　なりすまし攻撃評価の手順

なりすまし攻撃実験は TREMBA および S1,S2,S3 の 3 つの Autoencoder を用いて実施した．なりすまし攻撃評価にあたっては，攻撃用画像を攻撃対象モデルに入力し，損失関数として式 3 を用いて攻撃用画像の元のラベル y とは異なる特定のラベル t となるような摂動 δ を探索した．条件を満たす摂動 δ が見つかった場合に攻撃成功とし，50,000 回の探索を行っても攻撃が成功しない場合には攻撃失敗とした．なお，特定のラベル t は全攻撃画像において共通のラベルを用いた．攻撃成功確率および平均クエリ数は回避攻撃評価と同様の手順で算出した．

4.4　回避攻撃実験の結果

4.2 節の手順に従い，TREMBA による攻撃画像による攻撃実験を行った結果，攻撃成功確率は 100%，つまり全ての画像で回避攻撃に成功した．また，平均クエリ数は 84.9 回であった．

4.5　なりすまし攻撃実験の結果

TREMBA および S1,S2,S3 の 3 つの Autoencoder によるなりすまし攻撃実験の結果を表 2 に示す．表 2 から，TREMBA によるなりすまし攻撃成功確率が 85.71% と他の提案方式と比較して低いことがわかる．一方，提案した 3 つの Autoencoder を用いた攻撃方式 S1,S2,S3 の攻撃成功確率を見てみると，特に S2 において 96.15% と高い成功確率を示していることがわかる．また，平均クエリ回数については，S1,S2,S3 ともに TREMBA よりも少ない平均クエリ回数を達成しており，成功確率が最も高い S2 に関しては TREMBA と比較して約 65% のクエリ回数での攻撃が可能であることが確認できた．

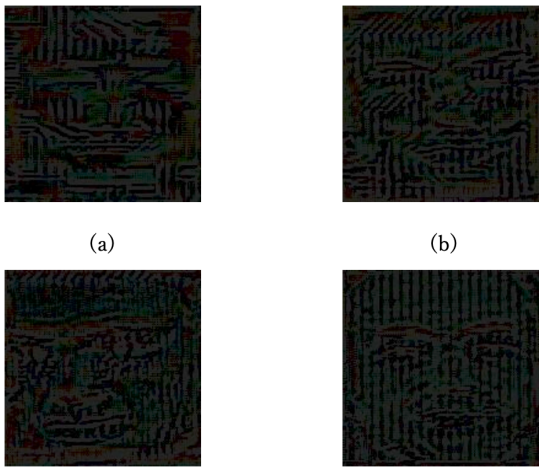
ここで，ネットワークの深さに着目すれば，本実験では，TREMBA，S1, S2, S3 の順番にネットワークの深さが浅くなるように設定した．最もネットワーク構造の深い TREMBA の攻撃成功確率が最も低かったことから，顔認証を対象とする場合はネットワーク構造が浅い Autoencoder を用いることの有効性が示されたが，一方で，最もネットワーク構造の浅い S3 よりも S2 の攻撃成功確率が高いことから，ネットワーク構造を浅くするほど攻撃成功確率が高くなるわけでもなく，また平均クエリ数の面でも，ネットワークを浅くする程平均クエリ数が増えるとは限らないことがわかる．

次に，中間層における埋め込みベクトル z の次元数に着目して S1,S2,S3 を比較する．Huang ら [6] は，潜在ベクトル z 上で A.E. を探索することは A.E. が存在する低次元ベクトル上を探索することと同様であると述べている．これらをふまえれば，TREMBA の Autoencoder が用いる潜在ベクトルの次元数 800 は顔認証における A.E. の探索には次元が小さすぎるため，A.E. の探索に失敗することが多かったと考えられる．その一方，S3 では探索空間が広すぎるために，A.E. の探索に必要なクエリ数が増大したとも考えられる．以上から，顔認証に対して有効な攻撃ネットワークの構成にあたっては，ネットワークの深さだけでなく，中間層における埋め込みベクトル z の次元数を考慮する必要があることが示唆された．

5.　議　　論

5.1　攻撃の転送性に関する検証

Huang ら [6] は Autoencoder を用いることで顔認証システムに対して有効な A.E. を作成し，他の攻撃対象ネットワークに対しても誤認証を引き起こすことが可能であることを示した．図 2(a)~(d) は本実験で得られた摂動の例であり，この例から人間の顔の外形が表現されていることが見てとれる．このことから



(a) (b)

(c) (d)

図 2: (a)(b)(c)(d) はそれぞれ TREMBA，S1，S2，S3 方式によって作成された摂動

も，本実験で作成した Autoencoder が攻撃対象の特徴，つまり人間の顔を模した摂動を取得することに成功しており，これが攻撃成功確率の向上に寄与していると考えられる．本研究においては，顔認証システムに有効な攻撃ネットワークの構成と，その有効性の検証を行ったが，一方で 2. 1. 1 節で述べた A.E. の転送性に基づき，生成した A.E. を異なるネットワークへの攻撃に用いた場合の A.E. の有効性に関する検証は行っていない．この検証は今後の課題である．

5. 2 A.E. への対策が施された顔認証システムの攻撃

Huang ら [6] は A.E. への対策 [18] が施されている認証システムの突破にも成功している．実際に市販の顔認証モデルでは様々な A.E. 対策が行われることが想定される．そのため，本研究の提案手法が，A.E. への対策が施されている顔認証システムに対して，有効であるか検証することは，今後の課題である．

6. おわりに

本研究では，顔認証における Black-box 型 A.E. 作成法について，攻撃対象ネットワークに対するクエリ回数を効率化しつつ攻撃成功確率を増大させる攻撃ネットワークの構成法について検討を行った．実験結果から，顔認証への攻撃ネットワークの深度および中間層の埋め込みベクトルの次元が攻撃成功確率およびクエリ数に大きく影響することを示唆する結果が得られたが，本稿では代表的な 3 つのネットワークから得られた結果を示したのみで，最適な構成法の確立には至っていない．このため，今後はネットワークの深度や埋め込みベクトルの次元を含む適切な指標を検討していくことで，より高い精度の攻撃を可能とする手法が確立できる可能性がある．また一方で，検討した指標を用いることで，攻撃対象ネットワークの Black-box 型の攻撃に対する耐性を推定し，対策を施すための手法を検討していくことも今後の課題である．

文　献

[1] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018.

[2] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *European Conference on Computer Vision*, pp. 19–37. Springer, 2020.

[3] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples, 2016.

[4] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, Nov 2017.

[5] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks, 2020.

[6] Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *International Conference on Learning Representations*, 2020.

[7] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors, 2019.

[8] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization, 2019.

[9] Thomas Brunner, Frederik Diehl, Michael Truong Le, and Alois Knoll. Guessing smart: Biased sampling for efficient black-box adversarial attacks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019.

[10] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017.

[11] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information, 2018.

[12] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, and Jürgen Schmidhuber. Natural evolution strategies, 2011.

[13] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. NATTACK: A STRONG AND UNIVERSAL GAUSSIAN BLACK-BOX ADVERSARIAL ATTACK, 2019.

[14] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch, 2014.

[15] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2017.

[16] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.