Computer Security Symposium 2020
26 - 29 October 2020

敵対的サンプルに対する顔認証のロバスト化手法

天田 拓磨[1,a)]　柿崎 和也[1]　Seng Pei Liew[1,†1]　古川 潤[2]　荒木 俊則[1]　Joseph Keshet[3]

概要：深層学習ベースの顔認証器を敵対的サンプルに対してロバストにする手法を提案する．敵対的サンプルは，機械学習モデルが誤識別を引き起こすように作為的に作られた入力である．顔認証は多くの場合，特徴量抽出器が用いられているが，同様に敵対的サンプルに対して脆弱であるため対策が必要である．しかし，これまでに提案されてきた防御技術の中に特徴量抽出器向けの手法はほとんど検討されていない．敵対的サンプルに対する分類器の防御手法の一つに，多様的なモデルのアンサンブルを用いる手法がある．我々はその手法を直接的に特徴量抽出器に適用したところ，モデルのロバスト性を高める効果がないことが分かった．そこで，我々は各クラスの特徴量を代表するベクトルを導入し，これをモデル間で共有する手法を提案する．共通的な特徴量の代表ベクトルを用いることで敵対的サンプルに対するモデルのロバスト性を高めることができることを実験的に確認した．

Improving Adversarial Robustness on Face Recognition Systems

Takuma Amada[1,a)]　Kazuya Kakizaki[1]　Seng Pei Liew[1,†1]　Jun Furukawa[2]　Toshinori Araki[1]
Joseph Keshet[3]

Abstract: We consider how to make deep learning-based face recognition robust against adversarial examples. An adversarial example is a maliciously crafted input that machine learning models misclassify it while humans do not. A large number of studies have proposed methods for protecting classifiers from adversarial examples. One of the most successful methods among them is to prepare an ensemble of models and promote diversity among them. We directly applied this successful method to feature extractors and found that it does not work at all, unlike to classifiers. Then, we proposed a method that synchronizes the direction of features among models and promotes the diversity of features compared to the synchronized directions. We experimented that our method of promoting diversity under the synchronization for feature extractors prevents adversarial examples significantly.

1. 導入

Deep neural netrworks（DNNs）は様々な認証タスクにおいて人間の能力を超えるほどのパフォーマンスを発揮するため，様々な分野のコア技術として利用されている [10], [20], [28]．顔認証は DNNs が利用されているアプリケーションの一つで [20], [27]，出入国管理やスマートフォンの認証まで広く活用されている．しかし，DNNs は敵対的サンプルに対して脆弱であるということが発見され [29]，顔認証技術に対しても脅威となりうる．敵対的サンプルは人間には知覚できない不正な摂動を載せた入力であり，機械学習モデルに誤識別をもたらす．敵対的サンプルの生成手法について，機械学習モデルのパラメータを用いて生成する手法を white-box 攻撃といい，パラメータ等を必要としない生成手法を black-box 攻撃という．Sharif らは顔認証において，物理的な眼鏡を付与する black-box 攻撃により機械学習モデルを騙すことができることを示した [25]．この結果から，敵対的摂動を付与した眼鏡をかけた不正者が，顔認証による出入国管理ゲートを通過できる可能性があるといえる．

[1] 日本電気株式会社
NEC
[2] NEC Israel Research Center
NEC Israel Research Center
[3] Bar-Ilan University
Bar-Ilan University
[†1] 現在，LINE 株式会社
Presently with LINE Corporation
[a)] t-amada@nec.com

敵対的サンプルに対する防御手法は今までに多数提案されてきたが，多くの手法は破られており攻撃と防御のいたちごっことなっている．敵対的サンプルに対する DNNs の防御手法のうち，有望な方法にアンサンブルモデルを使う方法がある [1], [6], [18]．特に，Adaptive Diversity Promoting（ADP）[18] はアンサンブルを構成するモデルの多様性を高めることで，アンサンブルのロバスト性を高める手法である．ADP は分類器向けの敵対的サンプルに対してロバスト性を高めることができる．

2. 準備

2.1 DNNs による顔認証

顔認証タスクは主に 2 つのカテゴリに分類される．一つは closed-set face classification と呼ばれる顔分類であり，もう一つは open-set face recognition と呼ばれる顔照合である．顔分類においては，入力の顔画像は事前に登録された顔画像の何れかと同一アイデンティティに属するという仮定がある．すなわち，これは複数クラスへの分類タスクである．一方で顔照合は任意の顔画像のペア（入力の顔画像と事前に登録された顔画像）を比較し，そのペアが同一のアイデンティティに属するか否かを決定する．本論文では，我々は顔照合に焦点を当てる．

近年の顔認証システムでは，入力画像を低次元の特徴量空間に写像する DNN ベースの特徴量抽出器がしばしば用いられる [20], [27]．2 入力の照合においては，写像された 2 つの特徴量ベクトルの類似度を算出し比較する．類似度には，Euclidean distance や cosine similarity が用いられる．算出した類似度が一定の閾値よりも高い場合に 2 入力は同一アイデンティティであると判定し，閾値を下回る場合は異なるアイデンティティであると判定する．

DNN ベースの特徴量抽出器の学習には 2 つの方法がある．一つは特徴量抽出部分の出力後に全結合層と softmax 層を追加して分類器とし，分類タスクとして学習する方法である [20], [27]．学習時は分類器として学習し，顔認証の運用時には学習された特徴量抽出部のみを用いることとなる．もう一つは，triplet loss を用いて特徴量抽出部分を直接的に学習する方法である（これは，距離学習とも呼ばれる）[23]．距離学習で用いられる triplets は，anchor サンプル，anchor と同じアイデンティティに属する positive サンプル，anchor と異なるアイデンティティに属する negative サンプルの 3 つからなる．学習には anochor-positive の特徴量ベクトル間の距離を近づけ，anochor-negative の特徴量ベクトル間の距離遠ざけるような損失関数を設定する．距離学習は triplet のサンプリングに学習の安定性が依存するため，一般的に学習が難しいと言われている．本稿では，前者の分類器としての学習アプローチを前提とする．

顔認証器として特徴量抽出器を用いる場合，新しいアイデンティティの顔画像を登録するときに，抽出される画像の特徴量が他のアイデンティティの画像とは距離が遠くなるように，特徴量抽出器を学習する必要がある．多クラス分類タスクとして特徴量抽出器を学習することは効率的であるが，照合タスクにおける各アイデンティティの抽出される特徴量のクラス間分散を大きくすることができない．そこで，分類タスクにおいて特徴量のクラス間分散を大きくする手法が提案されている [7], [15], [30].

2.2 Angular Margin Penalty

Angular Margin Penalty [7] は特徴量ベクトルのクラス間分散を大きくする手法の一つである．Angular Margin Penalty は特徴量抽出後の全結合層を変更し，正解ラベルに対応する予測確率が一定以上のマージンをもって，正解以外のラベルに対応する予測確率と乖離を持たせる．学習を進めると，抽出される特徴量の分布は同一クラス内の分散が小さくなり，異なるクラス間の分散が大きくなる．入力を x, それに対する正解ラベルを y とし，訓練対象のモデルパラメータを ϕ としたとき，モデルの出力 $g(x,\phi) \in \mathbb{R}^n$ に対する cross-entropy loss を $\mathcal{L}_{CE}(x,y)$ と表す．モデル内の d 次元の特徴量を抽出した層の出力を $f(x,\phi) \in \mathbb{R}^d$ としたときに，$f(x,\phi)$ に全結合層を追加した出力が $g(x,\phi)$ となり，これは n クラスの分類タスクとなる．$j = 1,\ldots,n$ について，$W_j \in \mathbb{R}^d$ と $b_j \in \mathbb{R}$ を最後の全結合層における重みベクトル，バイアスベクトルとしたときに cross-entropy loss は

$$\begin{aligned}\mathcal{L}_{CE}(x,y) &= {}^T 1_y \cdot \log g(x) \\ &= \log \frac{e^{W_y \cdot f(x)+b_y}}{\sum_{\ell=1}^{n} e^{W_\ell \cdot f(x)+b_\ell}}.\end{aligned}$$

と表記できる．上記の数式を含め本稿では $f(x,\phi)$ を $f(x)$ と省略して表記する．$\theta(x,\ell)$ を W_ℓ と $f(x) \in \mathbb{R}^d$ の 2 ベクトルのなす角度とすると，$\theta(x,\ell)$ は以下のように表すことができる．

$$\cos\theta(x,\ell) = \frac{W_\ell \cdot f(x)}{|W_\ell||f(x)|}.$$

$f(x)$ と w_j を $\tilde{f}(x) = \frac{f(x)}{|f(x)|}$ と $\tilde{W}_\ell = \frac{W_\ell}{|W_\ell|}$ のように L_2-正規化すると，以下ように新しい損失関数 $\tilde{\mathcal{L}}_{CE}$ を定義できる．

$$\tilde{\mathcal{L}}_{CE}(x,y) = \log \frac{e^{\sigma\cos\theta(x,y)+b_y}}{\sum_{\ell=1}^{n} e^{\sigma\cos\theta(x,\ell)+b_\ell}}$$

σ は softmax 関数の smoothness を調節するハイパーパラメータである．

簡単のために $b_\ell = 0$ とし [15]，Angular Margin Penalty のハイパーパラメータ μ としたとき，Angular Mergin Penalty を含む loss 関数 $\mathcal{L}_{ARC,\sigma,\mu}$ は以下のように表すことができる [7] .

$$\begin{aligned}&\mathcal{L}_{ARC,\sigma,\mu}(x,y) \\ &= \log \frac{e^{\sigma\cos\theta((x,y)+\mu)}}{e^{\sigma\cos\theta((x,y)+\mu)} + \sum_{\ell\in\{1,\ldots,n\}\setminus y} e^{\sigma\cos\theta(x,\ell)}} \quad (1)\end{aligned}$$

本稿ではこの loss で構築されるモデルを single model とし，実験において他のモデルとの比較対象とする．

2.3 敵対的サンプル

2.3.1 分類器に対する敵対的サンプル

DNNs は様々なタスクにおいて高い精度を出す一方で，敵対的サンプルに対して脆弱である [4], [8], [19]．敵対的サンプル x_{adv} は DNNs の入力 x_s に摂動 δ を加えて作られたサンプル（すなわち，$x_{adv} = x_s + \delta$）であり，DNNs に誤識別を引き起こす．多くの既存研究では人間には知覚できないほど微小な摂動を探索する．Fast Gradient Signed Method (FGSM)[8] はモデルの出力の確率分布が攻撃のターゲットとなるサンプルの出力の確率分布に勾配法を用いて近づける．Basic Iterative Method (BIM) は FGSM の拡張で，摂動のサイズに制約を加えつつ FGSM を繰り返し適用してより正確な摂動を求める [14]．Projected Gradient Descent (PGD) は BIM と同様の手法で摂動を探索する [16]．BIM との違いは初期点であり，BIM は初期点を x_s とし，PGD は x_s から一定距離内にあるランダムな点を初期点とする．Carlini & Wagner (CW) 攻撃は最も小さいサイズの摂動を探索する [4]．上述した BIM, PGD, CW の手法は敵対的サンプルを生成する有名でかつ強力な手法である．

2.3.2 特徴量抽出器における敵対的サンプル

Rozsa らはネットワークの中間表現を用いて，攻撃のターゲットに近づけるように入力画像に繰り返し摂動を載せていく方法を提案した（LOTS と呼ばれる）[21]．LOTS を特徴量抽出器の出力層に適用することで，特徴量抽出器に対して敵対的サンプルを生成することができる．LOTS のフレームワークは BIM,PGD,CW の手法を適用することができる．

2.4 敵対的サンプルに対する防御手法

敵対的サンプルに対してモデルをロバストにする手法はこれまでに数多く提案されてきた．ロバスト化手法には，敵対的訓練 [8], [14]，統計解析 [33]，アンサンブルモデル [1], [6], [13], [18] などがある．しかし，これらの手法のほとんどが，より強力な攻撃によってすでに破られてる [2], [3]．

Russakovsky らは，異なるモデルのアンサンブルが画像分類タスクにおいて汎化性能を高めることを示した [22]．その後，アンサンブルモデルは汎化性能を高めるだけでなく，敵対的サンプルに対するロバスト性も高めることが示された．Pang らはアンサンブルモデルの訓練時に，non-maximal predictions（分類器の出力ベクトルのうち，正解ラベルに対応する要素を除いたベクトル）をモデル間で多様的になるような正則化を用いることで敵対的サンプルに対してロバストになることを示した [18]．これは，Adaptive Diversity Promoting (ADP) 手法と呼ばれる．

アンサンブルモデルを用いた既存の防御手法は，分類器においてロバスト性を高めることが示され，且つ防御対象とする分類器のクラスはそれほど多くない．それらの防御手法は多クラスの分類器や特徴量抽出器に対して直接的に適用可能であるが，特徴量抽出器のロバスト性を高めるかどうかは明らかになっていない．

2.5 脅威モデル

本稿では以下のような状況の攻撃者を想定する．

攻撃者はモデルに完全にアクセスすることができる．これは white-box 設定であり，攻撃者はネットワーク構造や訓練済みパラメータ，全てのハイパーパラメータを知ることができる．また，攻撃者は入力値のデジタルデータを扱うことができ，特徴量抽出器に入力した場合の出力値をデジタルデータとして得ることができる．ただし，モデル学習時に用いられた情報（学習用データセットなど）は攻撃者は知らないものとする．これは顔認証器の運用において考えられうるシナリオあり，実際に本稿で扱う攻撃手法である LOTS 手法における BIM, PGD, CW は，これらの学習時の情報を使用しない．

本稿では white-box 設定で LOTS 手法をベースに敵対的サンプルを生成する．white-box 設定で生成される敵対的サンプルは，ターゲットとなるモデルを高い確率で欺くことができる．そこでモデルの脆弱性評価指標には，攻撃が成功するために必要な摂動のサイズを用いる．すなわち攻撃を成功させるために，より大きな摂動が必要であるほどモデルは堅牢であるといえる．

3. 提案手法

3.1 Adaptive Diversity Promoting (ADP) の適用

アンサンブルモデルを用いて顔認証器のロバスト性を高めるために，ADP 手法を Angular Margin Penalty に適用した学習を検討する．Pang らが提案した ADP 手法 [18] の概要を以下に示す．

アンサンブルは K 個のモデルから構成されており，それぞれのモデルは n クラス分類をするものとする．したがって Angular Margin Penalty の損失関数における全てのパラメータ $g(x), f(x), \tilde{f}(x), W_\ell, \tilde{W}_\ell, \theta(x,\ell)$ は K 個となる．

アンサンブルの予測を $\mathcal{G}$ とすると，これは各モデルの予測の平均であり，

$$\mathcal{G}_j(x) = \frac{1}{K}\sum_{k=1}^{K} g_{k,j}(x). \tag{2}$$

となる．また，$\{\mathcal{G}_j\}_{j=1,\ldots,n}$ の分布の Shannon エントロピーは，

$$\mathcal{H}(\mathcal{G}(x)) = -\sum_{j=1}^{N} \mathcal{G}_j(x)\log(\mathcal{G}_j(x)).$$

と表せる．

入力のペア (x, y) に対するアンサンブルモデルの損失関数 $\mathcal{L}_{EARC,\sigma,\mu}$ は k 番目のモデルの Angular Margin Penalty の損失 $\mathcal{L}_{CE}^{k}$ の総和であり，

$$\mathcal{L}_{EARC,\sigma,\mu}(x, y) = \sum_{k\in[K]} \mathcal{L}_{ARC,\sigma,\mu}^{k}(x, y) \qquad (3)$$

である．

ここで，$g_{k,\setminus y} \in \mathbb{R}^{n-1}$ を $g_k \in \mathbb{R}^n$ から正解ラベル y に対応する y 番目の要素を除いた $n-1$ 次元ベクトルとする．$\tilde{g}_{k,\setminus y} \in \mathbb{R}^{n-1}$ を L_2 正規化した $g_{k,\setminus y} \in \mathbb{R}^{n-1}$ とし，$(n-1) \times K$ 行列を $M(x, y) = (\tilde{g}_{1,\setminus y}(x), \ldots, \tilde{g}_{K,\setminus y}(x)) \in \mathbb{R}^{(n-1)\times K}$ とする．このとき，non-maximal prediction の多様性指標である ensemble diversity $\mathbb{ED}$ を

$$\mathbb{ED}(x, y) = \det({}^{T}M(x, y) \cdot M(x, y)).$$

と定義する．上記の演算子"$\cdot$"は $K \times (n-1)$ 行列と $(n-1) \times K$ 行列の行列積である．ここで $\mathbb{ED}$ はベクトル群 $\{\tilde{g}_{1,\setminus y}(x), \ldots, \tilde{g}_{K,\setminus y}(x)\}$ の張る立体の体積を表す．

ADP 手法の正則化 $\mathsf{ADP}_{\alpha,\beta}(x, y)$ は，

$$\mathsf{ADP}_{\alpha,\beta}(x, y) = \alpha \cdot \mathcal{H}(\mathcal{G}(x)) + \beta \cdot \log(\mathbb{ED}(x, y))$$

となる．α と β はハイパーパラメータである．ADP 手法における学習の損失関数は正則化項を用い，

$$\mathcal{L}_{EARC-ADP,\sigma,\mu,\alpha,\beta} = (\mathcal{L}_{EARC,\sigma,\mu}(x, y) - \mathsf{ADP}_{\alpha,\beta}(x, y)).$$

となる．

損失関数 $\mathcal{L}_{EARC-ADP,\sigma,\mu,\alpha,\beta}$ を用いて訓練された特徴量抽出器を $(\tilde{f}_k \in \mathbb{R}^d)_{k=1,\ldots,K}$ とする．特徴量のアンサンブル $\mathcal{F}$ を複数の特徴量抽出器の出力の平均として，

$$\mathcal{F}(x) = \frac{1}{K}\sum_{k=1}^{N} \tilde{f}_k(x). \qquad (4)$$

と定義する．特徴量抽出器のアンサンブルを用いた認証タスクにおける類似度指標には，入力 x に対する特徴量のアンサンブル $\mathcal{F}(x)$ と別の入力 x' に対する特徴量のアンサンブル $\mathcal{F}(x')$ との Euclidean distance を用いる．

3.2 Feature Diversity Promotion

ADP 手法においては，アンサンブルモデルの non-maximal prediction を多様的にする．しかしながら，攻撃者はモデルの prediction よりもむしろ特徴量を用いて敵対的サンプルを生成する．したがって，我々はアンサンブル特徴量を多様化する．ADP 手法と同様に，入力 x に対するアンサンブル特徴量の多様性 $\mathbb{ED}_{feat}$ は測ることができる．$\tilde{F}(x) = (\tilde{f}_1(x), \ldots, \tilde{f}_K(x)) \in \mathbb{R}^{d\times K}$ を $d \times K$ 行列とする．このとき，アンサンブル特徴量の多様性は

$$\mathbb{ED}_{feat}(x) = \det({}^{T}\tilde{F}(x) \cdot \tilde{F}(x))$$

と表せる．

したがって，以下の正則化を用いて，アンサンブル特徴量を多様化することができる．

$$\mathsf{FDP}_{\gamma}(x) = \gamma \log(\mathbb{ED}_{feat}(x))$$

重みの係数 γ はハイパーパラメータである．ADP 手法とは異なり，正則化項に Shannon エントロピーを含まない．

Feature Diversity Promotion（FDP）手法のモデルの学習における損失関数を以下のように表す．

$$\mathcal{L}_{EARC-FDP,\sigma,\mu,\gamma} = (\mathcal{L}_{EARC,\sigma,\mu}(x, y) - \mathsf{FDP}_{\gamma}(x)).$$

3.3 事前実験

ここで，正則化項である $\mathsf{ADP}_{\alpha,\beta}(x, y)$ 及び $\mathsf{FDP}_{\gamma}(x)$ の効果を確認する実験の結果を示す．

Single Model: 損失関数に $\mathcal{L}_{ARC,\sigma,\mu}$ を用い，正則化項なく学習された単一の特徴量抽出器．

Baseline: 損失関数に $\mathcal{L}_{EARC,\sigma,\mu}$ を用い，正則化項なく学習された複数の特徴量抽出器．

ADP: 損失関数に $\mathcal{L}_{EARC-ADP,\sigma,\mu,\alpha,\beta}$ を用いて学習された複数の特徴量抽出器．

FDP: 損失関数に $\mathcal{L}_{EARC-FDP,\sigma,\mu,\gamma}$ を用いて学習された複数の特徴量抽出器．

single model, baseline, ADP, FDP によって学習されたそれぞれのモデルの認証の正確性は，ROC カーブを示している図 5 から分かるようにすべて同程度といえる．また，敵対的サンプルに対するネットワークの正確性を表している図 6, 7, 8 から，ADP 及び FDP はネットワークのロバスト性を高めるとは言えないことがわかる．

図 1, 2, 3 は，baseline, ADP, FDP それぞれについて t-SNE によって特徴量を写像した可視化の図である．同じ色のプロットは同じクラスに属することを意味し，色の深度は違いは異なるモデルを意味している．これらの図から，ADP は baseline と比べて特徴量を多様化できていないことがわかる．また，FDP は ADP と比べて特徴量を多様化しているが，baseline と比べるとそれほど多様化できているとは言えない．どちらの手法も，ネットワークの敵対的サンプルに対するロバスト性を高めないといえる．

3.4 事前実験における問題

顔認証において，高い true acceptance rate と低い false acceptance rate を達成するために，抽出される特徴量ベクトルは以下の 2 つの条件を満たす必要がある [7], [15], [23], [31]：

- 高い true acceptance rate のために，特徴量ベクトルのクラス内分散が小さい．
- 低い false acceptance rate のために，特徴量ベクトルのクラス間分散が大きい．

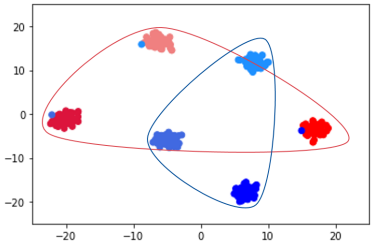

図 1 t-SNE visualization of feature mappings without regularizer

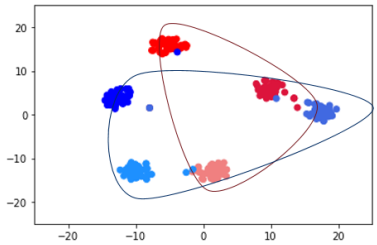

図 2 t-SNE visualization of feature mappings with ADP

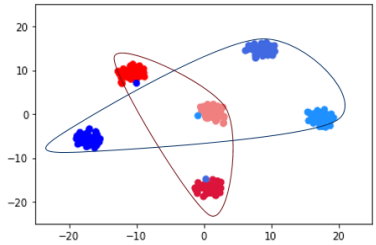

図 3 t-SNE visualization of feature mappings with FDP

アンサンブル特徴量の多様化は上記の条件と相反するが，我々の実験ではモデルのパフォーマンスを低下させないという結果が得られた．次に，このアプローチにおける敵対的サンプルに対するロバスト性について述べる．

Angular Margin Penalty の損失関数は全ての特徴量ベクトル $\{\tilde{f}_k(x)\}_{k=1,\dots,K}$ と重みベクトル $\{W_\ell^k\}_{k=1,\dots,K;\ell=1,\dots,n}$ が d 次元の単位球面上にある必要がある．学習においては，重みベクトル W_y^k が k 番目のモデルの中で異なるクラス間で互いに距離を取るように，且つ同一のクラス間で可能な限り距離が近くなるようにパラメータを更新する．x から生成する敵対的サンプル $x' = x + \delta$ は，$f_k(x)$ が W_y^k と近くなるが，$f_k(x')$ は $y \neq y'$ である $W_{y'}^k$ に近くなる．

アンサンブルモデルを用いてモデル間の多様性を高めることは，各モデルが出力する特徴量に対して摂動の影響の受け方を多様的にすることである．ADP と FDP の両方の手法は特徴量の摂動に対する影響を多様的にする可能性がある．しかし，モデル間で重みベクトルは独立に学習されるため，摂動を加えることでモデル毎に同様の方向に特徴がずれていく可能性がある．例えば，x に摂動を加えることで $f_1(x)$ が W_2^1 の方向に移動したとき，$f_2(x)$ も W_2^1 に移動する可能性がある．このような状況では，ADP や FDP 等の手法で出力をモデル間で多様的にしたとしても，敵対的サンプルに対するロバスト性を高める効果はないことがわかる．そこで，各クラスを代表する重みベクトルをモデル間で共有し，且つ出力を多様的にする正則化をする必要がある．

3.5 解決方法

上記の問題を解決する方法は，モデルの最終層の重みベクトル $\{W_\ell^k\}_{(k,\ell)\in\{1,\dots,K\}\times\{1,\dots,n\}}$ を全てのモデルで共有することである．すなわち，すべての k について，W_ℓ^k を W_ℓ に統一する．我々の解決策では，ラベル y に対する全ての特徴量が W_y に近づくようにする．このとき，アンサンブルモデルに対して攻撃を成功させるために，攻撃者は全てのモデルをの特徴量が W_y の向きから $W_{y'}$ の向きへ遷移するような摂動を探索する必要がある．仮に摂動に対して特徴量の受ける影響がモデル間で異なっていれば，攻撃者は全てのモデルの特徴量ベクトルを同様の向きに動かすことが困難である．モデル間で共有する最終層の重みベクトルを Shared Repreentative Vectors (SRV) と呼ぶこととする．

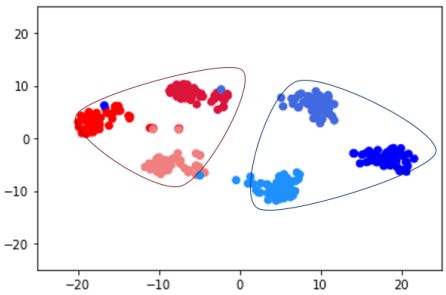

図 4 t-SNE visualization of feature mappings with SRV-FDP

SRV である W_ℓ と特徴量 $f_k(x) \in \mathbb{R}^d$ との角度を $\psi_k(x,\ell)$ とすると，

$$\cos\psi_k(x,\ell) = \frac{W_\ell \cdot f_k(x)}{|W_\ell||f_k(x)|}$$

と表せる．

SRV を導入した損失関数 $\mathcal{L}_{EARC-SRV,\sigma,\mu}$ は，

$$\begin{aligned}&\mathcal{L}_{EARC-SRV,\sigma,\mu}(x,y)\\ &= \sum_{k=1}^{K} \log \frac{e^{\sigma\cos(\psi_k(x,y)+\mu)}}{e^{\sigma\cos(\psi_k(x,y)+\mu)} + \sum_{\ell\in\{1,\dots,n\}\setminus y} e^{\sigma\cos\psi_k(x,\ell)}}.\end{aligned}$$

と表せる．以下の損失関数に従って学習される特徴量抽出器が敵対的サンプルに対してロバストであると期待する．

$$\mathcal{L}_{EARC-SRV-FDP,\sigma,\mu,\gamma}(x,y) \tag{5}$$

$$= \mathcal{L}_{EARC-SRV,\sigma,\mu}(x,y) - \mathsf{FDP}_\gamma(x). \tag{6}$$

図 4 は式 5 の損失関数にしたがって学習されたアンサンブルモデルの特徴量の t-SNE による可視化である．図から，特徴量は我々が想定してた通りの分布となった．

4. 実験

本章では，既存手法，提案手法，提案手法に至るまでのいくつかの実験バリエーションについて，設定を変えて顔認証のパフォーマンスを評価した．事前実験で示したように，既存の ADP 手法と FDP 手法はロバスト性を高める効果がなかった．モデルの最終層の重みベクトルをアンサンブルのモデル間で共有したときに，ADP 手法や FDP 手

法にロバスト性を高める効果を持たせることができるかを検証する．

4.1 実装の詳細

Dataset

学習プロセスは全て文献 [7] にしたがって行った．学習と照合用のデータには *emore dataset*[7], [9] を用いた．*emore dataset* は学習データに合計で 5.8M の顔画像が含まれており，85K のアイデンティティがある．また，学習の最中でモデルの照合のパフォーマンスを測定するために，LFW[11], AgeDB-30[17], 及び CFP-FP[24] を用いた．顔画像の前処理としては，MTCNN[32] を用いて facial landmark 検出を行い，アラインメント行った．また，画像を (112×112) にリサイズした．

4.2 ネットワークとハイパーパラメータ

特徴量抽出のネットワークには，広く用いられている CNN アーキテクチャである MobileFacenet[5] を用いた．最後の convolution 層の後に，512 次元の特徴量を得るために，BN [12]-Dropout [26]-FC-BN の構造を取り入れた．これらは文献 [7] と同じ実験設定である．

また，文献 [31] に従い，$\sigma = 64$，angular margin $\mu = 0.5$ とした．学習時のミニバッチサイズは 256 に設定し，モデル 3 つから構成されるアンサンブルを NVIDIA Tesla V100 (32GB) GPU 上で学習した．学習率は初期値を 10^{-3} とし，12, 15, 18 エポック終了時に 1/10 に減衰させた．訓練は 20 エポックで終了とした．

ADP 手法におけるハイパーパラメータは $(\alpha, \beta) = (2.0, 0.5), (2.0, 10.0)$，及び $(2.0, 50.0)$ を実験し，FDP 手法におけるハイパーパラメータは $\gamma = 1.0, 10.0$, 及び 50.0 を実験した．敵対的サンプルの生成には LOTS のフレームワークで BIM, PGD, CW を用いた．

学習したアンサンブルモデルの種類は以下のとおりである．

Original Angular Margin Penalty (Single model):

$\mathcal{L}_{ARC,\sigma,\mu}(x, y)$.

An ensemble of AMP (Baseline):

$\mathcal{L}_{EARC,\sigma,\mu}(x, y)$.

AMP with ADP (ADP):

$$\mathcal{L}_{EARC-ADP,\sigma,\mu,\alpha,\beta}(x, y) = \mathcal{L}_{EARC,\sigma,\mu}(x, y) - \mathsf{ADP}_{\alpha,\beta}(x, y).$$

AMP with Feature Diversity Promotion (FDP):

$$\mathcal{L}_{EARC-FDP,\sigma,\mu,\gamma}(x, y) = \mathcal{L}_{EARC,\sigma,\mu}(x, y) - \mathsf{FDP}_{\gamma}(x).$$

AMP with Shared representative vector (SRV):

$\mathcal{L}_{EARC-SRV,\sigma,\mu}(x, y)$.

SRV with ADP (SRV+ADP):

$$\mathcal{L}_{EARC-SRV-ADP,\sigma,\mu,\alpha,\beta}(x, y) = \mathcal{L}_{EARC-SRV,\sigma,\mu}(x, y) - \mathsf{ADP}_{\alpha,\beta}(x, y).$$

Proposed (SRV + FDP):

$$\mathcal{L}_{EARC-SRV-FDP,\sigma,\mu,\gamma}(x, y) = \mathcal{L}_{EARC-SRV,\sigma,\mu}(x, y) - \mathsf{FDP}_{\alpha,\beta}(x, y).$$

method	LFW	CFP-FP	AgeDB-30
single model	99.30	89.60	94.22
Baseline	99.12	89.71	94.15
ADP	98.90	86.20	90.52
FDP	**99.40**	89.94	94.13
SRV	99.26	**90.94**	94.93
SRV+ADP	99.38	86.62	93.98
SRV+FDP	**99.40**	89.97	**95.15**

表 1 Accuracy of verifications by different methods

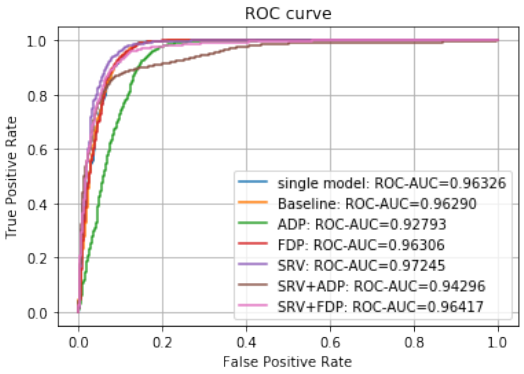

図 5 The ROC curves on 1000 pairs of test samples of VGG2 dataset. Each model is trained on *emore* training data.

4.3 正常サンプルに対するパフォーマンス

正常系のデータについて，顔認証の認証精度を表 1 に示す．この時のハイパーパラメータは $(\alpha, \beta) = (2.0, 0.5)$ で，$\gamma = 10.0$ である．表より，Baseline と SRV と FDP は single model と同程度のパフォーマンスであることがわかる提案手法である，SRV-FDP は常に single model に優る一方で，ADP 手法を採用した手法である ADP と SRV-ADP は常に single model に劣る結果となった．

また，VGG2 データセットにおいても認証精度を測定した．図 5 の ROC カーブに示すように，ADP を除く全てのモデルで同程度の認証精度となった．

4.4 敵対的サンプルに対するロバスト性

4.4.1 攻撃手法

敵対的サンプルを生成するために特徴量抽出器に対して，LOTS[21] のフレームワークを適用した．LOTS は入力 x_s に対するモデルの中間表現を攻撃のターゲット x_t の中間表現に近づけるように調節する手法である．これを実現するために，x_s の中間表現と x_t の中間表現の Eucldean 距離を用い，勾配法により x_s を動かしていく．このとき，勾配は $\nabla_{x_s} |f(x_s) - f(x_t)|_2$ である．この勾配は BIM 攻撃

や PGD 攻撃，さらに CW 攻撃に用いることができる．

各攻撃の詳細は以下である．BIM 攻撃ではサイズ $\varepsilon = 0.001$ の摂動を攻撃が成功するまで載せ続けた．分類器に対する BIM は摂動のサイズが一定以下にになるように制約を加えているが，本稿での評価では摂動サイズに対して制約は持たせないこととした．PGD 攻撃では，オリジナル画像と敵対的画像の距離が L_∞ ノルムで 10%以下に制約して，摂動を載せ続けた．CW 攻撃では iteration step を 1,000 とし，binary search step を 9 と設定し，摂動サイズが最小の敵対的サンプルを探索した．摂動のアップデートには Adam optimizer を用い，x_s と x_t の L_2 ノルム距離が一定の閾値を下回るまで摂動を載せ続けた．閾値については，各モデル毎に認証における F-1 スコアが最大となるときのものを用いた．

4.4.2 攻撃成功率の指標

顔認証のアンサンブルのロバスト性を測るために，ϵ-attack success rate を以下のように定義する．

$$\epsilon Acc = \frac{|\{x_{adv}|x_{adv} \in AX; |x_{adv} - x_s|_2 < \epsilon\}|}{|AX|} \quad (7)$$

ただし，x_s は正常サンプルであり x_{adv} は x_s から生成された敵対的サンプルである．AX は white-box 設定で生成した，攻撃が成功する敵対的サンプルの集合である．この指標は全ての正常サンプルに対して摂動サイズが ϵ 以下である敵対的サンプルの割合を意味する．特徴量抽出器を欺くためにより大きな摂動を必要とするならば，その特徴量抽出器はよりロバストであるといえる．ϵAcc を測定するために，VGG2 データセットから異なるアイデンティティの画像のペアをランダムに 500 個サンプリングし，上記の 3 手法で敵対的サンプルを生成した．

4.4.3 結果

図 6, 7, 及び 8 は BIM, PGD, CW の ϵAcc を表している．BIM と CW の結果から，以下の事実を言うことができる．

- baseline と single model は同程度のロバスト性であり，ADP と FDP は顕著にロバスト性を高めない．
- SRV を導入することでロバスト性が向上する．この現象は，SRV のおかげでアンサンブルの各モデルが互いをより区別可能になったことから生じると考察する．
- ADP を SRV に導入することは，モデルのロバスト性を高めない．ADP は SRV に対して効果的ではない．
- SRV に FDP を組み合わせることで，SRV 単体よりロバストになる．この現象は元々の仮定であった FDP が効果的であるということの理由付けとなったが，モデル間で特徴量ベクトルの基準を設ける必要があった．我々の SRV により，これが解決された．

次に，PGD について考察する．SRV と FDP を組み合わせた提案方式は敵対的サンプルに対して最もロバストであるという結果になった．他の手法は，BIM や CW と似た結果となった．

5. 結論

単純に ADP や提案の一つである Feature Diversity Promoting (FDP) を特徴量抽出器に導入するだけでは，敵対的サンプルに対するロバスト性を高めることができないことを確認した．そこで，アンサンブルの各モデルが出力する特徴量の基準となる Shared Representative Vectors (SRV) を我々は提案した．SRV は FDP によるロバスト化の効果を高めることを実験的に確認した．結論として，アンサンブルの特徴量抽出器に対する多様性向上の方法を明確にした．また，アンサンブルの分類器向けの多様性向上の方法と，アンサンブルの特徴量抽出器向けの方法は異なるという興味深い性質を発見した．

参考文献

[1] Abbasi, M. and Gagné, C.: Robustness to Adversarial Examples through an Ensemble of Specialists, *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings* (2017).

[2] Athalye, A., Carlini, N. and Wagner, D. A.: Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 274–283 (2018).

[3] Carlini, N. and Wagner, D. A.: Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods, *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pp. 3–14 (2017).

[4] Carlini, N. and Wagner, D. A.: Towards Evaluating the Robustness of Neural Networks, *IEEE Symposium on Security and Privacy*, IEEE Computer Society, pp. 39–57 (2017).

[5] Chen, S., Liu, Y., Gao, X. and Han, Z.: MobileFaceNets: Efficient CNNs for Accurate Real-time Face Verification on Mobile Devices, *CoRR*, Vol. abs/1804.07573 (2018).

[6] Dabouei, A., Soleymani, S., Taherkhani, F., Dawson, J. and Nasrabadi, N. M.: Exploiting Joint Robustness to Adversarial Perturbations, *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).

[7] Deng, J., Guo, J., Xue, N. and Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition, *CVPR*, Computer Vision Foundation / IEEE, pp. 4690–4699 (2019).

[8] Goodfellow, I. J., Shlens, J. and Szegedy, C.: Explaining and Harnessing Adversarial Examples, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

[9] Guo, Y., Zhang, L., Hu, Y., He, X. and Gao, J.: MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition, *ECCV (3)*, Lecture Notes in Computer Science, Vol. 9907, Springer, pp. 87–102 (2016).

[10] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE

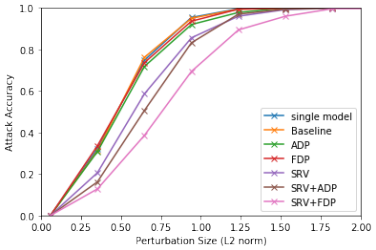

図 6 LOTS via BIM (on *emore* dataset)

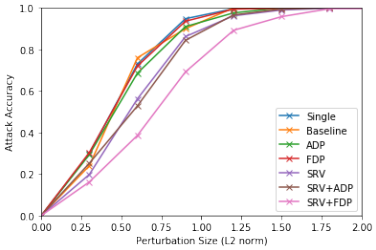

図 7 LOTS via CW (on *emore* dataset)

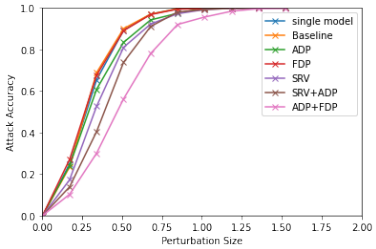

図 8 LOTS via PGD (on *emore* dataset)

Computer Society, pp. 770–778 (2016).

[11] Huang, G. B., Ramesh, M., Berg, T. and Learned-Miller, E.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical Report 07-49, University of Massachusetts, Amherst (2007).

[12] Ioffe, S. and Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 448–456 (2015).

[13] Kariyappa, S. and Qureshi, M. K.: Improving Adversarial Robustness of Ensembles with Diversity Training, *CoRR*, Vol. abs/1901.09981 (online), available from ⟨http://arxiv.org/abs/1901.09981⟩ (2019).

[14] Kurakin, A., Goodfellow, I. J. and Bengio, S.: Adversarial Machine Learning at Scale, *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (2017).

[15] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B. and Song, L.: SphereFace: Deep Hypersphere Embedding for Face Recognition, *CVPR*, IEEE Computer Society, pp. 6738–6746 (2017).

[16] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A.: Towards Deep Learning Models Resistant to Adversarial Attacks, *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* (2018).

[17] Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I. and Zafeiriou, S.: AgeDB: The First Manually Collected, In-the-Wild Age Database, *CVPR Workshops*, IEEE Computer Society, pp. 1997–2005 (2017).

[18] Pang, T., Xu, K., Du, C., Chen, N. and Zhu, J.: Improving Adversarial Robustness via Promoting Ensemble Diversity, *ICML*, Proceedings of Machine Learning Research, Vol. 97, PMLR, pp. 4970–4979 (2019).

[19] Papernot, N., McDaniel, P. D., Jha, S., Fredrikson, M., Celik, Z. B. and Swami, A.: The Limitations of Deep Learning in Adversarial Settings, *EuroS&P*, IEEE, pp. 372–387 (2016).

[20] Parkhi, O. M., Vedaldi, A. and Zisserman, A.: Deep Face Recognition, *Proceedings of the British Machine Vision Conference 2015, BMVC 2015, Swansea, UK, September 7-10, 2015* (Xie, X., Jones, M. W. and Tam, G. K. L., eds.), BMVA Press, pp. 41.1–41.12 (2015).

[21] Rozsa, A., Günther, M. and Boult, T. E.: LOTS about attacking deep features, *IJCB*, IEEE, pp. 168–176 (2017).

[22] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C. and Li, F.: ImageNet Large Scale Visual Recognition Challenge, *Int. J. Comput. Vis.*, Vol. 115, No. 3, pp. 211–252 (2015).

[23] Schroff, F., Kalenichenko, D. and Philbin, J.: FaceNet: A unified embedding for face recognition and clustering, *CVPR*, IEEE Computer Society, pp. 815–823 (2015).

[24] Sengupta, S., Chen, J., Castillo, C. D., Patel, V. M., Chellappa, R. and Jacobs, D. W.: Frontal to profile face verification in the wild, *WACV*, IEEE Computer Society, pp. 1–9 (2016).

[25] Sharif, M., Bhagavatula, S., Bauer, L. and Reiter, M. K.: Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016* (Weippl, E. R., Katzenbeisser, S., Kruegel, C., Myers, A. C. and Halevi, S., eds.), ACM, pp. 1528–1540 (2016).

[26] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.*, Vol. 15, No. 1, pp. 1929–1958 (2014).

[27] Sun, Y., Chen, Y., Wang, X. and Tang, X.: Deep Learning Face Representation by Joint Identification-Verification, *NIPS*, pp. 1988–1996 (2014).

[28] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: Going deeper with convolutions, *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, IEEE Computer Society, pp. 1–9 (2015).

[29] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J. and Fergus, R.: Intriguing properties of neural networks, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings* (Bengio, Y. and LeCun, Y., eds.) (2014).

[30] Wang, F., Cheng, J., Liu, W. and Liu, H.: Additive Margin Softmax for Face Verification, *IEEE Signal Process. Lett.*, Vol. 25, No. 7, pp. 926–930 (2018).

[31] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z. and Liu, W.: CosFace: Large Margin Cosine Loss for Deep Face Recognition, *CVPR*, IEEE Computer Society, pp. 5265–5274 (2018).

[32] Zhang, K., Zhang, Z., Li, Z. and Qiao, Y.: Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks, *CoRR*, Vol. abs/1604.02878 (2016).

[33] Zheng, Z. and Hong, P.: Robust Detection of Adversarial Attacks by Modeling the Intrinsic Properties of Deep Neural Networks, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 7924–7933 (2018).