2022 年 8 月 10 日(水)

思考と言語究室　中間審査会資料

顔検出器を攻撃することによる個人の写真のプライバシー保護

D2120005　　張重陽

1. はじめに

21 世紀初頭に深層学習手法が提案され、人工知能研究は日々加速的に進歩してきている。とりわけ画像認識に関する技術は、人間の認識能力をも凌駕する域に達し、社会に多くの恩恵をもたらしている。しかしながらその半面、敵対的生成ネットワークを利用したフェイススワッピング技術などは、利便性のみならず、以下のような社会的リスクを生み出し始めた。

・政治家の顔を変え虚偽情報を公表。人々の間でパニックを引き起こす。

・他人の写真を使い、冗談や侮辱などをする。

・他人の写真を使って詐欺をする。

本研究では、この問題を解決するために、個人の写真のプライバシー(ユーザの顔の権利と利益)の保護を目的として、顔検出器の検出能力削減手法を提案する。

2. 関連研究

Deepfakes アルゴリズムは顔を変える手法であり、これを用いて faceswap プロジェクトでは、社会的に人気のある人物の顔を変えている[1]。一方、MTCNN[2]、S3FD[3]および SSD[4]は、同ロジェクトで使用されている顔検出アルゴリズムであり、本研究ではこれらを研究対象（攻撃ターゲット）とする。また、FGSM と PGD[5]は、ホワイトボックス攻撃アルゴリズムであり、既知モデルのグラディエントを取得し、モデルの精度を攻撃（劣化）できるので、本研究はこれらの手法も利用することとした。

3. 本研究の概要

本研究は、有効攻撃範囲、黒線構造、およびマルチスケール摂動融合の敵対的攻撃の 3 つの実験で構成されています。

ターゲット検出モデルを攻撃する場合、多くの研究者は画像内のターゲットが位置する領域を直接攻撃して攻撃を実行します。しかし、非対象部位の検出確率を向上させることで目的を達成しようとする研究はまだいくつかあります[6]。そこで、顔検出敵対的攻撃の有効範囲のアイデアと根拠を提供するために、簡単な実験を行いました。

現在、多くの攻撃方法は、モデルによってフィードバックされた勾配値を変更するために摂動を使用していますが、この方法では、摂動を減らして画像を鮮明にするという点で不十分です。摂動の場所と数は制御不能であり、摂動値の選択も継続的な試行錯誤の複雑なプロセスであるため、冗長な摂動が存在しています。冗長な摂動を減らすために、CNN の特徴抽出の連続性を破ることにより、上記の 3 つの顔検出器を攻撃するための黒線構造を提案しました。

敵対的攻撃の一般的な手段を使用して、3 つの顔検出器を攻撃します。SSD と S3FD は簡単に攻撃を成功させることができますが、MTCNN を攻撃する過程で、顔が摂動によって完全に覆われていても、検出器は効果的で正確なバウンディングボックスをフィードバックすることができます。そこで、MTCNN の特性に従って、マルチスケール摂動融合敵対的攻撃方法を提案します。

4. 実験

4.1 有効攻撃範囲

図 1 に示すように、MTCNN の p-net を使用して画像を大まかに検出し、顔と顔以外の部分を取得します。

図 1 P-net は MTCNN の 3 つのネットワークの最初のネットワークであり、大まかに検出されたバウンディングボックスを 2 番目のネットワーク「R-net」に送信します。P-net が顔を検出できない場合、すべての MTCNN ネットワークは顔を検出できません。

図 2 に示すように、スタイルトランスファー「neural style transfer」を使用して、顔の特徴が顔以外の部分に追加されます。

図 2 スタイルトランスファーは、スタイルとコンテンツの両方を抽出することができ、本研究は顔の特徴を含むコンテンツの一部を使用します。

図 3 に示すように、MTCNN の p-net を再度使用して新しい画像を検出すると、顔以外の部分の検出率は上げましたが、顔は検出されたままであることがわかります。

図 3 顔の特徴は引き続き入力として使用され、R-net に渡されます。

4.2 黒線構造

攻撃方法

a. MTCNN を使用して画像の顔検出を実行し、検出フレームと顔の特徴の座標を取得します。
b. 検出フレームの四隅、口の隅、および 2 つ目を黒線で接続します。

図 4 に示すように、CLEBA と FFHQ データセットに対し、顔を黒線で覆った画像。具体的には、検証データとして、CELEBA データセットから 10,000 枚の画像がランダムにサンプリングされ、顔はそれぞれピクセル幅 6、8、10 の黒線で覆われていました。FFHQ データセットから 6000 枚の画像がランダムにサンプリングされ、顔はそれぞれピクセル幅 4、6、8 の黒線で覆われていました。

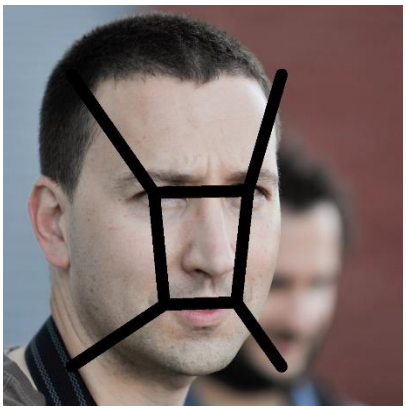

図 4 効果図

実験結果

表 1 と表 2 は検出される確率を示しています。

表 1 元 FFHQ データセットと、4、6、および 8 ピクセル幅の黒線が追加されたデータセットで顔を検出できる 3 つの顔検出アルゴリズムの結果を示しています。

FFHQ	No Noise	4 pixels	6 pixels	8 pixels
MTCNN	99.96%	11.7%	8.8%	7.2%
S3FD	99.7%	58.7%	23.7%	9%
SSD	99.93%	22.4%	9.7%	4.3%

表 2 元 CelebA データセットと、6、8、および 10 ピクセル幅の黒線が追加されたデータセットで顔を検出できる 3 つの顔検出アルゴリズムの結果を示しています。

CelebA	No Noise	6 pixels	8 pixels	10 pixels
MTCNN	99.79%	9.08%	7.49%	6.15%
S3FD	99.67%	65.5%	49.4%	32.1%
SSD	99.71%	24.9%	15.08%	9.47%

4.3　マルチスケール摂動融合敵対的攻撃

図 5 に示すように、MTCNN は画像ピラミッドを使用して画像を前処理し、ネットワークの次の層の入力として各スケールで検出結果を結合します。画像ピラミッドとは、基本的に画像を一定の比率で縮小する処理方法です。

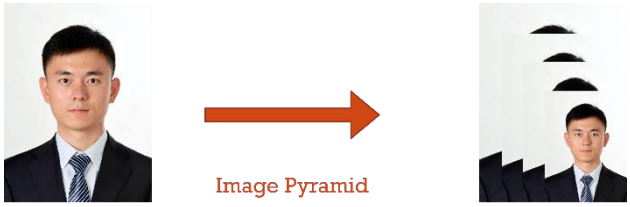

図 5 画像ピラミッド

画像ピラミッドは MTCNN にとって非常に重要です。これにより、さまざまなサイズの画像がさまざまなバウンディングボックスを提供できるようになります。これが、デジタル ドメインで MTCNN を攻撃できない理由です。

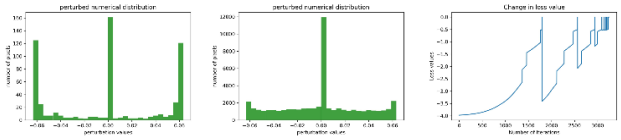

図 6 画像の元のサイズは 128 x 128 で、縮小すると 0.1 倍になります

Algorithm 1 By scaling the perturbation to find adversarial examples that are resistant to size changes.

Input: Original image, I_o;
The size by which the image will be scaled, S;
P-net of MTCNN face detection system, N_p;
Control the threshold of the detection box, t.
Output: Valid attack samples for each size.

Initialize An array with the same size as I_o and a value of zero, used to store the final perturbation, PI_o.
for each $s \in S$ **do**
  $I_s = I_o$.resize(s, bilinear interpolation)
  $PI_s = I_s$.data(value=0) // store temporary perturbations
  while *True* **do**
    prob $\leftarrow N_p(I_s)$
    loss $\leftarrow$ createLoss(prob) // loss is less than or equal to zero
    boxes $\leftarrow$ findBox(prob, s, t)
    sr $\leftarrow$ searchRate()
    pertur $\leftarrow$ sr $\times$ I_s.grad.sign()
    if loss $\geq -t$ or boxes is None **then**
      temp = PI_s.resize(I_o.size, bilinear interpolation)
      I_s= (temp + I_o).resize(I_s.size, bilinear interpolation)
      if loss $\geq -t$ or boxes is None in the next loop **then**
        PI_o += temp
        break
      else
        continue
      end if
    end if
    I_s[boxes] += pertur[boxes] // modify the data within the bounding boxes
    PI_s[boxes] += pertur[boxes]
  end while
end for
return $I_o + PI_o = 0$

図 7　アルゴリズム

MTCNN を攻撃する過程で、各サイズでの画像の効果的な摂動を考慮する必要があります。ただし、ズーム処理で画像がぼやけるのを避けるために、摂動のみがスケーリングされます。図 7 のアルゴリズムには 2 つのループが含まれており、外側のループを使用してさまざまなスケールを反復処理し、内側のループを使用して勾配ベースの摂動を検索します。

図 6 は、補間前後の摂動の分布と、内側のループの下での摂動の収束を示しています。損失値の各低下は摂動補間プロセスです。

5. 讨论

実験 4.1 は、顔以外の領域への摂動の追加が冗長な動作であることを証明しています。

黒線構造は一般的な敵対攻撃とは異なる手法で、物理攻撃におけるダイレクトマスキングに似ています。実験によりその有効性が確認されており、ピクセル幅が大きくなるにつれてパフォーマンスが向上します。MTCNN、S3FD、SSD がすべて CNN 構造体

を使用しているため、CNN の特別な抽出方法により、顔検出アルゴリズムは顔のキーポイント間の特徴に注目するようになっています。黒線は CNN の抽出された特徴の連続性を中断するため、顔情報を効果的に統合することができないため、3 つの顔検出器を同時に無効にすることができます。

図 8　左から右へ、画像の元のサイズは 128x128、178x218、300x232、3840x2160 です

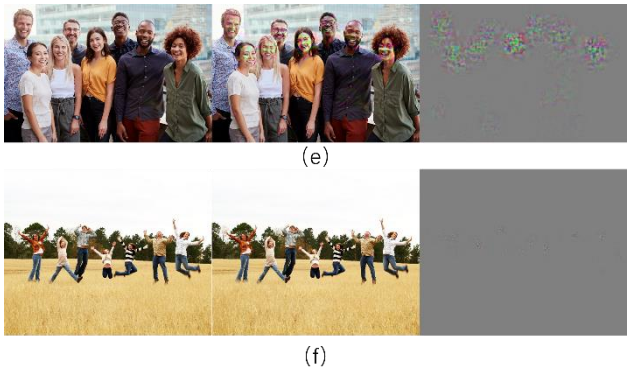

図 9　2 つの画像の元のサイズは 1300x867, 1024x820 です。

図 7 と図 8 は、マルチスケール摂動融合攻撃法の有効性を示しており、それぞれ元の画像、敵対的な例、摂動画像を示しています。

表 3 構造類似性比較結果。 画像の元のサイズをデータとして使用します。

	(a)	(b)	(c)	(d)	(e)	(f)
LPIPS	0.012	0.014	0.029	0.054	0.037	0.018
SSIM	0.968	0.970	0.955	0.971	0.943	0.978

表 3 は、上記 6 つの画像の構造類似性比較の結果を示しています。LPIPIS の結果が 0 に近いほど類似しており、SSIM の結果が 1 に近いほど類似しています。

6. おわりに

上記の実験を通じて、顔と非顔の検出が独立していることを実証し、CNN 特徴抽出の連続性を破ることができる黒い線構造を提案し、デジタル ドメインで MTCNN を攻撃する方法を見つけます。

これらの実験により、入学時の研究目標である、faceswap が画像内の顔を検出できなくなり、顔の変化が起こらないようにするという目標を達成しました。もちろん、研究が深まるにつれて、解決しなければならない問題がいくつか見つかりました。例えば、デジタルドメインではMTCNN と他の顔検出器を同時に攻撃できない敵対的攻撃方法は存在せず；マルチスケール摂動融合攻撃が高解像度画像を攻撃すると、追加された摂動が人間の目で検出される等。

今後は、2 つの研究方向に焦点を当てます。 CNN を中断して連続的な特徴を抽出できる摂動生成手法を開発し、画像ピラミッドの敵対的攻撃に対する耐性を利用して効果的な防御手法を開発します。

参考文献

[1] https://faceswap.dev/
[2] Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks
[3] S3FD: Single Shot Scale-invariant Face Detector
[4] SSD: Single Shot MultiBox Detector
[5] Towards Deep Learning Models Resistant to Adversarial Attacks
[6] hiding faces in plain sight: Disrupting ai face synthesis with adversarial perturbations.

発表実績

1. Multi-scale perturbation fusion adversarial attack on MTCNN face detection system, **CY. Zhang**, Y. Qi, H. Kameda, CISCE 2022, (Accepted), (2022-June).
2. A New filtering method of images/movies against Deep Fake by drawing lines between eyes and mouths, **CY. Zhang**, H. Kameda, Journal of Computers, Under review
3. Animal Exercise: A New Evaluation Method, Y. Qi, **CY. Zhang**, H. Kameda, Journal of Computer Science Research, Vol.04, Issue 02, pp.24-30 (2022-April).
4. Historical Summary and Future Development Analysis of Animal Exercise, Y. Qi, **CY. Zhang**, H. Kameda, ICERI 2021 (Accepted), (2021-Nov).
5. Motion transfer in crawling stance, Y. Qi, **CY. Zhang**, H. Kameda, Journal of AI, Under review