4Q-02

Deepfake を破壊する摂動の転移性調査と効率的な最適化手法の検討

加藤 義道 † 福原 吉博 † 森島 繁生 ‡

† 早稲田大学 ‡ 早稲田大学理工学術院総合研究所

1. はじめに

Deepfake は深層学習を用いたメディア合成技術である．これにより，人の印象を悪くするような悪意ある合成動画が作成され問題になっている．これを解決するために，人が認識できない微弱な摂動を用いて DNN 変換モデルを破壊する手法が注目されている．既存手法では，最適化した変換モデルに対しては効率的な破壊が可能だが，異なる変換モデルに対する摂動の転移性調査は行われていなかった．この技術を実用化する場合，こうした転移性が重要となると考えられ，どの変換モデルが使用されてもその出力が破壊出来なければいけない．そのためには，一つの変換モデルと最適化手法によって生成された摂動で任意の変換モデルの出力を破壊できるのが理想である．よって我々は，複数の変換モデル間での摂動転移性を網羅的に調査した．既存手法では摂動を大きくすることである程度出力の破壊を確認したが，大きな摂動を加えることで画像の知覚品質が低下した．これを踏まえて，大きな摂動を加えても画像の知覚品質が劣化しないような手法を検討した．

2. 関連研究

2.1 Deepfake Generation

Deepfake は，深層学習を用いて簡単に写実的な合成メディアを生成できる技術である．これにより，特定の人の顔を編集し，本人の名誉を毀損するような悪意ある合成動画が拡散される恐れがある．深層学習によって顔画像の編集を行う手法 [1–4] はこれまでに数多く提案されている．StarGAN [1] は，髪色，性別など複数のドメインを単一の生成モデルで編集する．StarGANv2 [2] は参照画像のスタイルを抽出し変換を行う．GANimation [3] は，cGAN によって顔の表情を編集する．Simswap [4] は，参照画像の顔を入力画像に移植する．

2.2 Disrupting Deepfakes

Deepfake による被害を防止するための手段として，画像に敵対的摂動を加えることで，変換モデルの出力を破壊する手法 [5] が提案されており，PDG Attack [6] を用いて最適化された摂動によって変換モデル [1, 3] を破壊することに成功している．しかし，一つの変換モデルに対して最適化された摂動が，他の変換モデルに転移し出力を破壊するかどうかは調査されていなかった．

2.3 Adversarial Attack

モデルの出力を破壊する為に，入力画像 $\boldsymbol{x}$ に敵対的摂動 $\boldsymbol{\eta}$ を加える手法 [6, 7] は数多く提案されている．

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{\eta} \tag{1}$$

摂動を一定の大きさに制限しつつ，摂動を加えた入力 $\tilde{\boldsymbol{x}}$ を変換した出力と基準との距離関数を最大化するように摂動を最適化する．

$$\max_{\boldsymbol{\eta}} L(\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta}), \boldsymbol{r}), \quad \text{subject to } ||\boldsymbol{\eta}||_\infty \leq \boldsymbol{\epsilon} \tag{2}$$

ここで，$\boldsymbol{G}(\cdot)$ は変換モデルの出力を表し，基準 $\boldsymbol{r}$ は，摂動を加えていない本来の出力としている．PGD Attack では，オリジナルの入力 $\boldsymbol{x}$ に大きさ $\boldsymbol{\epsilon}$ の範囲でランダムな摂動を加えたものを初期値 $\boldsymbol{x_0}$ として反復的な更新を行う．t 回目における更新は次式のように表せる．

$$\tilde{\boldsymbol{x}}_t = \text{clip}(\tilde{\boldsymbol{x}}_{t-1} + \boldsymbol{a}\ \text{sign}[\nabla_{\tilde{\boldsymbol{x}}} L(\boldsymbol{G}(\tilde{\boldsymbol{x}}_{t-1}), \boldsymbol{r})]) \tag{3}$$

ここで $\boldsymbol{a}$ はステップサイズを表し，また各ステップで $||\tilde{\boldsymbol{x}} - \boldsymbol{x}||_\infty \leq \boldsymbol{\epsilon}$ となるように clip 関数を用いている．

PGD Attack のほかにも，摂動を加えた画像の知覚品質を向上させる手法として Shadow Attack [7] がある．最適化のための目的関数は次式で表される．

$$\begin{aligned}\max_{\eta} [&L(\boldsymbol{G}(\boldsymbol{x}+\boldsymbol{\eta}), \boldsymbol{r}) \\ &- \lambda_c C(\boldsymbol{\eta}) - \lambda_{tv} TV(\boldsymbol{\eta}) - \lambda_s Dissim(\boldsymbol{\eta})]\end{aligned} \tag{4}$$

(2) 式に加えて，摂動を加えた入力の知覚品質を向上させるための 3 つの項が存在する．$C(\boldsymbol{\eta})$ は，各色チャンネルの平均値の変化を制限し，画像のカラーバランスの極端な変化を抑制する．$TV(\boldsymbol{\eta})$ は，摂動の全変動量を小さくすることでより滑らかで自然な表現を可能にする．$Dissim(\boldsymbol{\eta})$ は，各色チャンネルが似たような値をとる摂動を促進させ，画像のカラーバランスを変えずに画素を暗くしたり明るくしたりすることができる．

図 1 摂動転移性の概要図．モデル A に対して最適化された摂動はモデル A の出力を破壊できるが，モデル B に対して最適化された摂動ではモデル A の出力を破壊できない．

3. Disrupting Deepfakes の転移性調査

3.1 実験概要

特定の変換モデルを破壊する摂動が他の変換モデルを破壊することができるか調査した．摂動の生成には既存手法で使用されている PGD Attack を使用した．使

Investigation of Transferability of Perturbations that Disrupt Deepfake and Efficient Optimization Method:
Gido Kato†, Yoshihiro Fukuhara†, and Shigeo Morishima‡ (†Waseda University, ‡Waseda Research Institute for Science and Engineering)

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

図 2　各変換モデルから StarGAN [1] への摂動の転移

用したモデルは StarGAN, StarGANv2, GANimation, Simswap の 4 つである.

3.2　結果

4 つの変換モデルのいずれにおいても, 最適化した変換モデルの破壊が確認でき, 変換モデルごとに破壊のされ方が異なることを確認した. しかし, 図 2(上) のように最適化されていない変換モデルに対しては視覚的な破壊は見られず, 転移性が低いことが分かった. また, 摂動を大きくしていくと, 最適化されていない変換モデルであってもある程度の破壊効果が現れたが, その分入力画像の知覚品質が低下することが確認された.

4.　Shadow Attack を用いた摂動生成

4.1　実験概要

前項で, 最適化された摂動の大きさの範囲を大きくしていった場合, 最適化されてない変換モデルであってもある程度の破壊効果が確認された. そこで, 摂動を大きくしても入力画像の知覚品質が低下しない最適化手法によって変換モデルの破壊を試みた. 本項では, Shadow Attack を用いて変換モデルの破壊を試みた. また, PGD Attack と同様に, 摂動の転移性を調査した.

4.2　結果

図 3　転移性の比較 (StarGANv2 → StarGAN)

PGD Attack と同様に最適化したモデルの破壊が確認できた. しかし, 最適化されていない変換モデルに対しては, 図 2(下) のように目に見える出力の破壊は見られず, 摂動の転移は確認できなかった. 図 3 は, StarGANv2 で摂動を最適化し, StarGAN で変換を行った結果を表している. それぞれの最適化手法で摂動の大きさを変えた結果を示しているが, Shadow Attack で最適化された摂動では破壊はほとんど確認できず, PGD Attack の場合, 摂動が大きいときはある程度破壊が確認できるが, その分入力画像の知覚品質が低下していることが分かる.

4.3　考察

今回の実験で, 2 つの最適化手法を用いて変換モデルを破壊するような摂動を生成したが, いずれも摂動の転移は確認できなかった. また, 今回使用した 4 つの変換モデルでは, その破壊結果がどれも異なっていた. これらのことから, 摂動の転移は最適化手法に寄与するのではなく, 変換モデルの構造や特徴表現の類似度に寄与する可能性が考えられる.

5.　おわりに

本稿では, PGD Attack を用いて最適化された摂動の転移性を網羅的に調査し, 摂動の大きさの範囲を大きくした場合にのみ, 最適化していない変換モデルの出力がある程度破壊されることを確認した. それを踏まえて, 摂動の大きさの範囲を大きくしても入力画像の知覚品質が低下しない最適化手法として Shadow Attack を検討した. その結果, 最適化した変換モデルに対しては破壊効果が確認できたが, 最適化していないモデルに対する摂動の転移は確認できなかった. この結果を踏まえた今後の展望として, 摂動の転移性は最適化手法に寄与するのではなく, 変換モデルの構造や特徴表現の類似度に寄与する可能性を考慮した調査を行いたい.

謝辞

この研究は, JST 未来社会創造事業 (JPMJMI19B2) および JSPS 科研費 (19H01129, 19H04137, 21H05054) の補助を受けた.

参考文献

[1] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8188–8197, 2020.

[3] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[4] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2003–2011, 2020.

[5] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. 2020.

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

[7] Amin Ghiasi, Ali Shafahi, and Tom Goldstein. Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates. In *International Conference on Learning Representations*, 2020.

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.