6U-02

敵対的最適化と距離学習を用いたDeepfake検出

大竹 ひな † 福原 吉博 † 久保谷 善記 † 森島 繁生 ‡
† 早稲田大学 ‡ 早稲田大学理工学術院総合研究所

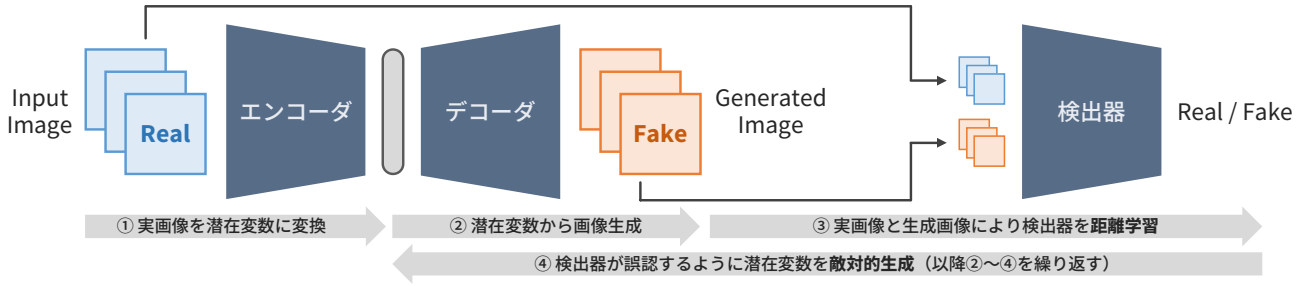

図 1: 提案手法の概要図

1. はじめに

Deepfake [1] は深層学習を用いて映像や音声を編集，合成する技術である．この技術を悪用したフェイクニュースやなりすましは深刻な社会問題であり，Deepfake の検出は情報の信頼性を担保する上で重要な課題である．一方で，近年の画像生成技術の発達に伴い，アーティファクトが微細で検出が困難な Deepfake が増加している．これに対して，既存手法では検出が困難な Deepfake を事前定義された操作により生成し，学習データとして使用することで検出器の精度向上を試みる手法が提案されているが，これは未学習のデータに対して検出精度が低下するという課題がある．本研究では，敵対的生成を使用することで検出器にとって検出が困難な Deepfake を自動で生成し，それらに対して距離学習を行う手法（図 1）を提案する．敵対的生成と距離学習を組み合わせることで，様々な種類の Deepfake に対して汎化した検出器の獲得を目指す．

2. 事前準備

本章では，提案手法で使用する GAN inversion，Adversarial attack，距離学習について説明する．

2.1 GAN inversion

生成モデル $G : \mathcal{Z} \to \mathcal{X}$ によって潜在変数 $\boldsymbol{z} \in \mathcal{Z}$ から生成されたデータ点を $\boldsymbol{x} \in \mathcal{X}$ とする．このとき，$\boldsymbol{x}$ の G に関するインバージョン $\boldsymbol{z}^* \in \mathcal{Z}$ を次のように定義する．

$$\boldsymbol{z}^* = \arg\min_{\boldsymbol{z}} \ell(G(\boldsymbol{z}), \boldsymbol{x}) \tag{1}$$

ただし，ここで $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ は適当な距離関数である．特に G として Generative Adversarial Network (GAN) を用いるとき GAN inversion と呼ぶ．$\boldsymbol{z}^*$ を推定する方法として様々な手法が提案されているが，主要なものの 1 つにニューラルネットワークを用いてインバージョンを行い，エンコーダ $E : \mathcal{X} \to \mathcal{Z}$ を学習する手法がある．

2.2 Adversarial attack

K-クラス分類器 $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^K$ に対して，損失 L を増加させるようにデータ点 $\boldsymbol{x} \in \mathcal{X}$ を更新したものは敵対的サンプル $\boldsymbol{x}'$ と呼ばれる．敵対的サンプルの生成手法の 1 つに次式で表される PGD [2] がある．

$$\begin{cases} \boldsymbol{x}'^{(0)} = P_{\mathcal{X},S_{\boldsymbol{x}}}(\boldsymbol{x} + \zeta) \\ \boldsymbol{x}'^{(n+1)} = P_{\mathcal{X},S_{\boldsymbol{x}}}\left(\boldsymbol{x}'^{(n)} + \eta \cdot \nabla_{\boldsymbol{x}} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}'^{(n)}), t_{\boldsymbol{x}})\right) \end{cases} \tag{2}$$

ただし，$t_{\boldsymbol{x}}$ は $\boldsymbol{x}$ のラベル，$S_{\boldsymbol{x}}$ は $\boldsymbol{x}$ に関する敵対的サンプルの候補となるデータ点の集合，$\eta > 0$ はステップサイズ，$P_{\mathcal{X},S_{\boldsymbol{x}}} : \mathbb{R}^d \to \mathbb{R}^d$ は，入力空間 $\mathcal{X}$ から $\mathcal{X} \cap S_{\boldsymbol{x}}$ 内の最近傍点への射影であり，$\zeta \in \mathbb{R}^d$ は初期化のために加える微小なベクトルである．

2.3 距離学習

距離学習はデータ間の距離や類似性を学習する手法である．距離学習を使用すると，分類境界面付近へのデータの密集を防ぎ，識別的な特徴空間を獲得できるため，学習データの分布外となるデータに対しても検出精度が向上する．最も基本的な距離学習の手法として任意の 2 つのデータ点の組に対して，intra-class の場合は距離を最小化し，inter-class の場合は距離を最大化するコントラスティブロス [3] があり，次式のように定義される．

$$\begin{aligned} L(\boldsymbol{x}_1, \boldsymbol{x}_2, t_1, t_2) = {} & \mathbb{1}_{t_1 = t_2} \left[\mathcal{M}^2(g_{\boldsymbol{w}}(\boldsymbol{x}_1), g_{\boldsymbol{w}}(\boldsymbol{x}_2))\right] \\ & + \mathbb{1}_{t_1 \neq t_2} \left[\max(0, \alpha - \mathcal{M}^2(g_{\boldsymbol{w}}(\boldsymbol{x}_1), g_{\boldsymbol{w}}(\boldsymbol{x}_2))\right] \end{aligned} \tag{3}$$

3. 提案手法

本章では，提案手法を用いて Deepfake 検出器 f_θ を学習する方法について説明する．検出器の汎化性能向上のため，検出器が本物と誤認しやすい fake 画像を敵対的に生成し，学習データとして使用する．また，識別力の高い特徴空間を獲得するために検出器の特徴空間で距離学習を行う．提案手法を用いた損失の計算方法をアルゴリズム 1 に示す．

3.1 GAN inversion で潜在変数を推定

実画像 $\boldsymbol{x} \in \mathcal{X}$ の潜在変数 $\boldsymbol{z} \in \mathcal{Z}$ を GAN inversion を用いて推定する．本稿ではエンコーダベースの手法である HyperInverter[4] を採用した．

$$\boldsymbol{z} = E(\boldsymbol{x}) \tag{4}$$

3.2 潜在空間で敵対的サンプルを生成

3.1 で推定した潜在変数を GAN のデコーダに入力することで，$\boldsymbol{x}$ を再構成した fake 画像 $G(\boldsymbol{z})$ を生成し，更に分類器 f_θ に代入することで分類結果を得る．ここで，$G \circ f_\theta : \mathcal{Z} \to \mathbb{R}^2$ を 1 つの合成関数とみなし，adversarial attack の手法を適用することで，潜在空間における敵対的サンプル $\boldsymbol{z}'$ を求める．$\boldsymbol{z}'$ から生成される fake 画像 $\boldsymbol{x}' = G(\boldsymbol{z}')$ は分類器 f_θ にとって $G(\boldsymbol{z})$ よりも分類が難しいデータ点となっている．提案手法では adversarial attack の手法として PGD を使用した．

Algorithm 1 提案手法の損失の計算方法

Input: 検出器 $f_{\boldsymbol{\theta}}$，実画像の集合 $\mathcal{X}$，GAN inversion のエンコーダ E 及びデコーダ G，PGD attack の摂動の最大値 ε 及びステップサイズ η 及びステップ数 s，損失の線型結合係数 λ

Output: 損失 L

for $\boldsymbol{x}_i \in \mathcal{X}$ **do**
  $\boldsymbol{z}_i \in \mathcal{Z} \leftarrow E(\boldsymbol{x}_i)$
  $\boldsymbol{z}_i'^{(0)} \leftarrow P_{\mathcal{Z},S_{\boldsymbol{z}_i}}(\boldsymbol{z}_i + \zeta)$
  for $j = 0 \ldots s-1$ **do**
    $\boldsymbol{z}_i'^{(j+1)} \leftarrow P_{\mathcal{Z},S_{\boldsymbol{z}}}\left(\boldsymbol{z}_i'^{(j)} + \eta \cdot \nabla_{\boldsymbol{z}} L(f_{\boldsymbol{\theta}}(\boldsymbol{z}_i'^{(j)}), t_{\boldsymbol{z}})\right)$
  end for
  $\boldsymbol{x}_i' \leftarrow G(\boldsymbol{z}_i'^{(s)})$
end for
$L \leftarrow \sum_{\boldsymbol{x}_i \in \mathcal{X}} \Big[\lambda \cdot \Big\{L_{\mathrm{CE}}(\boldsymbol{x}_i, t_{\mathrm{real}}) + L_{\mathrm{CE}}(\boldsymbol{x}_i', t_{\mathrm{fake}})\Big\} + (1-\lambda) \cdot L_{\mathrm{M}}(\boldsymbol{x}_i, \boldsymbol{x}_i')\Big]$

3.3 損失の計算

敵対的に生成した fake 画像と実画像を用いて 2 クラス分類器の学習を行う．学習の際の損失には，クラス分類のためのクロスエントロピーロス L_{CE} と距離学習のためのコントラスティブロス L_{M} の線形和を使用する式 (5)．ただし，N はバッチサイズである．

$$L = \sum_{i=1}^{N} \Big[\lambda \cdot \Big\{L_{\mathrm{CE}}(\boldsymbol{x}_i, t_{\mathrm{real}}) + L_{\mathrm{CE}}(\boldsymbol{x}_i', t_{\mathrm{fake}})\Big\} + (1-\lambda) \cdot L_{\mathrm{M}}(\boldsymbol{x}_i, \boldsymbol{x}_i')\Big] \quad (5)$$

4. 評価実験

4.1 実験条件

本実験では Deepfake 検出において広く利用されている FaceForensics++ [5] をモデルの訓練及び評価に使用した．ただし，今回の実験では全体のデータの 10%のみを使用して学習を行った．モデル訓練時のデータ拡張は先行研究 [6] に従い，色調変換，周波数変換，拡大縮小，平行移動を用いた．

Deepfake 検出器のモデルには EffcientNet-b4 を使用した．最適化には SAM を使用し，バッチサイズは 16，学習率は 0.001 に設定し，学習率が 75 エポック以降で線形に減衰するように 100 エポック学習を行った．PGD のパラメータは，ステップ数 s を 10 とし，摂動の最大値 ε は 4 に設定した．ステップサイズ η を先行研究 [7] に従い，$\eta = \varepsilon/\sqrt{s}$ とした．また，ノルムには L_2 を用いた．実際に生成された敵対的 Deepfake の例を図 2 に示す．

評価には FaceForensics++ [5] のテストセットの動画を使用した．実画像のデータには，各動画からサンプリングしたフレームに対して MediaPipe による顔検出を行い，フレームごとに最大サイズの顔画像のみを 380×380 にリサイズしたものを使用した．fake 画像のデータには，評価用の実画像データを GAN inversion を用いて再構成したものを使用した．また，評価尺度として検出器の分類精度を採用した．

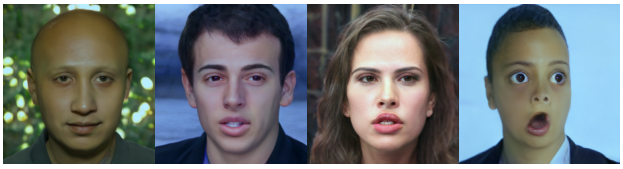

図 2: 敵対的 Deepfake の生成例

表 1: 実験結果

手法	λ	精度 [%]
SBI [6]	-	86.8
ベースライン	0.0	66.3
提案手法 (敵対的学習)	1.0	77.2
提案手法 (敵対的学習 + 距離学習)	0.9	59.2

4.2 実験結果

評価実験の結果を表 1 に示す．ベースラインはクロスエントロピーロスを用いた通常のクラス分類として学習を行った場合である．また，先行研究である SBI の精度を比較のために示しているが，SBI は FaceForensics++の全データを学習に使用しているため，提案手法とは実験条件が異なっている．

提案手法を用いて敵対的学習を行った場合は，ベースラインと比較して精度が 10.9%向上し，有効性が確認された．一方で，敵対的学習と距離学習を組み合わせた場合は，精度が 7.1%低下した．学習過程を分析すると，式 (5) のコントラスティブロス L_{M} が先に最適化され，学習の後半までクロスエントロピーロス L_{CE} が減少しない傾向が見られた．これは，敵対的 Deepfake を学習に用いた事で，クラス分類の学習が距離学習と比較して相対的に困難になっていることが原因と考えられる．したがって，両損失の線形和を最適化するのではなく交互に最適化することで，精度低下を防ぐことが可能であると推察される．

5. おわりに

本稿では，敵対的最適化と距離学習を用いた Deepfake 検出を提案した．評価実験の結果，敵対的最適化の有効性が確認された．一方で，距離学習の導入では期待した結果は得られなかった．今後は，距離学習の損失の最適化方法の改善を行うとともに，全データを使用した評価実験を行い，他のデータセットに対しても提案手法が有効かを確認する予定である．

謝辞 本研究は，JSPS 科研費 (21H05054) の補助を受けています．

参考文献

[1] R. Chen *et al.*: "SimSwap: An Efficient Framework For High Fidelity Face Swapping," *Proceedings of the 28th ACM International Conference on Multimedia*, 2003–2011, 2020.

[2] M. Aleksander *et al.*: "Towards Deep Learning Models Resistant to Adversarial Attacks," *International Conference on Learning Representations*, 1–19kaitenai, 2018.

[3] R. Hadsell *et al.*: "Dimensionality reduction by learning an invariant mapping," *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Vol. 2., 1735–1742, 2006.

[4] T. Dinh *et al.*: "Hyperinverter: Improving stylegan inversion via hypernetwork," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11389–11398, 2022.

[5] A. Rossler *et al.*: "Faceforensics++: Learning to detect manipulated facial images," *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11, 2019.

[6] K. Shiohara *et al.*: "Detecting deepfakes with self-blended images," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18720–18729, 2022.

[7] D. Kang *et al.*: "Transfer of adversarial robustness between perturbation types," *arXiv preprint arXiv:1905.01034*, 18720–18729, 2019.