L-35

deepfake 検出に向けた Fine-Tuning 型識別機の開発

Development of a Fine-Tuning-Based Detector for Deepfake Detection

○白井光[1]，岸本誠也[2]，大貫進一郎[2]
*Hikaru Shirai[1]，Seiya Kishimoto[2]，Shinichiro Ohnuki[2]

Abstract: In recent years, advances in generative AI have dramatically improved the accuracy of deepfakes technology that alters human faces and voices to create convincing fakes. When developing detectors for deepfakes, utilizing deep learning with large datasets enables feature extraction from generated images that are difficult for humans to discern. Furthermore, through retraining, it becomes possible to adapt to evolving manipulation techniques. This report describes a method to expand the adaptability of a transfer-learned deepfake detection model. This is achieved by gradually adding new training data: original images, real images of multiple individuals not previously trained on, and deepfake-manipulated images, one person at a time.

近年，生成 AI の進歩により，deepfake と呼ばれる人間の顔や声を別物に改ざんする技術の精度が飛躍的に向上している．このような背景から多様な偽造手法，人物に対応可能な汎用性の高い deepfake 検出器が必要とされている．検出器製作にあたり，大量のデータを用いる深層学習(Deep Learning)により，人間に判断困難な生成画像の特徴抽出に加え，再学習を行うことで進歩する加工技術に対応することが可能である．本報告では，転移学習済みの deepfake 検出モデルに対し，学習を行っていない人物の実画像である original 画像と deepfake 加工を施した画像を複数の人物について 1 人ずつ段階的に追加学習することで適応対象の拡大を図る．この際，製作した deepfake 検出モデルが人物の特徴を学習しているように学習過程や特徴抽出範囲の可視化を通して評価を行う．

学習済みのネットワークモデルである ResNet-50[1]に original と deepfake 画像を各 6389 枚用いた転移学習を行い[2]，特定の人物に対して正確に deepfake 検出可能なモデルを製作した．このモデルを基盤に Fine-Tuning を行う．以下の式(1)で表すことができる学習強度を一定にし，Fine-Tuning を行うことで検出に偏りのない汎用性の高い検出器の製作が可能となる．

$$\text{学習強度} = \text{データ数} \times \text{学習率} \times \text{エポック数} \tag{1}$$

Fine-Tuning したモデルの顔領域に対する認識精度を評価するために，損失関数の収束傾向と特徴抽出領域の可視化を行う Grad-CAM[3]を用い検討する．図 1 は損失関数の収束傾向を示し，過学習による破綻は見られず精度向上の推移を確認することができる．また，初期モデルと比較し学習回数に比例して識別性能を保てる対象人数の増加を確認した．次に，検出器が deepfake 検出を行う際，判断の重きを置いている特徴抽出領域の可視化を Grad-CAM[3]を用いて評価する．図 2 より検出器が顔領域の特徴抽出を行い判別していることが確認出来る．これらの結果を用いて顔領域に特徴抽出領域を持ち，判別可能な人物数を拡大したモデルを作成し，deepfake 検出器の検討を行う．

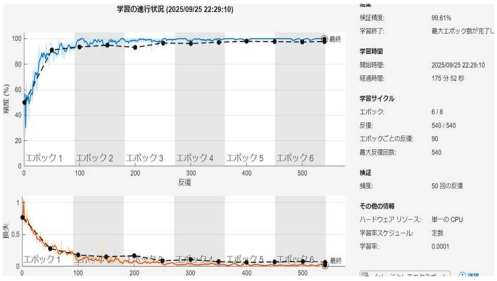

Figure 1. 学習強度の対する学習過程

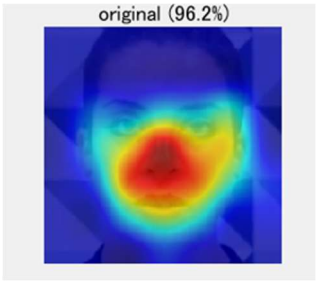

Figure 2. 特徴抽出領域の可視化

参考文献
[1] Kaiming H, et al.: “Deep Residual Learning for Image Recognition”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.
[2] Rössler A. et al.: “FaceForensics++: Learning to Detect Manipulated Facial Images”, IEEE International Conference on Computer Vision(ICCV), 2019.
[3] 西村ら: Grad-CAM を用いた画像認識 AI の特徴分析の試み, ソフトウェア工学の基礎ワークショップ論文集, Vol.29, pp. 211-212，2022 年.

1：日大理工・学部・電気，2：日大理工・教員・電気