TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名: フェイクメディア生成と検出
英文:Generation and detection of fake media

利用課題責任者 越前 功
First name Surname Isao Echizen

所属: 国立情報学研究所
Affiliation: National Institute of Informatics
URL: https://research.nii.ac.jp/~iechizen/official/

邦文抄録（300 字程度）

本プロジェクトでは，ディープフェイクに代表されるフェイク顔画像・映像の検出モデルに関して，Computer Vision（CV）タスクを行う最新の Vision Transformer（ViT）を活用することで，従来のニューラルモデル（ConvNets）より真贋判定の精度向上を目指すことを目的とする．10 種類以上の大規模ディープフェイクデータセットを用いた真贋判定精度評価を行い，自己教師付き学習を用いて学習された最先端かつ商用利用可能な Vision Transformer（ViT）である DINOv2 の判定精度は，従来のニューラルモデル（ConvNets）である EVA-CLIP モデルの判定精度を上回った．本プロジェクトで得た知見は，国内外で実利用が進んでいるフェイク顔画像・映像の真贋判定サービスの改善に有益なものと考えられる．

英文抄録（100 words 程度）

This project aims to improve the detection accuracy of fake face images and videos, represented by deepfakes, by utilizing the latest Vision Transformer (ViT) for Computer Vision (CV) tasks, thereby surpassing the performance of conventional neural models (ConvNets). We conducted the detection accuracy evaluations using more than 10 large-scale deepfake datasets. The detection accuracy of DINOv2, a state-of-the-art and commercially available Vision Transformer (ViT) trained using self-supervised learning, surpassed that of the EVA-CLIP model, a conventional neural network model (ConvNets). The insights gained from this project are expected to be beneficial for improving fake face image and video detection services, which are increasingly being adopted both domestically and internationally.

Keywords: Deepfake, 生成 AI, フェイクメディア, AI セキュリティ, なりすまし

背景と目的

顔，身体，音声，自然言語などの人間由来の情報を AI が学習し，本物と見紛うシンセティックメディアの生成が可能になりつつある．シンセティックメディアは，コミュニケーション分野やエンターテイメント分野を始めとした様々な用途で活用されている．一方で，シンセティックメディアの負の側面として，詐欺や思考誘導，世論操作を行う目的で，犯罪組織やテロ組織が，フェイク映像，フェイク音声，フェイク文書といったフェイクメディアを，AI を用いて生成・拡散させており，世界的な脅威となっている．

申請者らは，Deepfake が社会問題化する以前から Media Clone（造語．AI により本物と見紛うように加工・生成されたメディア）の拡散と生成を防止するプロジェクトを開始しており，2018 年にフェイク顔映像を AI を用いて検出する手法を世界で初めて提案した．さらに 2019 年に未知のフェイク顔映像にも頑健な検出手や，フェイク検出と改ざん領域の推定を同時に行う手法（5-years highest impact award 受賞）を提案し，Deepfake detection と呼ばれる新たな研究分野を創成した．

本プロジェクトでは，深刻化する多種多様なフェイクメディアの出現に対して，Computer Vision（CV）タスクを行う最新の Vision Transformer（ViT）を活用することで，従来モデルより真贋判定の精度向上を目指すことを目的とする．

本プロジェクトでは，深刻化する多種多様なフェイクメディアの出現に対して，Computer Vision（CV）タスクを行う最新の Vision Transformer（ViT）を活用することで，従来モデルと比較してフェイク顔画像・映像の真贋判定精度の向上を実現した．このプロジェクトで得た知見は，国内外で実利用が進んでいるフェイク顔画像・映像の真贋判定サービスの改善に有益なものと考えられる．

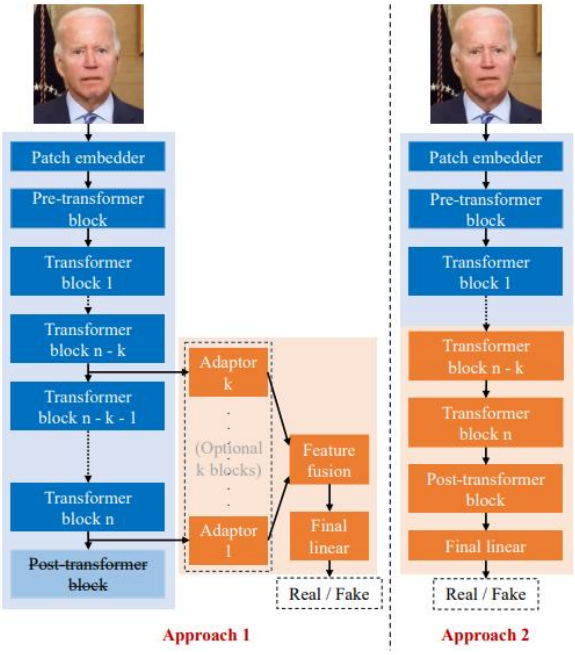

図 1：検討対象の 2 つのアプローチの概要．青色のブロックは frozen block をオレンジのブロックは，fine-tuned block を表す．

概要

フェイク顔画像・映像を検出するディープフェイク検出器を対象に，自己教師付き学習を用いて学習された最先端かつ商用利用可能な Vision Transformer（ViT）である DINOv2 を活用し，実用的なディープフェイク検知 AI 基盤モデルを開発する．Deepfake detection 分野で国際的にスタンダードな評価指標である分類精度（Accuracy）の等価誤り率（EER）の 2 つにおいて，従来のニューラルモデル（ConvNets）との比較評価を行った．

具体的には，顔のディープフェイク検出における事前学習済み ViT である DINOv2 の使用について，以下の 2 つのアプローチから比較研究を実施した（**図 1** 参照）：**Approach 1**：凍結したバックボーンをマルチレベル特徴抽出器として使用することと，**Approach 2**：最終的な変換ブロックを部分的に微調整（ファインチューニング）することである．その結果，**Approach 2** が有効であり，10 種類以上の大規模ディープフェイクデータセットを用いた真贋判定精度評価において，3 億パラメータの DINOv2 モデル（DINOv2-L14-Reg）が Accuracy 96%，EER 4%を達成した．一方，従来のニューラルモデル（ConvNets）である EVA-CLIP モデル（EVA02-CLIP-L-14）は，Accuracy 93%，EER 7%であった．

本評価で得られた知見は，国内外で実利用が進んでいるフェイク顔画像・映像の真贋判定サービスの改善に有益なものと考えられる．

まとめ、今後の課題

本プロジェクトでは，ディープフェイクに代表されるフェイク顔画像・映像の検出モデルに関して，Computer Vision（CV）タスクを行う最新の Vision Transformer（ViT）を活用することで，従来のニューラルモデル（ConvNets）より真贋判定の精度向上が可能なことを示した．今後の課題として，顔以外のオブジェクトや，音声，テキストといったモダリティを対象とした検討や，2 値判定のみならず，どの部分が AI により改変されたかといったローカライゼーションを行うモデルの検討が必要と考える．