TSUBAME 共同利用　平成 30 年度　学術利用　成果報告書

利用課題名：ニューラルネットワークに基づく生成音声と画像の識別
英文：Generated speech and image recognition using neural network

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://www.nii.ac.jp/

邦文抄録　近年、深層学習の発展により合成画像と合成音声は自然画像と自然音声にほぼ区別できなくなってきた。様々な有益なサービスを受けるようになった一方、悪用された場合、社会安全に大きく影響する懸念がある。そこで、本研究は合成画像と合成音声の品質をさらに改善した上、よりロバストなフェイク画像とフェイク音声を検知するアルゴリズムに取り組む。本稿は画像変換技術を用いて合成したフェイクビデオを検知するアルゴリズム及びその検知性能を報告する。

英文抄録　Recently fake image, fake video, and fake audio have achieved very high quality because development of advanced deep learning algorithms and it is difficult for human to distinguish the synthesized image, video, and audio samples from natural ones. In order to mitigate the threats from such fake samples, our work aims to develop an algorithm to detect these fake samples. In this report, we show our initial result of fake video detection.

Keywords: deep learning, fake image, fake video, fake audio, detection

背景と目的

IoT の進展により、画像や音声のようなデータの収集が容易になった。さらに、深層学習の発展により、収集したデータを用いてターゲットの高品質な顔画像や音声信号を合成できるようになった。このような技術が悪用された場合、社会に大きな悪影響を与える懸念がある。深層学習によりサイバー空間上で合成されたフェイク顔画像やフェイク音声と，現実に存在する顔画像や音声を識別する技術の確立は喫緊の課題である。そこで、本研究はフェイク画像、フェイクビデオ及びフェイク音声の検知アルゴリズムの開発を目的として行っている。

本研究は検討の初期段階であるため、本稿は主にフェイクビデオの検知アルゴリズムの仕組み及びその結果を報告する。

概要

フェイクビデオを検知するため、本研究は capsule neural network という新しく提案したニューラルネットワークを利用する。Capsule neural network は従来の convolutional neural network （CNN)とは異なり、特徴量を抽出すると共に、各特徴間の幾何情報を保つことが可能である。そのため、入力は微小な歪みだけであっても、出力は大きなコンフィデンスを得ることが可能となる。図 1 に提案の capsule network に基づくフェイクビデオの検知パイプラインを示す。提案法はまずビデオの各フレームを VGG という画像分類ネットワークに入力し、コンパクトな画像特徴量を抽出する。次に capsule network を用いて歪みを図り、真偽ビデオに関するコンフィデンススコアを出力する。最後に閾値と比較し、フェイクビデオかどうかを判別する。

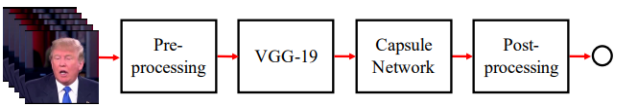

図 1　提案のフェイクビデオ検知手法の処理流れ

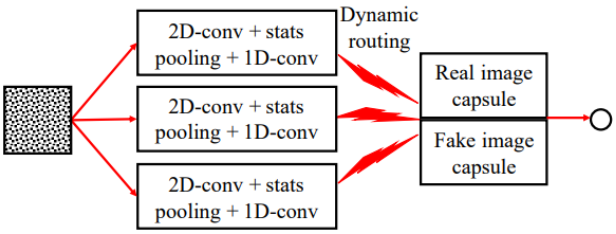

図 2　Capsule network の構成

図 2 は capsule network の具体的な構成を示す。コンボリューションチャンネルを 3 つ用いて、画像特徴量をカプセルにする。次に、dynamic routing アルゴリズムを用いて、類似のカプセル同士の合意を取って出力を決める。出力層は 2 つのカプセルにより構成され、それぞれは自然画像と偽画像を表す。最後に、softmax 関数を用いてカプセル要素の平均スコアを計算する。

結果および考察

実験はプレイバックビデオ、顔スワッピング(deepfake により生成)、facial reenactment (face2face により生成)の 3 種類のフェイクビデオを識別する実験を行った。

Capsule network を利用する提案法を用いてプレイバックビデオを検知する場合、検知精度は従来の CNN と同等レベルの 100%を達した。また、顔スワッピングは従来の 92%から 96%までに向上した。Facial reenactment の検知精度も従来の 98%から 99%までに向上した。

まとめ、今後の課題

本研究は capsule neural network を用いてフェイクビデオの検知を行った。検知精度は従来の CNN を利用する方法より高い精度を得られた。

今後はフェイクビデオだけではなく、フェイク音声の検知も行う予定である。そして、ビデオを強く圧縮しても、音声と顔や唇の動きは高い相関があるため、音声と画像情報を同時に利用するフェイクビデオやフェイク音声の検知も行う。