解除困難なディープフェイクプロテクト信号の開発

所属：　会津大学　コンピュータ理工学部　コンピュータ理工学科

助成対象者：富岡　洋一

概要

人工知能（AI）の発展に伴い，ある人物の顔を同一人物の別の表情、あるいは、別の人物に加工する「ディープフェイク」という改ざん技術が進歩し、ディープフェイクを悪用した犯罪への対抗手段が求められている。画像に不可視の微弱な信号（保護信号）を加えることでディープフェイク生成を妨害する手法が提案されているが、微弱な信号は画像を加工することで解除できる問題がある。そこで、本研究では、従来よりも解除困難である保護信号の開発を目指し、違和感の無い範囲で顔を再構成しながら保護信号を埋め込む技術と平滑化等の画像加工では除去が難しい保護信号を生成する技術を開発した。

abstract

With the development of artificial intelligence (AI), a tampering technique called ″deepfake,″ in which the face of one person is processed into another expression of the same person or a different person, has advanced, and there is a need for countermeasures against crimes that exploit deepfake. Methods to interfere with deepfake generation by adding an invisible weak signal (protection signal) to an image have been proposed, but there is a problem the weak signal can be easily removed by processing the image. Therefore, this study aimed to develop a protective signal that is more difficult to remove than conventional methods and developed a technique for embedding a protective signal while reconstructing a face so that people do not feel strange and a

technique for generating a protective signal that is difficult to remove by smoothing or other image processing.

研究内容

＜背景＞

ディープラーニングをはじめとする人工知能（AI）の発展に伴い，ある人物の顔を同一人物の別の表情、あるいは、別の人物に加工する「ディープフェイク」という改ざん技術が進歩し、ディープフェイクを悪用した犯罪への対抗手段が求められている。今日では、スマートフォンのアプリでも手軽にディープフェイク画像を作成できるようになるなど、技術的な敷居が下がるとともに、人工知能技術の進歩により今後もディープフェイク画像の品質が改善していくことが予想される。ディープフェイクは個人のプライバシーを侵害するだけではなく、政治的あるいは企業内で高い地位を持つ人物のディープフェイク動画が株価操作やビジネス詐欺に悪用される危険性もある。このため、ディープフェイク技術は国、企業、個人間の信頼関係を揺るがし、産業発展の大きな妨げとなることが危惧される。今後はウィルス対策だけではなく、ディープフェイクという次世代のサイバー犯罪に対抗する新しいセキュリティ対策が求められる。

文献[1]等、多くの研究者らによりディープフェイクにより加工された画像・動画（以下、まとめて画像と呼ぶ）を検出する手法は研究されている。しかし、これらの手法では、元画像の悪用を未然に防ぐことはできず、悪用された時点で個人のプライバシーや企業イメージの侵害を防げないことが問題となる。この問題の解決のためには、所有者が画像をウェブ上に公開する前に、ディープフェイクで悪用できないようにその画像を保護することが必要不可欠である。例えば、既存研究として、ディープフェイクに悪用することを防止する微弱な信号を元画像に加えるディープフェイクプロテクト技術が提案されている（文献[2]等）。プロテクト信号を加えた画像から生成されるディープフェイク画像は低品質になるため、画像の悪用を防ぐ効果が期待できる。しかし、既存のプロテクト信号は微弱であるため画像の加工により、比較的容易に解除できる問題がある。

＜目的＞

本研究では、画像圧縮や画像再構成で容易に解除できないプロテクト信号を確立することを目的とする。

＜結果＞

本研究では、従来よりも解除困難である保護信号の開発を目指し、違和感の無い範囲で顔を再構成しながら保護信号を埋め込む技術(ILVR-A)と平滑化等の画像加工では除去が難しい保護信号を生成する技術(PGD-Trap)を開発した。

提案手法は従来手法である Projected Gradient Descent (PGD)[3]と Iterative Latent Variable Refinement (ILVR)[4]に基づいている。PGD では、画像と Deepfake 生成器を入力とし、Deepfake 生成器の出力が低品質画像に近づくように画像に保護信号を加える。画像の品質を劣化させないために、保護信号は微弱な信号（摂動信号）にする必要がある。このため、平滑化などの処理により用意に削除できる問題がある。ILVR は画像データにノイズを加えて徐々に破壊する拡散過程を学習し、ノイズから逆拡散過程により徐々に画像を復元・生成する拡散モデルを拡張した手法である。ILVR では逆拡散過

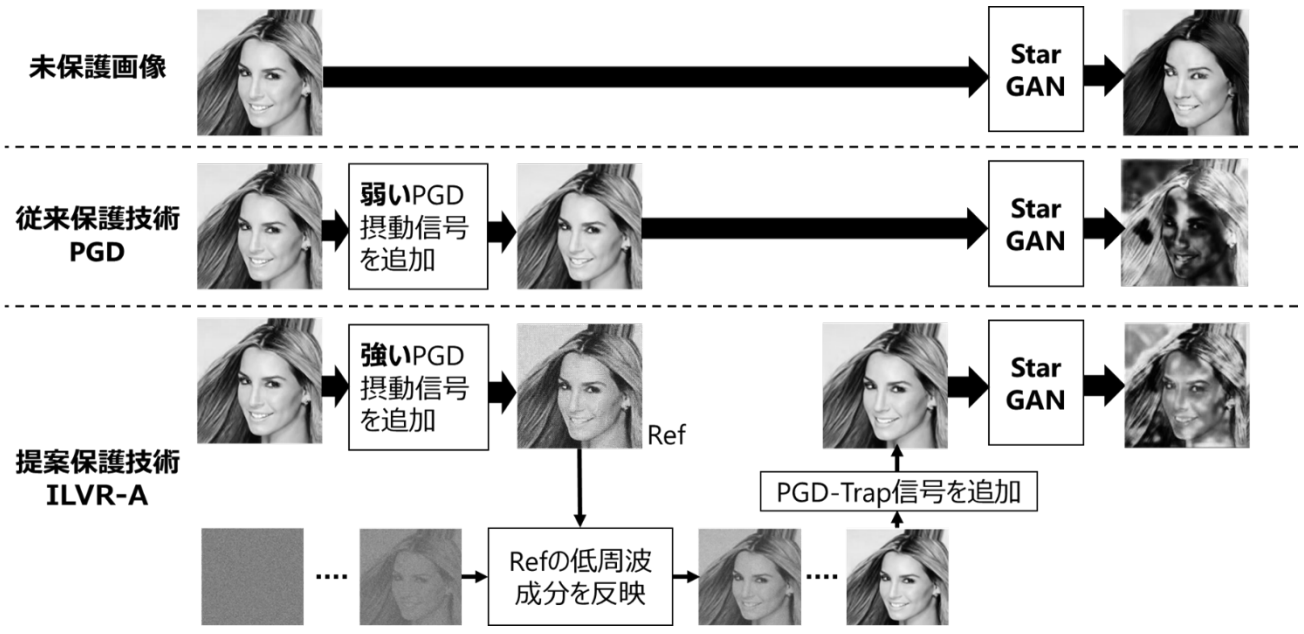

図 1 PGD と提案手法の処理フロー。処理対象の図は CelebA データセット(文献[5])の画像を使用。

程の各繰り返しにおいて、生成過程の画像に対して、参照画像の低周波成分を反映することで参照画像に類似した画像を生成する。

図 1 に示すように、提案の ILVR-A では ILVR の参照画像として、PGD に基づき強い摂動信号を加えた画像を用いることで、対象のディープフェイク生成器に対する保護信号を埋め込んだ自然な画像を生成する。生成された画像は元画像とは細部は異なる可能性があるが、視覚的に類似した画像となる。更に、ディープフェイク生成器に対する妨害効果を高めるため、PGD-Trap と呼ばれる技術を提案した。PGD-Trap は PGD に基づき生成した摂動信号θに対して、ガウシアンフィルタGの逆フィルタを適用した$G^{-1}(\theta)$を対象画像に加える技術である。特に、輝度値が低い部分が強く、輝度値が高い部分は弱くなるように輝度値に基づいて$G^{-1}(\theta)$を画像中に加える。PGD-Trap はガウシアンフィルタ等の平滑化処理を適用することで、ディープフェイク生成器に対する敵対的な保護信号が復元されることを狙った信号である。

図 2 に PGD と提案手法で生成した保護画像を StarGAN に入力したときの結果、図 3 にそれらの保護画像に標準偏差 0.8 のガウシアンフィルタを適用し、StarGAN に入力したときの結果を示す。この例では、StarGAN により髪色を黒色に変更する操作を妨害することを狙っている。図 2 では、提案手法、PGD ともにディープフェイクの妨害に成功しているが、図 3 ではガウシアンフィルタにより PGD の妨害が弱まっていることがわかる。一方、提案手法では、ガウシアンフィルタが適用された場合でもディープフェイクを妨害できてい

図 2 保護信号付きの画像の StarGAN 出力。評価には CelebA データセット(文献[5])の画像を使用。

図 3　保護信号付きの画像に対して標準偏差 0.8 のガウシアンフィルタを適用した場合の StarGAN 出力。評価には CelebA データセット(文献[5])の画像を使用。

ることが確認できた。

ILVR -A において、より自然な画像を生成するためには、自然な画像とそうでない画像を定量的に評価することが必要となる。そこで、本研究では、画像加工に頑健でかつ顔の各パーツに特化したディープフェイク検出器を活用し、加工された画像に対する本物らしさ、偽物らしさの定量的な評価に取り組んだ。また、ディープフェイク生成器とディープフェイク検出器を交互に敵対的に学習させる敵対的生成ネットワークを２段階に用いて、不自然な画像をより自然な画像に復元する生成器と不自然な画像と自然な画像と判別する判別器を同時に構築した。この手法では、１段階目の画像復元では、目の位置や大きさのずれなどの大きな修正を行うことに焦点を当て、２段階目の画像復元では目の色の統一やノイズの軽減といったより細かい画像復元に焦点を当てた。これらの画像復元用の生成器と共に学習した検出器により画像の自然性を評価したところ、１段階目よりも２段階目の検出器の方が、人の目で見たときの自然性の評価と近い結果が得られた。

＜今後＞

本研究では PGD-Trap がガウシアンフィルタに対する対抗手段となることを確認できた。今後は、様々な画像加工に対する逆変換となる関数を近似する

モデルを構築するなどして、多様な画像加工に頑健な PGD-Trap を実現することを狙っている。また、本研究での自然性の定量的評価に関する検討結果を踏まえ、人間の感覚に近い自然性評価を行える評価器を構築し、IVLR-A の画像生成過程に組み込むことで、保護信号を埋め込みつつ元画像よりもさらに品質の高い画像を生成する手法を開発することを目指す。

引用文献

[1] Nirkin, Y.; Wolf, L.; Keller, Y.; Hassner, T. DeepFake detection based on discrepancies between faces and their context. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 6111-6121, vol. 44, 2021.

[2] Wang, R.; Huang, Z.; Chen, Z.; Liu, L.; Chen, J.;Wang, L. Anti-Forgery: Towards a Stealthy and Robust DeepFake Disruption Attack via Adversarial Perceptual-aware Perturbations. arXiv preprint arXiv:2206.00477 2022.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” arXiv preprint arXiv:1706.06083, 2017.

[4] CHOI, Jooyoung, et al. ILVR: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938, 2021.

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision, pp. 3730-3738, 2015.

本助成に関わる成果物

[論文発表]

[J1] Kawabe A, Haga R, Tomioka Y, Shin J, Okuyama Y. A Dynamic Ensemble Selection of Deepfake Detectors Specialized for Individual Face Parts. Electronics. 2023; 12(18):3932. https://doi.org/10.3390/electronics12183932