ISSN 2186-5647

2-5

Image-to-Image 変換を用いた生成 AI モデルに対する AI 画像検出器による評価

日大生産工(院)　○工藤 妃奈乃　　日大生産工　栃窪 孝也

1. まえがき

近年、生成 AI の発達とともに、AI 生成画像も人間の目では判別が難しいほどの画像生成が可能となっている。文章によるテキストや元となる画像を与えることで、誰でも簡単に画像生成を行える反面、AI 生成画像によるデマ・偽情報の拡散、性的ディープフェイクによる被害が増加している。2024 年 11 月の米大統領選では、アフリカ系アメリカ人の有権者に共和党への投票を促すため、ドナルド・トランプ前大統領の支持者らが、AI で偽物の画像を生成・拡散していると BBC 調査報道番組「パノラマ」が明らかにした [1]。また近年では、子供の学校行事や卒業アルバムの写真が、AI で偽画像を生成する「ディープフェイク」の技術で偽の性的画像に加工され、SNS で拡散される被害が増加している [2]。このような問題に対して、AI 生成画像検出器を用いて Text-to-Image 変換によって生成された AI 生成画像の検出を行う方法が研究されている [3]。

本研究では、画像を入力として生成する Image-to-Image 変換を用いた画像生成 AI に対し、AI 生成画像検出ツールを用いて AI 生成画像の検出を行った場合の検出精度について評価する。画像生成 AI は、Pictor、MyEdit を用いて、部分消去、部分置換の 2 種類の Image-to-Image 変換を行い、画像生成を行う。これらの画像を検出する AI 生成画像検出器として、ILLUMINARTY、Hive AI Detector、Sightengine、isgen.ai、Decopy AI を用いて評価を行う。

2. 準備

2.1 画像生成AI

画像生成 AI とは、テキストによる指示（プロンプト）や既存の画像などを入力として与えることで、その条件に基づいた新しい画像を生成できる生成 AI の一分野である。Image-to-Image 変換は、テキストのみを入力として生成する Text-to-Image 変換と異なり、既存の画像とテキストプロンプトを入力として新たな画像を生成する技術である。この技術により、テキストのみでは伝えにくかった細かい雰囲気や色味などが再現しやすくなった。本研究では、画像生成 AI として以下の 2 つを使用した。

(1) Pictor[4]

Pictor とは、Permission Inc.によって開発された iPhone プラットフォームで利用可能なマルチメディアアプリケーションである。Pictor は、AI イラスト・画像生成 AI、消しゴムマジック、背景透過、部分置き換え、高画質化などが一体化となっている。画像生成処理には画像生成 AI である Stable Diffusion 等を利用している。

(2) MyEdit[5]

CyberLink Corp.によって開発されたブラウザ上で画像編集・音声編集が可能な生成 AI 技術を搭載した、オンライン写真・音声編集サイトである。MyEdit は AI 画像生成、高画質化、オブジェクト除去、背景透過・切り抜き、AI イラスト化、ピンボケ・手ぶれ補正、画像ノイズ除去、音声編集機能、文字起こし機能、ボイスチェンジャー、音声ノイズ除去、効果音作成など様々なツールがある。MyEdit は画像生成機能に Stable Diffusion である SDXL を導入している。

2.2 AI生成画像検出器

画像の微妙な不規則性を識別することで生成 AI によって人工的に生成されたものか、人間が作成したものかを推定する機器である。機械学習モデルを使用し、大量の画像が学習されている。本研究では、AI 画像検出器として、以下の 5 つを使用した。

(1) ILLUMINARTY[6]

画像やテキストを特定することに焦点を当てた AI 画像検出器である。ILLUMINARTY は、画像については、コンピュータビジョンアルゴリズムを組み合わせて、画像が生成 AI モデルによって生成された可能性を提供する。テキストについては NLP アルゴリズムを組み合わせて、テキストが AI モデルによって生成された可能性を提供する。

(2) Hive AI Detector[7]

人工知能を活用してコンテンツの分析や判定を行うプラットフォームである。画像、動画、テキスト、音声を分析し、明確な信頼度スコアを返し、真偽や品質を判定することができる。

(3) Sightengine[8]

画像、動画、テキストが生成 AI によって作られたも

Evaluation of AI Image Detectors for Generative AI Models Using Image-to-Image Conversion

Hinano KUDO and Kouya TOCHIKUBO

のかを解析し提供するクラウド型サービスである。

(4) isgen.ai[9)]

画像、テキストのコンテンツの認証を行う検出ツールである。多言語に対応しており、日本語を含む様々な言語のテキストを分析することができる。

(5) Decopy AI[10)]

AI 生成または偽の画像を検出する AI 画像検出ツールである。学習に使用したデータセットには、Midjourney、Stable Diffusion、DALL-E、Flux など様々な AI モデルによって生成された人工画像を含む約 1000 万枚の画像が含まれている。

3. AI画像検出器による評価

3.1 使用画像

画像生成 AI である Pictor と MyEdit を用い、それぞれ「部分消去」と「部分置換」での画像生成を行う。部分消去では、画像の消したい部分を選択し、物体を消去した画像を生成する。部分置換では、画像内のオブジェクトで置き換える対象となるものを選択し、テキストプロンプトにて置き換えたいものを指示し、画像生成を行う。テキストプロンプトは、生成後に置き換えたい人やモノを入力とする。画像枚数は、それぞれ 17 枚ずつ生成した。また、Image-to-Image 変換の基となる画像はフリー素材サイトで収集した。生成した Pictor と MyEdit の画像の例を図 1,2 に示す。図 1,2 はどちらもエッフェル塔の画像を入力(元画像)としたものである。部分消去ではエッフェル塔を消去した。部分置換ではエッフェル塔を置換対象としテキストプロンプトは「東京タワー」として、エッフェル塔を東京タワーに置き換えた画像を生成した。

(a)元画像　(b)部分消去　(c)部分置換

図1. PictorのAI生成画像

(a)元画像　(b)部分消去　(c)部分置換

図2. MyEditのAI生成画像

3.2 評価方法

AI 画像検出器である ILLUMINARTY、Hive AI Detector、Sightengine、isgen.ai、Decopy AI を用いて、Pictor と MyEdit で生成した AI 生成画像を一枚ずつ検出する。検出結果は、画像が「画像生成 AI で生成されたものか」という AI 確率で評価される。どの AI 画像検出器も AI 確率はパーセンテージで出力され、100%に近いほど画像生成 AI で生成された偽画像であると判断され、0%に近いほど人間が作成した画像に近いものと判断される。

4. 検出結果

各画像生成 AI で生成された画像の検出結果を図 3～6 に示す。

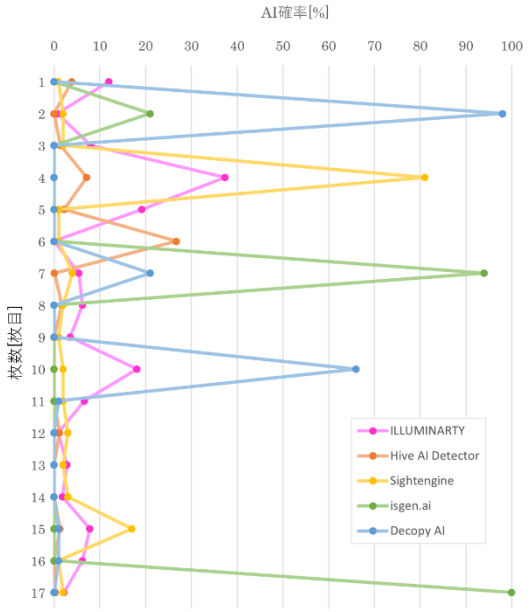

図3. Pictor/部分消去の検出結果

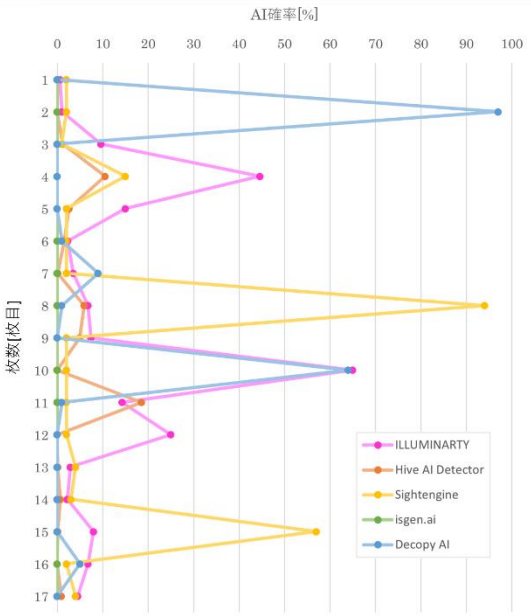

図4. Pictor/部分置換の検出結果

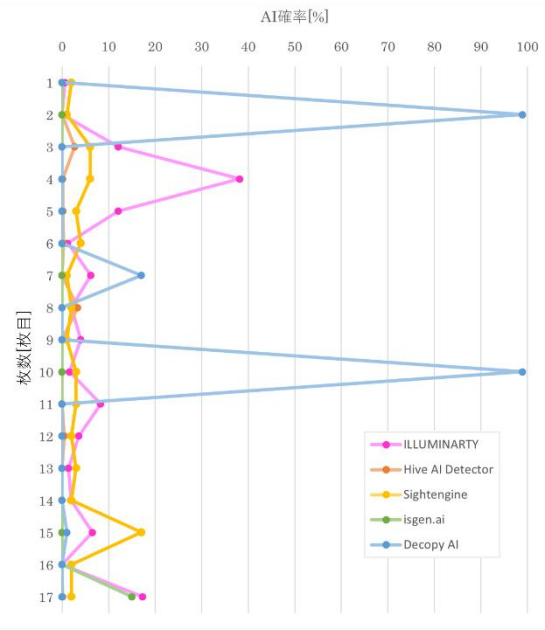

図5. MyEdit/部分消去の検出結果

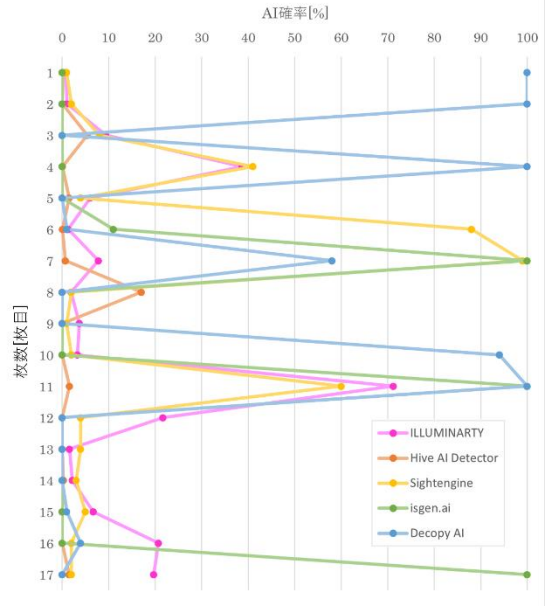

図6. MyEdit/部分置換の検出結果

検出結果は全体的に低く、ほとんどの検出結果が、AI 確率 20%以下に属していていた。また、全体の検出結果うち、AI 確率が 50%を超えるものは 24 枚であり、そのうち 90%を超えるものは 16 枚であった。これらの結果から、Decopy AI が他の検出器と比較して特に高い AI 確率を示す傾向があることがわかる。MyEdit による部分置換画像は他の結果と比較して、AI 確率が高く、50%以上が 13 枚、そのうち 90%以上が 9 枚であった。一方、Pictor による部分置換の画像では isgen.ai において 15 枚が検出が不可能であった。また、MyEdit による部分消去画像では ILLUMINARTY で 2 枚が検出不可となった。全体として Image-to-Image 変換を用いた AI 生成画像の検出結果は低い AI 確率を示した。

次に特徴的な画像とその結果について、観察する。階段に座る女性たちを元画像とし、真ん中の女性を部分消去したものである。それらの画像を図 7 に示す。

(a)元画像　　(b)消去部分の選択

(c)Pictor　　(d)MyEdit

図 7. 部分消去した画像

Pictor で生成した画像は、消去部分がぼやけており不自然な画像となっている。一方、MyEdit で生成した画像は元画像で見えていなかった部分が補完されており、より自然な仕上がりとなっていた。これらの画像の検出結果では AI 確率が 50%を超えたものは 1 枚のみであった。それは、Pictor による AI 生成画像を Sightengine で検出したもので、その AI 確率は 81%あった。ILLUMINARTY での検出結果は、Pictor・MyEdit のどちらの結果も約 38%前後であったが、いずれも低い値であった。他の検出結果は 10%未満であった。視覚的判断と検出結果では、相反した結果が得られた。

次に、東京の街並みを部分置換したものである。図 8 にそれらの画像を示す。

(a)元画像

(b)Pictor

(c)MyEdit

図 8. 部分置換した画像

画像生成時のテキストプロンプトには「街を破壊しながら歩くゴジラ」と用い、東京の街にゴジラが出現したような画像を生成した。この画像は一般的な人間の視覚的判断においても AI によって生成されたものであると識別できる画像である。検出結果では、MyEdit で生成した画像が Sightengine では 99%、isgen.ai では 100%と高い確率を示し、Decopy AI でも 58%と比較的高い検出結果であった。しかし Pictor およびその他の MyEdit による生成画像は 10%以下と低い検出結果となった。同一の画像とテキストを用いて、Pictor・MyEdit で生成した画像の場合でも、検出結果には大きな差が確認された。

5. まとめ

Image-to-Image変換を用いて生成したAI生成画像の検出結果は、全体として低い結果となった。画像の大部分を消去・置換した場合には、検出器によっては高い確率を得られたものの、低い検出結果となるものも確認できた。視覚的に明らかに不自然な画像であっても検出器によっては「AIでない」と識別される確率のほうが高いことから、Image-to-Image変換を用いたAI生成画像をAI画像検出器で検出することは困難であるといえる。また、画像検出器において、画像によって検出が不可能であった。このことから、画像の特徴やファイルサイズによって検出の可否が左右されると考えられる。

参考文献

1) Marianna Spring(BBC), Trump supporters target black voters with faked AI images, 2024, https://www.bbc.com/news/world-us-canada-68440150, (参照 2025-10-03)
2) 村上喬亮 (読売新聞), 卒業アルバム加工した偽の性的画像SNS拡散、小中高生らが作成…警視庁がAIサイト調査, 2025, https://www.yomiuri.co.jp/national/20250831-OYT1T50010/, (参照 2025-10-03)
3) 井口駿治, 稲葉宏幸, “各種生成AIモデルに対するAI生成画像検出ツールの性能比較に関する調査研究”, コンピュータセキュリティシンポジウム2024論文集, (2024), pp1295-1299.
4) Permission Inc., AIイラスト,画像生成,消しゴムマジック-Pictor, 2025, https://play.google.com/store/apps/details?id=jp.co.permission.pictor&hl=ja, (参照 2025-10-04)
5) CyberLink Corp., サイバーリンク、MyEditのAI画像生成機能にSDXLを導入、AI置き換え機能、AI人物背景機能、ボイスチェンジャーに日本語のプロファイルを20種類追加, 2024, https://jp.cyberlink.com/jpn/press_room/view_5043.html, (参照 2025-10-04)
6) aihinto, Illuminarty, 2024, https://aihinto.com/product/illuminarty/, (参照 2025-10-04)
7) Hive Moderation, AI-Generated Content Detection, 2025, https://hivemoderation.com/ai-generated-content-detection, (参照 2025-10-04)
8) Sightengine, Content Moderation and Image Analysis you can rely on, 2025, https://sightengine.com/, (参照 2025-10-04)
9) isgen.ai, 最も正確なAI画像検出器, 2025, https://isgen.ai/ja/AI%E7%94%BB%E5%83%8F%E6%A4%9C%E5%87%BA%E5%99%A8, (参照 2025-10-04)
10) Decopy AI, 無料AI画像検出 : AI生成画像を識別, 2025, https://decopy.ai/jp/ai-image-detector/, (参照 2025-10-04)