説明可能AIを用いた実写画像・AI生成画像の判別

Distinction between Real Photos and AI-Generated Images Using eXplainable AI

河合 青空[1)]
指導教員　青木 輝勝[2)]

1）東京工科大学大学院　バイオ・情報メディア研究科　コンピュータサイエンス専攻　青木研究室
2）東京工科大学　コンピュータサイエンス学部　コンピュータサイエンス学科　青木研究室

本研究では，生成AIによって生成された画像と実写画像を判別するために，XAI(eXplainable Artificial Intelligence)を用いた手法を提案する．具体的には，XAIを用いて判別結果の要因となる特徴を視覚化することで判別の根拠を明確化し，視覚化された特徴を活用した新たな判別手法の構築を最終目標とする．

説明可能AI，AI生成画像判別，画像特徴量，深層学習

１．はじめに

近年，生成系AIの発展により，世の中の利便性が向上している．しかしながら，著しく性能が向上した生成系AIは，虚偽の情報を拡散する恐れがある．最近では，ドナルド・トランプ氏が逮捕されている情景を描いた偽画像や，ペンタゴン付近で爆発が起きている情景を描いた偽画像の誤情報などが拡散されることで世の中に混乱を招いた[1][2]．これらの誤情報の拡散は，AIによって生成された画像と実写画像の判別が難しいことに起因している．したがって，生成AIによって生成された画像を実写画像判別するための手法の発展が必要不可欠である．

そこで，本研究では，実写画像・AI生成画像の判別器にXAIを導入することで判別器が着目している領域を視覚化し，判別結果の要因となった特徴を明確にすることを目的とする．また，視覚化した特徴を活用した手法の実装を最終目標とする．

２.XAIを用いた生成画像の判別に関する既存研究

Birdらは，実画像にデータセットCIFAR-10，生成画像にSD(Stable Diffusion) V1.4で生成した画像を用いたデータセットを構築し，CNN(Convolutional Neural Network)を用いた2値分類を実施した．また，XAIによって判別の要因となった領域を視覚化した[3]．しかし，判別要因がSD1.4で生成された画像に限られていること，低解像度画像(32×32画素)で構築されていることが問題である．

３．提案手法

前章で示した問題点を改善するために，実写画像としてImageNetデータセット，生成画像として8つの生成AIで生成した画像を用いて構築されているデータセットGenImage[4]を使用した2値分類を生成AIごとに実施し，XAIによって判別結果の要因となった特徴を明確にする．

３.１．GenImage

GenImageは約268万枚の画像を含んだデータセットであり，133万枚の本物画像と135万枚の生成画像で構成されている．実画像は，1000種類のカテゴリ(クラス)がある大規模なデータセットImageNetを使用している．生成画像は，BigGAN，GLIDE，VQDM，SD V1.4，SD V1.5，ADM，Midjourney，Wukongの8つの生成モデルを使用し，生成している．

３.２．Grad-CAMによる要因分析

本研究では，XAIとしてGrad-CAM(Gradient-weighted Class Activation Mapping)を使用する[5]．Grad-CAMはCNNの最後の畳み込み層に流入

する勾配情報を利用し，Convolution 層の特徴マップを重み付けて可視化する手法であり，下記の式で表される．

$$\alpha_k^C = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (1)$$

$$L_{Grad-CAM}^C = ReLU\left(\sum_k \alpha_k^C A^k\right) \quad (2)$$

ここで，y^c を最終出力，A^k は特徴マップのチャンネル k，I,j は特徴マップの幅と高さ，Z は要素数を表す．

4．実験

本研究では，全結合層のみに変更を加えた ResNet50 を用いて 2 値分類を実施する[6]．学習に使用した生成モデルは，SD V1.5 を除く 7 つのモデルであり，モデルごとの学習に 324000 枚の画像を用いる．また，テストに使用する画像は，実画像 6000 枚，生成画像 6000 枚である．

表 1 に ResNet50 を用いて 2 値分類した際の正解精度を示す．横軸は学習に使用した生成 AI，縦軸はテストに使用した生成 AI を表す．学習に使用した生成モデルに対する判別器の精度は高いが，学習に使用していない生成モデルに対しては精度が低い問題を示した．

図 1 に Grad-CAM を用いて判別結果の要因を視覚化した結果を示す．1 行目は Midjourney で生成した画像，2 行目は SD V1.4 で生成した画像を表し，列は ResNet50 を学習する際に使用した生成 AI を表す．また，判別結果の画像は，学習させた AI によって正確に分類された画像のみを表示している．判別要因の視覚化により，誤った判断をした AI では，正しい判断をした AI と比較して，注目領域が限定的であることを示した．

表 1 ResNet50 を用いた精度比較

	SD V1.4	Midjourney	GLIDE	VQDM	ADM	Wukong	BigGAN
SD V1.4	98.23%	59.50%	53.99%	50.27%	49.23%	97.26%	50.46%
Midjourney	66.96%	91.88%	75.99%	51.22%	54.33%	62.39%	57.23%

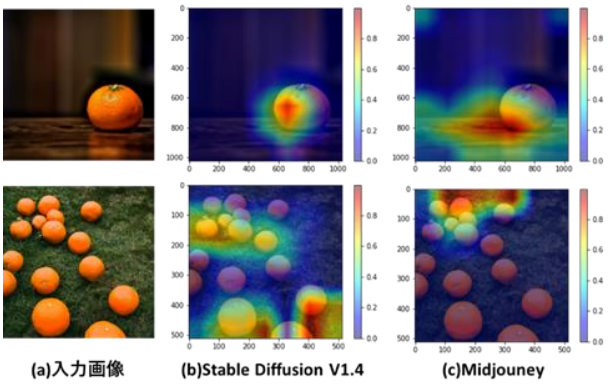

図 1 XAI による判別要因の視覚化

5．おわりに

本研究では，生成画像判別に使用される大規模データセットを用いて，判別結果の要因分析を行うことを提案した．今後の目標は，要因分析の結果を使用した手法の構築である．

参考文献

[1] Donie O'Sullivan, Jon Passantino, "'Verified' Twitter accounts share fake image of 'explosion' near Pentagon, causing confusion", https://edition.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html, (2024 年 7 月 10 日閲覧)

[2] Kayleen Devlin, Joshua Cheetham, "Fake Trump arrest photos: How to spot an AI-generated image", https://www.bbc.com/news/world-us-canada-65069316, (2024 年 7 月 10 日閲覧)

[3] Jordan J. Bird, Ahmad Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images", IEEE Access, vol. 12, pp.15642-15650, (2024)

[4] Mingjian Zhu, Hanting Chen, Qiangyu Yan, et al., "GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image", In Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23), Article 3398, 77771–77782, (2024)

[5] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", IEEE International Conference on Computer Vision (ICCV), 2017

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al., "Deep Residual Learning for Image Recognition". IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016